H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.0

nestor

vwh

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Korrektur), Martina Kerzel (Bildbearbeitung) und Jörn Tietgen (Layout und Formatierung des Gesamttextes) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte dieses Buchs stehen auch als Onlineversion (http://nestor.sub.uni-goettingen.de/handbuch/) sowie über den Göttinger Universitätskatalog (http://www.sub.uni-goettingen.de) zur Verfügung.
Die digitale Version 2.0 steht unter folgender Creative-Commons-Lizenz:
„Attribution-Noncommercial-Share Alike 3.0 Unported“
http://creativecommons.org/licenses/by-nc-sa/3.0/

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2009
www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide

Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-940317-48-3

URL für nestor Handbuch (Version 2.0): urn:nbn:de:0008-2009073109
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2009073109

Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.

Vorwort
Heike Neuroth

1 Einführung
Hans Liegmann (†), Heike Neuroth

2 State of the Art
2.1 Einführung Kap.2:1
Regine Scheffel
2.2 LZA-Aktivitäten in Deutschland aus dem Blickwinkel von nestor Kap.2:2
Mathias Jehn und Sabine Schrimpf
2.3 Bibliotheken Kap.2:6
Mathias Jehn und Sabine Schrimpf
2.4 Archive Kap.2:9
Christian Keitel
2.5 Museum Kap.2:16
Winfried Bergmeyer

3 Rahmenbedingungen für die LZA digitaler Objekte
3.1 Einführung Kap.3:1
Stefan Strathmann
3.2 Nationale Preservation Policy Kap.3:3
Stefan Strathmann
3.3 Institutionelle Preservation Policy Kap.3:6
Stefan Strathmann

3.4 Verantwortlichkeiten.....Kap.3:10
Natascha Schumann

3.5 Auswahlkriterien.....Kap.3:15
Andrea Hänger, Karsten Huth und Heidrun Wiesenmüller

4 Das Referenzmodell OAIS – Open Archival Information System

4.1 Einführung Kap.4:1
Achim Oßwald

4.2 Das Referenzmodell OAIS.....Kap.4:3
Nils Brübach
Bearbeiter: Manuela Queitsch, Hans Liegmann (†), Achim Oßwald

5 Vertrauenswürdigkeit von digitalen Langzeitarchiven

5.1 Einführung.....Kap.5:1
Susanne Dobratz und Astrid Schoger

5.2 Grundkonzepte der Vertrauenswürdigkeit und Sicherheit Kap.5:2
Susanne Dobratz und Astrid Schoger

5.3 Praktische Sicherheitskonzepte.....Kap.5:8
Siegfried Hackel, Tobias Schäfer und Wolf Zimmer

5.4 Kriterienkataloge für vertrauenswürdige digitale Langzeitarchive Kap.5:19
Susanne Dobratz und Astrid Schoger

6 Metadatenstandards im Bereich der digitalen LZA

6.1 Einführung.....Kap.6:1
Mathias Jehn

6.2 Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten.....Kap.6:3
Markus Enders

6.3 PREMIS.....Kap.6:9
Olaf Brandt

6.4 LMER.....Kap.6:14
Tobias Steinke

6.5 MIX.....Kap.6:17
Tobias Steinke

7 Formate

7.1 Einführung.....Kap.7:1
Jens Ludwig

7.2 Digitale Objekte und Formate.....Kap.7:3
Stefan E. Funk

7.3 Auswahlkriterien.....Kap.7:9
Jens Ludwig

7.4 Formatcharakterisierung.....Kap.7:13
Stefan E. Funk und Matthias Neubauer

7.5 File Format Registries.....Kap.7:19
Andreas Aschenbrenner und Thomas Wollschläger

8 Digitale Erhaltungsstrategien

8.1 Einführung.....Kap.8:1
Stefan E. Funk

8.2 Bitstream Preservation.....Kap.8:3
Dagmar Ullrich

8.3 Migration Kap.8:10
Stefan E. Funk

8.4 Emulation....Kap.8:16
Stefan E. Funk

8.5 Computermuseum....Kap.8:24
Karsten Huth

8.6 Mikroverfilmung....Kap.8:32
Christian Keitel

9 Access

9.1 Einführung....Kap.9:1
Karsten Huth

9.2 Workflows für den Objektzugriff Kap.9:3
Dirk von Suchodoletz

9.3 Retrieval....Kap.9:19
Matthias Neubauer

9.4 Persistent Identifier (PI) – ein Überblick Kap.9:22
Kathrin Schroeder

9.4.1 Der Uniform Resource Name (URN) Kap.9:46
Christa Schöning-Walter

9.4.2 Der Digital Objekt Identifier (DOI)....Kap.9:57
Jan Brase

10 Hardware

10.1 Einführung....Kap.10:1
Stefan Strathmann

10.2 Hardware-Environment....Kap.10:3
Dagmar Ullrich

10.3 Digitale Speichermedien Kap.10:6
Dagmar Ullrich

11 Speichersysteme mit Langzeitarchivierungsanspruch

11.1 Einführung.....Kap.11:1
Heike Neuroth

11.2 Repository Systeme – Archivsoftware zum Herunterladen Kap.11:3
Andreas Aschenbrenner

11.3 Speichersysteme mit Langzeitarchivierungsanspruch Kap.11:7
Karsten Huth, Kathrin Schroeder und Natascha Schumann

12 Technischer Workflow

12.1 Einführende Bemerkungen und Begriffsklärungen Kap.12:1
Reinhard Altenhöner

12.2 Workflow in der Langzeitarchivierung: Methode und Herangehensweise Kap.12:5
Reinhard Altenhöner

12.3 Technisches Workflowmanagement in der Praxis: Erfahrungen und Ergebnisse.....Kap.12:9
Reinhard Altenhöner

12.4 Systematische Planung von Digitaler Langzeitarchivierung..... Kap.12:14
Carmen Heister, Hannes Kulovits, Christoph Becker und Andreas Rauber

13 Tools

13.1 Einführung.....Kap.13:1
Stefan Strathmann

13.2 Plato Kap.13:3
Carmen Heister, Hannes Kulovits, Andreas Rauber

13.3 Das JSTOR/Harvard Object Validation Environment (JHOVE) Kap.13:20
Stefan E. Funk

13.4 Die kopal Library for Retrieval and Ingest (koLibRI) Kap.13:27
Stefan E. Funk

14 Geschäftsmodelle

14.1 Einführung Kap.14:1
Achim Oßwald

14.2 Kosten Kap.14:3
Thomas Wollschläger und Frank Dickmann

14.3 Service- und Lizenzmodelle Kap.14:9
Thomas Wollschläger und Frank Dickmann

15 Organisation

15.1 Einführung Kap.15:1
Sven Vlaeminck

15.2 Perspektiven der Beschreibung Kap.15:6
Christian Keitel

16 Recht

16.1 Einführung Kap.16:1
Mathias Jehn

16.2 Rechtliche Aspekte Kap.16:3
Arne Upmeier

16.3 Langzeitarchivierung wissenschaftlicher Primärdaten Kap.16:14
Gerald Spindler und Tobias Hillegeist

17 Vorgehensweise für ausgewählte Objekttypen

17.1 Einführung.....Kap.17:1
Regine Scheffel

17.2 Textdokumente.....Kap.17:3
Karsten Huth

17.3 Bilddokumente.....Kap.17:8
Markus Enders

17.4 Multimedia/Komplexe Applikationen.....Kap.17:19
Winfried Bergmeyer

17.5 Video.....Kap.17:25
Dietrich Sauter

17.6 Audio.....Kap.17:58
Winfried Bergmeyer

17.7 Langzeitarchivierung und -bereitstellung im E-Learning-Kontext.....Kap.17:63
Tobias Möller-Walsdorf

17.8 Interaktive digitale Objekte.....Kap.17:69
Dirk von Suchodoletz

17.9 Web-Archivierung zur Langzeiterhaltung von Internet-Dokumenten.....Kap.17:88
Andreas Rauber und Hans Liegmann (†)

17.10 Digitale Forschungsdaten.....Kap.17:104
Jens Klump

17.11 Computerspiele.....Kap.17:116
Karsten Huth

17.12 E-Mail-Archivierung.....Kap.17:131
Karin Schwarz

18 Praxisbeispiele

18.1 Einführung Kap.18:1
Regine Scheffel

18.2 Langzeitarchivierung von elektronischen Publikationen durch die Deutsche Nationalbibliothek Kap.18:3
Maren Brodersen und Sabine Schrimpf

18.3 Langzeitarchivierung eines digitalen Bildarchivs – Projekt zum Aufbau eines Langzeitarchivs für hochaufgelöste digitale Bilddateien der Staatsgalerie Stuttgart am BSZ Kap.18:13
Werner Schweibenz und Stefan Wolf

18.4 ARNE – Archivierung von Netzressourcen des Deutschen Bundestages..... Kap.18:22
Angela Ullmann

19 Qualifizierung im Themenbereich „Langzeitarchivierung digitaler Objekte“.....Kap.19:1

Regine Scheffel, Achim Oßwald und Heike Neuroth

Anhang

Herausgeberverzeichnis Kap.20:2
Autorenverzeichnis Kap.20:5
Akronym- und Abkürzungsverzeichnis..... Kap.20:11

Vorwort

Stellen Sie sich vor: Wir befinden uns im Jahre 2030 irgendwo in Deutschland. Irgendwo? Nein, bei Ihnen in der guten Stube, wo Sie Ihren Enkelkindern stolz von Ihrer Weltumsegelung aus dem Jahr 2010 berichten. Untermalen möchten Sie Ihre Geschichte gerne mit anschaulichem Bildmaterial und zeitgenössischer Musik.

Diese hatte damals wesentlich zur Mythen- und Legendenbildung im Freundeskreis beigetragen, seitdem genießen Sie den Ruf eines unerschrockenen Helden. Nun ist es an der Zeit, diese kleine Geschichte lebendig zu halten und sie der nächsten Generation, nämlich Ihren Enkelkindern, weiterzugeben.

Doch Ihr GODD (Global Omnipresent Digital Device) weigert sich, die aufwändig erstellte Videoschau überhaupt zu lesen. Ganz im Gegenteil, Ihnen wird lapidar mitgeteilt, dass es sich um veraltete Technik handelt, die nicht länger unterstützt wird. Sie möchten sich bitte an einen „Datenarchäologen" Ihres Vertrauens wenden.

Aber was ist nun eigentlich ein „Datenarchäologe"? Ein Datenarchäologe stellt nicht mehr lesbare Daten wieder her, um sie wieder nutzbar zu machen. Er - oder sie - kommt zum Einsatz, wenn die Havarie schon erfolgt ist. Doch soweit soll es nicht kommen. Deshalb benötigt man Experten wie den „Digital Curator" oder den „Digital Preservation Specialist", der dafür sorgt, dass bereits bei der Entstehung digitaler Daten perspektivisch ihre langfristige Erhal-

tung berücksichtigt wird. Er – oder sie – ist in der Lage eine Institution bei der Entwicklung ihrer Langzeitarchivierungsstrategie für die erzeugten Daten zu unterstützen oder Entwicklungen in einem vertrauenswürdigen digitalen Langzeitarchivsystem zu planen und durchzuführen.

Glücklicher als Sie mit Ihren privaten digitalen Daten sind da die Astronomen, wenn sie nach Daten von Himmels-Beobachtungen fahnden, die bereits Jahrzehnte zurückliegen. Obwohl die Bild- und Datenarchive dieser Beobachtungen in vielfältigen und sehr unterschiedlichen Formaten abgespeichert wurden, gibt es immer die Möglichkeit, über geeignete Interface-Verfahren die Originaldaten zu lesen und zu interpretieren.

Dies ist der Fall, weil durch das sogenannte Virtuelle Observatorium weltweit die Archive für astronomische Beobachtungen vernetzt und immer in den neuesten digitalen Formaten zugänglich sind, seien es digitale Aufnahmen von Asteroiden, Planetenbewegungen, der Milchstrasse oder auch Simulationen des Urknalls. Selbst Photoplatten von Beginn des 20. Jahrhunderts wurden systematisch digitalisiert und stehen zur Wiederverwendung bereit. So sind ältere und neue digitale Daten und Bilder gemeinsam nutzbar und gewähren einen Blick in das Universum, der sich über weit mehr Wellenlängen erstreckt als die Sinne des Menschen allein wahrnehmen können.

Wir freuen uns, Ihnen mit dem nestor Handbuch „Eine kleine Enzyklopädie der digitalen Langzeitarchivierung“ den aktuellen Wissensstand über die Langzeitarchivierung digitaler Objekte im Überblick sowie aus vielen Teilbereichen nun auch in gedruckter Form präsentieren zu können.

Schon seit Frühjahr 2007 ist das Handbuch in digitaler Version unter http://nestor.sub.uni-goettingen.de/handbuch/ verfügbar und seitdem in mehreren Intervallen aktualisiert worden. Die nun vorliegende Version 2.0 – hier gedruckt und unter o.g. URL auch weiterhin entgeltfrei herunterladbar – wurde neu strukturiert, um neue Themenfelder ergänzt und bislang schon vorhandene Beiträge wurden, wo fachlich geboten, überarbeitet.

Aus seiner Entstehung ergibt sich eine gewisse Heterogenität der einzelnen Kapitel untereinander, z.B. bezüglich der Ausführlichkeit des behandelten Themas oder des Schreibstils. Der Herausgeberkreis hat nicht primär das Ziel verfolgt, dies redaktionell lektorierend auszugleichen oder ein insgesamt kohärentes Gesamtwerk vorzulegen. Vielmehr geht es ihm darum, der deutschsprachigen Gemeinschaft eine möglichst aktuelle „Kleine Enzyklopädie der digitalen Langzeitarchivierung“ anbieten zu können.

Die parallel verfügbare entgeltfreie, digitale Version des Handbuchs wird bei Bedarf aktualisiert und erweitert, eine zweite Druckauflage ist bereits geplant.

Gerne nehmen wir Ihre Anregungen auf und berücksichtigen sie bei zukünftigen Aktualisierungen!

Unser Dank gilt insbesondere den Autorinnen und Autoren, ohne die es nur bei der Idee eines solchen Handbuches geblieben wäre. Mein Dank gilt aber auch den Mitherausgebern dieser Ausgabe, durch deren engagiertes Stimulieren und „Bändigen" der Autoren die vielen Beiträge erst zu einem Gesamtwerk zusammengeführt werden konnten.

Zusammen mit allen Beteiligten hoffe ich, dass dieses Handbuch Ihnen hilfreiche Anregungen und Anleitungen zu einem erfolgreichen Einstieg in die Theorie und Praxis der Langzeitarchivierung digitaler Objekte bietet!

Heike Neuroth

1 Einführung

Hans Liegmann (†), Heike Neuroth

Die digitale Welt, eine ständig wachsende Herausforderung

Die Überlieferung des kulturellen Erbes, traditionell eine der Aufgaben von Bibliotheken, Archiven und Museen, ist durch die Einführung digitaler Medien und innovativer Informationstechnologien deutlich anspruchsvoller geworden. In der heutigen Zeit werden zunehmend mehr Informationen (nur) digital erstellt und veröffentlicht. Diese digitalen Informationen, die Güter des Informations- und Wissenszeitalters, sind einerseits wertvolle kulturelle und wissenschaftliche Ressourcen, andererseits sind sie z.B. durch die Kurzlebigkeit vieler Formate sehr vergänglich. Die Datenträger sind ebenso der Alterung unterworfen wie die Datenformate oder die zur Darstellung notwendige Hard- und Software. Um langfristig die Nutzbarkeit der digitalen Güter sicherzustellen, muss schon frühzeitig Vorsorge getroffen werden. Es müssen Strategien zur digitalen Langzeitarchivierung entwickelt und umgesetzt werden.

Die Menge und die Heterogenität der Informationen, die originär in digitaler Form vorliegen, wachsen beständig an. In großem Umfang werden ursprüng-

lich analog vorliegende Daten digitalisiert (z.B. Google Print Projekt[1]), um den Benutzerzugriff über Datennetze zu vereinfachen. Im Tagesgeschäft von Behörden, Institutionen und Unternehmen werden digitale Akten produziert, für die kein analoges Äquivalent mehr zur Verfügung steht. Sowohl die wissenschaftliche Fachkommunikation wie der alltägliche Informationsaustausch sind ohne die Vermittlung von Daten in digitaler Form nicht mehr vorstellbar.

Mit der Menge der ausschließlich digital vorliegenden Information wächst unmittelbar auch ihre Relevanz als Bestandteil unserer kulturellen und wissenschaftlichen Überlieferung sowie die Bedeutung ihrer dauerhaften Verfügbarkeit für Wissenschaft und Forschung. Denn das in der „scientific community" erarbeitete Wissen muss, soll es der Forschung dienen, langfristig verfügbar gehalten werden, da der Wissenschaftsprozess immer wieder eine Neubewertung langfristig archivierter Fakten erforderlich macht. Die Langzeitarchivierung digitaler Ressourcen ist daher eine wesentliche Bedingung für die Konkurrenzfähigkeit des Bildungs- und Wissenschaftssystems und der Wirtschaft. In Deutschland existiert eine Reihe von Institutionen (Archive, Bibliotheken, Museen), die sich in einer dezentralen und arbeitsteiligen Struktur dieser Aufgabe widmen.

Im Hinblick auf die heutige Situation, in der Autoren und wissenschaftliche Institutionen (Universitäten, Forschungsinstitute, Akademien) mehr und mehr selbst die Veröffentlichung und Verbreitung von digitalen Publikationen übernehmen, erscheint auch weiterhin ein verteilter Ansatz angemessen, der jedoch um neue Verantwortliche, die an der „neuen" Publikationskette beteiligt sind, erweitert werden muss.

Langzeitarchivierung im digitalen Kontext

„Langzeitarchivierung" meint in diesem Zusammenhang mehr als die Erfüllung gesetzlicher Vorgaben über Zeitspannen, während der steuerlich relevante tabellarisch strukturierte Daten verfügbar gehalten werden müssen. „Langzeit" ist die Umschreibung eines nicht näher fixierten Zeitraumes, währenddessen wesentliche, nicht vorhersehbare technologische und soziokulturelle Veränderungen eintreten; Veränderungen, die sowohl die Gestalt als auch die Nutzungssituation digitaler Ressourcen in rasanten Entwicklungszyklen vollständig umwälzen können. Es gilt also, jeweils geeignete Strategien für bestimmte digitale Sammlungen zu entwickeln, die je nach Bedarf und zukünftigem Nutzungszenarium die langfristige Verfügbarkeit und Nachnutzung der digitalen

1 http://print.google.com
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Objekte sicherstellen. Dabei spielen nach bisheriger Erfahrung das Nutzerinteresse der Auf- und Abwärtskompatibilität alter und neuer Systemumgebungen nur dann eine Rolle, wenn dies dem Anbieter für die Positionierung am Markt erforderlich erscheint. „Langzeit" bedeutet für die Bestandserhaltung digitaler Ressourcen nicht die Abgabe einer Garantieerklärung über fünf oder fünfzig Jahre, sondern die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können.

Der Bedeutungsinhalt von „Archivierung" müsste hier nicht näher präzisiert werden, wäre er nicht im allgemeinen Sprachgebrauch mit der fortschreitenden Anwendung der Informationstechnik seines Sinnes nahezu entleert worden. „Archivieren" bedeutet zumindest für Archive, Museen und Bibliotheken mehr als nur die dauerhafte Speicherung digitaler Informationen auf einem Datenträger. Vielmehr schließt es die Erhaltung der dauerhaften Verfügbarkeit und damit eine Nachnutzung und Interpretierbarkeit der digitalen Ressourcen mit ein.

Substanzerhaltung

Eines von zwei Teilzielen eines Bestandserhaltungskonzeptes für digitale Ressourcen ist die unversehrte und unverfälschte Bewahrung des digitalen Datenstroms: die Substanzerhaltung der Dateninhalte, aus denen digitale Objekte physikalisch bestehen. Erfolgreich ist dieses Teilziel dann, wenn die aus heterogenen Quellen stammenden und auf unterschiedlichsten Trägern vorliegenden Objekte möglichst früh von ihren originalem Träger getrennt und in ein homogenes Speichersystem überführt werden. Die verantwortliche archivierende Institution wird vorzugsweise ein funktional autonomes Teilsystem einrichten, dessen vorrangige Aufgabe die Substanzerhaltung digitaler Ressourcen ist. Wichtige Bestandteile dieses Systems sind (teil-)automatisierte Kontrollmechanismen, die den kontinuierlichen systeminternen Datentransfer überwachen. Die kurze Halbwertszeit technischer Plattformen macht auch vor diesem System nicht halt und zwingt zum laufenden Wechsel von Datenträgergenerationen und der damit möglicherweise verbundenen Migration der Datenbestände.

Dauerhafte Substanzerhaltung ist nicht möglich, wenn die Datensubstanz untrennbar an einen Datenträger und damit an dessen Schicksal gebunden ist. Technische Maßnahmen zum Schutz der Verwertungsrechte (z.B. Kopierschutzverfahren) führen typischerweise mittelfristig solche Konfliktsituationen herbei. Ein digitales Archiv wird in Zukunft im eigenen Interesse Verantwortung nur für solche digitalen Ressourcen übernehmen, deren Datensubstanz es

voraussichtlich erhalten kann. Ein objektspezifischer „Archivierungsstatus“ ist in dieser Situation zur Herstellung von Transparenz hilfreich.

Erhaltung der Benutzbarkeit

Substanzerhaltung ist nur eine der Voraussetzungen, um die Verfügbarkeit und Benutzbarkeit digitaler Ressourcen in Zukunft zu gewährleisten. „Erhaltung der Benutzbarkeit“ digitaler Ressourcen ist eine um ein Vielfaches komplexere Aufgabenstellung als die Erhaltung der Datensubstanz. Folgen wir dem Szenario eines „Depotsystems für digitale Objekte“, in dem Datenströme sicher gespeichert und über die Veränderungen der technischen Umgebung hinweg aufbewahrt werden, so steht der Benutzer/die Benutzerin der Zukunft gleichwohl vor einem Problem. Er oder sie ist ohne weitere Unterstützung nicht in der Lage den archivierten Datenstrom zu interpretieren, da die erforderlichen technischen Nutzungsumgebungen (Betriebssysteme, Anwendungsprogramme) längst nicht mehr verfügbar sind. Zur Lösung dieses Problems werden unterschiedliche Strategien diskutiert, prototypisch implementiert und erprobt.

Der Ansatz, Systemumgebungen in Hard- und Software-Museen zu konservieren und ständig verfügbar zu halten, wird nicht ernsthaft verfolgt. Dagegen ist die Anwendung von Migrationsverfahren bereits für die Substanzerhaltung digitaler Daten erprobt, wenn es um einfachere Datenstrukturen oder den Generationswechsel von Datenträgertypen geht. Komplexe digitale Objekte entziehen sich jedoch der Migrationsstrategie, da der für viele Einzelfälle zu erbringende Aufwand unkalkulierbar ist. Aus diesem Grund wird mit Verfahren experimentiert, deren Ziel es ist, Systemumgebungen lauffähig nachzubilden (Emulation). Es werden mehrere Ansätze verfolgt, unter denen die Anwendung formalisierter Beschreibungen von Objektstrukturen und Präsentationsumgebungen eine besondere Rolle einnimmt.

Die bisher genannten Ansätze spielen durchgängig erst zu einem späten Zeitpunkt eine Rolle, zu dem das digitale Objekt mit seinen für die Belange der Langzeitarchivierung günstigen oder weniger günstigen Eigenschaften bereits fertig gestellt ist. Darüber hinaus wirken einige wichtige Initiativen darauf hin, bereits im Entstehungsprozess digitaler Objekte die Verwendung langzeitstabiler Datenformate und offener Standards zu fördern. Welche der genannten Strategien auch angewandt wird, die Erhaltung der Benutzbarkeit und damit der Interpretierbarkeit wird nicht unbedingt mit der Erhaltung der ursprünglichen Ausprägung des „originalen“ Objektes korrespondieren. Es wird erforderlich sein, die Bemühungen auf die Kernfunktionen (so genannte „significant pro-

perties“) digitaler Objekte zu konzentrieren, vordringlich auf das, was ihren wesentlichen Informationsgehalt ausmacht.

Technische Metadaten

Die Erhebung und die strukturierte Speicherung technischer Metadaten ist eine wichtige Voraussetzung für die automatisierte Verwaltung und Bearbeitung digitaler Objekte im Interesse ihrer Langzeitarchivierung. Zu den hier relevanten Metadaten gehören z.B. Informationen über die zur Benutzung notwendigen Systemvoraussetzungen hinsichtlich Hardware und Software sowie die eindeutige Bezeichnung und Dokumentation der Datenformate, in denen die Ressource vorliegt. Spätestens zum Zeitpunkt der Archivierung sollte jedes digitale Objekt über einen eindeutigen, beständigen Identifikator (persistent identifier) verfügen, der es unabhängig vom Speicherort über Systemgrenzen und Systemwechsel hinweg identifiziert und dauerhaft nachweisbar macht. Tools, die zurzeit weltweit entwickelt werden, können dabei behilflich sein, Formate beim Ingest-Prozess (Importvorgang in ein Archivsystem) zu validieren und mit notwendigen technischen Metadaten anzureichern. Ein viel versprechender Ansatz ist das JHOVE Werkzeug[2], das zum Beispiel Antworten auf folgende Fragen gibt: Welches Format hat mein digitales Objekt? Das digitale Objekt „behauptet“ das Format x zu haben, stimmt dies?

Ohne die Beschreibung eines digitalen Objektes mit technischen Metadaten dürften Strategien zur Langzeitarchivierung wie Migration oder Emulation nahezu unmöglich bzw. deutlich kostenintensiver werden.

Vertrauenswürdige digitale Archive

Digitale Archive stehen erst am Beginn der Entwicklung, während Archive für traditionelles Schriftgut über Jahrhunderte hinweg Vertrauen in den Umfang und die Qualität ihrer Aufgabenwahrnehmung schaffen konnten. Es werden deshalb Anstrengungen unternommen, allgemein akzeptierte Leistungskriterien für vertrauenswürdige digitale Archive aufzustellen (vgl. Kap. 5), die bis zur Entwicklung eines Zertifizierungsverfahrens reichen. Die Konformität zum OAIS-Referenzmodell spielt dabei ebenso eine wichtige Rolle, wie die Beständigkeit der institutionellen Struktur, von der das Archiv betrieben wird. Es wird erwartet, dass Arbeitsmethoden und Leistungen der Öffentlichkeit präsentiert werden, sodass aus dem möglichen Vergleich zwischen inhaltlichem Auftrag

2 JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove/

und tatsächlicher Ausführung eine Vertrauensbasis sowohl aus Nutzersicht als auch im Interesse eines arbeitsteiligen kooperativen Systems entstehen kann.

Wichtig in diesem Zusammenhang ist auch die Wahrung der Integrität und Authentizität eines digitalen Objektes. Nur wenn sichergestellt werden kann, dass das digitale Objekt zum Beispiel inhaltlich nicht verändert wurde, kann man mit der Ressource vertrauensvoll arbeiten.

Verteilte Verantwortung bei der Langzeitarchivierung digitaler Ressourcen

National

Hinsichtlich der Überlegungen zur Langzeitarchivierung digitaler Quellen in Deutschland muss das Ziel sein, eine Kooperationsstruktur zu entwickeln, die entsprechend den Strukturen im analogen Bereich die Bewahrung und Verfügbarkeit aller digitalen Ressourcen gewährleistet. Diese Strukturen müssen alle Ressourcen die in Deutschland, in deutscher Sprache oder über Deutschland erschienen sind berücksichtigen. Die Bewahrung und dauerhafte Verfügbarkeit der wichtigsten Objekte jedes Fachgebiets muss organisiert werden, unabhängig davon, ob es sich um Texte, Fakten, Bilder, Multimedia handelt.
Das Auffinden der Materialien soll dem interessierten Nutzer ohne besondere Detailkenntnisse möglich sein, d.h. ein weiteres Ziel einer angestrebten Kooperationsstruktur beinhaltet, die Verfügbarkeit durch Zugangsportale sicher zu stellen und die Nutzer dorthin zu lenken, wo die Materialien liegen. Dabei müssen selbstverständlich Zugriffsrechte, Kosten u.a. durch entsprechende Mechanismen (z.B. Bezahlsysteme) berücksichtigt werden.

Beim Aufbau einer solchen Struktur sind vor allem die Bibliotheken, Archive und Museen gefordert. In Deutschland müssen in ein entstehendes Kompetenznetzwerk Langzeitarchivierung aber auch die Produzenten digitaler Ressourcen, d.h. Verlage, Universitäten, Forschungseinrichtungen, Wissenschaftler sowie technische Dienstleister wie Rechen-, Daten- und Medienzentren und Großdatenbankbetreiber einbezogen werden.

Internationale Beispiele

Ein Blick ins Ausland bestärkt den kooperativen Ansatz. In Großbritannien ist im Jahr 2001 die Digital Preservation Coalition (DPC) mit dem Ziel initiiert worden, die Herausforderungen der Langzeitarchivierung und -verfügbarkeit digitaler Quellen aufzugreifen und die Langzeitverfügbarkeit des digitalen Erbes in nationaler und internationaler Zusammenarbeit zu sichern. Die DPC versteht sich als ein Forum, welches Informationen über den gegenwärtigen

Forschungsstand sowie Ansätze aus der Praxis digitaler Langzeitarchivierung dokumentiert und weiterverbreitet. Die Teilnahme an der DPC ist über verschiedene Formen der Mitgliedschaft möglich.

In den USA ist im Jahr 2000 ein Programm zum Aufbau einer nationalen digitalen Informationsinfrastruktur und ein Programm für die Langzeitverfügbarkeit digitaler Ressourcen in der Library of Congress (LoC) verabschiedet worden. Die Aufgaben werden in Kooperation mit Vertretern aus anderen Bibliotheken und der Forschung sowie kommerziellen Einrichtungen gelöst. Darüber hinaus hat die LoC in Folge ihrer Jubiläumskonferenz im Jahre 2000 einen Aktionsplan aufgestellt, um Strategien zum Management von Netzpublikationen durch Bibliothekskataloge und Metadatenanwendungen zu entwickeln. Der Ansatz einer koordinierten nationalen Infrastruktur, auch unter den Rahmenbedingungen einer äußerst leistungsfähigen Nationalbibliothek wie der LoC, bestätigt die allgemeine Einschätzung, dass zentralistische Lösungsansätze den künftigen Aufgaben nicht gerecht werden können.

Im Archivbereich wird die Frage der Langzeitverfügbarkeit digitaler Archivalien in internationalen Projekten angegangen. Besonders zu erwähnen ist das Projekt ERPANET, das ebenfalls den Aufbau eines Kompetenznetzwerks mittels einer Kooperationsplattform zum Ziel hatte. InterPares ist ein weiteres internationales Archivprojekt, welches sich mit konkreten Strategien und Verfahren der Langzeitverfügbarkeit digitaler Archivalien befasst. Die Zielsetzungen der Projekte aus dem Archivbereich verdeutlichen, wie ähnlich die Herausforderungen der digitalen Welt für alle Informationsanbieter und Bewahrer des kulturellen Erbes sind und lassen Synergieeffekte erwarten.

Ein umfassender Aufgabenbereich von Museen ist das fotografische Dokumentieren und Verfahren von Referenzbildern für Museumsobjekte. Die Sicherung der Langzeitverfügbarkeit der digitalen Bilder ist eine essentielle Aufgabe aller Museen. Im Bereich des Museumswesens muss der Aufbau von Arbeitsstrukturen, die über einzelne Häuser hinausreichen, jedoch erst noch nachhaltig aufgebaut werden.

Rechtsfragen

Im Zusammenhang mit der Langzeitarchivierung und -verfügbarkeit digitaler Ressourcen sind urheberrechtlich vor allem folgende Fragestellungen relevant:

- Rechte zur Durchführung notwendiger Eingriffe in die Gestalt der elektronischen Ressourcen im Interesse der Langzeiterhaltung,

- Einschränkungen durch Digital Rights Management Systeme (z.B. Kopierschutz),
- Konditionen des Zugriffs auf die archivierten Ressourcen und deren Nutzung.

Die EU-Richtlinie zur Harmonisierung des Urheberrechts in Europa greift diese Fragestellungen alle auf; die Umsetzung in nationales Recht muss aber in vielen Ländern, darunter auch Deutschland, noch erfolgen. Erste Schritte sind in dem „ersten Korb" des neuen deutschen Urheberrechtsgesetzes erfolgt.

Wissenschaftliche Forschungsdaten

Die Langzeitarchivierung wissenschaftlicher Primär- und Forschungsdaten spielt eine immer größere Rolle. Spätestens seit einigen „Manipulations-Skandalen" (zum Beispiel Süd-Korea im Frühjahr 2008) ist klar geworden, dass auch Forschungsdaten langfristig verfügbar gehalten werden müssen. Verschiedene Stimmen aus wissenschaftlichen Disziplinen, sowohl Geistes- als auch Naturwissenschaften, wünschen sich eine dauerhafte Speicherung und einen langfristigen Zugriff auf ihr wissenschaftliches Kapital.

Weiterhin fordern verschiedene Förderer und andere Institutionen im Sinne „guter wissenschaftlicher Praxis" (DFG) dauerhafte Strategien, wie folgende Beispiele zeigen:

- DFG, Empfehlung 7[3]
- OECD[4]
- Und ganz aktuell die EU[5] mit folgendem Zitat:

„Die Europäische Kommission hat am 10. April 2008 die 'Empfehlungen zum Umgang mit geistigem Eigentum bei Wissenstransfertätigkeiten und für einen Praxiskodex für Hochschulen und andere öffentliche Forschungseinrichtungen' herausgegeben. Zu diesem Thema war bereits im ersten Halbjahr 2007 unter der deutschen Ratspräsidentschaft ein Eckpunktepapier mit dem Titel 'Initiative zu einer Charta zum Umgang mit geistigem Eigentum an öffentlichen Forschungseinrichtungen und Hochschulen' ausgearbeitet worden."

3 http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf
4 http://www.oecd.org/dataoecd/9/61/38500813.pdf
5 http://ec.europa.eu/invest-in-research/pdf/ip_recommendation_de.pdf

Es gibt zurzeit in Deutschland konkrete Überlegungen, wie es gelingen kann, gemeinsam mit den Wissenschaftlern eine gute Praxis bezüglich des Umgangs mit Forschungsdaten zu entwickeln. Die beinhaltet auch (aber nicht nur) die Veröffentlichung von Forschungsdaten.

Interessante Fragen in diesem Zusammenhang sind zum Beispiel, wem die Forschungsdaten eigentlich gehören (dem Wissenschaftler, der Hochschule, der Öffentlichkeit), was Forschungsdaten eigentlich sind - hier gibt es bestimmte fachspezifische Unterschiede, welche Forschungsdaten langfristig aufbewahrt werden müssen - eine fachliche Selektion kann nur in enger Kooperation mit dem Wissenschaftler erfolgen, und wer für die Beschreibungen z.B. die Lieferung von technischen und deskriptiven Metadaten zuständig ist.

Im Juni 2008 hat sich die Schwerpunktinitiative „Digitale Information" der Allianz-Partnerorganisationen gegründet[6]. Vertreten in dieser Allianz sind zum Beispiel die Alexander von Humboldt-Stiftung, die Deutsche Forschungsgemeinschaft, die Helmholtz-Gemeinschaft deutscher Forschungszentren, die Hochschulrektorenkonferenz, die Leibnitz-Gemeinschaft, die Max-Planck-Gesellschaft und der Wissenschaftsrat. Ziel ist es „Wissenschaftlerinnen und Wissenschaftler mit der bestmöglichen Informationsinfrastruktur auszustatten, die sie für ihre Forschung brauchen … Im digitalen Zeitalter bedeutet das die digitale und für den Nutzer möglichst entgelt- und barrierefreie Verfügbarkeit von Publikationen, Primärdaten der Forschung und virtuellen Forschungs- und Kommunikationsumgebungen. Es gilt daher eine nachhaltige integrierte digitale Forschungsumgebung zu schaffen, in der jeder Forschende von überall in Deutschland auf das gesamte publizierte Wissen und die relevanten Forschungsprimärdaten zugreifen kann."
Die Allianz Partner haben sich auf folgende Schwerpunktaktivitäten geeinigt:

1. Nationale Lizenzierungen
2. Open Access
3. Nationale Hosting-Strategie
4. Forschungsprimärdaten
5. Virtuelle Forschungsumgebungen
6. Rechtliche Rahmenbedingungen

Die Arbeitsgruppe „Forschungsprimärdaten" hat sich im Oktober 2008 unter dem Vorsitz der Deutschen Forschungsgemeinschaft und Helmholtz-Gemeinschaft deutscher Forschungszentren gegründet und erarbeitet zurzeit ein

6 http://www.dfg.de/aktuelles_presse/das_neueste/download/pm_allianz_digitale_information_details_080612.pdf

Positionspapier „Grundsätze zum Umgang mit Forschungsdaten“ und einen Maßnahmenkatalog. Dabei werden insbesondere die Nachnutzung von Forschungsdaten, die Berücksichtigung der Begebenheiten in den unterschiedlichen Fachdisziplinen, die Verstärkung der wissenschaftlichen Anerkennung bei der Publikation von Forschungsdaten, die Lehre und Qualifizierung in diesem Bereich, die Einhaltung von (fachspezifischen) Standards und die Entwicklung geeigneter Infrastrukturen hervorgehoben.

2 State of the Art

2.1 Einführung

Regine Scheffel

„State of the Art" - unbescheiden und stolz klingt dieser Titel. Und in der Tat haben gerade Bibliotheken, aber auch Archive und Museen wichtige Voraussetzungen geschaffen und Ergebnisse erzielt im Kampf gegen ein drohendes „Dark Age" des Verlusts digitalen Kulturguts. Dennoch bleibt – vor allem bei den Museen – noch viel Entwicklungsbedarf. Dass das Potential dazu vorhanden ist, zeigen die folgenden Aufsätze im Überblick.

Besonders wichtig ist dieses Resümee des Erreichten aus dem Blickwinkel von nestor zu dem Zeitpunkt, zu dem das Kompetenzzentrum Langzeitarchivierung vom Projekt in den Status dauerhafter Regelaufgaben bei den Partnern wechselt.

2.2 LZA-Aktivitäten in Deutschland aus dem Blickwinkel von nestor

Mathias Jehn und Sabine Schrimpf

Die Herausforderung der digitalen Langzeitarchivierung betrifft alle Gedächtnisorganisationen - Bibliotheken, Archive, Museen - und kann effektiv und bezahlbar nur kooperativ bewältigt werden. Aus diesem Gedanken heraus wurde 2003 in Deutschland das Kompetenznetzwerk für digitale Langzeitarchivierung „nestor" mit den Arbeitsschwerpunkten Qualifizierung, Standardisierung, Vernetzung gegründet.

Bibliotheken, Archive und Museen stellen gemeinsam das wissenschaftliche, juristisch-administrative und kulturelle Gedächtnis einer Stadt, eines Landes, einer Nation dar. Neben ihrer Verantwortung für die Erhaltung physisch vorhandener Originale tritt seit einigen Jahren zunehmend die Verantwortung für die langfristige Bewahrung digitaler Informationen. Dies können elektronische Akten, digitale Publikationen, nachträglich angefertigte Digitalisate von anderen Kulturmedien, Informationsdatenbanken oder sonstige digitale Medien sein. Der Gesetzgeber hat den wachsenden Stellenwert digitaler Informationen anerkannt, indem er z.B. im Bibliotheksbereich den Sammelauftrag der Deutschen Nationalbibliothek auf digitale Medien ausgeweitet hat. Im Archivbereich erstreckt sich die Zuständigkeit ohnehin auf alle archivwürdigen Unterlagen, digitale Objekte fallen implizit darunter. Im Museumsbereich gibt es keine gesetzlichen Regelungen, aber auch hier gewinnen digitale Objekte zunehmend an Bedeutung.

Für alle Gedächtnisorganisationen stellt die dauerhafte Bewahrung von Zugänglichkeit und Nutzbarkeit digitaler Ressourcen eine enorme Herausforderung dar: So muss das digital publizierte Wissen auch unter den Bedingungen eines ständig stattfindenden Technologiewandels langfristig nutzbar und verfügbar gehalten werden. Der digitalen Langzeitarchivierung kommt hierbei eine Schlüsselrolle zu. Letztlich stellt sie eine wesentliche Bedingung für die Konkurrenzfähigkeit des Bildungs- und Wissenschaftssystems und damit mittelbar auch für die wirtschaftliche Leistungsfähigkeit eines jeweiligen Landes dar.

Die dauerhafte Lesbarkeit von elektronischen Medien ist insbesondere durch den schnellen technischen Wandel von Datenträgern und -formaten sowie durch die permanente Veränderung und Weiterentwicklung der für die Nutzung notwendigen Anwendungsprogramme gefährdet. Neben technischen Lösungen sind auch organisatorische Anstrengungen nötig – Zuständigkeiten und

Verantwortlichkeiten müssen gegebenenfalls überdacht und neue Absprachen getroffen werden. Dies zieht finanzielle Aufwände nach sich: Sobald einmal mit der Langzeitarchivierung begonnen wird, muss die langfristige Finanzierung gewährleistet sein. Zwar ist heute immer noch unklar, wie sich die Kosten in der Zukunft entwickeln werden, jedoch ist es sicher, dass einerseits große Geldsummen für den Aufbau und Betrieb von Langzeitarchivierungssystemen benötigt werden, andererseits der finanzielle Spielraum für den öffentlich-rechtlichen Bereich begrenzt sein wird. Es sind daher Strategien nötig, wie Gedächtnisorganisationen mit den begrenzten Mitteln die besten Ergebnisse erzielen können.

Auf Grund der komplexen und innovativen Herausforderungen, die mit dem Thema digitale Langzeitarchivierung verbunden sind, werden Langzeitarchivierungsvorhaben meist im Rahmen von Forschungsprojekten, häufig im kooperativen Projektverbund angegangen.

Seit 2004 sind in Deutschland eine Reihe von technischen Archivlösungen für die langfristige Bewahrung digitaler Informationen entwickelt worden, z.B. kopal, BABS, Digitales Archiv u.a. (siehe Kapitel 11 „Speichersysteme"). Neben der Entwicklung kompletter Archivsystem-Lösungen befassen sich zahlreiche Institutionen in unterschiedlichen Projekten mit weiteren Aspekten der digitalen Langzeitarchivierung, deren Themen von Strategiebildung hinsichtlich Langzeitarchivierung bis zur Entwicklung von Langzeitarchivierungswerkzeugen reichen. nestor bündelt alle derartigen Projekte in Deutschland, im deutschsprachigen Raum sowie die mit Beteiligung deutscher Partner auf der nestor-Homepage.[1] Aus dem Gedanken heraus, dass die Aufgabe der digitalen Langzeitarchivierung nur kooperativ zu bewältigen ist, wurde 2003 nestor, das Kompetenznetzwerk für digitale Langzeitarchivierung in Deutschland, gegründet. nestor ist das Akronym der englischen Übersetzung des Projekttitels: „Network of Expertise in long-term STOrage and availability of digital Resources in Germany".[2]

Ein kurzer Blick zurück: In Deutschland wurde die Problematik „digitale Langzeitarchivierung" zum ersten Mal 1995 in einem Positionspapier „Elektronische Publikationen" der Deutschen Forschungsgemeinschaft (DFG) aufgegriffen und als Aufgabenbereich der Virtuellen Fachbibliotheken benannt. In Anbetracht sowohl des Umfangs der Aufgabe als auch der föderalen Struktur Deutschlands mit der Verantwortlichkeit seiner Bundesländer für Wissenschaft und Kultur, war es folgerichtig, dass der Ansatz zu einer erfolgreichen Lösung

1 http://www.langzeitarchivierung.de/modules.php?op=modload&name=PagEd&file=index&page_id=16
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

2 Siehe: http://www.langzeitarchivierung.de

dieser Probleme nur ein kooperativer sein konnte. Aus der gemeinsamen Arbeit an konzeptionellen Fragen der künftigen Entwicklung digitaler Bibliotheken im Rahmen des vom Bundesministeriums für Bildung und Forschung (BMBF) getragenen Projektes „digital library konzepte“ ist eine Initiativgruppe Langzeitarchivierung hervorgegangen, deren Arbeitsplan im Rahmen eines sechsmonatigen Folgeprojekts im Jahre 2002 auf zwei Workshops ausgewählten Experten des Informationswesens zur Diskussion gestellt wurden. Diese „Initialzündung“ für eine kooperative Lösung der Langzeitarchivierung digitaler Ressourcen resultierte in einem Papier mit Abschlussempfehlungen für zentrale Komponenten einer kooperativen digitalen Langzeiterhaltungsstrategie für Deutschland. In den Jahren 2003 bis 2009 förderte das BMBF das Projekt nestor zum Aufbau eines nationalen Kompetenznetzwerks zur Langzeitarchivierung und Langzeitverfügbarkeit digitaler Objekte. Es bündelt die in Deutschland identifizierbaren Kompetenzen und koordiniert Kontakte zu entsprechenden Initiativen und Fachgruppen. Mit der Einrichtung von nestor soll gemeinsam den Herausforderungen der Langzeitarchivierung – unter Einbeziehung der „Produzenten“ digitaler Ressourcen, d.h. Verlage, Universitäten, Forschungseinrichtungen, Behörden, Wissenschaftler sowie technischer Dienstleister wie Rechen-, Daten- und Medienzentren und Großdatenbankbetreiber – begegnet werden. Die gemeinsame Fragestellung betrifft die dauerhafte Erhaltung sowohl genuin digitaler Objekte als auch retrodigitalisierter Ressourcen sowie die nachhaltige Verfügbarkeit dieser Informationen für spätere Generationen.
Arbeitsschwerpunkte von nestor sind:

1. Qualifizierung: In nestor wurde ein großer Aus- und Weiterbildungsbedarf im Bereich des noch neuen Aufgabenfeldes „digitale Langzeitarchivierung“ erkannt und zielgerichtete Qualifizierungsangebote entwickelt. Dazu gehören themen- und communityspezifische Workshops, die jährliche Spring bzw. Summer School und das nestor Handbuch. In Zusammenarbeit mit weiteren Hochschulpartnern und der Archivschule Marburg entwickelt nestor ein Aus- und Fortbildungsangebot sowie konkrete e-Tutorials für den Einsatz in der Lehre (s.a. nestor Handbuch Kap. 19).
2. Standardisierung: Die Verständigung auf Standards im Bereich der digitalen Langzeitarchivierung ist unbedingt erforderlich. Diese sollten in Übereinstimmung mit den sich aktuell im internationalen Rahmen abzeichnenden Standardisierungsinitiativen erarbeitet werden. Zu diesem Zweck kooperiert nestor u.a. mit dem DIN (NABD 15, Arbeitsausschuss „Schriftgutverwaltung und Langzeitverfügbarkeit digitaler Informationsobjekte“ im Normausschuss Bibliotheks- und Dokumentations-

wesen[3]). Die im DIN NABD 15 versammelten Experten erarbeiten aktiv nationale Standards und bringen sich in die Erarbeitung internationaler Standards ein.

3. Vernetzung: nestor bietet ein Forum für die Diskussion über Zuständigkeiten und die Etablierung von effektiven und effizienten Kooperationsstrukturen in Deutschland. Zur Vernetzung der relevanten Akteure und Aktivitäten dienen u.a. die nestor-Informationsdatenbanken, die Arbeitsgruppen, Seminare und Workshops. Ein wichtiges Ergebnis der ersten nestor-Projektphase war die Verabschiedung gemeinsamer Richtlinien: nestor hat in einem „Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland" die notwendigen Anstrengungen von politischen Entscheidungsträgern, Urhebern, Verlegern, Hard- und Softwareherstellern sowie kulturellen und wissenschaftlichen Gedächtnisorganisationen zusammengestellt, um die Rahmenbedingungen einer nationalen Langzeitarchivierungs-Policy abzustecken.[4]

Mittlerweile verteilen sich in nestor die notwendigen Fachkompetenzen für den Aufgabenkomplex „Langzeitarchivierung digitaler Ressourcen" über ein breites Spektrum von Personen, die in vielen Institutionen, Organisationen und Wirtschaftsunternehmen tätig sind. nestor bringt so die Experten der Langzeitarchivierung zusammen und fördert den Austausch von Informationen, die Entwicklung von Standards sowie die Nutzung von Synergieeffekten. Alle Sparten der Gedächtnisinstitutionen werden bei der Herausforderung unterstützt, die Bewahrung und Verfügbarkeit aller digitalen Ressourcen selbst zu gewährleisten, die Bewahrung und Verfügbarkeit der wichtigsten Objekte jedes Fachgebiets zu organisieren sowie schließlich die Bewahrung und Verfügbarkeit digitaler Archivalien garantieren zu können. Auch nach Ende der Projektförderung in 2009 wird nestor von den Partnern im Kooperationsverbund als das Kompetenznetzwerk für digitale Langzeitarchivierung in Deutschland fortgeführt.

---

3 http://www.nabd.din.de/gremien/NA+009-00-15+AA/de/54774796.html

4 Siehe: http://www.langzeitarchivierung.de/downloads/memo2006.pdf

2.3 Bibliotheken

Mathias Jehn und Sabine Schrimpf

Für Bibliotheken gehört der Umgang mit elektronischen Ressourcen zu den größten Herausforderungen des 21. Jahrhunderts. Die Sammlung, Erschließung und dauerhafte Aufbewahrung elektronischer Ressourcen erweitert das Aufgabenfeld von Bibliotheken heutzutage enorm. Auch mit dem Aufbau von Langzeitspeichern müssen Bibliotheken sich auseinandersetzen.

Für die Bibliotheken gehört der Umgang mit elektronischen Ressourcen angesichts der sich gegenwärtig vollziehenden Veränderungen in der Informationsgesellschaft zu den größten Herausforderungen des 21. Jahrhunderts. Zwar ist die jeweilige Sichtweise auf digitale Informationen je nach Bibliothekstyp und -aufgabe traditionell sehr unterschiedlich, jedoch hat in den letzten Jahren ein Prozess intensiven Nachdenkens darüber eingesetzt, welche gemeinsamen Wege beschritten werden müssen, um dem bibliothekarischen Auftrag auch in Zukunft gerecht zu werden. Für die langfristige, zuverlässige Archivierung elektronischer Ressourcen sind mittlerweile unterschiedliche Lösungsansätze vorhanden, aber noch ist nicht die abschließende Lösung für die Herausforderungen der Langzeitarchivierung gefunden. Dazu gehören die Sicherung sowohl der Datenströme als auch des Zugriffs und der Lesbarkeit der in ihnen enthaltenen Informationen und deren dauerhafte Nutzbarkeit, also die Erschließung und Bereitstellung. Alle Bibliotheken sind sich darüber einig, dass unter dem wachsenden Druck betriebswirtschaftlichen Denkens keine Institution allein alle digitalen Ressourcen dauerhaft archivieren kann, sondern dass geeignete nationale Kooperations- und Austauschmodelle greifen müssen. In diesem Kontext stehen die Themenfelder „Netzpublikationen", „Langzeitspeicher" und „nationales Vorgehen" im Zentrum der aktuellen Diskussion:

1. *Erweiterter Sammelauftrag:* Seit Mitte der 1990er Jahre nimmt die Bedeutung originär digitaler Publikationen stetig zu. Zahlreiche Verlage veröffentlichen wissenschaftliche Zeitschriften - besonders im naturwissenschaftlichen Bereich - auch oder ausschließlich in digitaler Form. Die zunehmende Bedeutung von Netzpublikationen erweitert das Aufgabenspektrum der Bibliotheken und befördert die organisatorischen und technischen Anstrengungen zur Sicherung und langfristigen Nutzbarkeit digitaler Objekte. Auf Empfehlung der Kultusministerkonferenz (KMK) wird von den Universitäten seit 1998 zunehmend die Veröffentlichung von Promotions- und Habilitationsarbeiten in digitaler Form

akzeptiert. Pflichtexemplar- und Sondersammelgebietsbibliotheken haben in den vergangenen Jahren Kompetenzen bei der Sammlung und Bearbeitung digitaler Medien aufgebaut. Im Juni 2006 wurde das Gesetz über die Deutsche Nationalbibliothek verabschiedet; seitdem sind elektronische Veröffentlichungen in den Sammelauftrag der Deutschen Nationalbibliothek einbezogen. Nach der Novellierung des Bundesgesetzes wurden die Pflichtexemplargesetze für Bibliotheken in bislang zwei Bundesländern entsprechend ausgeweitet. Für das Sammeln elektronischer Publikationen bietet sich das sogenannte „Drei-Varianten-Vorgehen" an: 1. Direkte Kooperation mit den Ablieferern oder Kooperation mit aggregierenden Partnern wie regionalen Pflichtexemplarbibliotheken oder zentralen Fachbibliotheken hinsichtlich der Sammlung einzeln identifizierbarer Online-Publikationen. 2. Implementierung einer generell nutzbaren Schnittstelle auf der Website für die Ablieferung einzeln identifizierbarer Netzpublikationen in standardisierten Verfahren. 3. Erprobung von Harvesting-Methoden für die Sammlung bzw. den Abruf definierter Domainbereiche.

2. Aufbau eines Langzeitspeichers: Die Sammlung der Netzpublikationen macht den Aufbau gewaltiger Datenspeicher erforderlich. Dies setzt neue Formen der Zusammenarbeit in Deutschland voraus. Allein die bloße Datenspeicherung genügt nicht; große Datenmengen müssen verwaltet werden, um adressierbar und nutzbar zu bleiben. Zudem müssen Prozesse entwickelt werden, die den „Import" neuer Daten in den Datenspeicher regeln. Darüber hinaus muss für die künftige Migration, Emulation oder Konversion der Daten zum Zweck der Langzeitarchivierung Vorsorge getroffen werden. Die Nutzbarkeit sollte gewährleistet sein, auch wenn Hard- und Softwareumgebungen und Benutzungstools technisch veralten und eine weitere Nutzbarkeit der ursprünglichen Form verhindern. All diese Fragen wurden seit 2004 von der Deutschen Nationalbibliothek zusammen mit den Partnern Staats- und Universitätsbibliothek Göttingen, IBM und Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen im Projekt kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen[5]) bearbeitet. Zur dauerhaften Adressierung der Online-Objekte vergibt die Deutsche Nationalbibliothek persistente Identifikatoren in Form eines URN (Uniform Resource Name), der anders als eine Web-URL dauerhaft adressierbar und damit zitierbar bleibt.

5 Siehe: http://kopal.langzeitarchivierung.de

3. *Errichtung eines kooperativen Netzwerks*: Die notwendige Steuerung, Koordination, Forschung und Entwicklung für eine leistungsfähige Langzeitarchivierung fand in Deutschland in der Vergangenheit nur in geringem Umfang statt. Aus diesem Grund hat sich im Jahr 2003 mit dem Projekt nestor (Network of Expertise in long-term Storage and availability of digital Ressources in Germany) erstmals ein nationales Kompetenznetzwerk gebildet, um den immer dringender werdenden Herausforderungen der Langzeitarchivierung gemeinsam zu begegnen.[6]

Eine wesentliche Vorbedingung für die Etablierung einer Archivierungsstruktur für elektronische Ressourcen in Deutschland ist die Stärkung der öffentlichen Bewusstseinsbildung für die Relevanz der Langzeitarchivierung elektronischer Ressourcen. Derzeit kommen die entscheidenden Entwicklungen auf diesem Gebiet vor allem aus dem angloamerikanischen Raum (USA, England, Australien). Um in Zukunft die Anschlussfähigkeit der Archivierungsaktivitäten an diese Entwicklungen zu gewährleisten und diese vor dem Hintergrund der spezifischen bibliothekarischen Bedürfnisse und Gegebenheiten der deutschen Informationslandschaft mitzugestalten, wird eine intensivere Kooperation und eine noch stärkere Partizipation der Bibliotheken an diesen Initiativen notwendig sein.

6 Siehe: http://www.langzeitarchivierung.de

2.4 Archive

Christian Keitel

Die digitale Revolution fordert die klassischen Archive in zwei Bereichen heraus: Zum einen bedürfen die nun digital übernommenen Objekte ständiger Aufmerksamkeit und Pflege. Es genügt nicht mehr sie in einem Regal abzulegen und über Findbücher nachweisbar zu halten. Stattdessen müssen der Lebenslauf (Lifecycle) eines Objekts und mit ihm die Phasen der Bewertung, Übernahme, Aufbereitung, Archivierung und Benutzung erneut überdacht werden. Zum anderen müssen die Archive bereits vor dem Zeitpunkt der Bewertung aktiv werden, um ihren Aufgaben auch künftig nachkommen zu können. Während in den angelsächsischen Ländern die Archive seit jeher auch für die Schriftgutverwaltung der abgebenden Stellen (Behörden, Unternehmen...) zuständig sind, ist die Aufgabe des Recordsmanagements für die deutschen Archive neu.

Recordsmanagement

Grundbücher, Register und Akten werden in immer mehr Behörden elektronisch geführt. Auch die Geschäftsprozesse in Kirchen, Unternehmen und Verbänden werden immer öfter digital abgewickelt. So gut wie immer wird dabei ein neues IT-System in Betrieb genommen. Bereits zu diesem Zeitpunkt sollten die späteren Phasen im Lebenszyklus der Dokumente bedacht werden, die Archive sollten also an der Einführung beteiligt werden, um wenigstens die Anbietung und den Export der im System zu produzierenden Unterlagen zu gewährleisten. Neben der Definition von Schnittstellen ist dabei über geeignete Formate und die Ausgestaltung von Löschroutinen zu sprechen. Vor einer Löschung sollte stets die Anbietung der Unterlagen an das zuständige Archiv erfolgen. Bei einem weitergehenden Anspruch kann das Archiv auch versuchen, in der Behörde auf eine authentische und integre Schriftgutverwaltung hinzuwirken. Als Standards im Bereich der Schriftgutverwaltung können genannt werden: DOMEA (Deutschland), GEVER (Schweiz), ELAK (Österreich), NOARK (Norwegen), MoReq (EU, angelsächsisch geprägt) und die ISO 15489. In Australien soll sich jedes in der Behörde entstehende Dokument über eine spezielle Nummer eindeutig dieser Einrichtung zuweisen lassen (AGLS). Ebenfalls sehr weit ausgearbeitet ist das VERS-Konzept aus der australischen Provinz Victoria. In Deutschland sind in diesem Bereich die im Auftrag der Archivreferentenkonferenz arbeitenden AG „Elektronische Systeme in Justiz und Verwaltung“ (AG ESys) und die Bundeskonferenz der Kommunalarchive beim

Deutschen Städtetag tätig. Die Kolleginnen und Kollegen haben allgemeine Empfehlungen und konkrete Aussonderungskonzepte für einzelne IT-Systeme erarbeitet und sich an der Erarbeitung übergreifender Schnittstellen wie XDOMEA und XJUSTIZ beteiligt.

Bewertung

Seit jeher können Archive nur einen Bruchteil der in den abgebenden Stellen verwahrten Unterlagen übernehmen. Die Auswahl der archivwürdigen digitalen Unterlagen weicht teilweise von der archivischen Bewertung papierener Unterlagen ab. Gemein ist beiden Prozessen der Versuch, vielfältig interpretierbare aussagekräftige Unterlagen zu ermitteln. Dienstreiseanträge sind auch dann nicht archivwürdig, wenn sie in digitaler Form vorliegen. Andererseits ermöglichen digitale Unterlagen neue Formen der Informationssuche und -aggregierung. Es kann daher sinnvoll sein, in manchen Bereichen ganze Datenbanken zu übernehmen, aus denen bisher mangels Auswertbarkeit nur wenige oder keine Papierakten ins Archiv übernommen wurden. Letzten Endes müssen papierene und digitale Unterlagen auf ihre Informationsgehalte und die Benutzungsmöglichkeiten hin gemeinsam bewertet werden. Bei personenbezogenen Unterlagen kann beispielsweise zunächst zwischen den Benutzungszielen (1) Grundinformationen zu jeder Person, (2) statistischer Auswertbarkeit, (3) umfassender Information zu einer „zeittypischen" oder (4) einer „berühmten" Person und (5) Rekonstruktion des Behördenhandelns unterschieden werden.[7] In einem zweiten Schritt kann dann überlegt werden, wie diese Benutzungsziele am besten abgebildet werden können. Für die ersten beiden Benutzungsziele kommen in erster Linie Datenbanken in Frage, während es für die sich anschließenden Benutzungsziele der vollständigen Personalakten bedarf, die jedoch zumeist noch auf Papier geführt werden. Bei zu bewertenden Datenbanken ist wiederum ein Abgleich zwischen den zu erwartenden Informationsmöglichkeiten und dem dafür erforderlichen Erhaltungsaufwand vorzunehmen. Gerade bei sehr umfangreichen Datenbanken kann es nötig sein, nur einige Tabellen auszuwählen. Die Bewertung bezieht sich somit nicht mehr (wie bei Papierakten) auf bereits formierte Einheiten, sie geht darüber hinaus und formiert erst die Einheiten, die für die künftigen Benutzer aufzubewahren sind.

7 Ernst et al. (2008).

Übernahme und Aufbereitung

Abhängig von den bei der Systemeinführung erfolgten Absprachen bekommen die Archive im günstigsten Fall sämtliche Daten in archivfähiger Form angeboten, im schlechtesten müssen sie sich selbst um den Export und die spätere Umwandlung in taugliche Formate sowie deren Beschreibung bemühen. Die meisten Archive setzen auf das Migrationskonzept, benötigen also eine entsprechend aufwändige Aufbereitung der Daten. In der Archivwelt werden drei Migrationszeitpunkte diskutiert:

- Migration unmittelbar nach der Erstellung (z.B. VERS-Konzept),
- Migration nach Ablauf einer Transferfrist (DOMEA-Konzept),
- Migration bei der Übernahme (z.B. australisches Nationalarchiv).

Die Migration der Unterlagen verändert den Bitstream der Dateien und verhindert daher den Einsatz der elektronischen Signatur in den Archiven. Auf der anderen Seite können gerade die Unterlagen, die von Behörden übernommen werden, elektronisch signiert im Archiv ankommen, da sie rechtserheblicher Natur sind. In diesem Fall muss das Archiv die Signatur auf ihre Gültigkeit hin überprüfen und dies entsprechend dokumentieren. Die Glaubwürdigkeit der Dokumente im Archiv wird dann auf anderem Weg erhalten (s. Kapitel Vertrauenswürdigkeit von digitalen Langzeitarchiven).

In zunehmendem Maß stehen für die Aufbereitung kleine Tools zur Verfügung, die v.a. von angelsächsischen Archiven als Open Source Software veröffentlicht werden, z.B. JHOVE (Harvard University), DROID (National Archives, Kew) und XENA (National Archives of Australia). In Deutschland wurden bislang Tools zur Vereinheitlichung der angebotenen Metadaten (StandardArchivierungsModul - SAM, Bundesarchiv), und als Open Source das Tool IngestList zur Dokumentation und Validierung der Übernahmen (Landesarchiv Baden-Württemberg) entwickelt. Für die Webarchivierung liegen vom Archiv für soziale Demokratie und vom Bundestagsarchiv zwei Tools vor. Instrumente zur Auswahl geeigneter Formate haben der Arbeitskreis Elektronische Archivierung des Verbands der Wirtschaftsarchive (AKEA) und die schweizerische Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST) entwickelt.[8] Die KOST wurde 2005 auf der Grundlage einer Strategiestudie der schweizerischen Archive[9] eingerichtet, sie soll kooperative Antworten auf die digitalen Herausforderungen finden.

---

8 Gutzmann (2007) und Katalog archivischer Dateiformate (KaD).

9 Schärli et al. (2002).

Archivierung

Ende des letzten Jahrhunderts wurde im angelsächsischen Raum das Konzept der *postcustodial option* diskutiert. Danach sollten die datenerzeugenden Stellen die Unterlagen bei festgestellter Archivwürdigkeit unbefristet selbst aufbewahren. Den Archiven würde dann die Aufgabe der Bewertung und die Kontrolle über die Speicherung und Zugänglichkeit der Daten zufallen. Dieses Konzept wird seit einigen Jahren nicht mehr diskutiert, mit dem australischen Nationalarchiv hat sich 2000 auch ein ehemaliger Fürsprecher wieder der klassischen Übernahme und Archivierung zugewandt. Die deutschen Archive diskutieren neben der Eigenarchivierung auch die Möglichkeit, die Daten physisch in einem Rechenzentrum abzulegen (z.B. Landesarchiv Niedersachsen). Inzwischen wird in dieser Diskussion zunehmend zwischen der physischen Speicherung (Bitstream-Preservation) und darauf aufbauend dem Erhalt der Informationen in einem für die langzeitige Archivierung geeigneten Repository unterschieden. Das Bundesarchiv hat bei der Wiedervereinigung zahlreiche Altdaten der DDR übernommen und baut derzeit ein Digitales Archiv auf.[10] Das Landesarchiv Baden-Württemberg hat mit dem Digitalen Magazin DIMAG ebenfalls ein Repository entwickelt.[11] Beide Systeme beruhen zwar auf einer Datenbank, sie speichern jedoch die Archivierungspakete außerhalb von ihr ab. Eine Rekonstruktion der Inhalte kann somit auch ohne die jeweilige Repository-Software erfolgen. Parallel dazu wurden entsprechende Metadatenkonzepte entwickelt.[12] Neben der Speicherung müssen die digitalen Unterlagen auch in ein zu entwickelndes Verhältnis mit den herkömmlichen papierenen Archivalien gesetzt werden, zumal auf absehbare Zeit viele Unterlagen weder rein digitaler noch ausschließlich analoger sondern hybrider Natur sein werden. Das Landesarchiv Baden-Württemberg hat hierzu ein an PREMIS angelehntes Repräsentationenmodell entwickelt.

Benutzung

Archive bergen im Regelfall Unikate, die nicht ersetzt und daher nur im Lesesaal benutzt werden können. Nachdem digitale Archivalien weder den Begriff des Originals noch eine Bindung an einen Träger kennen, können diese Ar-

10 Huth (2008).

11 Keitel et al. (2007).

12 Vgl. das XBARCH-Konzept des Bundesarchivs, s. Huth (2008); für das Landesarchiv Baden-Württemberg s. Metadaten für die Archivierung digitaler Unterlagen; zum Repräsentationenmodell s.a. Keitel et al. (2007).

chivalien auch in einem geschützten Intranet oder im Internet benutzt werden. Benutzungsmöglichkeiten über das Internet bieten derzeit die National Archives, Kew (National Digital Archive of Datasets, NDAD) und die NARA, Washington an (Access to Archival Databases, AAD).[13] Das dänische Reichsarchiv hat Ende 2008 das Tool Sofia vorgestellt, das eine Benutzung im Archiv selbst ermöglicht.[14]

Zusammenfassend sind die deutschen Archive im Bereich des Recordsmanagements gut aufgestellt. Zentrale Fragen der elektronischen Archivierung werden seit 1997 im Arbeitskreis „Archivierung von Unterlagen aus digitalen Systemen" diskutiert. Eine statistische Auswertung der gehaltenen Vorträge zeigt, dass hier zunächst die Systemeinführung im Mittelpunkt stand. Seit 2006 wurden immer mehr Berichte über die sich anschließenden Phasen im Lebenszyklus gehalten. Dennoch kann in den Bereichen der Übernahme, Archivierung und Benutzung auch noch 2009 ein Vorsprung der angelsächsischen Archive und hier insbesondere der Nationalarchive konstatiert werden.[15]

Quellen und Literatur

Literatur

Albrecht Ernst et al. (2008): *Überlieferungsbildung bei personenbezogenen Unterlagen*, in: Archivar 2008 (61), S. 275 - 278.

Gutzmann, Ulrike et al. (2007): *Praktische Lösungsansätze zur Archivierung digitaler Unterlagen: „Langzeitarchivierung" und dauerhafte Sicherung der digitalen Überlieferung*. In: Archivar 2007 (60), S. 322-329. Auch in: Archiv und Wirtschaft 40/1 (2007), S. 20-27 und http://www.wirtschaftsarchive.de/akea/handreichung.htm

Huth, Karsten (2008): *PREMIS in the Federal Archives Germany*. In: http://www.loc.gov/standards/premis/premis_tut_Berlin-final.ppt

Keitel, Christian/Lang, Rolf/Naumann, Kai (2007): *Konzeption und Aufbau eines Digitalen Archivs: von der Skizze zum Prototypen*. In: Ernst, Katharina (Hg.): Erfahrungen mit der Übernahme digitaler Daten. Elfte Tagung des AK „Archivierung von Unterlagen aus digitalen Systemen" vom 20./21. März 2007, Stuttgart, S. 36-41.

13 NDAD: http://www.ndad.nationalarchives.gov.uk/; AAD: http://aad.archives.gov/aad/

14 Yoneyama (2008).

15 Keitel (2009).

Keitel, Christian (2009): *Elektronische Archivierung in Deutschland. Eine Bestandsaufnahme*. Erscheint in: Für die Zukunft sichern! Bestandserhaltung analoger und digitaler Unterlagen, 78. Deutscher Archivtag 2008 in Erfurt, Tagungsdokumentation zum Deutschen Archivtag Bd. 13, Fulda 2009.

Schärli, Thomas et al. (2002): *Gesamtschweizerische Strategie zur dauerhaften Archivierung von Unterlagen aus elektronischen Systemen*. In: http://www.vsa-aas.org/index.php?id=110&L=0

Yoneyama, Jun Petersen (2008): *Creating access to electronic records: Two approaches*. In: http://www.dlm2008.com/img/pdf/yoneyama_ab_gb.pdf

Quellen

AAD, Access to Archival Databases (o.J.): http://aad.archives.gov/aad/

AGLS, Australian Government Locator Service (o.J.): http://www.naa.gov.au/records-management/create-capture-describe/describe/agls/index.aspx

DOMEA, Dokumenten-Management und elektronische Archivierung 2.1 (2005): http://www.verwaltung-innovativ.de/cln_047/nn_684678/DE/Organisation/domea__konzept/domea__konzept__node.html?__nnn=true

DROID, Digital Record Object Identification (o.J.): http://droid.sourceforge.net/wiki/index.php/Introduction

ELAK (o.J.): http://www.digitales.oesterreich.gv.at/site/5286/default.aspx

GEVER (o.J.): http://www.isb.admin.ch/themen/architektur/00078/index.html?lang=de

IngestList : ab Mai 2009 unter http://ingestlist.sourgeforge.net .

ISO 15489 (2001): http://www.landesarchiv-bw.de/sixcms/media.php/25/ISO_DIN_15489.pdf, http://www.whitefoot-forward.com/iso_15489-2.pdf

JHOVE, JSTOR/Harvard Object Validation Environment (o.J.): http://hul.harvard.edu/jhove/

KaD, Katalog archivischer Dateiformate, (o.J.): http://www.kost-ceco.ch/wiki/whelp/KaD/index.html

Metadaten für die Archivierung digitaler Unterlagen (2008): http://www.landesarchiv-bw.de/sixcms/media.php/25/konzeption_metadaten10.pdf

MoReq, Model Requirements Specification for the Management of Electronic Records 2.0 (2008): http://www.moreq2.eu/

NDAD, National Digital Archive of Datasets (o.J.): http://www.ndad.nationalarchives.gov.uk/

NOARK, Norsk arkivsystem (o.J.): http://riksarkivet.no/arkivverket/lover/elarkiv/noark-4.html

SAM, StandardArchivierungsModul (o.J.): http://www.bundesarchiv.de/service/behoerdenberatung/01435/index.html

VERS, Victorian Electronic Records Strategy (o.J.): http://www.prov.vic.gov.au/vers/vers/default.htm

XBARCH: s.o. Huth (2008).

XDOMEA (o.J.): http://www.koopa.de/produkte/xdomea2.html
XENA (o.J.): http://xena.sourceforge.net/

XJUSTIZ (o.J.): http://www.xjustiz.de/

2.5 Museum

Winfried Bergmeyer

Langzeitarchivierung digitaler Daten bedeutet für Museen eine neue Aufgabe, die auf Grund der heterogenen Sammlungsbestände und der vielfältigen Aktivitäten in den Bereichen Sammeln, Erhalten, Forschen und Vermitteln ein breites Spektrum an Anforderungen an die Institutionen stellt. Als Teil des kulturellen Erbes sind die Museen in der Verantwortung, nicht nur ihre digitalen Sammlungsobjekte bzw. Digitalisate sondern auch ihre Forschungs- und Vermittlungstätigkeiten zu bewahren.

Im Jahre 2007 gab es 6.197 Museen und museale Sammlungen in Deutschland.[16] Die Spannbreite der Sammlungskonzepte umfasst Werke der bildenden Kunst, historische und volkskundliche Objekte, technische Denkmäler bis hin zu Spezialsammlungen. Diese Vielfalt spiegelt sich auch in den Arbeitsaufgaben der Institution wieder: Sammeln, Bewahren, Forschen und Vermitteln als zentrale Aufgaben der Museen[17] erfordern zahlreiche, stark ausdifferenzierte Aktivitäten. Für diese Zwecke erzeugen Museen unterschiedlichste Arten von Informationen und dies zunehmend in digitaler Form.

Im Folgenden soll ein kurzer Überblick die Szenarien in Museen, in denen digitale Daten produziert und bewahrt werden, vorstellen. Nicht alle Museen decken das komplette Spektrum ab, aber es zeigt die mögliche Bandbreite der in diesem Rahmen entstehenden digitalen Objekte.

Digitalisate von Sammlungsgegenständen

Die Digitalisierung von Sammlungsgegenständen wird mit unterschiedlichsten Zielsetzungen durchgeführt. In der Regel wird Flachware, wie Zeichnungen, Bilder oder Drucke, zum Zweck der Publikation digitalisiert, zunehmend werden aber auch Tondokumente digitalisiert und dreidimensionale Kopien erzeugt. Diese digitalen Abbilder finden in Publikationen wie Internetauftritten, interaktiven Applikationen im Rahmen von Ausstellungen o. a. Verwendung. Hierbei sollte aus konservatorischen Gründen die Belastung für den Sammlungsgegenstand durch das Digitalisierungsverfahren so gering wie möglich

16 Staatliche Museen zu Berlin – Preußischer Kulturbesitz, Institut für Museumsforschung (Hrsg.): Statistische Gesamterhebung an den Museen der Bundesrepublik Deutschland für das Jahr 2006, Materialien aus dem Institut für Museumskunde, Heft 62, Berlin 2008.

17 Siehe dazu die ICOM Statuten: http://icom.museum/statutes.html#2

gehalten werden. Viele Objekte sind lichtempfindlich oder aus anderen Gründen fragil und dürfen der Belastung durch Fotografieren oder Scannen nur in Ausnahmefällen ausgesetzt werden. Aus diesem Grund sollten die Aspekte der Langzeitarchivierung in diesen Fällen bereits vor der Digitalisierungsmaßnahme eingebracht werden.[18] So ist die Scanauflösung so hoch wie möglich anzusetzen, der Farbraum und das Dateiformat auszuwählen und der Vorgang entsprechend von Dokumentationsrichtlinien festzuhalten, um die notwendigen Metadaten, die im Rahmen der digitalen Langzeitarchivierung notwendig werden, verfügbar zu haben.

Die Restaurierung bildet in vielen größeren Museen einen eigenen Bereich, dessen Aufgabe die Sicherung des Erhaltes der musealen Objekte ist. Die neuen Medien bieten Restauratoren und Wissenschaftlern zahlreiche Möglichkeiten ihre Arbeit effizienter zu gestalten. Neben den digitalen Restaurierungsberichten bildet die Technik der virtuellen Rekonstruktion eine Methode, fragmentarisch erhaltene museale Objekte ohne Beeinträchtigung des realen Objektes zu ergänzen. Durch die Nutzung virtueller Abbilder als Ersatz beispielsweise für die Vorauswahl von Objekten im Zuge einer Ausstellungsvorbereitung kann die mechanische und klimatische Belastung der Originale reduziert und somit deren Erhaltung gesichert werden. Objekte aus fragilen Materialien unterliegen oft einem nur hinauszuzögernden Verfallsprozess, so dass hochauflösende digitale Scans hier eine konservatorische Alternative in der Nutzung der Objekte (beispielsweise für Druckgraphiken in Kupferstichkabinetten) bieten. Digitalisate ersetzen natürlich nicht die realen Objekte, können aber im Falle des Verlustes zumindest umfangreiche Informationen enthalten und Visualisierungen bereitstellen. Diese Aufgabe können sie allerdings nur bei entsprechender Langzeitarchivierung erfüllen.

Digitale Sammlungsobjekte

Spätestens seit der Entwicklung der Video-Kunst ist eine Abhängigkeit zwischen Kunstwerken und elektronischen Medien gegeben. Die Nutzung digitaler Medien in der Kunst stellt die Museen vor neue Herausforderungen. Hierbei geht es nicht allein um die Konservierung von Datenströmen, sondern auch von komplexen Installationen mit entsprechender Hardware. Die künstlerische Wirkung von Video-Installationen wird häufig durch die spezifische Wiederga-

18 Im Rahmen des Minerva-Projektes sind hierzu Handreichungen entstanden. Siehe dazu das G*ood Practice Handbuch für Digitalisierungsprojekte.* In: http://www.minervaeurope.org/publications/gphandbuch.htm

be-Hardware bestimmt.[19] Projekte wie z.B. das Erl King-Projekt von Grahame Weinbren und Roberta Friedman[20] aus den Jahren 1982-1985 basieren mit ihrer eigens programmierten Software auf speziellen Computern und Peripheriegeräten. Die Langzeitarchivierung digitaler Medienkunst ist eine Aufgabe, die auf Grund ihrer Komplexität zahlreiche unterschiedliche Konzepte hervorgebracht hat. Der Ansatz, den Künstler/die Künstlerin in den Prozess der Erstellung von digitalen Archivierungskonzepten einzubinden, ist dabei richtungsweisend. In Absprache mit ihm/ihr sollte geklärt werden, wie das Verhältnis zwischen physischer Präsentationsumgebung (Hardware, Software) und inhaltlichem Konzept zu gewichten ist. Erst danach kann entschieden werden, welche Archivierungskonzepte gewählt werden sollten. Die statische Konservierung beinhaltet die Aufbewahrung (und Pflege) von Hard- und Software, also des kompletten Systems und ist die aufwändigste, technisch komplexeste und sicherlich nicht für alle Institutionen realisierbare Methode. Die Migration der Daten von einem Dateiformat in ein anderes (aktuelles) oder die Emulation von Hard- und Software-Umgebungen sind alternative Konzepte zur Langzeitarchivierung.[21] Unabhängig von der gewählten Methode ist aber die Forderung nach Archivierung von Metainformationen, die zu diesem Kunstwerk, seiner Entstehung, seiner Rezeptionen und Provenienz in Beziehung stehen, zu berücksichtigen und die entsprechenden Metadaten sind zu erfassen.[22]

Sammlungsdokumentation

Zu den originären Aufgaben eines Museums gehört das Sammlungsmanagement, das neben der wissenschaftlichen Inventarisation auch zahlreiche administrative Bereiche umfasst. Die digitale Inventarisation hat seit den 1990er Jahren

---

19 Hanhardt, John G.: *Nam June Paik, TV Garden, 1974*, in: Depocas, Alain/Ippolito, Jon/ Jones, Caitlin (Hrsg.)(2003): The Variable Media Approach - permanence through change, New York, S. 70 – 77.

20 Rothenberg, Jeff/Grahame Weinbren/Roberta Friedman, *The Erl King, 1982–85*, in: Depocas, Alain; Ippolito, Jon; Jones, Caitlin (Hrsg.)(2003): *The Variable Media Approach - Permanence through Change.* New York, S. 101 – 107. Ders.: *Renewing The Erl King*, January 2006. In: http://bampfa.berkeley.edu/about/ErlKingReport.pdf

21 Rothenberg, Jeff: *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* In: http://www.clir.org/PUBS/reports/rothenberg/contents.html (15.02.2009). Er fordert die Einbindung digitaler Informationen in die Emulatoren, die es ermöglichen, originäre Abspielumgebungen zu rekonstruieren. Leider ist dieser Vorschlag bislang noch nicht umgesetzt worden.

22 Rinehart, Richard: *The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Centura.* In: http://switch.sjsu.edu/web/v6n1/article_a.htm

Einzug in große und mittlere Institutionen gehalten und ist integraler Bestandteil der täglichen Museumsarbeit geworden.[23] Sie bildet eine wesentliche Voraussetzung für die Nutzung von Sammlungen und deren Objekten. Zur Bewahrung des Wissens über die musealen Objekte ist die Erhaltung der Metadaten und ihrer Struktur notwendig. Um hier eine Langzeitverfügbarkeit zu gewährleisten sind Standards im Bereich der Ontologien, Thesauri oder kontrollierte Vokabularien unabdingbar. Als bekanntestes Metadaten-Schema gilt *Dublin Core,*[24] das von den meisten Anbietern unterstützt wird. Mit dem Datenaustauschformat *museumdat,*[25] basierend auf dem von J. Paul Getty Trust zusammen mit ARTstor entwickelten *CDWA Lite*[26] sowie dem *CIDOC-CRM,*[27] gibt es weitere Ansätze zur Standardisierung bzw. zum Austausch von komplexeren Metadaten. Die zahlreichen unterschiedlichen terminologischen Ressourcen zur Erschließung bedürfen ebenso einer Standardisierung, um sammlungsübergreifendes Retrieval zu erlauben. Eine Vielzahl von Software-Herstellern bietet Lösungen für kleine bis große Institutionen an. Schon 1998 wurde ein Software-Vergleich zur Museumsdokumentation erarbeitet. Das Thema der Langzeitarchivierung war hier allerdings noch nicht Bestandteil der überprüften Kriterien.[28] Die wichtigsten Anbieter sind mittlerweile in der Lage Schnittstellen für Metadaten nach DC und museumdat sowie Web-Services für Vokabulare zu nutzen[29].

Präsentationen

Museen sind Orte des offenen Zugangs zur kulturellen, technologischen oder politischen Geschichte und Gegenwart. Sie vermitteln der interessierten Öffentlichkeit wissenschaftliche Informationen und verwenden dabei zunehmend die Möglichkeiten der neuen Medien. In diesem Bereich erfreut sich moderne

23 Im Jahr 2000 haben nach eigenen Angaben drei Viertel aller an einer Befragung teilnehmenden deutschen Museen digitale Sammlungsdaten. Witthaut, Dirk unter Mitarbeit von Zierer, Andrea/Dettmers, Arno/Rohde-Enslin, Stefan (2004): *Digitalisierung und Erhalt von Digitalisaten in deutschen Museen*, nestor-Materialien 2, Berlin, S. 25.

24 http://dublincore.org/

25 Nähere Informationen zu museumdat unter: http://www.museumsbund.de/cms/index.php?id=603

26 http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html

27 http://cidoc.ics.forth.gr/; http://www.museumsbund.de/cms/fileadmin/fg_doku/publikationen/CIDOC_CRM-Datenaustausch.pdf

28 1998 wurde ein Vergleich zahlreicher Museumsdokumentations-Software von Deutschen Museumsbund durchgeführt: http://www.museumsbund.de/cms/index.php?id=261&L=0&STIL=0

29 Ein erster Ansatz ist dabei die Bereitstellung unterschiedlicher Vokabularien, wie dies z.B. im Projekt *museumvok* erfolgt. http://museum.zib.de/museumsvokabular/

Informationstechnologie in Form von Terminalanwendungen, Internet-Auftritten und elektronischen Publikationen zunehmend größerer Beliebtheit.[30] In diesem Rahmen werden Anwendungen genutzt, die sich unterschiedlicher und zum Teil kombinierter Medientypen (Audio, Video, Animationen etc.) bedienen.

Dem Wunsch nach Bereitstellung von Sammlungsinformationen für eine breite Nutzerschicht nachkommend, entstehen zurzeit Portale, die dies auf nationaler[31] und europäischer Ebene[32] ermöglichen werden. Die Informationsrecherche über diese Portale erfolgt durch Web-Harvesting, dessen Voraussetzung die Existenz von museumseigenen Internetpräsenzen mit recherchierbaren Sammlungsbeständen ist. Sollen diese Informationen dauerhaft verfügbar sein, müssen auch die Digitalisate, Metadaten und die Applikation selbst nutzbar gehalten werden.

Forschungsdaten

Neben der Bewahrung und Vermittlung ist die Forschung in Museen ein weiteres Tätigkeitsfeld, in dem digitale Daten produziert werden. Die Ergebnisse werden in Form von Datenbanken, elektronischen Publikationen aber auch als virtuelle Rekonstruktionen oder Simulationen präsentiert. Sie bilden mittlerweile einen Teil unseres wissenschaftlich-kulturellen Erbes und sind somit langfristig zu bewahren.

Museen als Teil des kulturellen Erbes

Es stellt sich natürlich die Frage, ob und in welcher Form alle diese oben angeführten digitalen Daten langfristig bewahrt werden müssen. Museen sammeln und bewahren Zeugnisse unserer Kultur. Diese Objekte werden entsprechend einer gesellschaftlichen Übereinkunft als solche definiert. Aber auch die Institution Museum selbst ist Teil unseres kulturellen Erbes und dies nicht nur auf

30 Hünnekens, Annette (2002): *Expanded Museum. Kulturelle Erinnerung und virtuelle Realitäten*. Bielefeld.

31 Zur im Aufbau befindlichen Deutschen Digitalen Bibliothek siehe: http://www.bundesregierung.de/Webs/Breg/DE/Bundesregierung/BeauftragterfuerKulturundMedien/Medienpolitik/DeutscheDigitaleBibliothek/deutsche-digitale-bibliothek.html
Das BAM-Portal ist bereits seit einigen Jahren online. http://www.bam-portal.de/

32 Am 20. November 2008 ging eine erste Version der europäischen Bibliothek (Europeana) online. http://www.europeana.eu/portal/

Grund ihrer Sammlungen, sondern auch auf Grund der Sammlungskonzepte, der Ausstellungen, der Vermittlung und der Forschung.

Es ist üblich Ausstellungen zu dokumentieren oder Forschungsergebnisse zu archivieren und somit den Umgang mit den Informationen und Objekten zu erhalten. Bislang geschah dies überwiegend in analoger Form in Berichten und dem Erhalt von Ausstellungskatalogen. Interessenten konnten zu einem späteren Zeitpunkt mit Hilfe dieser Dokumente die Ausstellung und deren Inhalt rekonstruieren. Im digitalen Zeitalter erfolgt dies mittels Textverarbeitungsprogrammen, digitaler Fotografie oder digitalen Videoaufzeichnungen. Als Bestandteil aktueller Ausstellungen werden z.B. Terminalanwendungen häufig nach deren Ende nicht weiter gepflegt und damit der Möglichkeit einer weiteren oder späteren Nutzung entzogen. Als Teil der Vermittlungsgeschichte oder in Form einer Nachnutzung in anderen Bereichen sollten auch sie unter Beachtung von festgelegten Auswahlkriterien bewahrt werden. Die Komplexität und Vielfältigkeit der verwendeten Medien (Fotos, Audiodaten, Videos) dieser Anwendungen erfordert dabei jeweils individuelle Konzepte. Vergleichbar der digitalen Kunst ist besonderer Wert auf eine umfangreiche Dokumentation zu legen, in der Programmierungsrichtlinien, Hardware-Anforderungen, Installationsvorgaben und Bedienungsanleitungen gesichert werden.

Konzepte zur Langzeitarchivierung digitaler Daten

Museen sehen sich also mit einer Reihe unterschiedlicher Medien- und Objekttypen im Rahmen der Bewahrung digitaler Daten konfrontiert. Dies trifft sowohl auf kleine als auch große Institutionen zu. Die Komplexität der in den Museen anfallenden digitalen Daten erfordert von den Institutionen ein jeweils individuell für ihre Sammlung definiertes Konzept zur Langzeitarchivierung. Allein durch die unterschiedlichen Institutionsgrößen - von ehrenamtlich betreuten Museen bis hin zu großen Häusern – ist die Vorstellung eines universell anwendbaren Konzepts zur Langzeitarchivierung undenkbar. Personelle, finanzielle und technische Ressourcen sind in den Institutionen in unterschiedlichem Umfang vorhanden. Darüber hinaus sind die digitalen Bestände, die zu erhalten sind, sehr verschieden. Sinnvoll wären hier skalierbare Konzepte, die auf Basis bestehender Standards und Empfehlungen den unterschiedlichen Ressourcenpools der Institutionen gerecht werden.

In Anlehnung an das Konzept des Canadian Heritage Information Network[33] sind die notwendigen Maßnahmen für die Erhaltung digitaler Objekte in Museen in zwei Teile aufzugliedern. Der erste Teil beschreibt die von den einzelnen Institutionen durchzuführenden Maßnahmen, der zweite Teil diejenigen, die nur durch übergeordnete Institutionen oder Kooperationen umzusetzen sind.

Maßnahmen in den Museen

Erstellung eines institutionellen Konzeptes

Auf Basis des Leitbildes ist zu definieren, welche Aufgaben der Langzeitarchivierung digitaler Daten die Institution im Rahmen der Erhaltung des kulturellen Erbes zu übernehmen hat. Dazu gehören neben der Beachtung des Sammlungskonzeptes auch die Bereiche Forschung und Vermittlung.

Bestandsaufnahme des vorhandenen digitalen Materials

Zu den grundlegenden Vorarbeiten für ein Konzept gehört die Bestandsaufnahme der digitalen Daten, der vorhanden Medientypen, der Speichermedien und Dateiformate.

Auswahl der Dateiformate und Speichermedien

Um eine effektive Langzeitarchivierung gewährleisten zu können, sollten so wenige unterschiedliche Dateiformate und Speichermedien im Rahmen des Archivierungsprozesses Verwendung finden wie möglich. Dies vereinfacht einerseits die Kontrolle der Obsoleszens, andererseits den Aufwand für das Refreshing (Kopieren der Daten auf neue Speichermedien).

Klärung der Rechtesituation

Es ist in jedem Einzelfall darauf zu achten, dass das Museum im Besitz der notwendigen Rechte für das Kopieren oder Migrieren der digitalen Daten sowie deren spätere Nutzung ist.

Wahl eines Metadatenstandards

Für Erhaltung und Nutzung der Daten ist es von elementarer Bedeutung, dass die technischen Informationen (z.B. Dateiformat- und –version, Digitalisierungsvorgaben oder verwendete Programme) sowie die inhaltlichen und admi-

33 Yeung, Tim Au (2004): *Digital Preservation: Best Practice for Museums*. In: http://www.chin.gc.ca/English/Pdf/Digital_Content/Digital_Preservation/digital_preservation.pdf

nistrativen Informationen erhalten bleiben. Die Wahl eines solchen Standards bedeutet gleichzeitig die Festlegung der Informationen, die unbedingt für eine Aufnahme in den Erhaltungsprozess notwendig sind.

Erstellung von Auswahlkriterien

Auf Basis dieser Informationen kann ein Kriterienkatalog für die Auswahl der Daten erstellt werden, die aktiv erhalten werden sollen. Dies erfordert ein Umdenken im Umgang mit Objekten und Informationen, weil nicht in den Prozess der Langzeiterhaltung aufgenommene digitale Daten auf Grund der Obsoleszens von Speichermedien und -technologien sowie durch veraltete Datenformate mittelfristig nicht mehr nutzbar sein werden. Nutzbare Dachbodenfunde wird es nicht mehr geben.[34] Dieser Kriterienkatalog ist zudem für die Anforderungen bei der Erstellung neuer digitaler Daten im Hause, aber auch für die Beauftragung externer Produzenten maßgeblich.

Auswahl des Personals und der Zuständigkeiten

Für die effektive und zuverlässige Durchführung des Prozesses der Langzeitarchivierung digitaler Daten ist es notwendig, das Personal für die einzelnen Aufgaben und Zuständigkeitsbereiche zu bestimmen und zu qualifizieren. Der komplexe Workflow bedarf nicht nur entsprechender Handlungsanweisungen sondern auch Verantwortlichkeiten.

Maßnahmen durch Kooperationen oder übergreifend arbeitende Institutionen

Technology Watch

Um Obsolenszensen bei Speichertechnologien, Dateiformaten oder auch Metadatenschemata vorzubeugen ist die permanente Beobachtung aktueller Entwicklungen notwendig. Entsprechende Warnungen sind dann an die einzelnen Museen weiterzuleiten.

34 Dazu N. Beagrie: „A digital resource which is not selected for active preservation treatment at an early stage will very likely be lost or unuseable in the near future". Jones, Maggie/ Beagrie, Niels (2002): *Preservation Management of Digital Materials: A Handbook*. In: http://www.dpconline.org/graphics/handbook/

Aufbau eines Netzwerkes zum Austausch und zur Abstimmung von Konzepten

Die Langzeitarchivierung digitaler Daten in Museen ist als neues Aufgabenfeld vom Austausch von Erfahrungen unter den Institutionen abhängig. Nur so können gemeinsame Konzepte und Kooperationsmöglichkeiten umgesetzt und Standards entwickelt werden.

Interessenvertretung

Die Stärkung des Bewußtseins für die Notwendigkeit des Erhaltes digitaler Daten innerhalb der Museumscommunity ist der erste Schritt, dem die Interessenvertretung für Belange der Langzeitarchivierung digitaler Daten auf politischer Ebene folgen muss. Dies ist nicht zuletzt angesichts der anfallenden Kosten dringend geboten.

Ausblick

Die Langzeitarchivierung digitaler Daten in Museen ist ein Prozess, dessen Durchführung sowohl zusätzliche technische, finanzielle und personelle Anforderungen als auch intellektuelle Herausforderungen beinhaltet. Die Museen in all ihrer Heterogenität bedürfen dazu zusätzlicher Resourcen. Es müssen die finanziellen Mittel bereit gestellt werden, um die notwendigen Investitionen zu tätigen. Zugleich müssen Arbeitsprozesse im Informationsmanagement der Museen effizienter gestaltet werden. Hierzu ist ein entsprechendes Bewußtsein in den Museen selbst, aber auch in den sie finanzierenden Institutionen zu wecken.

Zudem ist aber auch eine stärkere Einbindung der neuen Informationstechnologien in die universitäre Lehre und Ausbildung unabdingbar.[35] Dabei sollten weniger die technischen Grundlagen als vielmehr der intellektuelle Umgang mit diesen Medien in der wissenschaftlichen Forschung und bei der Vermittlung musealer Inhalte im Vordergrund stehen. In Zeiten, in denen das Web 2.0 unsere Kommunikation und die Produktion von kulturellen Zeugnissen revolutioniert, muss auch die Institution Museum auf die Veränderungen reagieren.

35 Diese Forderung wurde u.a. von T. Nagel bereits 2002 erhoben. Nagel, Tobias: *Umbruch oder Abbruch? – Beobachtungen zur Situation der EDV-gestützten Dokumentation in den Museen*, in: zeitenblicke 2 (2003), Nr. 1. http://www.zeitenblicke.historicum.net/2003/01/nagel/index.html (15.02.2009)

Literatur

Staatliche Museen zu Berlin – Preußischer Kulturbesitz, Institut für Museumsforschung (Hrsg.) (2008): *Statistische Gesamterhebung an den Museen der Bundesrepublik Deutschland für das Jahr 2005*, Materialien aus dem Institut für Museumskunde, Heft 62, Berlin 2008

Hünnekens, Annette (2002): *Expanded Museum. Kulturelle Erinnerung und virtuelle Realitäten*. Bielefeld.

Jones, Maggie/Beagrie, Niels (2002): *Preservation Management of Digital Materials: A Handbook*. In: http://www.dpconline.org/graphics/handbook/

Depocas, Alain/Ippolito, Jon/Jones, Caitlin (Hrsg.) (2003): *The Variable Media Approach - permanence through change*. New York.

Rinehart, Richard: *The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Century*, http://switch.sjsu.edu/web/v6n1/article_a.htm

Rothenberg, Jeff: *Avoiding Tecnological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. In: http://www.clir.org/PUBS/reports/rothenberg/contents.html

Witthaut, Dirk unter Mitarbeit von Zierer, Andrea/Dettmers, Arno/Rohde-Enslin, Stefan (2004): *Digitalisierung und Erhalt von Digitalisaten in deutschen Museen*, nestor-Materialien 2. Berlin.

Yeung, Tim Au (2004): *Digital Preservation: Best Practice for Museums*. In: http://www.chin.gc.ca/English/Pdf/Digital_Content/Digital_Preservation/digital_preservation.pdf

3 Rahmenbedingungen für die LZA digitaler Objekte

3.1 Einführung

Stefan Strathmann

Die Langzeitarchivierung digitaler Objekte bedarf umfangreicher und verbindlicher Regelungen, die eine geordnete und dauerhafte Bereitstellung des digitalen Kulturerbes ermöglichen.

Diese Regelungen werden mit dem Begriff Policy zusammengefasst; dieser englische Begriff entspricht in diesem Zusammenhang etwa den deutschen Begriffen „Rahmenbedingungen", „Grundsätze", „Richtlinien". Bei einer Preservation Policy handelt es sich um den Plan zur Bestandserhaltung. Im Gegensatz zu einer Strategie, die festlegt, wie die Erhaltung erfolgen soll, wird von der Policy festgelegt, was und warum etwas für wie lange erhalten werden soll[1].

---

1 Vgl.: Foot (2001), S. 1
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Die Preservation Policy ist also notwendige Grundlage für jede Preservation Strategie.

Diese Richtlinien sind nicht zeitlich befristet, sondern auf dauerhaften Bestand angelegt. D.h. sie sind, anders als beispielsweise Strategien zur Erhaltung digitaler Objekte, nicht an technischen Innovationszyklen oder politischen Veränderungen bzw. institutionellen Führungswechseln orientiert, sondern sollten langfristig Geltung haben.

Preservation Policies werden üblicherweise anhand ihres Geltungsbereiches unterschieden. Am geläufigsten sind nationale oder institutionelle Preservation Policies. Aber auch internationale Policies werden entwickelt und können maßgeblich zur Erarbeitung und Umsetzung nationaler Policies beitragen. Ein herausragendes Beispiel für eine internationale Policy ist die „Charta zur Bewahrung des digitalen Kulturerbes“[2], die am 17. Oktober 2003 auf der 32. Generalkonferenz der UNESCO verabschiedet wurde.

2 UNESCO (2003)

3.2 Nationale Preservation Policy

Stefan Strathmann

Eine nationale Preservation Policy bestimmt den Rahmen für die Bemühungen eines Staates zur Sicherung der digitalen kulturellen und wissenschaftlichen Überlieferung.

Eine solche Policy muss nicht in geschlossener Form eines Dokumentes vorliegen, vielmehr wird sie sich im Normalfall aus einer Vielzahl von Gesetzen, Bestimmungen, Vereinbarungen, Regeln etc. konstituieren.

Eine nationale Preservation Policy kann Regelungen zu sehr unterschiedlichen Fragen der digitalen Langzeitarchivierung umfassen; so finden sich typischerweise Aussagen zu verschiedenen Themenkomplexen:

- Generelles Bekenntnis, das digitale Erbe zu sichern
  Ausgangspunkt einer jeden Preservation Policy ist die verbindliche Aussage, digitale Objekte langfristig zu erhalten. Ein Staat, der den Langzeiterhalt digitaler Objekte als Aufgabe von nationaler Bedeutung erkannt hat, sollte diesem Interesse Ausdruck verleihen und so die daraus resultierenden Aktivitäten begründen und unterstützen.

- Verfügbarkeit und Zugriff
  Da die digitale Langzeitarchivierung kein Selbstzweck, sondern immer auf eine spätere Nutzung/Verfügbarkeit ausgerichtet ist, sollte dieser Bereich in einer nationalen Policy maßgeblich berücksichtigt werden. Die Rahmenbedingungen sollen eine spätere Nutzung ermöglichen.

- Rechtliche Rahmenbedingungen
  Die digitale Langzeitarchivierung ist in vielerlei Hinsicht von Rechtsfragen tangiert. Dies sollte seinen Niederschlag in allen relevanten Bereichen der Gesetzgebung finden. Hierzu gehören beispielsweise die Archivgesetze, Urheber- und Verwertungsrechte, Persönlichkeitsrechte etc.

- Finanzierung
  Eng verknüpft mit den rechtlichen Rahmenbedingungen sind auch die Fragen der Finanzierung digitaler Langzeitarchivierung. Hierzu gehört die langfristige Bereitstellung der Mittel, um die Langzeitarchivierung im gewünschten Umfang durchzuführen.

- Verantwortlichkeiten und Zuständigkeiten
  Bestandteil einer nationalen Preservation Policy sind auch Festlegungen bezüglich der Verantwortlichkeiten und Zuständigkeiten. In Deutschland beispielsweise sind die Zuständigkeiten von Bund, Ländern und Gemeinden zu berücksichtigen. Vorstellbar sind auch Aussagen zur Verantwortlichkeit für bestimmte Objekttypen (Webseiten, Archivgut, Wissenschaftliche Rohdaten, Doktorarbeiten) oder fachliche Inhalte (Wissenschaftliche Literatur bestimmter Fächer).

- Auswahlkriterien
  Es sollte festgelegt sein, welche digitalen Objekte bewahrt werden sollen. Hierbei sollte das ganze Spektrum digitaler Objekte berücksichtigt werden. Da der komplette Erhalt aller digitalen Objekte kaum sinnvoll und machbar ist, sind insbesondere transparente Entscheidungs- und Auswahlkriterien von großer Wichtigkeit.

- Sicherheit
  Der Anspruch an die Sicherheit (Integrität, Authentizität, Redundanz etc.) der digitalen Bestandserhaltung sollte in einer nationalen Policy berücksichtigt werden.

In vielen Staaten finden Diskussionen zur Entwicklung nationaler Policies statt. Da zur Entwicklung einer tragfähigen nationalen Policy ein breiter gesellschaftlicher, politischer und fachlicher Konsens notwendig ist, ist die Entwicklung ein sehr langwieriger und komplizierter Prozess, der bisher nur wenig greifbare Ergebnisse aufweisen kann. Ein Beispiel für eine niedergelegte generelle nationale Preservation Policy findet sich in Australien[3]. Ein weiteres Beispiel für einen Teil einer nationalen Preservation Policy ist das „Gesetz über die Deutsche Nationalbibliothek“[4] vom 22. Juni 2006, in dem der Sammelauftrag der DNB auf Medienwerke in unkörperlicher Form (d.h. u.a. Webseiten) ausgedehnt wird. Mit der Pflichtablieferungsverordnung[5] und bspw. dem Beschluß der Kultusministerkonferenz zur Abgabe amtlicher Veröffentlichungen an Bibliotheken[6] wurden die Grundsätze der digitalen Langzeitarchivierung in Deutschland weiter präzisiert. Diese Gesetze und Verordnungen sind selbstverständlich nicht die deutsche nationale Preservation Policy, es sind aber wichtige Bausteine

3 AMOL (1995)
4 DNBG (2006)
5 Pflichtablieferungsverordnung (2008)
6 KMK (2006)

zur Definition der Rahmenbedingungen der digitalen Langzeitarchivierung in Deutschland.

In Deutschland bemüht sich insbesondere nestor um die Entwicklung einer nationalen Preservation Policy. Zu diesem Zweck wurden von nestor mehrere Veranstaltungen (mit)organisiert, eine Expertise in Auftrag gegeben[7], eine Befragung zu den Auswahlkriterien und Sammelrichtlinien durchgeführt, sowie ein „Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland“[8] veröffentlicht, das sehr breit mit der Fachcommunity abgestimmt ist.

7 Hilf, Severiens (2006)

8 nestor (2006a)

3.3 Institutionelle Preservation Policy

Stefan Strathmann

Rahmenbedingungen und Grundsätze für die digitale Langzeitarchivierung müssen gemäß ihrer Dringlichkeit formuliert werden. Hierbei ist nicht nur der (inter)nationale, sondern auch der lokale und institutionsspezifische Rahmen zu berücksichtigen.

Jede mit dem Erhalt des digitalen wissenschaftlichen und kulturellen Erbe betraute Institution sollte die eigenen Grundsätze in einer institutionellen Preservation Policy festlegen. Diese Policy entspricht häufig einer Selbstverpflichtung, auch wenn weite Teile bspw. durch gesetzliche Anforderungen vorgegeben sind.

Eine solche Policy ist für die jeweiligen Institutionen dringend notwendig, um nach innen das Bewusstsein für die Aufgaben und Belange der digitalen Langzeitarchivierung zu schaffen und nach außen die für Vertrauenswürdigkeit notwendige Transparenz zu gewährleisten[9].

Da innerhalb einer einzelnen Institution die Abstimmungs- und Konsensfindungsprozesse häufig einfacher sind als auf nationaler oder internationaler Ebene, gibt es eine Reihe von Beispielen von institutionellen Preservation Policies[10]. Dennoch ist es bisher nicht der Regelfall, dass Gedächtnisorganisationen eine eigene Policy zum Erhalt ihrer digitalen Bestände formulieren.

Institutionelle Policies können sehr viel spezifischer an die Bedürfnisse der jeweiligen Institutionen angepasst werden, als das bei einer eher generalisierenden nationalen Policy der Fall ist. Aber auch hier ist zu bedenken, dass es sich um Leitlinien handelt, die nicht regelmäßig an das Alltagsgeschäft angepasst werden sollten, sondern dass sich vielmehr das Alltagsgeschäft an den in der Policy festgelegten Linien orientieren sollte.

Die institutionelle Preservation Policy bestimmt den Rahmen für die institutionelle Strategie zum Erhalt der digitalen Objekte. Sie sollte konkret am Zweck und Sammelauftrag der Institution ausgerichtet sein. Hierzu gehören sowohl der Sammlungsaufbau wie auch die Bedürfnisse der jeweiligen intendierten Nutzergruppen. Eine wissenschaftliche Bibliothek bspw. muss ihren Nutzern eine andere Sammlung und anderen Zugang zu dieser Sammlung zur Verfügung stellen als ein Stadtarchiv oder ein Museum.

---

9 Vgl.: nestor (2006b)

10 Vgl. bspw.: NAC (2001), OCLC (2006), PRO (2000), UKDA (2005)

Die in den Rahmenbedingungen spezifizierten Prinzipien des Sammlungsaufbaues sollten ggf. durch Hinweise auf Kooperationen und/oder Aufgabenteilungen ergänzt werden.

Ein weiterer zentraler Bestandteil der Rahmenbedingungen für die Erhaltung digitaler Objekte innerhalb einer Institution ist die Sicherstellung der finanziellen und personellen Ressourcen für den beabsichtigten Zeitraum der Langzeitarchivierung. Eine einmalige Anschubfinanzierung ist nicht ausreichend.

Da Institutionen häufig nur eine begrenzte Zeit ihren Aufgaben nachkommen, sollte eine institutionelle Policy auch auf die Eventualitäten einer Institutionsschließung o.ä. eingehen (Fallback-Strategie, Weitergabe der archivierten Objekte an andere Institutionen).

Nutzungsszenarien sind gleichfalls wichtige Bestandteile einer institutionellen Preservation Policy. Abhängig vom Zweck der Institution sollte eine generelle Festlegung erfolgen, was wem unter welchen Bedingungen und in welcher Form zur Nutzung überlassen wird.

Fragen der Sicherheit der Daten können ebenfalls in einer institutionellen Policy geregelt werden. Dies erfolgt häufig in Form von eigens hierzu erstellten Richtlinien-Dokumenten, die Bestandteil der institutionellen Policy sind (Richtlinien zum Datenschutz, zur Netzwerksicherheit, zur Computersicherheit, zum Katastrophenschutz etc.). Auch sollte der für die Zwecke der Institution benötigte Grad an Integrität und Authentizität der digitalen Objekte festgelegt werden. In diesem Zusammenhang kann auch das Maß der akzeptablen Informationsverluste, wie sie z.B. bei der Migration entstehen können, beschrieben werden.

In einigen institutionellen Preservation Policies[11] werden sehr detailliert die Dienste der Institution festgelegt und die Strategien zur Erhaltung der digitalen Objekte spezifiziert (Emulation, Migration, Storage-Technologie etc.). Dies bedeutet, dass diese Policies relativ häufig einer Revision unterzogen und den aktuellen technischen Anforderungen und Möglichkeiten angepasst werden müssen.

11 Vgl. bspw: OCLC 2006

Literatur

AMOL (1995): National Conservation and Preservation Policy. http://www.nla.gov.au/preserve/natpol.html

KMK (2006): Bericht zur Abgabe amtlicher Veröffentlichungen an Bibliotheken http://amtsdruckschriften.staatsbibliothek-berlin.de/downloads/abgabe_veroeffentl_an_bibliotheken060317.pdf

DNBG (2006): Gesetz über die Deutsche Nationalbibliothek (DNBG) http://217.160.60.235/BGBL/bgbl1f/bgbl106s1338.pdf

Foot (2001): Building Blocks for a Preservation Policy. http://www.bl.uk/npo/pdf/blocks.pdf

Hilf, Severiens (2006): Zur Entwicklung eines Beschreibungsprofils für eine nationale Langzeit-Archivierungs-Strategie - ein Beitrag aus der Sicht der Wissenschaften. http://nbn-resolving.de/urn:nbn:de:0008-20051114021

NAC (2001): National Archives of Canada: Preservation Policy http://www.collectionscanada.ca/preservation/1304/docs/preservationpolicy_e.pdf

nestor (2006a): Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland http://www.langzeitarchivierung.de/downloads/memo2006.pdf

nestor (2006b): Kriterienkatalog vertrauenswürdige digitale Langzeitarchive http://nbn-resolving.de/urn:nbn:de:0008-2006060710

OCLC (2006): OCLC Digital Archive Preservation Policy and Supporting Documentation http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf

Pflichtablieferungsverordnung (2008): Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek http://www.bgblportal.de/BGBL/bgbl1f/bgbl108s2013.pdf

PRO (2000): Public Record Office: Corporate policy on electronic records http://www.nationalarchives.gov.uk/documents/rm_corp_pol.pdf

UKDA (2005): UK Data Archive: Preservation Policy http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf

UNESCO (2003): Charta zur Bewahrung des digitalen Kulturerbes. http://www.unesco.de/444.html (Inoffizielle deutsche Arbeitsübersetzung der UNESCO-Kommissionen Deutschlands, Luxemburgs, Österreichs und der Schweiz)

Weitere Literatur findet sich u.a. im PADI Subject Gateway (http://www.nla.gov.au/padi/), in der nestor Informationsdatenbank (http://nestor.sub.uni-goettingen.de/nestor_on/index.php) und in der ERPANET Bibliography on Digital Preservation Policies (http://www.erpanet.org/assessments/ERPANETbibliography_Policies.pdf)

3.4 Verantwortlichkeiten

Natascha Schumann

Dieser Beitrag behandelt die verschiedenen Ebenen der Verantwortlichkeiten im Bereich der digitalen Langzeitarchivierung. Unterschiedliche Einrichtungen sind mit der Herausforderung konfrontiert, digitale Objekte verschiedenster Art langfristig zu erhalten und ihre Nutzbarkeit zu gewährleisten. Dabei kommen diverse Ausgangsbedingungen zum Tragen, neben gesetzlichen Regelungen können dies spezielle Sammelaufträge oder Absprachen sein. Eine übergreifende Verteilung von Zuständigkeiten auf nationaler Ebene fehlt bislang allerdings.

Die Herausforderungen der digitalen Langzeitarchivierung betreffen in großem Maße Einrichtungen, deren (gesetzlicher) Auftrag in der Erhaltung des kulturellen Erbes besteht. Archive, Museen und Bibliotheken müssen sicherstellen, dass die digitalen Objekte in Zukunft noch vorhanden und auch nutzbar sind. Innerhalb der Communities gibt es generelle Unterschiede bezüglich der Art und des Auftrags der zugehörigen Institutionen.

Bibliotheken

Im Bereich der wissenschaftlichen Bibliotheken unterscheidet man zwischen der National- und Landesbibliothek, Universitäts- und Hochschulbibliothek, Fach- und Spezialbibliothek. Die Aufgaben ergeben sich unter anderem aus dem Pflichtexemplarrecht, z.B. aus dem Gesetz über die Deutsche Nationalbibliothek (DNBG):[12] Letzteres bezieht elektronische Publikationen explizit mit in den Sammelauftrag ein. Auf Länderebene ist dies bislang nur teilweise der Fall.

Als zentrale Archivbibliothek und nationalbibliografisches Zentrum hat die Deutsche Nationalbibliothek die Aufgabe, sämtliche Werke, die in deutscher Sprache oder in Deutschland erschienen sind, zu sammeln, zu erschließen und langfristig verfügbar zu halten. Das Gesetz über die Deutsche Nationalbibliothek, im Jahr 2006 in Kraft getreten, erweiterte den Sammelauftrag explizit auf elektronische Publikationen. Die Pflichtablieferungsverordnung[13] von 2008 regelt die Einzelheiten. Somit ist für den Bereich der durch den Sammelauftrag der DNB abgedeckten elektronischen Publikationen die Langzeitarchivierung rechtlich festgeschrieben[14] (s. a. nestor Handbuch Kap. 2.3 und 18.4).

12 http://217.160.60.235/BGBL/bgbl1f/bgbl106s1338.pdf
13 http://www.bgblportal.de/BGBL/bgbl1f/bgbl108s2013.pdf
14 DNBG, §2, Absatz 1

Regional- und Landesbibliotheken haben den Auftrag, regionales Schrifttum zu sammeln und zu archivieren und erstellen eine entsprechende Bibliografie. Der Sammelauftrag ist durch das Pflichtexemplarrecht geregelt. Das bedeutet, ein Exemplar eines veröffentlichten Werkes muss an die zuständige Bibliothek abgeliefert werden. Der Sammelauftrag folgt dem geografischen Bezug. Bislang beziehen sich die entsprechenden Landesregelungen in erster Linie noch auf Printmedien und werden nur teilweise auch auf elektronische Publikationen angewendet. Im Moment (Februar 2009) gibt es nur in Thüringen und in Baden-Württemberg ein Pflichtexemplarrecht, welches explizit die Ablieferung von digitalen Pflichtexemplaren regelt.

Die Universitätsbibliotheken haben in erster Linie die Aufgabe, die Angehörigen der Hochschule mit der notwendigen Literatur zu versorgen. Hier gilt kein Pflichtexemplarrecht und die Auswahl richtet sich nach den thematischen Schwerpunkten der Einrichtung. Einige Universitätsbibliotheken, aber auch andere Bibliotheken sind gleichzeitig Sondersammelgebietsbibliotheken (SSG). Das bedeutet, zu einem bestimmten Schwerpunkt werden möglichst umfassend alle Publikationen in der entsprechenden Bibliothek gesammelt. Die Schwerpunkte sind in Absprache verteilt auf verschiedene Einrichtungen. Die SSG sind Teil des Programms zur überregionalen Literaturversorgung der Deutschen Forschungsgemeinschaft, mit deren Hilfe eine verteilte Infrastruktur hergestellt werden soll, die allen Wissenschaftlern den dauerhaften Zugriff auf diese Objekte sicherstellt.

Seit 2004 finanziert die DFG den Erwerb von Nationallizenzen.[15] Zur Gewährleistung der überregionalen wissenschaftlichen Literaturversorgung wurden die bei einzelnen Bibliotheken angesiedelten Sondersammelgebiete im Rahmen der Nationallizenzen auf elektronische Publikationen erweitert. Das bedeutet, dass der Zugang zu Online-Datenbanken gefördert wird. Die Lizenzen sind auf zeitlich unbefristete Nutzung ausgerichtet und beinhalten daher auch das Recht, technische Maßnahmen zur dauerhaften Erhaltung vorzunehmen. Der Zugang ist zunächst über die technische Infrastruktur des Lizenzgebers gesichert, ein physischer Datenträger wird dem Lizenznehmer ausgehändigt, wie es auf der Webseite der Nationallizenzen dazu heißt.

Fach- und Spezialbibliotheken sind Bibliotheken mit besonderem inhaltlichem Fokus, die in der Regel zu einer größeren Einrichtung gehören. Dabei kann es sich ebenso um wissenschaftliche Einrichtungen wie auch um Unternehmen handeln.

---

15 http://www.nationallizenzen.de/

Die zu archivierenden elektronischen Objekte im Bibliotheksbereich sind sehr heterogen, sowohl im Hinblick auf die (ggf. vorhandenen) physischen Datenträger als auch auf Dateiformate. Neben gängigen Textformaten wie beispielsweise PDF bzw. PDF/A werden, je nach Auftrag, auch interaktive Anwendungen, Musiktracks u.a. gesammelt.

Neben den eigentlichen Publikationen werden zunehmend auch die zugrunde liegenden Forschungsdaten als archivierungsrelevant betrachtet. Ein Zusammenschluss aus Wissenschaft, Förderern und Bibliotheken[16] beschäftigt sich mit den Fragen der Langzeitarchivierung und des Zugriffs auf die Daten sowie mit der Frage, welche Stakeholder welche Aufgabe übernehmen sollen.

Archive

Im Bereich der Archive existiert ebenfalls eine Aufgabenverteilung. Ein Archiv ist in der Regel für die historische Überlieferung einer Organisation zuständig, z.B. ein staatliches Archiv für ein Bundesland, ein kirchliches Archiv für eine Kirche, ein Unternehmensarchiv für eine konkrete Firma usw. Diese Zuständigkeit erstreckt sich zumeist auf alle Unterlagen, so der einschlägige Begriff aus den Archivgesetzen (s. u.), die in der abgebenden Stelle im Zuge der Geschäftserfüllung entstanden sind. Beispiele sind Akten, Datenbanken, Bilder, Filme etc. Nach Ablauf der Aufbewahrungsfristen müssen diese Unterlagen dem zuständigen Archiv angeboten werden. Dieses entscheidet dann über den historischen Wert, also darüber, was für künftige Generationen übernommen und archiviert werden soll.

Das Bundesarchiv mit dem Hauptsitz in Koblenz hat die Aufgabe, die Dokumente der obersten Bundesbehörden auszuwählen, zu erschließen und zu archivieren. Gesetzliche Grundlage dafür bildet das Bundesarchivgesetz[17]. Als Kriterium für die Archivierung gilt die Annahme, dass die ausgewählten Dokumente von „bleibendem Wert für die Erforschung oder das Verständnis der deutschen Geschichte, die Sicherung der berechtigten Belange der Bürger oder die Bereitstellung von Informationen für Gesetzgebung, Verwaltung oder Rechtsprechung“[18] sind.

Staats- und Landesarchive sind, wie der Name schon zeigt, staatliche Archive mit der Aufgabe, die relevanten Dokumente ihres Bundeslandes zu archivieren. Die Archivstruktur ist in den einzelnen Bundesländern unterschiedlich geregelt

16 http://oa.helmholtz.de/index.php?id=215

17 http://www.bundesarchiv.de/benutzung/rechtsgrundlagen/bundesarchivgesetz/index.html

18 ebd. , § 3

und es gelten die Archivgesetze des jeweiligen Landes. Wie bei allen anderen Archiven auch können die Akten aus den Behörden und Gerichten nur in Auswahl übernommen werden.

Weitere Archivarten sind zum Beispiel Kommunalarchive, Wirtschaftsarchive, Kirchenarchive, Film- oder Literaturarchive etc. Je nach ihrer Ausrichtung ist auch der jeweilige Aufgabenbereich ausgerichtet.

Die Archivierung elektronischer Akten bedeutet eine besondere Herausforderung und bedarf zusätzlicher Maßnahmen zur dauerhaften Sicherung. Während der Aufbewahrung in der Behörde bietet u.a. die elektronische Signatur eine Voraussetzung zur Gleichstellung mit herkömmlichen Papierdokumenten. Nach der Übernahme ins Archiv wird die Signatur geprüft und dokumentiert. Für die Archivierung selbst werden elektronische Signaturen nicht fortgeführt. Hier gelten andere Mechanismen zur Aufrechterhaltung der Glaubwürdigkeit der digitalen Dokumente (s. a. nestor Handbuch Kap. 2.4).

Museen

Auch im Museumsbereich gibt es unterschiedliche Formen von Einrichtungen mit verschiedenen Schwerpunkten und Aufgaben. Es bestehen sowohl zahlreiche Museen mit einem thematischen Schwerpunkt als auch mit regionalem Bezug.

Im Gegensatz zu Archiven und Bibliotheken ist die Bezeichnung Museum aber nicht geschützt und es gibt keine gesetzlichen Regelungen in Bezug auf die Aufgaben und den Auftrag eines bestimmten Museums. Viele Museen bestehen in der Rechtsform einer Stiftung.

Der Internationale Museumsrat ICOM[19] definiert ein Museum als „eine gemeinnützige, ständige, der Öffentlichkeit zugängliche Einrichtung im Dienst der Gesellschaft und ihrer Entwicklung, die zu Studien-, Bildungs- und Unterhaltungszwecken materielle Zeugnisse von Menschen und ihrer Umwelt beschafft, bewahrt, erforscht, bekannt macht und ausstellt". Im Jahr 2006 hat der Deutsche Museumsbund[20] gemeinsam mit ICOM „Standards für Museen"[21] vorgelegt, die zur Definition und Orientierung in der Museumslandschaft dienen sollen.

Auch im Museumsbereich stellt sich mehr und mehr die Frage nach der Erhaltung digitaler Objekte. Diese können recht unterschiedlicher Natur sein,

19 http://www.icom-deutschland.de/

20 http://www.museumsbund.de/cms/index.php

21 http://www.museumsbund.de/cms/fileadmin/geschaefts/dokumente/varia/Standards_fuer_Museen_2006.pdf

zum Beispiel als originär digital erstelltes Objekt oder als digitale Reproduktion oder auch in Form einer digitalen Zusatz- und Kontextinformation (s. a. nestor Handbuch Kap. 2.5).

Fazit

Die Verantwortlichkeiten für die Bewahrung unseres kulturellen Erbes sind für den nicht-digitalen Bereich zumindest teilweise geregelt. Dies hängt unter anderem vom Vorhandensein gesetzlicher Aufträge und Vorhaben ab. Erst in den letzten Jahren gerät die langfristige Verfügbarhaltung digitaler Objekte mehr in den Fokus. Diesbezügliche Regelungen sind in manchen Bereichen bereits vorhanden, z.B. durch das Gesetz über die Deutsche Nationalbibliothek.

Es bestehen bereits einige Kooperationsprojekte im Bereich der digitalen Langzeitarchivierung, diese beziehen sich aber in der Regel weniger auf die Aufteilung verschiedener Zuständigkeiten, sondern auf die gemeinsame Nutzung von Ressourcen. Als beispielhaftes Projekt sei hier auf das Baden-Württembergische Online-Archiv BOA[22] verwiesen, in dem verschiedene Partner die Sammlung, Erschließung und Langzeitarchivierung betreiben.

Verantwortlichkeiten können sich auf verschiedene Bereiche beziehen, z.B. auf die inhaltliche Auswahl der zu sammelnden digitalen Objekte oder auf verschiedene Arten von Objekten. Es kann auch überlegt werden, einzelne Arbeitsschritte bei der Langzeitarchivierung zu verteilen.

Es wäre wünschenswert, wenn eine überregionale Verteilung der Verantwortlichkeiten in Bezug auf die digitale Langzeitarchivierung weiter voranschreiten würde. Eine direkte Übertragung von herkömmlichen Regelungen auf die digitale Langzeitarchivierung erscheint nicht immer sinnvoll, da mit elektronischen Publikationen nicht nur andere Herausforderungen bestehen, sondern sich auch neue Chancen einer verteilten Infrastruktur bieten, wenn ein Objekt nicht länger an ein physisches Medium gebunden ist. Hier bedarf es über Absprachen hinaus entsprechender gesetzlicher Regelungen, z.B. in Form der Ausweitung von Landesgesetzen auf elektronische Publikationen. Dazu bedarf es aber auch der Einigung auf Standards im Bezug auf Schnittstellen und Formate.

22 http://www.boa-bw.de/

3.5 Auswahlkriterien

Andrea Hänger, Karsten Huth und Heidrun Wiesenmüller

Vertrauenswürdige digitale Langzeitarchive müssen für die Auswahl ihrer digitalen Objekte Kriterien entwickeln. Definierte und offen gelegte Auswahlkriterien unterstützen die praktische Arbeit, machen Nutzern, Produzenten und Trägern das Profil des Langzeitarchivs deutlich und sind eine wichtige Voraussetzung für den Aufbau kooperativer Netzwerke zur Langzeitarchivierung. Die Auswahlkriterien sind i.d.R. aus dem Gesamtauftrag der Institution abzuleiten. Als Ausgangspunkt dienen häufig bereits vorhandene Kriterien für analoge Objekte, die jedoch aufgrund der Besonderheiten digitaler Objekte überprüft und ggf. abgeändert werden müssen. Zu unterscheiden sind inhaltlich-fachliche Auswahlkriterien (z.B. die verwaltungstechnische, institutionelle oder räumliche Zuständigkeit) und formal-technische Auswahlkriterien, die die Lesbarkeit des Objekts im Archiv sichern sollen (z.B. das Vorliegen der Objekte in geeigneten Formaten). Spezifische Hinweise werden für den Bereich der Netzpublikationen gegeben, die eine für Bibliotheken besonders wichtige Gattung digitaler Objekte darstellen.

Allgemeines

Die Auswahl digitaler Objekte geschieht auf der Basis von definierten und auf die jeweilige Institution zugeschnittenen Kriterien – beispielsweise in Form von Sammelrichtlinien, Selektions- und Bewertungskriterien oder Kriterien für die Überlieferungsbildung. Im Bibliotheks- und Museumsbereich spricht man i.d.R. von Sammlungen, die aus den Sammelaktivitäten hervorgehen, im Archivbereich dagegen von Beständen, die das Resultat archivischer Bewertung darstellen. Der Begriff der Sammlung wird nur im Bereich des nicht-staatlichen Archivguts verwendet.

Bei digitalen Langzeitarchiven, die von öffentlichen Institutionen betrieben werden, sind die Auswahlkriterien i.d.R. aus dem Gesamtauftrag der Institution abzuleiten. In einigen Fällen gibt es auch gesetzliche Grundlagen – z.B. in den Archivgesetzen, die u.a. auch die formalen Zuständigkeiten staatlicher Archive regeln, oder den nationalen und regionalen Pflichtexemplargesetzen, welche Ablieferungspflichten an bestimmte Bibliotheken festlegen.

Festgelegte, dokumentierte und offen gelegte Auswahlkriterien sind in mehrfacher Hinsicht von zentraler Bedeutung für digitale Langzeitarchive: Als praktische Arbeitsanweisung für das eigene Personal unterstützen sie einen stringenten, von individuellen Vorlieben oder Abneigungen unabhängigen Aufbau der digitalen Sammlung bzw. der digitalen Bestände. Den Nutzern, aber auch den Produzenten bzw. Lieferanten der digitalen Objekte und der allgemeinen Öffentlichkeit machen sie das Profil der digitalen Sammlung bzw. der digitalen Bestände deutlich. Anhand der veröffentlichten Auswahlkriterien können beispielsweise Nutzer entscheiden, ob ein bestimmtes digitales Langzeitarchiv für ihre Zwecke die richtige Anlaufstelle ist oder nicht. Dasselbe gilt für Produzenten digitaler Objekte, soweit es keine gesetzlichen Ablieferungs- oder Anbietungspflichten gibt. Das Vorhandensein von Auswahlkriterien stellt deshalb auch einen wichtigen Aspekt von Vertrauenswürdigkeit dar.[23] Gegenüber den Trägern wird anhand der Auswahlkriterien belegt, dass die Sammelaktivitäten dem Auftrag der Institution entsprechen. Und schließlich spielen die jeweiligen Auswahlkriterien auch eine wichtige Rolle beim Aufbau von Netzwerken zur verteilten, kooperativen Langzeitarchivierung (beispielsweise im nationalen Rahmen).

Zumeist stellt die Aufnahme digitaler Objekte in die Sammlung bzw. die Bestände eine zusätzliche Aufgabe dar, die zu bestehenden Sammelaktivitäten bzw. Bewertungen für konventionelle Objekte hinzukommt. Viele Institutionen besitzen deshalb bereits Auswahlkriterien im analogen Bereich, die als Ausgangspunkt für entsprechende Richtlinien im digitalen Bereich dienen können. Mit Blick auf die Besonderheiten digitaler Objekte müssen diese freilich kritisch überprüft, abgeändert und erweitert werden. Dabei sind fünf Aspekte besonders zu beachten:

- *Spezielle Objekt- und Dokumenttypen:* Während sich für viele Arten von digitalen Objekten eine Entsprechung im konventionellen Bereich finden lässt, gibt es auch spezielle digitale Objekt- und Dokumenttypen, die in den Auswahlrichtlinien zu berücksichtigen sind. Beispielsweise besitzt eine E-Dissertation im PDF-Format ein analoges Pendant in der konventionellen, gedruckten Dissertation. Eine Entsprechung für originär digitale Objekte wie Websites oder Datenbanken lässt sich hingegen nicht in gleicher Weise finden. Deshalb ist eine Orientierung an vorhan-

23 Das Kriterium 1.1 im 'Kriterienkatalog Vertrauenswürdige Archive' lautet: „Das digitale Langzeitarchiv hat Kriterien für die Auswahl seiner digitalen Objekte entwickelt". Vgl. nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (2008), S. 11. Zur Vertrauenswürdigkeit digitaler Langzeitarchive allgemein s.u. Kap. 5.

denen konventionellen Auswahlkriterien hier nur bedingt möglich (nämlich nur für die inhaltlich-fachlichen Aspekte des Objektes).

- *Technische Anforderungen:* Anders als bei konventionellen Objekten spielen technische Anforderungen (z.B. das Dateiformat und die notwendige technische Umgebung zur Darstellung der Information) für die Abläufe im digitalen Langzeitarchiv eine wichtige Rolle. Sie sind deshalb in die Überlegungen mit einzubeziehen.
- *Veränderte Arbeitsabläufe:* Digitale Objekte sind unbeständiger als ihre papierenen Gegenstücke und weniger geduldig; sollen sie dauerhaft bewahrt werden, muss bereits bei ihrer Entstehung dafür gesorgt werden. Beispielsweise müssen Bibliotheken auf die Produzenten einwirken, damit diese ihre Publikationen in langzeitgeeigneter Form erstellen; ebenso müssen Archive bei den von ihnen zu betreuenden Behörden bereits bei der Einführung elektronischer Systeme präsent sein. Sollen Informationen aus Datenbanken oder Geoinformationssystemen archiviert werden, muss sichergestellt werden, dass vorhandene Daten bei Änderung nicht einfach überschrieben werden, sondern dass so genannte Historisierungen vorgenommen werden, die einen bestimmten Stand festhalten.
- *Unterschiedliche Mengengerüste:* Die Zahl und der Umfang der theoretisch auswahlfähigen digitalen Objekte liegt häufig in deutlich höheren Größenordnungen als bei entsprechenden analogen Objekten. Beispielsweise sind Netzpublikationen sehr viel leichter zu realisieren als entsprechende Printpublikationen, so dass ihre Zahl die der gedruckten Publikationen bei weitem übersteigt. Ebenso werden zum Beispiel Statistikdaten in der Papierwelt nur in aggregierter, d.h. zusammengefasster Form als Quartals- oder Jahresberichte übernommen. In digitaler Form können jedoch auch die Einzeldaten übernommen und den Nutzern in auswertbarer Form zur Verfügung gestellt werden.
- *Schwer zu bemessender Arbeitsaufwand:* Der Umgang mit konventionellen Objekten erfolgt über etablierte Kanäle und Geschäftsgänge, so dass Aufwände gut zu messen und zu bewerten sind. Der Aufwand zur Beschaffung, Erschließung, Bereitstellung und Langzeitarchivierung digitaler Objekte ist dagegen wegen fehlender Erfahrungswerte schwer abzuschätzen.

Die letzten beiden Punkte können u.U. dazu führen, dass Auswahlkriterien für digitale Objekte strenger gefasst werden müssen als für konventionelle Objekte, sofern nicht auf anderen Wegen – beispielsweise durch den Einsatz maschineller Methoden oder zusätzliches Personal – für Entlastung gesorgt werden kann. Die zusätzliche Berücksichtigung digitaler Objekte bei den Sammelaktivitäten bzw. bei der Bewertung kann außerdem Rückwirkungen auf die Aus-

wahlkriterien für konventionelle Objekte derselben Institution haben, indem etwa die beiden Segmente in ihrer Bedeutung für die Institution neu gegeneinander austariert werden müssen.

Die zu erarbeitenden Auswahlkriterien[24] können sowohl inhaltlich-fachlicher als auch formal-technischer Natur sein. Darüber hinaus können beispielsweise auch finanzielle sowie lizenz- und urheberrechtliche Aspekte in die Auswahlkriterien mit eingehen; die folgende Liste erhebt keinen Anspruch auf Vollständigkeit.

Inhaltlich-fachliche Auswahlkriterien

Aus inhaltlich-fachlicher Sicht kommen typischerweise drei Kriterien in Betracht:

- *Verwaltungstechnische, institutionelle oder räumliche Zuständigkeit*, z.B. eines Unternehmensarchivs für die Unterlagen des Unternehmens; eines Museums für Digitalisate eigener Bestände; des Dokumentenservers einer Universität für die dort entstandenen Hochschulschriften; einer Pflichtexemplarbibliothek für die im zugeordneten geographischen Raum veröffentlichten Publikationen.
  Leitfrage: Ist mein Archiv gemäß der institutionellen oder rechtlichen Vorgaben zur Übernahme des Objekts verpflichtet?
- *Inhaltliche Relevanz, ggf. in Verbindung mit einer Qualitätsbeurteilung,* z.B. thematisch in ein an einer Bibliothek gepflegtes Sondersammelgebiet fallend; zu einer Spezialsammlung an einem Museum passend; von historischem Wert für die zukünftige Forschung; von Bedeutung für die retrospektive Verwaltungskontrolle und für die Rechtssicherung der Bürger. Dazu gehört auch der Nachweis der Herkunft des Objekts aus seriöser und vertrauenswürdiger Quelle. Ggf. können weitere qualitative Kriterien angelegt werden, z.B. bei Prüfungsarbeiten die Empfehlung eines Hochschullehrers.
  Leitfragen: Ist das Objekt durch sein enthaltenes Wissen bzw. seine Ästhetik, Aussagekraft o.ä. wichtig für meine Institution? Kann das Objekt bei der Beantwortung von Fragen hilfreich sein, die an meine Institution gestellt werden? Ist das Objekt aufgrund seiner Herkunft, seiner Provenienz von bleibendem (z.B. historischem) Wert?

---

24 Vgl. zum Folgenden auch die Ergebnisse einer Umfrage zu den in verschiedenen Institutionen angewendeten Auswahlkriterien, die im Rahmen der ersten Phase des nestor-Projektes durchgeführt wurde: Blochmann (2005), S. 9-31.

- *Dokumentart*, z.B. spezifische Festlegungen für Akten, Seminararbeiten, Geschäftsberichte, Datenbanken, Websites etc.
  Leitfragen: Besitzt mein Archiv schon Bestände der gleichen Dokumentart? Verfüge ich über das nötige Fachwissen und die nötigen Arbeitsmittel zur Erschließung und Verzeichnung der Dokumentart?

Formal-technische Auswahlkriterien

Aus formal-technischer Sicht steht auf der obersten Ebene das folgende Kriterium:

- *Lesbarkeit des Objekts im Archiv,* z.B. die Prüfung, ob ein Objekt mit den verfügbaren technischen Mitteln (Hardware/Software) des Langzeitarchivs dargestellt werden kann. Darstellen heißt, dass die vom Objekt transportierte Information vom menschlichen Auge erkannt, gelesen und interpretiert werden kann.
  Leitfrage: Verfügt mein Archiv über die nötigen Kenntnisse, Geräte und Software, um das Objekt den Nutzern authentisch präsentieren zu können?

Aus diesem obersten formal-technischen Zielkriterium lassen sich weitere Unterkriterien ableiten:

- *Vorhandensein der notwendigen Hardware,* z.B. die Feststellung, ob ein einzelner Rechner oder ein ganzes Netzwerk benötigt wird; ob die Nutzung des Objekts an ein ganz spezielles Gerät gebunden ist usw. Außerdem muss geprüft werden, ob das Objekt mit den vorhandenen Geräten gespeichert und gelagert werden kann.
  Leitfragen: Verfügt mein Archiv über ein Gerät, mit dem ich das Objekt in authentischer Form darstellen und nutzen kann? Verfügt mein Archiv über Geräte, die das Objekt in geeigneter Form speichern können?
- *Vorhandensein der notwendigen Software,* z.B. die Feststellung, ob die Nutzung eines Objekts von einem bestimmten Betriebssystem, einem bestimmten Anzeigeprogramm oder sonstigen Einstellungen abhängig ist. Außerdem muss das Archiv über Software verfügen, die das Speichern und Auffinden des Objektes steuert und unterstützt.
  Leitfragen: Verfügt mein Archiv über alle Programme, mit denen ich das Objekt in authentischer Form darstellen und nutzen kann? Verfügt mein Archiv über Programme, die das Objekt in geeigneter Form speichern und wiederfinden können?
- *Vorliegen in geeigneten Formaten,* bevorzugt solchen, die normiert und stan-

dardisiert sind, und deren technische Spezifikationen veröffentlicht sind. Dateiformate sollten nicht von einem einzigen bestimmten Programm abhängig, sondern idealerweise weltweit verbreitet sein und von vielen genutzt werden. Je weniger Formate in einem Archiv zulässig sind, desto leichter kann auch das Vorhandensein der notwendigen Hard- und Software geprüft werden.
Leitfragen: Hat mein Archiv schon Objekte dieses Formats im Bestand? Sind die notwendigen Mittel und Kenntnisse zur Nutzung und Speicherung des Formats offen zugänglich und leicht verfügbar?

- *Vorhandensein geeigneten Personals,* z.B die Feststellung, ob die Mitarbeiterinnen und Mitarbeiter über das technische Fachwissen verfügen, das zur Nutzung und Speicherung des Objekts notwendig ist.
  Leitfragen: Habe ich Personal, dem ich aus technischer Sicht die Verantwortung für das Objekt anvertrauen kann? Verfüge ich über die Mittel, um Personal mit den entsprechenden Kenntnissen einzustellen oder um Dienstleister mit der Aufgabe zu betrauen?

Auswahlkriterien für Netzpublikationen

Eine für Bibliotheken besonders wichtige Gattung digitaler Objekte sind die so genannten *Netzpublikationen*, auch als „Medienwerke in unkörperlicher Form" bezeichnet und als „Darstellungen in öffentlichen Netzen"[25] definiert. Auch für diese gelten die oben dargestellten allgemeinen Auswahlkriterien, doch sollen im Folgenden noch einige spezielle Hinweise aus bibliothekarischer Sicht gegeben werden[26]. Dabei ist es nützlich, die Vielfalt der Netzpublikationen in zwei Basistypen zu unterteilen: In die Netzpublikationen mit Entsprechung in der Printwelt einerseits und die sog. Web-spezifischen Netzpublikationen andererseits.[27]

Bei den *Netzpublikationen mit Entsprechung in der Printwelt* lassen sich wiederum zwei Typen unterscheiden:

- *Druckbildähnliche Netzpublikationen,* welche ein weitgehend genaues elektronisches Abbild einer gedruckten Publikation darstellen, d.h. 'look and

25 Gesetz über die Deutsche Nationalbibliothek (2006), § 3, Abs. 3.

26 Auf andere Arten von Gedächtnisorganisationen ist die folgende Darstellung nicht zwingend übertragbar.

27 Für die folgenden Ausführungen vgl. Wiesenmüller et al. (2004), S. 1423-1437. Unbenommen bleibt, dass die im Folgenden genannten Typen von Netzpublikationen auch in Offline-Versionen vorkommen können.

feel' des gedruckten Vorbilds möglichst exakt nachahmen wollen und diesem bis hin zum äußeren Erscheinungsbild entsprechen (z.B. Titelblatt, festes Layout mit definierten Schriftarten und -größen, feste Zeilen- und Seitenumbrüche etc.).

- *Netzpublikationen mit verwandtem Publikationstyp in der Printwelt,* welche zwar keine Druckbildähnlichkeit aufweisen, jedoch einem aus der Printwelt bekannten Publikationstyp zugeordnet werden können, z.B. ein Lexikon im HTML-Format.

Bei der Erarbeitung von Auswahlkriterien für diese beiden Typen ist i.d.R. eine Orientierung an bereits vorhandenen Sammelrichtlinien für konventionelle Materialien möglich. Besondere Beachtung verdient dabei der durchaus nicht seltene Fall, dass zur jeweiligen Netzpublikation eine gedruckte Parallelausgabe vorliegt. Unter Abwägung des zusätzlichen Aufwandes einerseits und des möglichen Mehrwerts des digitalen Objekts andererseits ist festzulegen, ob in einem solchen Fall nur die konventionelle oder nur die digitale Version in das Archiv aufgenommen wird, oder ob beide Versionen gesammelt werden.

Zu den *Web-spezifischen Netzpublikationen* zählen beispielsweise Websites oder Blogs. Sie können keinem aus der Printwelt bekannten Publikationstyp zugeordnet werden, so dass eine Orientierung an bestehenden Sammelrichtlinien nur sehr bedingt möglich ist. Für diese Publikationstypen müssen daher neue Auswahlkriterien entwickelt werden.[28]

Der Umgang mit *Websites* wird dadurch erschwert, dass unterhalb der Website-Ebene häufig weitere Netzpublikationen – mit oder ohne Entsprechung in der Printwelt – liegen, die getrennt gesammelt, erschlossen und bereitgestellt werden können (z.B. ein Mitteilungsblatt auf der Website einer Institution). In den Auswahlkriterien muss also auch festgelegt sein, unter welchen Umständen (nur) die Website als Ganzes gesammelt wird, oder zusätzlich bzw. stattdessen auch darin integrierte Netzpublikationen in das Archiv aufgenommen werden sollen. Bei Websites, die immer wieder ergänzt, aktualisiert oder geändert

28 Auch Online-Datenbanken sind am ehesten den Web-spezifischen Netzpublikationen zuzuordnen, weil es in der Printwelt keinen Publikationstyp gibt, der in Funktionalität und Zugriffsmöglichkeiten mit ihnen vergleichbar ist. Ein grundsätzlicher Unterschied zu einem gedruckten Medium ist z.B., dass dessen gesamter Inhalt sequentiell gelesen werden kann, während bei einer Datenbank gemäß der jeweiligen Abfrage nur eine Teilmenge des Inhalts in lesbarer Form generiert wird. Was jedoch den in Online-Datenbanken präsentierten *Inhalt* angeht, so kann es natürlich durchaus Entsprechungen zu Produkten aus der Printwelt geben (z.B. sind in vielen Fällen gedruckte Bibliographien durch bibliographische Datenbanken abgelöst worden).

werden und deshalb in Zeitschnitten zu sammeln sind, muss jeweils auch das Speicherintervall festgelegt werden.

Bei der Erarbeitung von Auswahlkriterien für Websites sollte unterschieden werden *zwischen* solchen, welche Personen oder Körperschaften (inkl. Gebietskörperschaften, Ausstellungen, Messen etc.) repräsentieren, und *solchen*, die sich einem bestimmten Thema widmen – wobei freilich auch Mischformen möglich sind.

Bei *repräsentierenden Websites* setzen die Auswahlkriterien in erster Linie beim Urheber an: Ist die repräsentierte Person oder Körperschaft für mein Archiv relevant? Welche Arten von Personen und Körperschaften sollen schwerpunktmäßig gesammelt, welche ausgeschlossen werden?[29] Ein zusätzliches Kriterium können die auf der Website gebotenen Informationen sein, was sich am besten am Vorhandensein und an der Gewichtung typischer Elemente festmachen lässt: Beispielsweise könnten Websites, die umfangreiche Informationen zur repräsentierten Person oder Körperschaft, einen redaktionellen Teil und/oder integrierte Netzpublikationen bieten, mit höherer Priorität gesammelt werden als solche, die im wesentlichen nur Service- und Shop-Angebote beinhalten.

Bei thematischen Websites kommt neben der inhaltlichen Relevanz auch die Qualität als Auswahlkriterium in Frage. Zwar kann i.d.R. keine Prüfung auf Richtigkeit oder Vollständigkeit der gebotenen Information geleistet werden, doch können als Auswahlkriterien u.a. der Umfang, die Professionalität der Darbietung und die Pflege der Website herangezogen werden, außerdem natürlich der Urheber (z.B. Forschungsinstitut vs. Privatperson).

Detaillierte Sammelrichtlinien für Netzpublikationen, die als Anregung dienen können, sind beispielsweise im Rahmen des PANDORA-Projekts von der Australischen Nationalbibliothek erarbeitet und veröffentlicht worden.[30]

---

29 Die Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek (2008), § 9, Abs. 1, schließt beispielweise „lediglich privaten Zwecken dienende Websites" generell von der Ablieferungspflicht aus.

30 Vgl. National Library of Australia (2005).

Quellenangaben und weiterführende Literatur

Arbeitskreis Archivische Bewertung im VdA – Verband deutscher Archivarinnen und Archivare (Hrsg.) (2004): *Positionen des Arbeitskreises Archivische Bewertung im VdA – Verband deutscher Archivarinnen und Archivare zur archivischen Überlieferungsbildung: vom 15. Oktober 2004*

http://www.vda.archiv.net/texte/ak_bew_positionen2004.doc

Blochmann, Andrea (2005): *Langzeitarchivierung digitaler Ressourcen in Deutschland: Sammelaktivitäten und Auswahlkriterien* (nestor – Kompetenznetzwerk Langzeitarchivierung, AP 8.2). Version 1.0. Frankfurt am Main: nestor

http://www.langzeitarchivierung.de/downloads/nestor_ap82.pdf

Gesetz über die Deutsche Nationalbibliothek (2006) vom 22. Juni 2006. In: Bundesgesetzblatt 2006, I/29, 28.06.2006, S. 1338-1341

http://www.bgblportal.de/BGBL/bgbl1f/bgbl106s1338.pdf

National Library of Autralia (2005): *Online Australian publications: selection guidelines for archiving and preservation by the National Library of Australia*. Rev. August 2005. Canberra: National Library of Australia

http://pandora.nla.gov.au/selectionguidelines.html

nestor-Arbeitsgruppe Standards für Metadaten, Transfer von Objekten in digitale Langzeitarchive und Objektzugriff (Hrsg.) (2008): *Wege ins Archiv: ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv*. Version 1, Entwurf zur öffentlichen Kommentierung. (nestor-Materialien 10). Göttingen: nestor

http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_10.pdf
urn:nbn:de:0008-2008103009

nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2008): *Kriterienkatalog vertrauenswürdige digitale Langzeiarchive*. Version 1. (nestor-Materialien 8). Frankfurt am Main: nestor

http://nbn-resolving.de/urn:nbn:de:0008-2008021802
urn:nbn:de:0008-2008021802

Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek (2008) vom 17. Oktober 2008. In: Bundesgesetzblatt 2008, I/47, 22.10.2008, S. 2013-2015

http://www.bgblportal.de/BGBL/bgbl1f/bgbl108s2013.pdf

Wiesenmüller, Heidrun et al. (2004): *Auswahlkriterien für das Sammeln von Netzpublikationen im Rahmen des elektronischen Pflichtexemplars: Empfehlungen der Arbeitsgemeinschaft der Regionalbibliotheken*. In: Bibliotheksdienst 11. 2004 (Jg. 38), S. 1423-1444

http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte/heft9-1204/digitalebib1104.pdf

4 Das Referenzmodell OAIS – Open Archival Information System

4.1 Einführung

Achim Oßwald

Relativ selten in der Geschichte der Anwendung von IT-Verfahren ist es vorgekommen, dass ein Modell weltweit so rasch und kaum angezweifelt Akzeptanz erfahren hat wie das OAIS-Referenzmodell, das 2002 von der Data Archiving and Ingest Working Group des Consultative Committee for Space Data Systems (CCSDS) unter Federführung der NASA, der US-amerikanischen Raumfahrtbehörde, veröffentlicht wurde. Lange Zeit durfte dieses Referenzmodell in keiner Präsentation zum Thema LZA fehlen. Die orientierende, katalytische und in doppeltem Sinne normierende Wirkung dieses zum ISO-Standard erhobenen Modells auf die Diskussionen und den Austausch von konzeptionellen sowie praktisch realisierenden Überlegungen innerhalb der Gemeinschaft der Langzeitarchivierungsspezialisten kann nicht hoch genug eingeschätzt werden. Die Verständigung der Experten über ihre jeweiligen Lösungskonzepte kann fast immer auf zentrale Komponenten des Referenzmodells zurückgeführt werden. Solche Verständigung erleichtert die Kommunikation über Sprach- und Forschungsgruppengrenzen hinweg, ermöglicht die funktionale Zuordnung

von Neuentwicklungen und beschleunigt letzten Endes die Forschung und Entwicklung insgesamt. Ein gemeinsames Denk- und Referenzmodell kann allerdings auch Nachteile haben, die nicht unterschlagen werden sollen: Es kann einengen, kann dort als Innovationsbremse wirken, wo seine Vorgaben und seine Leistungs- bzw. Tragfähigkeit sich als kritisch erweisen. Auch deshalb findet in den letzten Jahren verstärkt eine Diskussion zur Überarbeitung des Modells bzw. der das Modell beschreibenden Dokumente statt.

Kapitel 4 gibt einen Überblick zu all diesen Aspekten, in dem es

- die Entwicklung des OAIS und seinen Ansatz darstellt und erläutert
- die Kernkomponenten Informationsobjekte und das Datenmodell konkretisiert
- das Funktionsmodell des OAIS skizziert und
- die Akzeptanz des OAIS-Modells begründet.

Das Kapitel berücksichtigt noch nicht die im Jahre 2006 federführend vom britischen Digital Curation Centre und der Digital Preservation Coalition vorgeschlagenen und im Dezember 2008 noch einmal kommentierten Vorschläge für Klarstellungen und Veränderungen des OAIS-Modells. Bis zum Redaktionsschluss des nestor Handbuches war nicht eindeutig, welche der Vorschläge tatsächlich, sowie wenn ja wie konkret und wann, umgesetzt werden würden.

4.2 Das Referenzmodell OAIS

Nils Brübach

Bearbeiter: Manuela Queitsch, Hans Liegmann (†), Achim Oßwald

[Überarbeitete Fassung eines Vortrags, gehalten auf der 6. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“ am 5./6. März 2002 in Dresden]

Das als ISO 14721 verabschiedete Referenzmodell „Open Archival Information System – OAIS“ beschreibt ein digitales Langzeitarchiv als eine Organisation, in dem Menschen und Systeme mit der Aufgabenstellung zusammenwirken, digitale Informationen dauerhaft über einen langen Zeitraum zu erhalten und einer definierten Nutzerschaft verfügbar zu machen.

Im folgenden Beitrag werden vier Ziele verfolgt: Erstens sollen die Entwicklung des OAIS, sein Konzept und sein Ansatz skizziert werden. Zweitens werden die wesentlichen Kernkomponenten des OAIS, nämlich die in ihm vorgesehenen Informationsobjekte bzw. Informationspakete und das ihnen zu Grunde liegende Datenmodell analysiert und vorgestellt, um drittens das Funktionsmodell des OAIS zu erläutern. Es ist ein besonderes Kennzeichen des OAIS, das bereits bei seiner Entwicklung nicht nur ein auf theoretischen Überlegungen fußendes Modell entwickelt wurde, sondern dass die Frage nach der Anwendbarkeit und deren Prüfung vorab an konkreten Anwendungsfällen mit in die Konzeption und Entwicklung einbezogen wurden. Deswegen wird im vierten Abschnitt kurz auf einige bereits existierende Anwendungsbeispiele des OAIS eingegangen: OAIS ist kein am „grünen Tisch“ auf Basis rein theoretischer Überlegungen entwickelter Ansatz, sondern für die Praxis entwickelt worden.

4.2.1 Die Entwicklung des OAIS und sein Ansatz

Das Open Archival Information System hat seine Wurzeln im Gebiet der Raumfahrt. Diese Tatsache ist nur auf den ersten Blick wirklich überraschend, wird aber verständlich, wenn man sich vor Augen führt, dass in diesem Bereich seit den sechziger Jahren elektronische Daten in großen Mengen angefallen sind - demzufolge die das klassische öffentliche Archivwesen jetzt beschäftigenden Fragen schon weit eher auftreten mussten. Federführend für die Entwicklung des OAIS, die seit dem Jahre 1997 betrieben wurde, war das „Consultative Committee for Space Data Systems“(CCSDS), eine Arbeitsgemeinschaft verschiedener Luft- und Raumfahrtorganisationen wie etwa der NASA, der ESA

oder der Deutschen Gesellschaft für Luft- und Raumfahrt unter Federführung der NASA. Beteiligt waren von archivischer Seite seit 1997 die amerikanische nationale Archivverwaltung (NARA) und die Research Libraries Group (RLG). Das OAIS wurde im Jahre 1999 erstmals als vollständige Textfassung in Form eines so genannten „Red Book" vorgelegt. Lou Reich und Don Sawyer von der CCSDS bzw. der NASA sind die Autoren der unterschiedlichen Textfassungen und hatten auch die Koordination der Arbeitsgruppe zur Erstellung des Textes inne. Im gleichen Jahr 1999, in dem das Red Book veröffentlicht und der internationalen Fachgemeinschaft der Archivarinnen und Archivare zur Diskussion gestellt wurde, wurde die Vorlage auch bei der ISO als internationaler Standard eingereicht. Er durchlief dort die üblichen Prüfungsverfahren. Der Text des Red Book wurde nach Ergänzung und Überarbeitung im Juni 2001 als ISO/DIS 14721 angenommen und zum 1. Januar 2002 in das Normenwerk der Internationalen Standardorganisation integriert.[1] Die Übernahme in das deutsche Normenwerk steht allerdings noch aus. Wir haben es also für diesen Bereich, ähnlich wie bei der ISO/DIN 15489 „Schriftgutverwaltung", erneut mit einem Standard zu tun und nicht nur mit einem Arbeitsdokument unter vielen. Allein schon das Abstimmungsverfahren und die nur wenigen vorgenommenen Änderungen am ursprünglichen Text des Red Book zeigen, wie ausgefeilt und wie weit entwickelt das Projekt bereits gediehen war, als es bei der ISO als Standard vorgelegt wurde. Dieses Arbeitsverfahren - mit Hilfe von Standards gesicherte Arbeitsergebnisse zu einer Art von „anwendungsbezogenem Allgemeingut" werden zu lassen - scheint sich im Bereich der Archivierung elektronischer Unterlagen immer stärker durchzusetzen: So wurde vom ISO TC 46 und TC 171 eine Untermenge des PDF-Formats (PDF/A = PDF/Archive) ein Standardisierungsprozess (ISO 19005-1. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF (PDF/A)) eingeleitet, der zur größeren Akzeptanz des Formats für die Langzeitarchivierung digitaler Dokumente führen soll.[2]

Das OAIS-Konzept ist ein Standard in Form eines Referenzmodells für ein dynamisches, erweiterungsfähiges Archivinformationssystem. Ganz bewusst

1 http://public.ccsds.org/publications/archive/650x0b1.pdf. CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS)*. Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2003.
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

2 Der Begriff "Langzeitarchivierung" wird als Äquivalent zum englischen Terminus "long-term preservation" verwendet. Er ist als technischer Begriff zu sehen, der darauf hin deuten soll, dass anders als bei der Archivierung im analogen Umfeld, die dauerhafte Aufbewahrung von digitalen Objekten eben nicht auch die dauerhafte Zugänglichkeit automatisch nach sich zieht.

versteht sich OAIS als ein offener Standard, aber auch als ein Modell, das den Anspruch der Allgemeingültigkeit verfolgt. Das hat zwei Konsequenzen:

1. OAIS verzichtet auf eine Beschränkung auf bestimmte Datentypen, Datenformate oder Systemarchitekturen (im technischen Sinne) und
2. OAIS will anwendungsfähig und skalierbar sein für eine Vielzahl bestimmter Institutionen und ihre jeweiligen Bedürfnisse.

Der Text des OAIS hat insgesamt sieben Abschnitte:

Kapitel 1 „Einführung" beschreibt die Zielsetzung, den Anwendungsrahmen, bestimmte Anwendungsregeln und stellt die notwendigen Begriffsdefinitionen voran.

In Kapitel 2 wird das Konzept des OAIS, d.h. die unterschiedlichen Typen von Informationen, die modellierbaren standardisierten Operationen und auch die Systemumgebung (im funktionalen Sinne) beschrieben.

Kapitel 3, eines der Kernkapitel, beschreibt die Tätigkeitsfelder eines OAIS-Betreibers.

Kapitel 4 ist den Datenmodellen gewidmet, die dem OAIS zugrunde liegen. Hier wird einerseits das Funktionsmodell beschrieben und andererseits die unterschiedlichen Informationspakete, ihre Verwendung und ihre Verknüpfung zu einem Informationsmodell.

Kapitel 5 ist der zweite Kernbereich, denn hier wird beschrieben, welche Operationen für eine dauerhafte Aufbewahrung digitaler Aufzeichnungen und für die Gewährleistung des Zugangs zu ihnen unverzichtbar sind.

Die heutige Archivlandschaft ist eine offene Archivlandschaft. Demzufolge widmet sich Kapitel 6 dem Betrieb eines Archivs nach OAIS-Standard in Kooperation mit anderen Archiven. So entscheidende Fragen wie die der technischen Kooperation, die Frage nach Funktion und Aufbau von Schnittstellen und eines gemeinsamen übergreifenden Managements verschiedener digitaler Archive werden hier angesprochen.

Der 7. Teil des Standards enthält die Anhänge, in denen Anwendungsprobeläufe beschrieben werden, auf andere Standards verwiesen wird, Modelle für Kooperationen skizziert und Entwicklungsmodelle für bestimmte Software-Lösungen zumindest angedeutet werden.[3] Auf diesen letzen Aspekt der „Interoperabilität" sei an dieser Stelle besonders hingewiesen. OAIS versteht sich nämlich nicht als eine geschlossene Lösung, sondern als ein offenes Informationssystem, das in jedem Fall und in jedem Stadium mit anderen Parallelsyste-

3 Gail M. Hogde: Best Practices for Digital Archiving. In D-LIB-Magazine, Vol.6 No.1, January 2000, S.8. http://www.dlib.org/dlib/january00/01hodge.html

men vernetzbar sein soll. Dadurch, dass OAIS sich selbst als Referenzmodell definiert, ist es auch offen für verschiedene technische Lösungsmöglichkeiten, die aber über den zentralen Punkt der funktionalen Interoperabilität aufeinander abgestimmt und miteinander verknüpfbar sein müssen.

Das Open Archival Information System beschreibt ein Informationsnetzwerk, das den Archivar und den Nutzer als Hauptkomponenten des digitalen Archivs versteht. Archivierung ist nicht an Maschinen delegierbar: Der Mensch hat im Sinne des OAIS die Verantwortung für die Sicherung von Informationen und deren Bereitstellung für eine bestimmte Nutzergruppe. Die Unterscheidung verschiedener Nutzergruppen (Designated Communities) ist eine Besonderheit des OAIS. Die Interoperabilität liegt nämlich nicht nur in technischer und in funktioneller Ebene, sondern eben auch darin, dass unterschiedliche Benutzergruppen unterschiedliche Anforderungen an elektronische Archive in der Gegenwart haben und in der Zukunft haben werden: Anforderungen, die heutige Entwicklergenerationen technischer Lösungen überhaupt nicht voraussehen können und bei denen das, was Archivierung eigentlich ausmacht - Sicherung von Authentizität und Integrität durch dauerhafte Stabilisierung und Zugänglichmachung von authentischen unikalen Kontexten - auch im digitalen Umfeld gewährleistet ist. Die Offenheit des OAIS ist also auf Zukunftsfähigkeit und auf Nachhaltigkeit ausgerichtet. Die heute im Rahmen des OAIS realisierten Lösungen sollen auch in der Zukunft verwendbar und in neue technische Realisierungen übertragbar sein. Das OAIS wird damit auch offen für neue Anforderungen an die Nutzung.

Das OAIS konzentriert sich auf die Langzeitaufbewahrung und Langzeitnutzbarhaltung hauptsächlich digitaler Aufzeichnungen und dies unter Berücksichtigung der sich verändernden Technologien. Wenn die Autoren des OAIS sich hauptsächlich auf digitale Aufzeichnungen konzentrieren, so verweisen sie doch darauf, dass in einem weiteren Sinne jedes digitale Archiv, das dem OAIS-Standard folgt, immer auch mit schon bestehenden, sich auf analoge Unterlagen konzentrierenden Archivlösungen verknüpfbar sein und dass diese Verknüpfung auch in der Zukunft erhalten bleiben muss. Das OAIS zeigt also Wege auf zur dauerhaften Sicherung digitaler Unterlagen in ihrem Kontext und den wechselseitigen Beziehungen zu analogem Schriftgut, die sich wandeln können: Die Gedächtnisorganisationen werden in Zukunft eben auch Papier enthalten müssen, es treten neue Aufzeichnungsformen hinzu, die die alten keineswegs vollständig verdrängen werden. Ebenso wie sich das noch vor wenigen Jahren propagierte „papierlose Büro“ als Hirngespinst erwiesen hat und, viel bescheidener, heute nur noch vom „papierarmen Büro“ gesprochen wird, sind Überlegungen zu einem vollständigen Medienbruch bei der Archivierung reali-

tätsfremd. Das OAIS berücksichtigt Bestehendes: Es ist gerade deshalb ein Modellansatz und ein Standard, der damit auch Einfluss auf zukünftige Arbeitsmethoden im Archiv nehmen wird. Es geht von den klassischen archivischen Arbeitsfeldern, Erfassen, Aussondern, Bewerten, Übernehmen, Erschließen, Erhalten und Zugänglichmachen aus, aber definiert sie in ihren Teilaufgaben und Arbeitsabläufen unter dem Blickwinkel der Bedürfnisse digitaler Archivierung neu. Im gewissen Sinne beantwortet der Text des OAIS die schon so häufig gestellte, aber bisher bestenfalls unbefriedigend beantwortete Frage nach dem zukünftigen Aufgabenspektrum von Gedächtnisorganisationen im digitalen Zeitalter. Auch die Frage danach, welche Funktionen automatisierbar sind, wird thematisiert. Hier liegt nicht zuletzt auch ein für Fragen der Aus- und Fortbildung interessanter Aspekt.

Das OAIS erhebt den Anspruch, auf jedes Archiv anwendbar zu sein, Archiv vom Begriff her bezieht sich hier ausdrücklich auf den Bereich der dauerhaften Aufbewahrung und langfristigen Zugangssicherung. Dabei wird auch kein Unterschied gemacht, ob die Archivierung organisationsintern bei den produzierenden Stellen selbst erfolgt, oder bei Organisationen, die digitale Objekte zur Archivierung übernehmen.

4.2.2 Die Kernkomponenten: Informationsobjekte und Datenmodell

Das OAIS unterscheidet zwischen drei so genannten Informationsobjekten, die miteinander in Verbindung stehen und sich aufeinander beziehen, aber entwickelt worden sind, um den unterschiedlichen Umgang und die unterschiedlichen Tätigkeiten bei der digitalen Archivierung besser beschreiben zu können. Das was Archive an digitalen Unterlagen übernehmen, heißt in der Terminologie des OAIS Submission Information Packages (SIP). Im Archiv selbst werden diese SIP vom Archiv durch Metainformationen ergänzt und umgeformt zu Archival Information Packages (AIP), die weiter verarbeitet werden und die im Kern die Form darstellen, in der die digitalen Informationen tatsächlich langfristig aufbewahrt werden. Zugänglich gemacht werden die AIPs über die so genannten Dissemination Information Packages (DIP), die für bestimmte Nutzergruppen je nach Vorliegen bestimmter rechtlicher Bedürfnisse generiert und zielgruppenorientiert zur Verfügung gestellt werden können. Dieser Ansatz ist im Vergleich zum klassischen Bestandserhaltung durchaus ungewöhnlich. Im Sinne des OAIS wird nämlich nicht ohne Veränderung das einfach aufbewahrt, was man übernimmt, sondern es wird zukünftig die Aufgabe der Verantwortlichen sein, sehr viel mehr noch als im Bereich der Archivierung von analogen

Unterlagen dafür zu sorgen, dass die Unterlagen überhaupt archivfähig sind. Die Umformung der SIPs zu Archival Information Packages kann z.B. darin bestehen, dass aus den mit übernommenen Objekten und den mitgelieferten Metadaten die zur Langzeiterhaltung notwendigen Metadaten generiert werden. Darüber hinaus sind die Formate, in denen ein SIP dem Archiv angeboten und von ihm übernommen wird, keinesfalls unbedingt identisch mit den Aufbewahrungsformaten, in denen die Archival Information Packages dann tatsächlich vorliegen. Sichergestellt sein muss die Bewahrung von Authentizität und Integrität auch mit Blick auf die rechtswahrende und rechtssichernde Funktion digitaler Archive. Ein AIP aus dem Jahre 2003 wird naturgemäß in einem ganz anderen Format und in einer ganz anderen Datenstruktur vorliegen, als das gleiche AIP etwa im Jahre 2010. Grundgedanke dieser Arbeit mit Informationspaketen ist es, dass Inhalte, Metadaten und - wo unverzichtbar - die entsprechenden Strukturen der digitalen Aufzeichnungen nachvollziehbar bzw. rekonstruierbar gehalten werden, unabhängig von den sich wandelnden technischen Gegebenheiten. Dies ist ein Aspekt, der eben auch auf die Benutzung der Unterlagen zielt. Die Dissemination Information Packages dienen der Nutzung und dem Zugang je nach den Bedürfnissen der jeweiligen Benutzergruppen und sind ganz gezielt für unterschiedliche Benutzer anzupassen und auch anpassbar zu erhalten. Gerade das ist für die klassische dauerhafte Bestandserhaltung in Archiven eine ungewöhnliche Vorstellung: dem Benutzer wird nicht mehr das vorgelegt, was im Magazin verwahrt wird, sondern aus dem, was verwahrt wird, werden Informationspakete generiert, die auf die Bedürfnisse der Kunden natürlich auch in Abhängigkeit von die Nutzung einschränkenden Rechten Betroffener oder Dritter zugeschnitten werden. Diese Umformung der AIPs in DIPs bezieht sich dabei keinesfalls ausschließlich auf die Veränderung der Datenformate, sondern eben auch auf die Bereitstellung von digitalen Informationen in Verbindung mit einer für den Benutzer besonders komfortablen Funktionalität. Hier wird im OAIS ein Ansatz aufgegriffen, der im Bereich der archivischen online-Findmittel verwendet wird. Die einzelnen Informationspakete werden im Rahmen des OAIS als digitale Objekte verstanden. Sie bestehen immer aus Daten und beschreibenden und ggf. ergänzenden, repräsentativen Zusatzinformationen.

Jedes Informationspaket enthält erstens inhaltliche Informationen (Content Information), die aus den übernommenen, ggf. aufbereiteten Ursprungsdaten und der beschreibenden Repräsentationsinformation bestehen, und zweitens so genannte „Informationen zur Beschreibung der Aufbewahrungsform“ (Preservation Description Information (PDI)), die erklären, was an Technik und welche Verfahren auf die Inhaltsinformation angewandt wurden, also wie sie ver-

ändert wurden und welche Technik und welche Verfahren benötigt werden, um sie zu sichern, sie eindeutig zu identifizieren, sie in ihren Kontext einzuordnen und für die Zukunft nutzbar zu machen. Die Preservation Description enthält Informationen, die die dauerhafte Aufbewahrung beschreibt, sie besteht wiederum aus vier Elementen.

Erstes Element ist die Provenienz, hier werden also die Quelle der Inhaltsinformation seit deren Ursprung und ihre weitere Entwicklung, also ihr Entstehungs- und Entwicklungsprozess, beschrieben.

Zweites Element ist der Kontext, wo die Verbindung einer konkreten Inhaltsinformation mit anderen Informationen außerhalb des jeweiligen Informationspakets nachvollziehbar gehalten wird.

Drittes Element sind Beziehungen (References), wo über ein System von eindeutigen Bezeichnern (unique identifiers) die Inhaltsinformationen mit den auf sie bezogenen Metadaten sowie anderen Inhaltsinformationen eindeutig identifizierbar und eindeutig unterscheidbar gemacht werden.

Viertes Element sind Informationen zur Stabilisierung (fixity), damit die Inhaltsinformationen vor nicht erfasster Veränderung bewahrt werden können.

4.2.3 Das Funktionsmodell des OAIS

Es sind sechs Aufgabenbereiche (vgl. Abbildung 1), die im Rahmen des skizzierten Standards beschrieben werden:

1. Datenübernahme (Ingest)
2. Datenaufbewahrung (Archival Storage)
3. Datenmanagement
4. Systemverwaltung
5. Planung der Langzeitarchivierung (Preservation Planning)
6. Zugriff (Access)

SIP Submission Information Package = die digitalen Ressourcen, welche die aufbewahrenden Institutionen übernehmen.

AIP Archival Information Package = vom Langzeitarchiv mit Metadaten ergänzte digitale Objekte. In dieser Form werden die digitalen Objekte langfristig aufbewahrt.

DIP Dissemination Information Package = in dieser Form werden die digitalen Objekte je nach rechtlichen Bedürfnissen generiert und zur Verfügung gestellt.

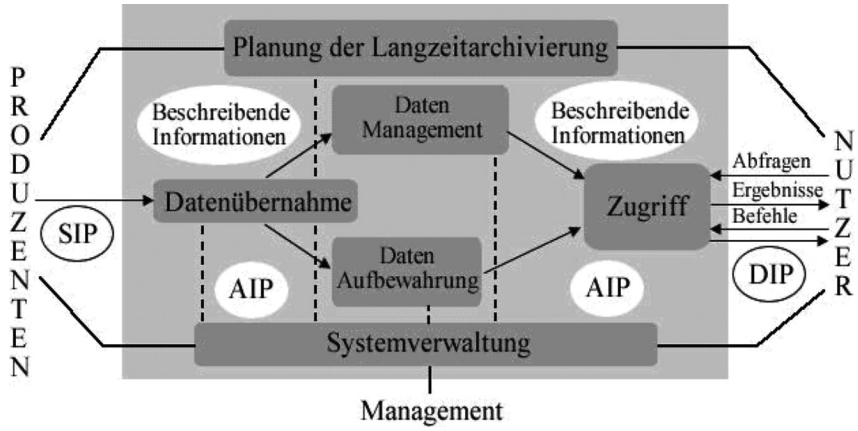


Abbildung 1: Das Funktionsmodell des OAIS

Im Bereich Ingest geht es um die Übernahme des digitalen Archivguts. Zunächst wird die Vorbereitung der Einlagerung im Archiv vorzunehmen sein, dazu gehört etwa auch die Bereitstellung der notwendigen technischen Kapazitäten und die Kontaktaufnahme mit dem Produzenten. Ein weiterer Aspekt, der ganz entscheidend ist, ist die Qualitätssicherung der Submission Information Packages, d.h. ihre Prüfung auf Lesbarkeit, Verständlichkeit und korrekten Kontext und dann die Herstellung der archivischen Informationspakete (AIP), die mit den Formaten und Standards des jeweils aufbewahrenden Archivs übereinstimmen. Der Analyse, Sicherung und ggf. Verbesserung der Datenqualität kommt im digitalen archivischen Vorfeld eine Schlüsselrolle zu, hier wird aber auch erstmalig verändernd eingegriffen. Das OAIS geht davon aus, dass digitale Archive aus ganz unterschiedlichen Systemumgebungen SIPs in einer Vielzahl von unterschiedlichen Formaten einfach übernehmen müssen und diese erst bei der digitalen Archivierung, also bei der Einlagerung ins digitale Magazin, zu nach einheitlichen Standards aufgebauten und zu generierenden AIPs umformen. Zum Bereich Übernahme gehört auch die Erstellung der notwendigen Erschließungsinformationen für die Erschließungsdatenbank des digitalen Archivs und erste planende Maßnahmen, die das regelmäßige Update des Datenspeichers und das dazu notwendige Datenmanagement organisieren.

Der zweite Teil „Achival Storage“ umfasst den digitalen Speicher, seine Organisation und seinen Aufbau im engeren Sinne. Hier werden die AIPs vom Übernahmebereich in Empfang genommen und eingelagert und es wird dafür gesorgt, dass regelmäßig gewartet und die Wiederauffindbarkeit der archivischen Informationspakete überprüft wird. Dazu gehört der Aufbau einer

technischen Lagerungshierarchie und die regelmäßige systematische Erneuerung der im jeweiligen Archiv standardisiert verwendeten Datenträger, sowie das so genannte Refreshing, d.h. die Überprüfung der verwendeten Datenträger auf ihre Lesbarkeit und die Verständlichkeit der gespeicherten AIP. In diesem Zusammenhang ist darauf zu verweisen, das OAIS ausdrücklich die Vorteile einer redundanten Archivierung auf zwei verschiedenen Informationsträgern hervorhebt.

Im Bereich Datenmanagement geht es um die Wartung und das Zugänglichhalten der Verzeichnungsinformationen und ihre kontinuierliche Ergänzung und Aufbereitung, dann aber auch das Verwalten verschiedener Archivdatenbanken und auch in diesem Bereich die Ausführung von verschiedenen Datenbank-Updates zur Sicherung von Lesbarkeit, Verständlichkeit und Nutzbarkeit.

Punkt vier umfasst das Management des OAIS. Management bezieht sich auf die Beziehungen zwischen Archivaren und Nutzern auf der einen Seite und dem Software/Hardware-System auf der anderen. Beschrieben werden alle Regelungen zur Zuständigkeit für die Arbeitsvorgänge im Archivssystem, wozu auch gehört, dass das, was automatisierbar ist, von den Vorgängen getrennt wird, die von Menschen erledigt werden müssen. Ebenso der Bereich Qualitätssicherung ist hier eingeordnet. Auch das Aushandeln von Verträgen zur Übergabe und zur Nutzung und die Prüfung der Informationspakete sowie das Unterhalten von jeweils verwendeten Hard- und Softwarelösungen gehörten natürlich zum Bereich des Managements im Open Archival Information System.

Der fünfte Teilbereich, der Bereich der Planung der Langzeitarchivierung im digitalen Archiv (Preservation Planning) befasst sich nicht nur mit der Sicherstellung des reibungslosen Informationszugangs in der Gegenwart, sondern ist vielmehr auf die Zukunft gerichtet. Es geht nämlich darum, Empfehlungen abzugeben, in welchen Zeitzyklen Updates vorgenommen werden müssen und in welchen Zyklen eine Migration der in einem Standardformat aufbewahrten elektronischen Aufzeichnungen in ein anderes neues Format vorgenommen werden müssen. Das heißt, eine ständige Überwachung im Bereich der Veränderung der Technologie gehört hier unabdingbar dazu. Aber auch der Blick auf den Benutzer und Veränderungen von Nutzungsgewohnheiten spielt hierbei eine Rolle. Preservation Planning umfasst demzufolge auch die Erstellung von Vorlagen (Templates) für die Information Packages und die Entwicklung einer Migrationsstrategie im Archiv.

Der sechste und abschließende Bereich Zugriff (Access) befasst sich mit der Unterstützung der Benutzer beim Auffinden der entsprechenden elektronischen Informationen. Hier werden Anfragen entgegengenommen, Zugangsberechtigungen koordiniert und dann den jeweiligen Benutzergruppen die für

sie nutzbaren Dissemination Information Packages, also Nutzungsinformationspakete, generiert und verteilt. Neben diesen fachlich ausgerichteten Aufgabenbereichen gehört natürlich auch ein Bereich der Verwaltung von OAIS als Gesamtsystem zum Betrieb und Unterhalt dazu, gewissermaßen die „Zentralabteilung" des digitalen Archivs. Besondere Bedeutung hat dabei die Verwaltung der OAIS-Software, die nötig ist, um das Archiv überhaupt betreiben zu können. Dazu gehören der Aufbau eines funktionstüchtigen, aber auch geschützten Netzwerks, und die regelmäßige Überprüfung und Verbesserung der Sicherheit des OAIS, um die in ihm enthaltenen Informationen vor unberechtigtem Zugang zu schützen.

Das OAIS setzt vollständig auf eine Migrationsstrategie als die derzeit von den Funktionen und der Technik her am besten beherrschbaren Strategie, selbst wenn es anderen Archivierungstechniken (z.B. Emulation) gegenüber offen ist. Migration wird im Sinne des OAIS in vier Bereiche systematisch zergliedert: erstens den Bereich des "Refreshment", des Wiederauffrischens mit dem Ziel, die Lesbarkeit der Datenträger zu sichern. Refreshment ist vor allen Dingen im Rahmen der AIPs, aber auch im Bereich der SIPs notwendig, damit überhaupt eine Übernahme möglich ist. Zum Refreshment tritt zweitens die „Replication", bei der regelmäßig der Kontext der verschiedenen Informationssysteme überprüft wird: Bestehende Verknüpfungen oder im Rahmen der Generierung von AIPs im Archiv hergestellte Verknüpfungen werden auf ihre Funktionstüchtigkeit und darauf überprüft, ob sie logisch schlüssig und verständlich sind. Ggf. ist drittens ein „Repackaging", also eine Art von digitaler Umbettung nötig, damit die bestehenden Verknüpfungen wieder funktionstüchtig sind oder ggf. neue Verknüpfungen erstellt werden (etwa dann, wenn vom Produzenten neue SIPs übernommen und zu AIPs umgeformt werden). Zum Schluss gehört auch die Transformation, d.h. die Übertragung auf neue, für einen bestimmten Zeitraum als tauglich erkannte Speichermedien, dazu. Hier wird im Rahmen des OAIS ein ganz zentraler Punkt angesprochen. Eine dauerhafte Lösung für die Langfristspeicherung, d.h. für die technische Sicherung der Zugänglichkeit wird auch in Zukunft nicht zu erwarten sein, sondern zur Archivierung digitaler Unterlagen wird es ab sofort gehören, immer mit den gegenwärtig zum technischen Standard gehörenden Informationsträgern leben zu müssen, die eine nur beschränkte Haltbarkeit haben und in Zukunft regelmäßig durch neue Formen von Informationsträgern ersetzt werden müssen. Es soll hier nur angedeutet werden, dass dieser Sachverhalt für eine Kostenplanung eines digitalen Archivs von entscheidender Bedeutung sein wird, weil nämlich neben eine Migration, die der Sicherung des Zugangs dient, auch eine solche treten wird, die durch technische Innovationen im Hard- und Softwarebereich und eine weitere

durch Veränderungen im Vorfeld des Archivs bedingt ist: Mit der Technik von gestern lassen sich digitale Objekte, die aus den gegenwärtigen Produktionssystemen stammen, nicht archivieren und langfristig zugänglich erhalten. Im Rahmen des OAIS verkennt man aber auch nicht, dass durch die skizzierte Migrationsstrategie Datenverluste möglich sind. Besonders im Bereich des Repackaging und der Transformation können diese Datenverluste auftreten. Man sieht aber im Augenblick noch keine realisierungsfähige technische Lösung, die diese Verluste vermeiden könnten.

4.2.4 Akzeptanz des OAIS-Modells

Das OAIS wird mittlerweile weltweit von Initiativen zur Langzeitarchivierung digitaler Ressourcen als Referenzmodell wahrgenommen und akzeptiert. Im Jahr 2002 wurde von der Niederländischen Nationalbibliothek in Den Haag der erste Prototyp eines digitalen Archivsystems (der gemeinsam mit IBM entwickelt wurde) in Dienst gestellt, das digitale Publikationen zugänglich halten soll. Dabei wurde das OAIS gezielt als Referenzmodell eingesetzt. Die Lösung ist großrechnerbasiert (IBM RS 6000S Winterhawk 2) und umfasst einen „Storage Server“ mit 3,4 Tbyte Kapazität, sowie ein System redundanter Speicherung auf Optischen Medien (3x 1,3 Tbyte Kapazität) und Bandspeicherbibliotheken mit insgesamt 12 Tbyte Kapazität.

Das nationale Datenarchiv Großbritanniens (NDAD) hat seine Routinen und Prozeduren auf das OAIS umgestellt, und auch das australische Nationalarchiv orientiert sich im Rahmen des PANDORA-Projektes am OAIS.

Das amerikanische Nationalarchiv (NARA) hat die OAIS-Modellierung als Grundlage für die groß angelegte Ausschreibung zur Entwicklung des ehrgeizigen ERA-Systems (Electronic Records Archives) verwendet.

Standardisierungsaktivitäten für technische Metadaten zur Langzeiterhaltung und Kriterien für vertrauenswürdige digitale Archive verwenden Terminologie, Objekt- und Funktionsmodell von OAIS.

Quellenangaben

CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS)*. Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2003 http://public.ccsds.org/publications/archive/650x0b1.pdf

Hogde, Gail M. (2002): Best Practices for Digital Archiving. In: D-LIB-Magazine, Vol.6 No.1, January 2000, S.8. http://www.dlib.org/dlib/january00/01hodge.html

5 Vertrauenswürdigkeit von digitalen Langzeitarchiven

5.1 Einführung

Susanne Dobratz und Astrid Schoger

Gedächtnisorganisationen wird gemeinhin beim Umgang mit ihren Beständen großes Vertrauen entgegengebracht - insbesondere beim Erhalt des ihnen anvertrauten kulturellen, wissenschaftlichen und gesellschaftlichen Erbes. Wissenschaftliche Erkenntnisse, historische Dokumente und kulturelle Leistungen liegen in stark zunehmendem Maße - und häufig ausschließlich - in digitaler Form vor. Diese spezielle Medienform stellt Gedächtnisorganisationen vor neue Herausforderungen und wirft die berechtigte Frage auf, ob sie auch im digitalen Bereich vertrauenswürdig handeln.

Das Konzept der Vertrauenswürdigkeit digitaler Langzeitarchive, die Kriterienkataloge, Checklisten und Werkzeuge helfen archivierenden Einrichtungen sowie Produzenten und Nutzern die Qualität und Nachhaltigkeit der Langzeitarchivierung zu bewerten und zu verbessern.

5.2 Grundkonzepte der Vertrauenswürdigkeit und Sicherheit

Susanne Dobratz und Astrid Schoger

Der Begriff der Vertrauenswürdigkeit digitaler Langzeitarchive wird von bewährten Konzepten der Vertrauenswürdigkeit von IT-Systemen abgeleitet. Im Internet Security Glossary[1] wird Vertrauenswürdigkeit (engl. trustworthiness) als die Eigenschaft eines Systems definiert, gemäß seinen Zielen und Spezifikationen zu operieren (d.h. es tut genau das, was es zu tun vorgibt) und dies auch in geeigneter Weise glaubhaft zu machen (z.B. durch eine formale Analyse). Die Common Criteria[2] führen Vertrauenswürdigkeit folgendermaßen ein:

„Werte" an deren Erhaltung „Eigentümer" Interesse haben, sind durch „Risiken"[3] bedroht. Zur Minimierung dieser Risiken werden „Gegenmaßnahmen" eingesetzt. Die Prüfung und Bewertung der eingesetzten Maßnahmen erbringt den Nachweis der „Vertrauenswürdigkeit". Vgl. dazu nachfolgende Abbildungen.

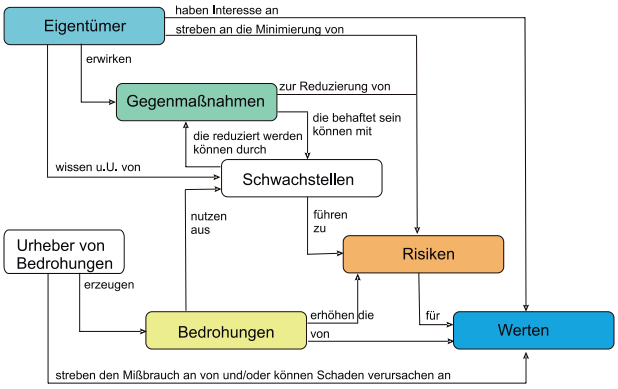


Abbildung 1: Konzept der Bedrohungen und Risiken gemäß Common Criteria (Quelle: BSI 2006)

1 Network Working Group (2000): "In discussing a system or system process or object, this Glossary (and industry usage) prefers the term „trusted" to describe a system that operates as expected, according to design and policy. When the trust can also be guaranteed in some convincing way, such as through formal analysis or code review, the system is termed „trustworthy";"

2 BSI (2006)

3 Vgl. dazu auch Howard (1998)

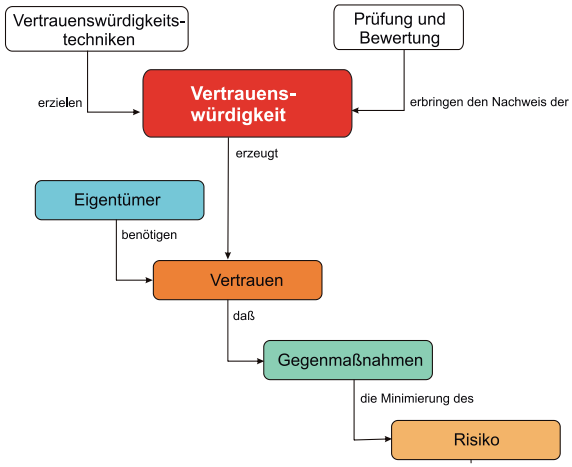


Abbildung 2: Vertrauenswürdigkeitskonzept gemäß den Common Criteria (Quelle: BSI 2006)

Ziel der digitalen Langzeitarchivierung ist der Erhalt der Informationen, die durch digitale Objekte repräsentiert sind.

Gemäß OAIS[4] wird unter einem digitalen Langzeitarchiv eine Organisation (bestehend aus Personen und technischen Systemen) verstanden, die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit zum Zwecke der Nutzung durch eine bestimmte Zielgruppe übernommen hat. Dabei bedeutet „Langzeit" über Veränderungen in der Technik (Soft- und Hardware) hinweg und auch unter Berücksichtigung möglicher Änderungen der Zielgruppe.

Aus dem Ziel der Langzeitarchivierung lassen sich folgende zentrale Aufgaben eines digitalen Langzeitarchivs ableiten: Aufnahme (Ingest), Archivablage (Archival Storage), Nutzung (Access); ferner unterstützende Aufgaben wie das Datenmanagement und die Administration des Gesamtsystems. Besondere Bedeutung kommt der strategischen Planung (Preservation Planning) und Durchführung der Langzeiterhaltungsmaßnahmen[5], die die Langzeitverfügbarkeit und Interpretierbarkeit (d.h. der Rekonstruierbarkeit der darin enthaltenen Informationen) sicherstellen, zu. [6]

Diese Aufgaben stellen die Grundlage für die Spezifikation von Anforderungen an digitale Langzeitarchive dar, wie bereits 2002 im Bericht[7] der RLG/

4 Vgl. dazu CCSDS (2002) sowie Kapitel 4 dieses Handbuchs
5 Vgl. dazu Kapitel 8 dieses Handbuchs
6 Vgl. dazu das Funktionsmodell des OAIS-Referenzmodells
7 RLG/OCLC (2002)

OCLC Working Group on Digital Archive Attributes aufgeführt. Die RLG-NARA Task Force on Digital Repository Certification hat 2007 (als Entwurf zur öffentlichen Kommentierung bereits 2006) eine Liste von Kriterien „Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC)“[8] erarbeitet, die ein vertrauenswürdiges digitales Langzeitarchiv erfüllen muss. Diese Liste dient der Orientierung beim Auf- und Ausbau digitaler Langzeitarchive und kann als Checkliste auch zur Selbstevaluierung sowie zum externen Audit eingesetzt werden.

nestor hat unter Berücksichtigung nationaler Ansätze und Arbeitsergebnisse wie des „DINI-Zertifikats für Dokumenten- und Publikationsserver“[9] sowie den oben genannten internationalen Arbeiten Kriterien entwickelt, die den speziellen Bedürfnissen der deutschen Gedächtnisorganisationen Rechnung tragen. Diese wurden zunächst im Sommer 2006 als Entwurf zur öffentlichen Kommentierung publiziert und dank der vielfältigen Rückmeldungen der Anwender gründlich überarbeitet und liegen nun in Version 2 vor.[10]

Eine Prüfung und Bewertung digitaler Langzeitarchive gemäß dieser Kriterienkataloge kann somit den Nachweis der Vertrauenswürdigkeit erbringen. Die Grundprinzipien der Kriterienkataloge sowie deren Inhalte werden in Kapitel 5.4 anhand des nestor-Kriterienkataloges genauer erörtert.

Im Sinne der Langzeitarchivierung stellen Informationen den zu erhaltenden „Wert“ dar. Informationen, die durch digitale Objekte repräsentiert werden, sind bedroht durch Einbußen in ihrer Integrität, Authentizität und Vertraulichkeit sowie den gänzlichen Verlust der Verfügbarkeit und Nutzbarkeit. Diese Eigenschaften bilden eine Teilmenge des Gesamtkonzeptes Sicherheit in der Informatik, wie sie u.a. in Steinmetz (2002) beschrieben sind:

- **Integrität**: sagt aus, ob die digitalen Objekte unverändert vorliegen,
- **Authentizität**: bezieht sich auf die Echtheit der digitalen Objekte, insbesondere den Aspekt der Nachweisbarkeit der Identität des Erstellers (Urhebers, Autors),
- **Vertraulichkeit**: bezieht sich darauf, dass unberechtigten Dritten kein Zugang zu den digitalen Objekten gewährleistet wird,
- **Verfügbarkeit**: bezieht sich auf den Aspekt der Zugänglichkeit zum digitalen Objekt.

8 RLG/NARA (2007)
9 DINI (2007)
10 nestor AG Vertrauenswürdige Archive - Zertifizierung (2008)

Im Rahmen des EU-Projektes DigitalPreservationEurope (DPE) in Zusammenarbeit mit dem Digital Curation Centre (DCC) wurde das Tool „Digital Repository Audit Method based on Risk Assessment (DRAMBORA)“[11] zur Selbstevaluierung entwickelt, das die Risikoanalyse als Methode einsetzt. Ausgehend von den Zielen eines digitalen Langzeitarchivs müssen zunächst die Aktivitäten spezifiziert und die damit verbundenen Werte identifiziert werden. In einem weiteren Schritt werden dann die Risiken aufgedeckt und die zu deren Minimierung eingesetzten Maßnahmen bewertet.

Somit wird ein anderer Weg zum Nachweis der Vertrauenswürdigkeit beschrieben.

Internationale Kooperation, Standardisierung und Zertifizierung – 10 gemeinsame Prinzipien

Bevor ein international abgestimmtes Zertifizierungsverfahren für digitale Langzeitarchive entwickelt werden kann, ist es zunächst wichtig, einen internationalen Konsens über die Evaluierungskriterien zu finden. Ferner müssen aus den Erfahrungen mit der Anwendung der Kriterienkataloge und Evaluierungstools Bewertungsmaßstäbe für unterschiedliche Typen von digitalen Langzeitarchiven ausgearbeitet werden.

Wesentliche Vertreter des Themas Vertrauenswürdigkeit auf internationaler Ebene - Center for Research Libraries (CRL), Digital Curation Centre (DCC), Projekt DigitalPreservationEurope (DPE) sowie nestor haben 10 gemeinsame Prinzipien[12] herausgearbeitet, die den oben genannten Kriterienkatalogen und Audit Checklisten zu Grunde liegen. Diese stellen die Grundlage der weiteren inhaltlichen Zusammenarbeit dar. Die 10 Kriterien lauten wie folgt[13]:

1. Das digitale Langzeitarchiv übernimmt die Verantwortung für die dauerhafte Erhaltung und kontinuierliche Pflege der digitalen Objekte für die identifizierten Zielgruppen.
2. Das digitale Langzeitarchiv belegt die organisatorische Beständigkeit (auch in den Bereichen Finanzierung, Personalausstattung, Prozesse), um seine Verantwortung zu erfüllen.
3. Das digitale Langzeitarchiv verfügt über die erforderlichen Rechte (per Vertrag oder Gesetz), um seine Verantwortung zu erfüllen.
4. Das digitale Langzeitarchiv besitzt ein effektives und effizientes Geflecht von Grundsätzen (policy).

11 DCC/DPE (2008)

12 CRL/DCC/DPE/nestor (2007)

13 nestor-Übersetzung

5. Das digitale Langzeitarchiv erwirbt und übernimmt digitale Objekte auf der Grundlage definierter Kriterien gemäß seinen Verpflichtungen und Fähigkeiten.
6. Das digitale Langzeitarchiv stellt die Integrität, Authentizität und Nutzbarkeit der dauerhaft aufbewahrten Objekte sicher.
7. Das digitale Langzeitarchiv dokumentiert alle Maßnahmen, die während des gesamten Lebenszyklus auf die digitalen Objekte angewendet werden, durch angemessene Metadaten.
8. Das digitale Langzeitarchiv übernimmt die Bereitstellung der digitalen Objekte.
9. Das digitale Langzeitarchiv verfolgt eine Strategie zur Planung und Durchführung von Langzeiterhaltungsmaßnahmen.
10. Das digitale Langzeitarchiv besitzt eine angemessene technische Infrastruktur zur dauerhaften Erhaltung und Sicherung der digitalen Objekte.

Sowohl die Kriterienkataloge als auch das Verfahren DRAMBORA werden zur Zeit der nationalen sowie internationalen Standardisierung zugeführt. Im Rahmen des DIN kümmert sich der neu gegründete Arbeitskreis „Vertrauenswürdige digitale Langzeitarchive“ im Rahmen des NABD15 „Schriftgutverwaltung und Langzeitverfügbarkeit digitaler Informationsobjekte“ um die Vorbereitung einer deutschen Norm für diesen Bereich. Dieser Arbeitskreis arbeitet eng zusammen mit den entsprechenden Ausschüssen der ISO und zwar dem TC46/SC11, der mit dem Entwurf „Risk assessment for records systems” die Normungsarbeit an DRAMBORA übernommen hat sowie dem TC20/SC13, der „TRAC: Trustworthy Repositories Audit and Certification: Criteria and Checklist“ einer Normierung zuführt.

Die Anwendung von Konzepten der IT-Sicherheit wie Hashfunktionen, Fingerprintingverfahren und digitalen Signaturen, kann bestimmte Risiken, die den Erhalt digitaler Objekte bedrohen, minimieren, insbesondere jene, welche die Integrität, Authentizität und Vertraulichkeit digitaler Objekte betreffen. Von besonderer Bedeutung für die Langzeitarchivierung ist der „Langzeit“-Aspekt, so dass bei allen eingesetzten Methoden die Nachhaltigkeit besonders geprüft werden muss. Diese Verfahren werden im nachfolgenden Kapitel dargestellt.

Literatur

Network Working Group (2000): *Internet Security Glossary. Request for Comments: 2828* http://www.ietf.org/rfc/rfc2828.txt

BSI Bundesamt für Sicherheit in der Informationstechnik (2006): *Gemeinsame Kriterien für die Prüfung und Bewertung der Sicherheit von Informationstechnik, Common Criteria* V 3.1, http://www.bsi.de/cc/

Howard, John D. / Longstaff, Thomas A. (1998): *A Common Language for Computer Security Incidents*. SANDIA Reports SAND98-8667. Albuquerque, New Mexico : Sandia National Laboratories http://www.cert.org/research/taxonomy_988667.pdf

CCSDS (Consultative Committee for Space Data Systems) (2002*): Reference Model for an Open Archival Information System (OAIS). Blue Book.* http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf entspricht ISO 14721:2003

RLG-NARA Task Force on Digital Repository and Certification, (2007): *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC).* http://www.crl.edu/PDF/trac.pdf

RLG/OCLC Working Group on Digital Archive Attributes (2002): *Trusted Digital Repositories: Attributes and Responsibilities*, http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

DINI Deutsche Initiative für Netzwerkinformation / AG Elektronisches Publizieren (2007): *DINI-Zertifikat für Dokumenten- und Publikationsservice 2007*. DINI-Schriften 3. http://nbn-resolving.de/urn:nbn:de:kobv:11-10079197

nestor Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (2008): *nestor-Kriterien: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive.* Version 2. Frankfurt am Main : nestor http://nbn-resolving.de/urn:nbn:de:0008-2008021802

Steinmetz, Ralf (2000) : *Multimedia-Technologie: Grundlagen, Komponenten und Systeme*, 3. Auflage , Berlin, Heidelberg, New York : Springer

DCC Digital Curation Centre / DPE Digital Preservation Europe (2008): *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), interactive,* http://www.repositoryaudit.eu/

CRL Center for Research Libraries / DCC Digital Curation Centre / DPE Digital Preservation Europe / nestor (2007): *Core Requirements for Digital Archives*, http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92

5.3 Praktische Sicherheitskonzepte

Siegfried Hackel, Tobias Schäfer und Wolf Zimmer

5.3.1 Hashverfahren und Fingerprinting

Ein wichtiger Bestandteil praktischer Sicherheitskonzepte zum Schutz der Integrität und Vertraulichkeit digitaler Daten sind Verschlüsselungsinfrastrukturen auf der Basis sogenannter kryptographisch sicherer Hashfunktionen. Mit Hilfe kryptographisch sicherer Hashfunktionen werden eindeutige digitale „Fingerabdrücke" von Datenobjekten berechnet und zusammen mit den Objekten versandt oder gesichert. Anhand eines solchen digitalen „Fingerabdrucks" ist der Empfänger oder Nutzer der digitalen Objekte in der Lage, die Integrität eines solchen Objektes zur prüfen, bzw. unautorisierte Modifikationen zu entdecken.

Hashfunktionen werden in der Informatik seit langem eingesetzt, bspw. um im Datenbankumfeld schnelle Such- und Zugriffsverfahren zu realisieren. Eine Hashfunktion ist eine mathematisch oder anderweitig definierte Funktion, die ein Eingabedatum variabler Länge aus einem Urbildbereich (auch als „Universum" bezeichnet) auf ein (in der Regel kürzeres) Ausgabedatum fester Länge (den Hashwert, engl. auch message digest) in einem Bildbereich abbildet. Das Ziel ist, einen „Fingerabdruck" der Eingabe zu erzeugen, die eine Aussage darüber erlaubt, ob eine bestimmte Eingabe aller Wahrscheinlichkeit nach mit dem Original übereinstimmt.

Da der Bildbereich in der Regel sehr viel kleiner ist als das abzubildende „Universum", können so genannte „Kollisionen" nicht ausgeschlossen werden. Eine Kollision wird beobachtet, wenn zwei unterschiedliche Datenobjekte des Universums auf den gleichen Hashwert abgebildet werden.

Für das Ziel, mit einer Hashfunktion einen Wert zu berechnen, der ein Datenobjekt eindeutig charakterisiert und damit die Überprüfung der Integrität von Daten ermöglicht, sind derartige Kollisionen natürlich alles andere als wünschenswert. Kryptographisch sichere Hashfunktionen H, die aus einem beliebig langen Wort M aus dem Universum von H einen Wert H(M), den Hashwert fester Länge erzeugen, sollen daher drei wesentliche Eigenschaften aufweisen:

1. die Hashfunktion besitzt die Eigenschaften einer effizienten Ein-Weg-Funktion, d.h. für alle M aus dem Universum von H ist der Funktionswert h = H(M) effizient berechenbar und es gibt kein effizientes Verfahren, um aus dem Hashwert h die Nachricht zu berechnen[14],
2. es ist - zumindest praktisch - unmöglich zu einem gegebenen Hashwert h = H(M) eine Nachricht M' zu finden, die zu dem gegebenen Hashwert passt (Urbildresistenz),
3. es ist - zumindest praktisch – unmöglich, zwei Nachrichten M und M' zu finden, die denselben Hashwert besitzen (Kollisionsresistenz).

Praktisch unmöglich bedeutet natürlich nicht praktisch ausgeschlossen, sondern bedeutet nicht mehr und nicht weniger, als dass es bspw. sehr schwierig ist, ein effizientes Verfahren zu finden, um zu einer gegebenen Nachricht M eine davon verschiedene Nachricht M' zu konstruieren, die denselben Hashwert liefert. Für digitale Objekte mit binären Zeichenvorräten Z = {0,1} lässt sich zeigen, dass für Hashfunktionen mit einem Wertbereich von 2n verschiedenen Hashwerten, beim zufälligen Ausprobieren von 2n/2 Paaren von verschiedenen Urbildern M und M' die Wahrscheinlichkeit einer Kollision schon größer als 50% ist.

Beim heutigen Stand der Technik werden Hashfunktionen mit Hashwerten der Länge n = 160 Bit als hinreichend stark angesehen.[15] Denn, selbst eine Schwäche in der Kollisionsresistenz, wie bereits im Jahre 2005 angekündigt[16], besagt zunächst einmal lediglich, dass ein Angreifer zwei verschiedene Nachrichten erzeugen kann, die denselben Hashwert besitzen. Solange aber keine Schwäche der Urbildresistenz gefunden wird, dürfte es für einen Angreifer mit einem gegebenen Hashwert und passendem Urbild immer noch schwer sein, ein zweites, davon verschiedenes Urbild zu finden, das zu diesem Hashwert passt.

Kern kryptographischer Hashfunktionen sind Folgen gleichartiger Kompressionsfunktionen K, durch die eine Eingabe M blockweise zu einem Has-

14 Obwohl die Ein-Weg-Funktionen in der Kryptographie eine wichtige Rolle spielen, ist nicht bekannt, ob sie im streng mathematischen Sinne eigentlich existieren, ihre Existenz ist schwer zu beweisen. Man begnügt sich daher zumeist mit Kandidaten, für die man die Eigenschaft zwar nicht formal bewiesen hat, für die aber derzeit noch keine effizienten Verfahren zur Berechnung der Umkehrfunktion bekannt sind.

15 Ein Rechner, der in der Lage ist, pro Sekunde den Hashwert zu einer Million Nachrichten zu berechnen, bräuchte 600.000 Jahre, um eine zweite Nachricht zu ermitteln, deren Hashwert mit einem vorgegebenen Hashwert der Länge 64 Bit übereinstimmt. Derselbe Rechner könnte allerdings in etwa einer Stunde irgendein Nachrichtenpaar mit gleichem Hashwert finden.

16 Schneier (2005)

hwert verarbeitet wird. Um Eingaben variabler Länge zu komprimieren, wendet man den Hashalgorithmus f iterierend an. Die Berechnung startet mit einem durch die Spezifikation des Hashalgorithmus festgelegten Initialwert $f(0):=I_0$. Anschließend gilt:

$$f(i) := K\,(f(i-1),M_i) \text{ mit } M = M_1, \ldots, M_n,\ i = 1, \ldots, n$$
$$H(M):= f(n) = h \text{ ist der Hashwert von } M$$

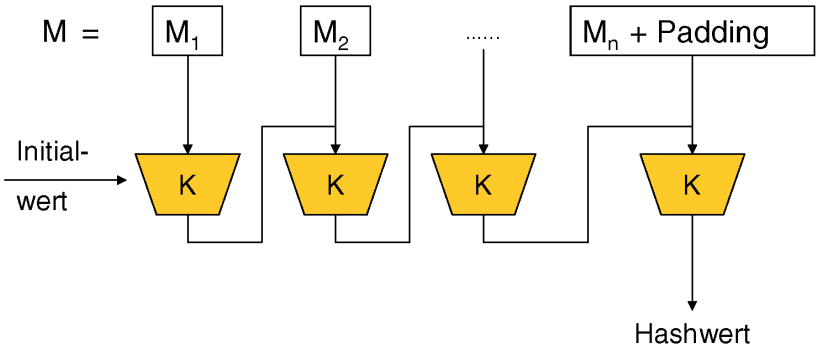


Abbildung 3: Allgemeine Arbeitsweise von Hashfunktionen (nach C. Eckert [17]*)*

Neben auf symmetrischen Blockchiffren, wie dem bereits 1981 durch das American National Standards Institute (ANSI) als Standard für den privaten Sektor anerkannten Data Encryption Standard (DES)[18], finden heute vor allem Hashfunktionen Verwendung, bei denen die Kompressionsfunktionen speziell für die Erzeugung von Hashwerten entwickelt wurden. Der bislang gebräuchlichste Algorithmus ist der Secure Hash Algorithm SHA-1 aus dem Jahre 1993.[19]

Der SHA-1 erzeugt Hashwerte von der Länge 160 Bits[20] und verwendet eine Blockgröße von 512 Bits, d.h. die Nachricht wird immer so aufgefüllt, dass die Länge ein Vielfaches von 512 Bit beträgt. Die Verarbeitung der 512-Bit Ein-

17 Eckert (2001)
18 Vgl. bspw. Schneier (1996)
19 Vgl. bspw. Schneier (1996)
20 Da nicht ausgeschlossen werden kann, dass mit der Entwicklung der Rechentechnik künftig auch Hashwerte von der Länge 160 Bit nicht mehr ausreichend kollisions- und urbildresistent sind, wird heute für sicherheitstechnisch besonders sensible Bereiche bereits der Einsatz der Nachfolger SHA-256, SHA-384 und SHA-512 mit Bit-Längen von jeweils 256, 384 oder 512 Bits empfohlen.

gabeblöcke erfolgt sequentiell, für einen Block benötigt SHA-1 insgesamt 80 Verarbeitungsschritte.

Merkle-Hashwertbäume

In der Kryptographie und Informatik werden Merkle-Bäume[21] eingesetzt, um über große Mengen von Daten oder Dokumenten einen zusammenfassenden Hashwert zu bilden.
Die Blätter eines Merkle-Baums sind die Hashwerte der Dokumente $H_i(d_i)$. Jeder Knoten im Baum wird als Hashwert $H(h_1 | h_2)$ seiner Kinder h_1 und h_2 gebildet. Dabei ist h_1 der Hashwert des Dokuments d_1 und h_2 der Hashwert des Dokuments d_2 und „|“ die Verkettung (Konkatenation) der beiden Hashwerte. Die Abbildung 4 zeigt ein Beispiel für einem binären Merkle-Hashbaum.

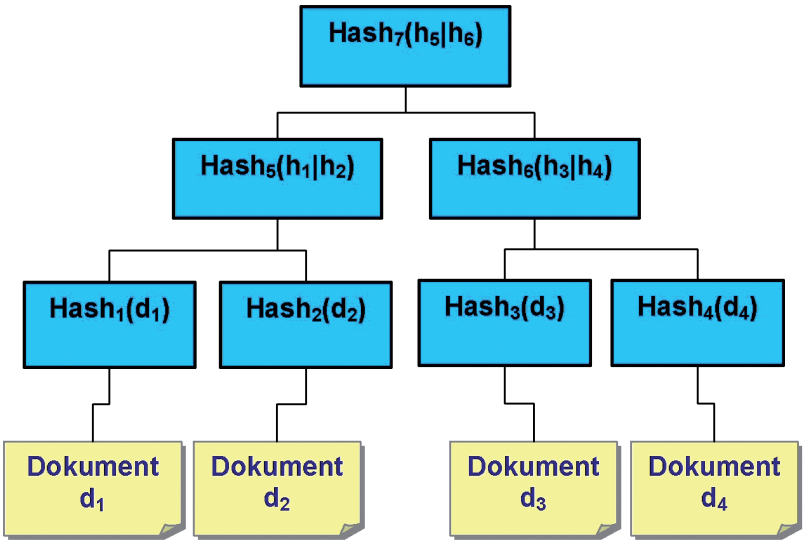


Abbildung 4: Binärer Merkle-Hashbaum.

Hash-Bäume können zum Nachweis der Integrität beliebiger Daten verwendet werden, die in oder zwischen Rechnern gespeichert, verarbeitet oder übertragen werden.

Die meisten Hash-Baum-Implementierungen sind binär, d.h. jeder Knoten besitzt zwei Kinder, es sind jedoch auch mehrere Kind-Knoten möglich.

21 Im Jahr 1979 von Ralph Merkle erfunden

Holt man für den obersten Hashwert (H_7 in Abbildung 4) einen Zeitstempeln ein, lässt sich beweisen, dass die Dokumente zum aktuellen Zeitpunkt existiert haben und seitdem nicht manipuliert wurden. Dieses Verfahren wurde im ArchiSig-Konzept[22] aufgegriffen und weiterentwickelt, um das Problem der nachlassenden Sicherheitseignung von Hash- und Signaturalgorithmen zu lösen. Dabei wird mit Hilfe von (akkreditierten) Zeitstempeln die vom deutschen Signaturgesetz geforderte Übersignatur von signierten Dokumenten vorgenommen.

5.3.2 Digitale Signatur

Elektronische Signaturen sind „Daten in elektronischer Form, die anderen elektronischen Daten beigefügt oder logisch mit ihnen verknüpft sind und die zur Authentifizierung im elektronischen Rechts- und Geschäftsverkehr dienen. Ihre Aufgabe ist die Identifizierung des Urhebers der Daten, d.h. der Nachweis, dass die Daten tatsächlich vom Urheber herrühren (Echtheitsfunktion) und dies vom Empfänger der Daten auch geprüft werden kann (Verifikationsfunktion). Beides lässt sich nach dem heutigen Stand der Technik zuverlässig am ehesten auf der Grundlage kryptographischer Authentifizierungssysteme, bestehend aus sicheren Verschlüsselungsalgorithmen sowie dazu passenden und personifizierten Verschlüsselungs-Schlüsseln (den so genannten Signaturschlüsseln) realisieren.

Die Rechtswirkungen, die an diese Authentifizierung geknüpft werden, bestimmen sich aus dem Sicherheitsniveau, das bei ihrer Verwendung notwendig vorausgesetzt wird. Dementsprechend unterscheidet das im Jahre 2001 vom deutschen Gesetzgeber veröffentlichte „Gesetz über Rahmenbedingungen für elektronische Signaturen und zur Änderung weiterer Vorschriften“[23], kurz Signaturgesetz (SigG), vier Stufen elektronischer Signaturen:

- „Einfache elektronische Signaturen“ gem. § 2 Nr. 1 SigG,
- „Fortgeschrittene elektronische Signaturen“ gem. § 2 Nr. 2 SigG,
- „Qualifizierte elektronische Signaturen“ gem. § 2 Nr. 3 SigG,
- „Qualifizierte elektronische Signaturen“ mit Anbieter-Akkreditierung gem. § 15 Abs. 1 SigG.

Mit Ausnahme der einfachen elektronischen Signaturen, denen es an einer verlässlichen Sicherheitsvorgabe völlig fehlt, wird das mit der Anwendung

22 Siehe www.archisig.de

23 BGBl I 876; BT-Drs 14/4662 und 14/5324

elektronischer Signaturen angestrebte Sicherheitsniveau grundsätzlich an vier Elementen festgemacht (§ 2 Nr. 2 SigG). Elektronische Signaturen müssen demnach

- ausschließlich dem Signaturschlüssel-Inhaber zugeordnet sein,
- die Identifizierung des Signaturschlüssel-Inhabers ermöglichen,
- mit Mitteln erzeugt werden, die der Signaturschlüssel-Inhaber unter seiner alleinigen Kontrolle halten kann und
- mit den Daten, auf die sie sich beziehen, so verknüpft sein, dass eine nachträgliche Veränderung der Daten erkannt werden kann.

Europaweit als Ersatz für die handschriftliche Unterschrift akzeptiert werden jedoch lediglich qualifizierte elektronische Signaturen. Für sie wird zusätzlich gefordert (§ 2 Nr. 3 SigG), dass sie

- auf einem zum Zeitpunkt ihrer Erzeugung gültigen qualifizierten Zertifikat beruhen und
- mit einer sicheren Signaturerstellungseinheit erzeugt werden.

Das Zertifikat übernimmt in diesem Fall die Authentizitätsfunktion, d.h. es bescheinigt die Identität der elektronisch unterschreibenden Person.[24] Sichere Signaturerstellungseinheiten sind nach dem Willen des Gesetzgebers Software- oder Hardwareeinheiten, die zur Speicherung und Anwendung des Signaturschlüssels dienen.[25]

Das Verfahren der digitalen Signatur basiert auf so genannten asymmetrischen kryptographischen Authentifizierungssystemen, bei denen jeder Teilnehmer ein kryptographisches Schlüsselpaar besitzt, bestehend aus einem ge-

---

24 Nach § 2 Nr. 6 SigG sind Zertifikate elektronische Bescheinigungen, mit denen Signaturschlüssel einer Person zugeordnet werden und die Identität einer Person bescheinigt wird. Für die Anwendung von Signaturverfahren von besonderer Bedeutung ist die Feststellung, dass „qualifizierte Zertifikate" nur auf natürliche Personen ausgestellt werden dürfen.

25 Das deutsche Signaturgesetz fordert, § 17 Abs. 1 SigG, dass sichere Signaturerstellungseinheiten vor unberechtigter Nutzung zu schützen sind. Nach § 15 Abs. 1 der Verordnung zur elektronischen Signatur (SigV) ist hierfür eine Identifikation „durch Besitz und Wissen oder durch Besitz und ein oder mehrere biometrische Merkmale" erforderlich. Da bislang keine Implementierungen biometrischer Verfahren bekannt sind, die die Anforderungen des Signaturgesetzes (vgl. Anlage 1 SigV) nachweislich erfüllen, werden für qualifizierte elektronische Signaturen in der Praxis immer Personal Identification Numbers (PIN) als Identifikationsdaten eingesetzt.

heimen privaten Schlüssel (private key, Kpriv) und einem öffentlichen Schlüssel (public key, Kpub).

Eine wesentliche Eigenschaft solcher asymmetrischer Authentifizierungssysteme ist, dass es praktisch unmöglich ist, den privaten Schlüssel aus dem öffentlichen Schlüssel herzuleiten, der öffentliche Schlüssel wird durch Anwendung einer sogenannten Einwegfunktion aus dem privaten Schlüssel berechnet. Der öffentliche Schlüssel kann daher in einem öffentlich zugänglichen Verzeichnis hinterlegt werden, ohne damit den privaten Schlüssel preiszugeben.

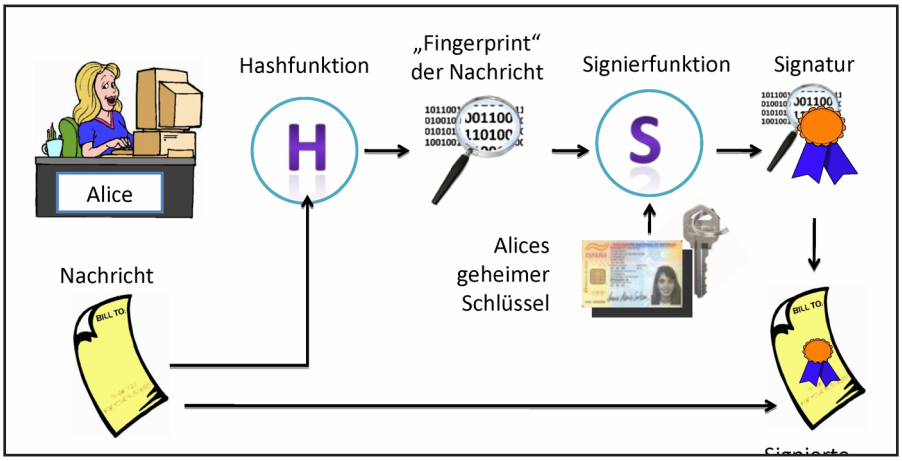


Abbildung 5: Digitale Signatur

Der Urheber, respektive Absender elektronischer Daten „unterschreibt“ nun seine Daten, indem er sie mit seinem geheimen, privaten Schlüssel verschlüsselt. Jeder, der die Daten empfängt, kann sie dann mit dem öffentlichen Schlüssel wieder entschlüsseln (s. Abbildung 5).

Unter der Voraussetzung, dass der öffentliche Schlüssel eindeutig und zuverlässig einer Person zugeordnet werden kann, bezeugt die Signatur folglich die Identität des Unterzeichners. Da die Signatur zudem das Ergebnis einer Verschlüsselungsoperation ist, sind die signierten Daten nachträglich auch nicht mehr veränderbar bzw. eine Änderung ist sofort erkennbar. Die Signatur kann auch nicht unautorisiert weiter verwendet werden, weil das Ergebnis der Verschlüsselungsoperation natürlich abhängig von den Daten ist. Geht man ferner davon aus, dass der private Signaturschlüssel nicht kompromittiert worden ist, kann der Absender der Daten die Urheberschaft auch nicht mehr zurückweisen, weil ausschließlich er selbst über den privaten Signaturschlüssel verfügt.

Technisch wäre natürlich eine Verschlüsselung der gesamten Daten (eines

Dokuments oder einer Nachricht) viel zu aufwändig. Aus diesem Grunde wird aus den Daten eine eindeutige Prüfsumme, ein Hashwert (s. dazu auch Kap. 5.3.1) erzeugt, dieser verschlüsselt („unterschrieben“) und den Originaldaten beigefügt. Der mit dem geheimen Schlüssel verschlüsselte Hashwert repräsentiert fortan die elektronische Signatur („Unterschrift“) der Originaldaten. Der Empfänger seinerseits bildet nach demselben Verfahren, d.h. mit demselben Hash-Algorithmus ebenfalls eine Prüfsumme aus den erhaltenen Daten und vergleicht sie mit der des Absenders. Sind die beiden Prüfsummen identisch, dann sind die Daten unverändert und stammen zuverlässig vom Inhaber des geheimen Schlüssels, denn nur er war in der Lage die Prüfsumme so zu verschlüsseln, dass sie mit dem zugehörigen öffentlichen Schlüssel auch entschlüsselt werden konnte.

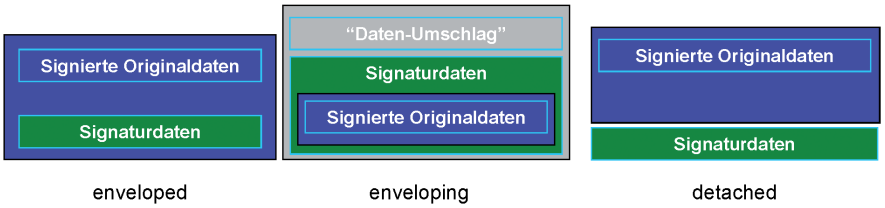


Abbildung 6: Hinzufügung der Signaturdaten

Die Hinzufügung der Signaturdaten zu den Originaldaten kann grundsätzlich auf folgende Weise geschehen (s. Abbildung 6):

- Enveloped („eingebettet“): die Signaturdaten sind als Element in den Originaldaten enthalten.
  Dieses Verfahren, auch als so genannte „Inbound-Signatur“ bezeichnet, wird vor allem bei der Signatur von PDF-Dokumenten und PDF-Formularen bspw. im Projekt ArchiSafe der Physikalisch-Technischen Bundesanstalt benutzt (s. a. Abb. 7).[26] Dabei werden die binären Signaturdaten direkt in das PDF-Dokument eingebettet und gemeinsam mit den Originaldaten im PDF-Format angezeigt. Mit dem neuen Adobe® Reader® (Version 8) ist der Empfänger der signierten Daten darüber hinaus imstande, unmittelbar eine Überprüfung der Integrität der angezeigten und signierten Daten vorzunehmen.
  Eingebettete Signaturen werden ebenso bei der Signatur von XML-Daten[27] verwendet und sollen zudem nun auch für den neuen XDOMEA

26 s. http://www.archisafe.de

27 1999 bis 2002 wurde der W3C-Standard für das Signieren von XML-Dokumenten am

Standard 2.0[28] spezifiziert werden. Da die Signatur eine binäre Zahlenfolge ist, lässt sie sich jedoch nicht direkt in ein XML-Dokument einbetten. Man codiert daher die binären Werte im Base64-Format (RFC 1521), um aus ihnen ASCII-lesbare Zeichen zu gewinnen. Die erhaltene Zeichendarstellung der Signatur findet sich schliesslich als SignatureValue in der XML-Signatur wieder[29].

- **Enveloping („umschließend"):** die Signaturdaten „umschließen" die Originaldaten. Diese Methode wird hauptsächlich für die Signatur von E-Mail Nachrichten oder reinen XML-Daten benutzt. Eine S/MIME Client-Anwendung, wie bspw. Microsoft Outlook, bettet in diesem Fall die Nachricht in einen signierten „Umschlag" ein.
- **Detached („getrennt"):** die Signaturdaten befinden sich außerhalb der Originaldaten in einer zusätzlichen, binären Signaturdatei. Diese Form, auch als „Outbound-Signatur" bezeichnet, wird standardmäßig für XML-Signaturen sowie die Signatur binärer Originaldaten eingesetzt. Ein separater Link in den Original-Daten oder zusätzlichen Beschreibungsdaten sorgt dann für die notwendige permanente Verknüpfung der Originaldaten mit den Signaturdaten.

Massachusetts Institute of Technology (MIT) entwickelt (XMLDSIG). Die XML Signatur Spezifikation (auch XMLDSig) definiert eine XML Syntax für digitale Signaturen. In ihrer Funktion ähnelt sie dem PKCS#7 Standard, ist aber leichter zu erweitern und auf das Signieren von XML Dokumenten spezialisiert. Sie findet Einsatz in vielen weiterführenden Web-Standards wie etwa SOAP, SAML oder dem deutschen OSCI. Mit XML Signaturen können Daten jeden Typs signiert werden. Dabei kann die XMLSignatur Bestandteil des XML Datenpakets sein (enveloped signature), die Daten können aber auch in die XML-Signatur selbst eingebettet sein (enveloping signature) oder mit einer URL adressiert werden (detached signature). Einer XML-Signatur ist immer mindestens eine Ressource zugeordnet, das heisst ein XML-Baum oder beliebige Binärdaten, auf die ein XML-Link verweist. Beim XML-Baum muss sichergestellt sein, dass es zu keinen Mehrdeutigkeiten kommt (zum Beispiel bezüglich der Reihenfolge der Attribute oder des verwendeten Zeichensatzes). Um dies erreichen zu können, ist eine so genannte Kanonisierung des Inhalts erforderlich. Dabei werden nach Maßgabe des Standards alle Elemente in der Reihenfolge ihres Auftretens aneinander gereiht und alle Attribute alphabetisch geordnet, so dass sich ein längerer UTF8-String ergibt (es gibt auch Methoden, die einen UTF16-String erzeugen). Aus diesem wird der eigentliche Hash-Wert gebildet beziehungsweise erzeugt man durch verschlüsseln den Signaturcode. So ist man wieder beim Standard-Verfahren für elektronische Signaturen (RFC 2437).

28 s. http://www.kbst.bund.de

29 Im Rahmen der Struktur eines XML-Dokuments lassen sich Subelemente explizit vom Signieren ausschliessen, so auch die Signatur selbst. Umgekehrt lassen sich beliebig viele Referenzen auflisten, die gemeinsam als Gesamtheit zu signieren sind.

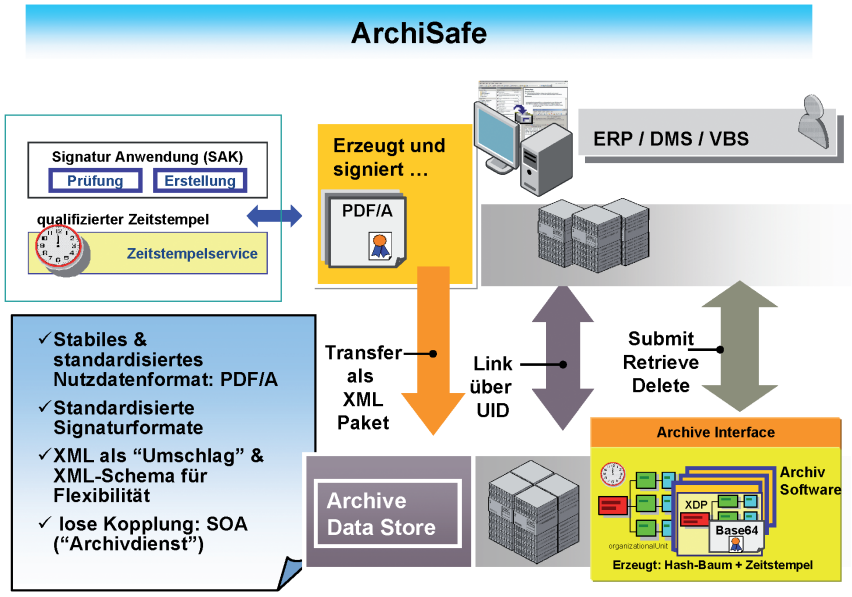


Abbildung 7: ArchiSafe – Rechts- und revisionssichere Langzeitspeicherung elektronischer Dokumente

Die Flexibilität der Hinzufügung von Signaturdaten zu Originaldaten basiert auf der als RFC 3852 – Cryptographic Message Syntax (CMS) im Juli 2004[30] durch die Internet Engineering Task Force (IETF) veröffentlichten Spezifikation sowie dem ursprünglich durch die RSA Laboratories veröffentlichten PKCS#7 (Public Key Cryptography Standard) Dokument in der Version 1.5. In beiden Dokumenten wird eine allgemeine Syntax beschrieben, nach der Daten durch kryptographische Maßnahmen wie digitale Signaturen oder Verschlüsselung geschützt, respektive Signaturdaten über das Internet ausgetauscht werden können. Die Syntax ist rekursiv, so dass Daten und Umschläge verschachtelt oder bereits chiffrierte Daten unterschrieben werden können. Die Syntax ermöglicht zudem, dass weitere Attribute wie z.B. Zeitstempel mit den Daten oder dem Nachrichteninhalt authentifiziert werden können und unterstützt eine Vielzahl von Architekturen für die Schlüsselverwaltung auf der Basis von elektronischen Zertifikaten.

30 Network Working Group (2004)

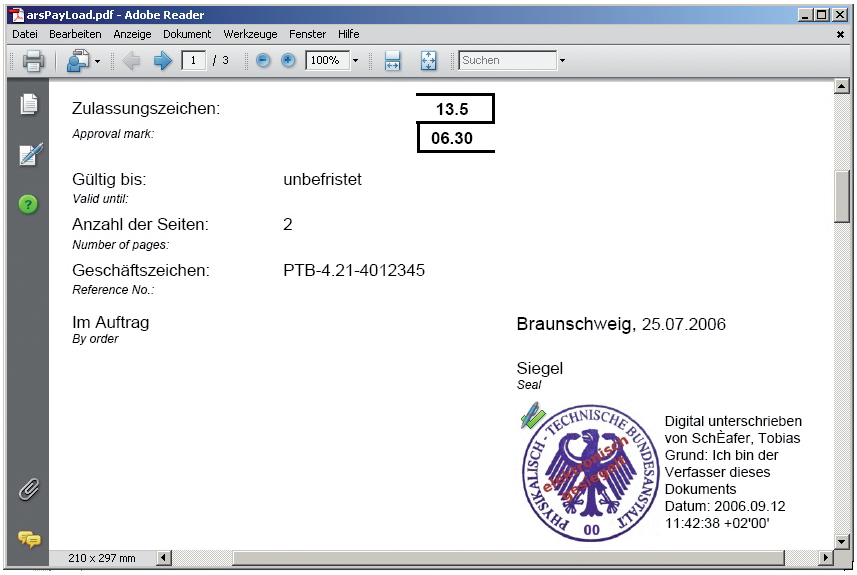


Abbildung 8: Digitale PDF-Signatur

Literatur

Eckert, Claudia (2001): *IT-Sicherheit: Konzepte – Verfahren – Protokolle*. München: Oldenbourg Wissenschaftsverlag

Network Working Group (2004): R. Hously: *Cryptographic Message Syntax (CMS) Request for Comment*s: 3852 http://www.ietf.org/rfc/rfc3852

Schneier, Bruce (1996): *Angewandte Kryptographie*, Bonn u.a.: Addison-Wesley Verl.

Schneier, Bruce (2005): *Schneier on Security. A blog covering security and security technology*. SHA-1 Broken (February 15, 2005) http://www.schneier.com/blog/archives/2005/02/sha1_broken.html

5.4 Kriterienkataloge für vertrauenswürdige digitale Langzeitarchive

Susanne Dobratz und Astrid Schoger

Verschiedene Organisationen und Initiativen (vgl. Kapitel 5.1) haben Anforderungen an vertrauenswürdige digitale Langzeitarchive formuliert. Diese Kriterien betreffen sowohl organisatorische als auch technische Aspekte, die erfüllt werden müssen, um der Aufgabe der Informationserhaltung gerecht werden zu können. Die Grundprinzipien der Kriterienkataloge sowie deren Inhalte werden nun in Kapitel 5.3 anhand des nestor-Kriterienkataloges genauer erörtert.

Grundprinzipien der Kriterienkataloge

Bei der Herleitung und für die Anwendung von Kriterien der Vertrauenswürdigkeit gelten folgende Grundprinzipien, die erstmals von der nestor Arbeitsgruppe „Vertrauenswürdige Archive – Zertifizierung" formuliert wurden:

Abstraktion: Ziel ist es, Kriterien zu formulieren, die für ein breites Spektrum digitaler Langzeitarchive angewendet werden können und über längere Zeit Gültigkeit behalten. Deshalb wird von relativ abstrakten Kriterien ausgegangen.

Dokumentation: Die Ziele, die Konzeption und Spezifikation sowie die Implementierung des digitalen Langzeitarchivs sind angemessen zu dokumentieren. Anhand der Dokumentation kann der Entwicklungsstand intern und extern bewertet werden. Eine frühzeitige Bewertung kann auch dazu dienen, Fehler durch eine ungeeignete Implementierung zu vermeiden. Insbesondere erlaubt es eine angemessene Dokumentation aller Stufen, die Schlüssigkeit eines digitalen Langzeitarchivs umfassend zu bewerten. Auch alle Qualitäts- und Sicherheitsnormen fordern eine angemessene Dokumentation.

Transparenz: Transparenz wird realisiert durch die Veröffentlichung geeigneter Teile der Dokumentation. Transparenz nach außen gegenüber Nutzern und Partnern ermöglicht diesen, selbst den Grad an Vertrauenswürdigkeit festzustellen. Transparenz gegenüber Produzenten und Lieferanten bietet diesen die Möglichkeit zu bewerten, wem sie ihre digitalen Objekte anvertrauen. Die Transparenz nach innen dokumentiert gegenüber den Betreibern, den Trägern, dem Management sowie den Mitarbeitern die angemessene Qualität des digitalen Langzeitarchivs und sichert die Nachvollziehbarkeit der Maßnahmen. Bei denjenigen Teilen der Dokumentation, die für die breite Öffentlichkeit nicht geeignet sind (z.B. Firmengeheimnisse, Informationen mit Sicherheitsbezug),

kann die Transparenz auf einen ausgewählten Kreis (z.B. zertifizierende Stelle) beschränkt werden. Durch das Prinzip der Transparenz wird Vertrauen aufgebaut, da es die unmittelbare Bewertung der Qualität eines digitalen Langzeitarchivs durch Interessierte zulässt.

Angemessenheit: Das Prinzip der Angemessenheit berücksichtigt die Tatsache, dass keine absoluten Maßstäbe möglich sind, sondern dass sich die Bewertung immer an den Zielen und Aufgaben des jeweiligen digitalen Langzeitarchivs ausrichtet. Die Kriterien müssen im Kontext der jeweiligen Archivierungsaufgabe gesehen werden. Deshalb können ggf. einzelne Kriterien irrelevant sein. Auch der notwendige Erfüllungsgrad eines Kriteriums kann – je nach den Zielen und Aufgaben des digitalen Langzeitarchivs – unterschiedlich ausfallen.

Bewertbarkeit: Die Kriterien decken die Aspekte ab, die im Rahmen einer Prüfung der Vertrauenswürdigkeit zu betrachten sind. Sie sind zwar abstrakt formuliert, verlangen aber vom betrachteten digitalen Archiv detaillierte Antworten. Aufgrund dieser Antworten ist eine Bewertung der durch das Archiv eingeleiteten Methoden und Maßnahmen zur Sicherung der digitalen Objekte möglich. Allerdings existieren zum Teil - insbesondere unter Langzeitaspekten - keine objektiv bewertbaren (messbaren) Merkmale. In diesen Fällen ist man auf Indikatoren angewiesen, die den Grad der Vertrauenswürdigkeit repräsentieren. Transparenz macht auch die Indikatoren für eine Bewertung zugänglich. Nicht objektiv bewertbar ist das Kriterium „Die Organisationsform ist angemessen". Hingegen ist zum Beispiel das Kriterium „Die Objekte sind eindeutig und dauerhaft identifiziert" bewertbar, da es dazu bereits Standards gibt.

Formaler Aufbau der Kriterienkataloge

Jedem abstrakt formulierten Kriterium ist eine Erläuterung beigegeben, die der Verständlichkeit dient. Beispiele aus unterschiedlichen Anwendungsbereichen sowie spezielle Literaturhinweise tragen zur Anwendbarkeit bei.

Inhaltlicher Aufbau der Kriterienkataloge

Die eigentlichen Kriterien sind in drei Bereiche gegliedert:

A. Ziele und Organisation (Organisatorischer Rahmen): Hier werden Anforderungen an den organisatorischen Rahmen gestellt, innerhalb dessen das digitale Langzeitarchiv operiert, d.h. seine Zielsetzung, die rechtlichen Bedingungen ebenso wie die personellen und finanziellen Ressourcen sowie die Organisationsform.

B. Umgang mit Objekten: Hier finden sich die Kriterien, die alle objektbezogenen Anforderungen während des Lebenszyklus der Objekte im digitalen Langzeitarchiv spezifizieren. Ausgangspunkt sind die im OAIS-Referenzmodell definierten zentralen Aufgaben: Aufnahme (Ingest), Archivablage (Archival Storage, inklusive Umsetzung der Langzeiterhaltungsmaßnahmen) und Nutzung (Access), die unterstützenden Prozesse Datenmanagement und Planung der Langzeiterhaltungsmaßnahmen (Preservation Planning).

C. Infrastruktur und Sicherheit: Die Kriterien in diesem Abschnitt betrachten die technischen Aspekte des Gesamtsystems sowie die Aspekte der Sicherheit.

A - Ziele und Organisation

Abbildung 9: Gliederung der Kriterien für vertrauenswürdige digitale Langzeitarchive

Um die eingesetzten Verfahren bewerten zu können, ist es notwendig, dass die Betreiberorganisation die Ziele und Rahmenbedingungen für den Betrieb des digitalen Langzeitarchivs spezifiziert, dokumentiert und veröffentlicht hat. Welche Objekte werden gesammelt? Für welche Nutzergruppe wird gesammelt? In welcher Form und unter welchen Konditionen sollen die digitalen Objekte den Nutzern bereitgestellt werden? Der Archivbetreiber muss weiterhin darlegen, auf welcher rechtlichen Basis er operiert, er muss entsprechende rechtliche Regelungen mit seinen Produzenten vereinbaren. Die Organisationsform der Betreiberorganisation und des digitalen Langzeitarchivs muss angemessen sein, d.h. sie muss für die Erfüllung der Aufgabe geeignet sein und geeignete Ent-

scheidungsstrukturen besitzen. Sowohl eine strategische als auch eine operative Planung sind vorzuweisen, ebenso Pläne zur Reaktion auf substantielle Veränderungen. So ist der Betrieb des digitalen Langzeitarchivs auch im Falle einer Auflösung der Betreiberorganisation sicherzustellen. Der Einsatz entsprechend qualifizierten Personals ist nachzuweisen. Ein weiterer wichtiger Aspekt ist die Durchführung eines Qualitätsmanagements.

A. Organisatorischer Rahmen

1. Das dLZA hat seine Ziele definiert.
1.1 Das dLZA hat Kriterien für die Auswahl seiner digitalen Objekte entwickelt.
1.2 Das dLZA übernimmt die Verantwortung für den dauerhaften Erhalt der durch die digitalen Objekte repräsentierten Informationen.
1.3 Das dLZA hat seine Zielgruppe(n) definiert.
2. Das dLZA ermöglicht seinen Zielgruppe(n) eine angemessene Nutzung der durch die digitalen Objekte repräsentierten Informationen.
2.1 Das dLZA ermöglicht seinen Zielgruppe(n) den Zugang zu den durch die digitalen Objekte repräsentierten Informationen.
2.2 Das dLZA stellt die Interpretierbarkeit der digitalen Objekte durch seine Zielgruppe(n) sicher.
3. Gesetzliche und vertragliche Regelungen werden eingehalten.
3.1 Es bestehen rechtliche Regelungen zwischen Produzenten und dem digitalen Langzeitarchiv.
3.2 Das dLZA handelt bei der Archivierung auf der Basis rechtlicher Regelungen.
3.3 Das dLZA handelt bei der Nutzung auf der Basis rechtlicher Regelungen.
4. Die Organisationsform ist für das dLZA angemessen.
4.1 Die Finanzierung des digitalen Langzeitarchivs ist sichergestellt.
4.2 Es steht Personal mit angemessener Qualifikation in ausreichendem Umfang zur Verfügung.
4.3 Es bestehen angemessene Organisationsstrukturen für das dLZA.
4.4 Das dLZA betreibt eine langfristige Planung.
4.5 Die Fortführung der festgelegten Aufgaben ist auch über das Bestehen des digitalen Langzeitarchivs hinaus sichergestellt.
5. Es wird ein angemessenes Qualitätsmanagement durchgeführt.
5.1 Alle Prozesse und Verantwortlichkeiten sind definiert.
5.2 Das dLZA dokumentiert alle seine Elemente nach einem definierten Verfahren.
5.3 Das dLZA reagiert auf substantielle Veränderungen.

B - Umgang mit Objekten

Der Informationserhalt im Hinblick auf die Nutzung wird als zentrale Aufgabe der Langzeitarchivierung definiert, die in Abhängigkeit von der Zielgruppe und deren Bedürfnissen geleistet werden muss. Nutzung setzt zum einen den Erhalt der Integrität, Authentizität und Vertraulichkeit sowie die Verfügbarkeit der digitalen Objekte voraus, zum anderen die Sicherstellung der Interpretierbarkeit der digitalen Objekte durch die Zielgruppe, um die darin enthaltene Information in einer geeigneten Weise zu rekonstruieren.

Daher werden in diesem Abschnitt zunächst die Anforderungen definiert, die sich auf die Erhaltung der Integrität und Authentizität der digitalen Objekte auf allen Stufen der Verarbeitung konzentrieren. Die Verarbeitungsstufen sind die im OAIS-Modell abgebildeten: Aufnahme, Archivablage und Nutzung sowie zur Sicherung der Interpretierbarkeit die Durchführung der technischen Langzeiterhaltungsmaßnahmen.[31] Die Übernahme der digitalen Objekte von den Produzenten erfolgt nach definierten Vorgaben. Für die Objekte muss spezifiziert sein, wie die Übergabepakete (Submission Information Packages, SIPs), die Archivpakete (Archival Information Packages, AIPs) und die Nutzungspakete (Dissemination Information Packages, DIPs) aussehen. Es muss eine Spezifikation geben, wie die Transformation der Informationspakete untereinander aussieht. Dazu sind die erhaltenswerten Kerneigenschaften zu spezifizieren. Das digitale Langzeitarchiv muss die technische Kontrolle über die digitalen Objekte besitzen, um diese Transformationen sowie speziell die Langzeiterhaltungsmaßnahmen durchführen zu können.

Das Datenmanagement muss dazu geeignet sein, die notwendigen Funktionalitäten des digitalen Langzeitarchivs zu gewährleisten. Die eindeutige und dauerhafte Identifikation[32] der Objekte und deren Beziehungen untereinander ist essentiell für deren Auffindbarkeit und Rekonstruierbarkeit. Das digitale Langzeitarchiv muss in ausreichendem Maße Metadaten[33] für eine formale, inhaltliche, strukturelle sowie für eine technische Beschreibung der digitalen Objekte erheben. Zudem sind Metadaten notwendig, die alle vom digitalen Langzeitarchiv vorgenommenen Veränderungen an den digitalen Objekten beinhalten. Entsprechende Nutzungsrechte und –bedingungen sind ebenfalls in Metadaten zu verzeichnen.

---

31 vgl. dazu auch Kapitel 8 dieses Handbuchs

32 vgl. dazu auch Kapitel 9.4 dieses Handbuchs

33 vgl. dazu auch Kapitel 6.2 dieses Handbuchs

Die nestor-Kriterien lauten im Detail:

B. Umgang mit Objekten

6. Das dLZA stellt die Integrität der digitalen Objekte auf allen Stufen der Verarbeitung sicher.
6.1 Aufnahme (Ingest): Das dLZA sichert die Integrität der digitalen Objekte.
6.2 Archivablage (Archival Storage): Das dLZA sichert die Integrität der digitalen Objekte.
6.3 Nutzung (Access): Das dLZA sichert die Integrität der digitalen Objekte.
7. Das dLZA stellt die Authentizität der digitalen Objekte und Metadaten auf allen Stufen der Verarbeitung sicher.
7.1 Aufnahme (Ingest): Das dLZA sichert die Authentizität der digitalen Objekte.
7.2 Archivablage (Archival Storage): Das dLZA sichert die Authentizität der digitalen Objekte.
7.3 Nutzung (Access): Das dLZA sichert die Authentizität der digitalen Objekte.
8. Das dLZA betreibt eine langfristige Planung seiner technischen Langzeiterhaltungsmaßnahmen.
9. Das dLZA übernimmt digitale Objekte von den Produzenten nach definierten Vorgaben.
9.1 Das dLZA spezifiziert seine Übergabeobjekte (Submission Information Packages, SIPs).
9.2 Das dLZA identifiziert, welche Eigenschaften der digitalen Objekte für den Erhalt von Informationen signifikant sind.
9.3 Das dLZA erhält die physische Kontrolle über die digitalen Objekte, um Lang-zeitarchivierungsmaßnahmen durchführen zu können.
10. Die Archivierung digitaler Objekte erfolgt nach definierten Vorgaben.
10.1 Das dLZA definiert seine Archivobjekte (Archival Information Packages, AIPs).
10.2 Das dLZA sorgt für eine Transformation der Übergabeobjekte in Archivobjekte.
10.3 Das dLZA gewährleistet die Speicherung und Lesbarkeit der Archivobjekte.
10.4 Das dLZA setzt Strategien zum Langzeiterhalt für jedes Archivobjekt um.
11. Das dLZA ermöglicht die Nutzung der digitalen Objekte nach definierten Vorgaben.
11.1 Das dLZA definiert seine Nutzungsobjekte (Dissemination Information Packages, DIPs).

11.2 Das dLZA gewährleistet eine Transformation der Archivobjekte in Nutzungsobjekte.
12. Das Datenmanagement ist dazu geeignet, die notwendigen Funktionalitäten des digitalen Langzeitarchivs zu gewährleisten.
12.1 Das dLZA identifiziert seine Objekte und deren Beziehungen eindeutig und dauerhaft.
12.2 Das dLZA erhebt in ausreichendem Maße Metadaten für eine formale und inhaltliche Beschreibung und Identifizierung der digitalen Objekte.
12.3 Das dLZA erhebt in ausreichendem Maße Metadaten zur strukturellen Beschreibung der digitalen Objekte.
12.4 Das dLZA erhebt in ausreichendem Maße Metadaten, die die vom Archiv vorgenommenen Veränderungen an den digitalen Objekten verzeichnen.
12.5 Das dLZA erhebt in ausreichendem Maße Metadaten zur technischen Beschreibung der digitalen Objekte.
12.6 Das dLZA erhebt in ausreichendem Maße Metadaten, die die entsprechenden Nutzungsrechte und –bedingungen verzeichnen.
12.7 Die Zuordnung der Metadaten zu den Objekten ist zu jeder Zeit gegeben.

C - Infrastruktur und Sicherheit

Ein digitales Langzeitarchiv muss eine angemessene IT-Infrastruktur besitzen, die in der Lage ist, die Objekte wie in Abschnitt B – Umgang mit Objekten zu bearbeiten. Es muss ein Sicherheitskonzept existieren und umgesetzt werden, sodass die Infrastruktur den Schutz des digitalen Langzeitarchivs und seiner digitalen Objekte gewährleisten kann.

Die nestor-Kriterien im Detail:
C. Infrastruktur und Sicherheit

13. Die IT-Infrastruktur ist angemessen.
13.1 Die IT-Infrastruktur setzt die Forderungen aus dem Umgang mit Objekten um.
13.2 Die IT-Infrastruktur setzt die Sicherheitsanforderungen des IT-Sicherheitskonzeptes um.
14. Die Infrastruktur gewährleistet den Schutz des digitalen Langzeitarchivs und seiner digitalen Objekte.

Der Weg zum vertrauenswürdigen digitalen Langzeitarchiv

Ein digitales Langzeitarchiv entsteht als komplexer Gesamtzusammenhang. Die Umsetzung der einzelnen Kriterien muss stets vor dem Hintergrund der Ziele des Gesamtsystems gesehen werden. Sowohl die Realisierung des digitalen Langzeitarchivs als Ganzes als auch die Erfüllung der einzelnen Kriterien läuft als Prozess in mehreren Stufen ab:

1. Konzeption
2. Planung und Spezifikation
3. Umsetzung und Implementierung
4. Evaluierung

Diese Stufen sind nicht als starres Phasenmodell zu betrachten. Vielmehr müssen sie im Zuge der ständigen Verbesserung regelmäßig wiederholt werden. Das Qualitätsmanagement überwacht diesen Entwicklungsprozess.

Die Kriterienkataloge und Checklisten sowie das Tool DRAMBORA können auf allen Stufen der Entwicklung zur Orientierung und Selbstevaluierung eingesetzt werden.

Darüber hinaus wurden auf der Grundlage von TRAC bereits mehrere externe Audits durchgeführt und DRAMBORA wurde bereits in einigen zum Teil von externen Experten begleiteten Selbstevaluierungen eingesetzt.

Literatur

Network Working Group (2000): R. Shirey: *Internet Security Glossary. Request for Comments: 2828* http://www.ietf.org/rfc/rfc2828.txt

BSI Bundesamt für Sicherheit in der Informationstechnik (2006): *Gemeinsame Kriterien für die Prüfung und Bewertung der Sicherheit von Informationstechnik, Common Criteria* V 3.1, http://www.bsi.de/cc/

Howard, John D. / Longstaff, Thomas A. (1998): *A Common Language for Computer Security Incidents*. SANDIA Reports SAND98-8667. Albuquerque, New Mexico : Sandia National Laboratories http://www.cert.org/research/taxonomy_988667.pdf

CCSDS (Consultative Committee for Space Data Systems) (2002*): Reference Model for an Open Archival Information System (OAIS). Blue Book.* http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf entspricht ISO 14721:2003

RLG-NARA Task Force on Digital Repository and Certification, (2007): *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC).* http://www.crl.edu/PDF/trac.pdf

RLG/OCLC Working Group on Digital Archive Attributes (2002): *Trusted Digital Repositories: Attributes and Responsibilities*, http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

DINI Deutsche Initiative für Netzwerkinformation / AG Elektronisches Publizieren (2007): *DINI-Zertifikat für Dokumenten- und Publikationsservice 2007*. DINI-Schriften 3. http://nbn-resolving.de/urn:nbn:de:kobv:11-10079197

nestor Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (2008): *nestor-Kriterien: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive.* Version 2. Frankfurt am Main : nestor http://nbn-resolving.de/urn:nbn:de:0008-2008021802

Steinmetz, Ralf (2000) : *Multimedia-Technologie: Grundlagen, Komponenten und Systeme*, 3. Auflage , Berlin, Heidelberg, New York : Springer

DCC Digital Curation Centre / DPE Digital Preservation Europe (2008): *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), interactive,* http://www.repositoryaudit.eu/

CRL Center for Research Libraries / DCC Digital Curation Centre / DPE Digital Preservation Europe / nestor (2007): *Core Requirements for Digital Archives*, http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92

6 Metadatenstandards im Bereich der digitalen LZA

6.1 Einführung

Mathias Jehn

Für den Erfolg der digitalen Langzeitarchivierung bilden Standards eine unabdingbare Voraussetzung für kompatible und interoperative Systeme aller Art. Sie werden für technische als auch für organisatorische Aspekte in der digitalen Langzeitarchivierung benötigt. Ein Standard kann in einem formalisierten oder nicht-formalisierten Regelwerk bzw. in einem sich ungeplant ergebenden Regelfall bestehen, beispielsweise in einer einzelnen Regel bzw. mehreren Regeln oder einer Norm. Standards fördern nicht nur die Wiederverwendbarkeit und Austauschbarkeit von Komponenten, sondern gewähren auch verlässliche Vorgaben für System- und Produktentwickler. Öffentlich verfügbare und realistisch umsetzbare Vorgaben sind Basis für konkurrierende Implementierungen und somit für einen funktionierenden Markt. Das notwendige Konsensprinzip erfordert dabei sehr aufwändige Abstimmungsprozesse und wirtschaftliche oder

sonstige Interessen führen teilweise zu konkurrierenden Inhalten oder unnötigem Umfang von Ansätzen. Die Abgrenzung von Inhalten und die zeitliche Synchronisation können zudem auch durch die Vielzahl der Standardisierungsorganisationen negativ beeinflusst werden. Auf jeden Fall ist das Prozedere der Standardisierung und der Aufbau der Standards sehr unterschiedlich. Die geforderte Offenheit von Standards ist nicht nur eine rein definitorische Angelegenheit, sondern kann weitgehende rechtliche und wirtschaftliche Konsequenzen haben. Versteckte Patente oder sonstige Hindernisse, die z.B. Mitbewerber bei einer Implementierung behindern, können sich nachteilig auf die Zuverlässigkeit und Wirtschaftlichkeit der Langzeitarchivierung auswirken. Vorteilhaft ist, dass sich die Standardisierungsorganisationen um mehr Transparenz und auch Einheitlichkeit bei der Behandlung und Darstellung von Rechten (Intellectual Property Rights – IPR) bemühen. Das folgende Kapitel präsentiert einige wesentliche Entwicklungen im Bereich der internationalen Standards und der Bemühungen, im Bereich der technischen Standards und der Metadatenstandards für die digitale Langzeitarchivierung zu entwickeln.

6.2 Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten

Markus Enders

Ausgehend von den Digitalisierungsaktivitäten der Bibliotheken Mitte der 1990er Jahre entstand die Notwendigkeit, die so entstandenen Dokumente umfassend zu beschreiben. Diese Beschreibung muss im Gegensatz zu den bis dahin üblichen Verfahrensweisen nicht nur einen Datensatz für das gesamte Dokument beinhalten, sondern außerdem einzelne Dokumentbestandteile und ihre Abhängigkeiten zueinander beschreiben. So lassen sich gewohnte Nutzungsmöglichkeiten eines Buches in die digitale Welt übertragen. Inhaltsverzeichnisse, Seitennummern sowie Verweise auf einzelne Bilder müssen durch ein solches Format zusammengehalten werden.

Zu diesem Zweck wurde im Rahmen des „Making Of Amerika" Projektes Ebind entwickelt[1]. Ebind selber war jedoch ausschließlich nur für Digitalisate von Büchern sinnvoll zu verwenden.

Um weitere Medientypen sowie unterschiedliche Metadatenformate einbinden zu können, haben sich Anforderungen an ein komplexes Objektformat ergeben. Dies setzt ein abstraktes Modell voraus, mit Hilfe dessen sich Dokumente flexibel modellieren lassen und als Container Format verschiedene Standards eingebunden werden können. Ein solches abstraktes Modell bildet die Basis von METS und wird durch das METS-XML-Schema beschrieben. Daher wird METS derzeit auch fast ausschließlich als XML serialisiert und in Form von Dateien gespeichert. Als Container Format ist es in der Lage weitere XML-Schema (so genannte Extension Schemas) zu integrieren.

Das METS Abstract Model

Das METS „Abstract Model" beinhaltet alle Objekte innerhalb eines METS Dokuments und beschreibt deren Verhältnis zueinander. Zentraler Bestandteil eines METS-Dokuments ist eine Struktur. Das entsprechende Element nennt sich daher structMap und ist als einziges Element im „Abstract Model" verpflichtend. Jedes METS Dokument muss ein solches Element besitzen. Unter Struktur wird in diesem Fall eine hierarchische Struktur mit nur einem Start-

1 O.V.: An Introduction to the Electronic Binding DTD (Ebind). http://sunsite.berkeley.edu/Ebind/
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

knoten verstanden. Eine Struktur kann also als Baum interpretiert werden. Der Wurzelknoten sowie jeder Ast wird als Struktureinheit bezeichnet. Jede Struktur muss über einen Wurzelknoten verfügen. In der Praxis kann diese verpflichtende Struktur bspw. die logische Struktur – also das Inhaltsverzeichnis einer Monographie speichern. Im Minimalfall wird dieses lediglich die Struktureinheit der Monographie umfassen, da der Wurzelknoten in dem Baum verpflichtend ist. Weitere Strukturen sind optional. Eine weitere Struktur könnte bspw. die physische Struktur des Dokuments sein. Die physische Struktur beschreibt bspw. aus der Exemplarsicht (gebundene Einheit mit Seiten als unterliegende Struktureinheiten).

Verknüpfungen zwischen zwei Struktureinheiten werden in einer separaten Sektion gespeichert. Das „Abstract Model" stellt dazu die structLink Sektion zur Verfügung, die optional genutzt werden kann. Jede Verknüpfung zwischen zwei Struktureinheiten wird in einem eigenen Element definiert.

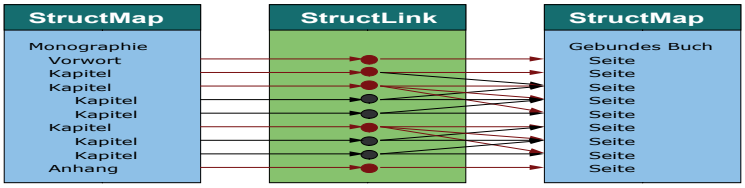


Abbildung 1: Verknüpfung von zwei Strukturen im Abstract-Model

Das „Abstract Model" macht allerdings keine Vorgaben, aus welcher strukturellen Perspektive ein Dokument beschrieben wird oder wie detailliert diese einzelnen Strukturen ausgearbeitet werden müssen.

Ferner berücksichtigt das „Abstract Model" auch Metadaten. Hierunter sind allerdings nicht nur bibliographische Metadaten zu verstehen. Vielmehr wird in deskriptive Metadaten (in der Descriptive Metadata Section) und administrative Metadaten (in der Administrative Metadata Section) unterschieden. Während die deskriptiven Metadaten bibliographische Informationen enthalten, werden Informationen zu Rechteinhabern, Nutzungsrechte, technische Informationen zu einzelnen Dateien oder Langzeitarchivierungsmetadaten in den administrativen Metadaten gespeichert. Für beide Metadatentypen können beliebige Schema, so genannte „Extension Schema" genutzt werden, die in der jeweiligen Sektion gespeichert werden. Auch die Referenzierung von Metadatensätzen ist möglich, sofern diese bspw. per URL zugänglich sind. Jede Datei sowie jeder Struktureinheit lässt sich mit entsprechenden Metadatensätzen versehen, wobei jeder Einheit mehrere Datensätze zugeordnet werden können. Als „Extensi-

on Schema“ können sowohl XML-Metadatenschema wie bspw. MARC XML, MODS, Dublin Core) sowie Binärdaten benutzt werden. Dies erlaubt auch die Integration gängiger bibliothekarischer Standards wie bspw. PICA-Datensätze.

Abbildung 2: Verweis auf Metadatensektionen im METS-Abstract-Model

Neben den Struktureinheiten und ihren zugehörigen Metadaten spielen auch Dateien bzw. Streams eine wesentliche Rolle, da letztlich in ihnen die durch das METS-Dokument beschriebenen Inhalte manifestiert/gespeichert sind. Eine Datei kann bspw. den Volltext einen Buches, die Audioaufnahme einer Rede oder eine gescannte Buchseite als Image enthalten. Entsprechende Daten können in ein METS-Dokument eingebunden werden (bspw. Base64 encoded in die METS-XML Datei eingefügt werden) oder aber mittels xlink referenziert werden. Ein METS-Dokument kann also als Container alle für ein Dokument notwendigen Dateien enthalten oder referenzieren, unabhängig davon, ob die Dateien lokal oder auf entfernten Servern vorhanden sind. Metadatensätze, die nicht in die METS Datei eingebunden sind, werden nicht als Datei betrachtet, sondern sind aus der entsprechenden Metadatensektion zu referenzieren.

Grundsätzlich müssen alle für ein METS-Dokument relevanten Dateien innerhalb der File-Sektion aufgeführt werden. Innerhalb der File-Sektion können Gruppen (File-Groups) von Dateien gebildet werden, wobei die Abgrenzungskriterien zwischen einzelnen Gruppen nicht durch das „Abstract Model“ definiert sind. Ja nach Modellierung lassen sich Dateien bspw. nach technischen Parametern (Auflösung oder Farbtiefe von Images), Anwendungszweck (Anzeige, Archivierung, Suche) oder sonstigen Eigenschaften (Durchlauf bestimmter Produktionsschritte) den einzelnen Gruppen zuordnen.

Das METS-Abstract-Model erlaubt das Speichern von administrativen Metadaten zu jeder Datei. Generelle, für jede Datei verfügbare technische Metadaten wie Dateigröße, Checksummen etc. lassen sich direkt in METS speichern. Für weiterführende Metadaten kann mit jeder Datei eine oder mehrere Administrative Metadatensektion(en) verknüpft werden, die bspw. Formatspezifische Metadaten enthalten (für Images könnten die Auflösungsinformationen, Informationen zur Farbtiefe etc. sein).

Abbildung 3: Administrative Metadata zu Dateien

Dateien sind darüber hinaus mit Struktureinheiten verknüpft. Die Struktureinheit, die eine einzelne Buchseite repräsentiert, kann somit mit einer einzelnen Datei, die ein Image dieser Seite beinhaltet, verknüpft werden. Das „METS-Abstract-Model" stellt hierzu eine N:M Verknüpfung bereit. Das bedeutet, dass eine Datei von mehreren Struktureinheiten (auch aus unterschiedlichen Strukturesektionen) aus verknüpft werden kann, genauso wie eine Struktureinheit mehrere Dateien verknüpfen kann. Im Ergebnis heißt das, dass der Struktureinheit vom Typ „Monographie" sämtliche Imagedateien eines gescannten Werkes direkt unterstellt sind.

Für die Verknüpfung von Dateien sieht das „METS-Abstract-Model" noch weitere Möglichkeiten vor. So lassen sich mehrere Verknüpfungen hinsichtlich ihrer Reihenfolge beim Abspielen bzw. Anzeigen bewerten. Dateien können entweder sequentiell angezeigt (Images eines digitalisierten Buches) oder auch parallel abgespielt (Audio- und Videodateien gleichen Inhalts) werden. Darüber hinaus kann nicht nur auf Dateien, sondern auch in Dateiobjekte hinein verlinkt werden. Diese Verlinkungen sind u.a. dann sinnvoll, wenn Einheiten beschrieben werden, die aus technischen Gründen nicht aus der Datei herausgetrennt werden können. Das können bestimmte Teile eines Images sein (bspw. einzelne Textspalten) oder aber konkrete zeitliche Abschnitte einer Audioaufnahme. In der Praxis lassen sich so einzelne Zeitabschnitte eines Streams markieren und bspw. mit inhaltlich identischen Abschnitten eines Rede-Manuskriptes taggen. Das METS-Dokument würde über die Struktureinheit eine Verbindung zwischen den unterschiedlichen Dateien herstellen.

Abbildung 4: Struktureinheit ist mit verschiedenen Dateien und Dateibereichen verknüpft

Das METS-Abstract-Model nutzt intensiv die Möglichkeit, einzelne Sektionen miteinander zu verknüpfen. Da METS überwiegend als XML realisiert ist, geschieht diese Verknüpfung über XML-Identifier. Jede Sektion verfügt über einen Identifier, der innerhalb des XML- Dokumentes eindeutig ist. Er dient als Ziel für die Verknüpfungen aus anderen Sektionen heraus. Aufgrund der XML-Serialisierung muß er den XML-ID Anforderungen genügen. Es muss bei Verwendung von weiteren Extension Schemas darauf geachtet werden, dass die Eindeutigkeit der Identifier aus dem unterschiedlichen Schema nicht gefährdet wird, da diese üblicherweise alle im gleichen Namensraum existieren.

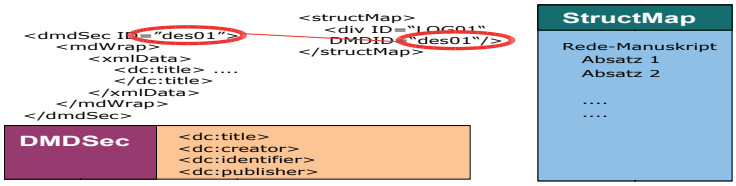


Abbildung 5: Unterschiedliche Sektionen mittels XML-IDs verknüpft

Dokumentation

Wie deutlich geworden ist, stellt das METS-Abstract-Model sowie des XML-Serialisierung als METS-XML Schema lediglich ein grobes Modell da, welches auf den jeweiligen Anwendungsfall angepasst werden muss. Die Verwendung von Extension Schema sollte genauso dokumentiert werden wie die Nutzung optionaler Elemente und Attribute in METS. Hierbei sollte vor allem auch die Transformation realer, im zu beschreibenden Dokument vorhandene Objekte in entsprechende METS-Objekte bzw. METS-Sektionen im Vordergrund stehen. Eine einzige Strukturesektion kann bspw. logische Einheiten (bspw. das Inhaltsverzeichnis eines Buches) umfassen als auch bestimmte physische Einheiten (bspw. einzelne Seiten) enthalten. Alternativ können jedoch bestimmte Einheiten in eine separate Strukturesektion ausgelagert werden. Das „Abstract

Model“ erlaubt diese Flexibilität. Eine Implementierung von METS für einen bestimmten Anwendungsfall muss dieses jedoch konkret festlegen.

Um die Dokumentation zu standardisieren wurde das METS-Profile Schema entwickelt. Es gibt eine Grobstrukturierung vor, die sicher stellt, dass alle wesentlichen Bereiche eines METS-Dokuments in der Dokumentation berücksichtigt werden. Die Dokumentation selber muss derzeit noch auf XML Basis erfolgen. Die so entstandene XML-Datei lässt sich jedoch anschliessend als HTML oder PDF konvertieren.

Um ein solches Profil auf der offiziellen METS-Homepage veröffentlichen zu können, wird es durch Mitglieder des METS-Editorial-Board verifiziert. Nur verifizierte METS-Profile werden veröffentlicht und stehen auf der Homepage zur Nachnutzung bereit. Sie können von anderen Institutionen adaptiert und modifiziert werden und somit erheblich zur Reduktion der Entwicklungszeit einer eigenen METS-Implementierung beitragen.

Fazit

Aufgrund der hohen Flexibilität des METS Abstract Models wird METS in einer großen Zahl unterschiedlicher Implementierungen für sehr verschiedene Dokumententypen genutzt. Neben der ursprünglichen Anwendung, digitalisierte Büchern zu beschreiben, existieren heute sowohl METS-Profile zur Webseitenbeschreibungen (Webarchivierung) sowie Audio- und Videodaten. Während in den ersten Jahren METS überwiegend zum Beschreiben komplexer Dokumente genutzt wurde, um diese dann mittels XSLTs oder DMS-Systeme verwalten und anzeigen zu können, wird METS heute gerade auch im Bereich der Langzeitarchivierung zur Beschreibung des Archival Information Packets (AIP) genutzt. METS ist heute für viele Bereiche, in denen komplexe Dokumente beschrieben werden müssen, ein De-facto-Standard und kann sowohl im universitären als auch im kommerziellen Umfeld eine große Zahl an Implementierungen vorweisen. Ein großer Teil derer ist im METS-Implementation Registry auf der METS-Homepage (http://www.loc.gov/mets) nachgewiesen.

6.3 PREMIS

Olaf Brandt

Das Akronym PREMIS löst sich in „PREservation Metadata: Implementation Strategies“ auf. PREMIS ist eine Initiative, welche die Entwicklung und Pflege des international anerkannten gleichnamigen PREMIS-Langzeitarchivierungsmetadatenstandards verantwortet. Sie wurde im Jahre 2003 von OCLC (Online Computer Library Center) und RLG (Research Library Group) ins Leben gerufen.

Langzeitarchivierungsmetadaten sind - vereinfacht ausgedrückt - strukturierte Informationen über digitale Objekte, ihre Kontexte, ihre Beziehungen und Verknüpfungen, welche die Prozesse der digitalen Langzeitarchivierung ermöglichen, unterstützen oder dokumentieren.

Das Hauptziel von PREMIS ist die Entwicklung von Empfehlungen, Vorschlägen und Best-Practices zur Implementierung von Langzeitarchivierungsmetadaten, d.h. die Fortentwicklung des Standards, sowie die Anbindung an weitere Standards. Die Fortentwicklung wird zurzeit vom PREMIS Editorial Committee geleistet. Das ist eine internationale Gruppe von Akteuren aus Gedächtnisorganisationen wie Archiven, Bibliotheken und Museen sowie der Privatwirtschaft. Die Arbeit von PREMIS baut auf den Ergebnissen der Preservation-Metadata Working-Group auf, die bereits 2001 die Entwicklung eines gemeinsamen Rahmenkonzeptes für Langzeitarchivierungsmetadaten vorantrieb.[2] Nach der Veröffentlichung des PREMIS Data Dictionaries der Version 1.0 im Jahr 2005[3] galt es zunächst Implementatierungen zu unterstützen und die Weiterentwicklung von PREMIS zu institutionalisieren. Dafür wurde eine PREMIS Managing Agency gegründet, welche an der Library of Congress angesiedelt ist.[4] Sie übernimmt in enger Abstimmung mit dem PREMIS Editorial Committee die Koordination von PREMIS im Hintergrund. Zu den Aufgaben gehören z.B. das Hosting und die Pflege der Webseite, die Planung und Durchführung von Maßnahmen für die PREMIS-Verbreitung und der Betrieb und die Moderation der PREMIS-Diskussionslisten. Das PREMIS Editorial Committee erarbeitet zusammen mit der Managing

2 Preservation Metadata Working Group (PMWG 2002) Framework: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

3 Abschlußbericht der PREMIS Arbeitsgruppe mit „Data Dictionary for Preservation Metadata“: http://www.oclc.org/research/projects/pmwg/premis-final.pdf

4 Webseite der PREMIS Maintenance Activity: http://www.loc.gov/standards/premis/

Agency die Ziele und die weitere Entwicklung von PREMIS. Das betrifft v.a. die Weiterentwicklung und Pflege des Data Dictionary und der XML-Schemas. Weiter sorgt das Editorial Committee für die Verbreitung des Wissens über PREMIS durch Vorträge und Publikationen. Die PREMIS Implementors Group ist über eine Mailingsliste und ein Wiki organisiert. Sie ist offen für jede Person oder Institution, die ein Interesse an digitaler Langzeitarchivierung oder PREMIS hat.

Wichtigste Neuerung des Jahres 2008 ist sicherlich die Veröffentlichung des PREMIS Data Dictionary für Langzeitarchivierungsmetadaten in Version 2.0 und des neu erarbeiteten generischen XML-Schemas.[5] Aber auch die Fortschritte bei der Implementierung von PREMIS und METS sind in ihrer Bedeutung sicherlich nicht zu unterschätzen. So ist PREMIS seit einiger Zeit ein offizielles Erweiterungsschema von METS.[6] Empfehlungen für die Implementierung von PREMIS und METS[7] finden ihren Niederschlag in fruchtbaren Diskussionen.[8] PREMIS hat sich in der Langzeitarchivierungscommunity einen festen Platz als Nachschlagewerk für Implementierungen von Langzeitarchivierungsmetadaten und als gemeinsames Austauschformat[9] erarbeitet.

Um einen ersten Einblick in die Welt von PREMIS zu bekommen, wird im nun folgenden Abschnitt eine Einführung in das PREMIS-Datenmodell gegeben.

Aufbau Datenmodell

Das PREMIS Datenmodell kennt fünf grundlegende Einheiten, sog. Entities:

- Intellectual Entities
- Object Entity
- Events Entity
- Rights Entity
- Agent Entity

---

5 Siehe dazu http://www.loc.gov/standards/premis/v2/premis-2-0.pdf und http://www.loc.gov/standards/premis/schemas.html

6 Siehe dazu http://www.loc.gov/standards/mets/mets-extenders.html

7 Siehe dazu http://www.loc.gov/premis/guidelines-premismets.pdf

8 Siehe dazu http://www.dlib.org/dlib/september08/dappert/09dappert.html

9 Siehe dazu http://www.library.cornell.edu/dlit/MathArc/web/resources/MathArc_metadataschema031a.doc oder auch in jüngster Zeit http://www.dlib.org/dlib/november08/caplan/11caplan.html

Entities sind abstrakte Klassen von 'Dingen', also z.B. „digitale Objekte“ oder „Agenten“. Die Eigenschaften von vier Entities werden im PREMIS Data Dictionary mit sog. Semantic Units (semantische Einheiten) näher beschrieben. Semantic Units sind die für die digitale Langzeitarchivierung relevanten Eigenschaften der Entities.

Intellectual Entities

Intellectual Entities sind als zusammenhängende Inhalte definiert, die als eine Einheit beschrieben werden. Sie stellen somit eine Idee dar, welche in analogen oder digitalen Manifestationen oder Repräsentationen vorliegen kann. Es könnte sich also sowohl um einen Zeitschriftenband handeln als auch um den digitalisierten Zeitschriftenband. Dieser kann wiederum weitere Intellectual Entities (z.B. Zeitschriftenausgaben oder Artikel) enthalten. Intellectual Entities werden im Data Dictionary nicht mit semantischen Einheiten beschrieben, da sie außerhalb des Fokus, Kerninformationen für die digitale Langzeitarchivierung bereitzustellen, liegen. Auf sie kann aber von Objekten verwiesen werden.

Object Entity

In der *Object Entitiy* werden die zu archivierenden Daten mit relevanten Informationen für das Management und die Planung von Langzeitarchivierungsprozessen beschrieben. Die Object Entity kann unterschiedliche digitale Objekte beschreiben: sogenannte Representations, Dateien und auch Bitstreams.

Eine *Representation* ist eine Instanz oder Manifestierung einer Intellektuellen Entität, realisiert oder enthält sie also. Eine Representation ist eine logisch-funktionale Einheit aus digitalen Daten oder Dateien und ihrer Strukturbeschreibung. Als Beispiel kann eine Webseite dienen, die aus mehreren einzelnen Dateien besteht. Ihre Struktur und die Beziehungen der einzelnen Elemente untereinander zu kennen ist essentiell für die langfristige, sinnvolle und komplette Darstellung dieser Webseite als Einheit. Beim gegebenen Beispiel einer Webseite müsste z.B. beschrieben werden, dass eine Einstiegsseite existiert, die auf bestimmte Art und Weise mehrere Unterseiten und andere Elemente (wie z.B. Grafikdateien) einbindet. Dateien werden im PREMIS Data Dictionary als „named and ordered sequence of bytes that is known by an operating system“ bezeichnet. *Bitstream* (Datenstrom) wird nur als in zu archivierenden Dateien enthaltener und adressierbarer Teil beschrieben. Ein Datenstrom kann nur durch Umwandlung oder Hinzufügung von weiteren Informationen zu einer Datei werden. Zu den beschreibbaren Informationen von Objekten gehören z.B. eindeutige Identifikatoren, Charakteristika der Daten wie Größe und Format, Beschrei-

bungen der Systemumgebungen (Software, Hardware), Beschreibungsmöglichkeiten der relevanten Eigenschaften der Objekte, sowie die Beziehungen zu anderen Objekten, Events und Rechteinformationen.

Event Entity

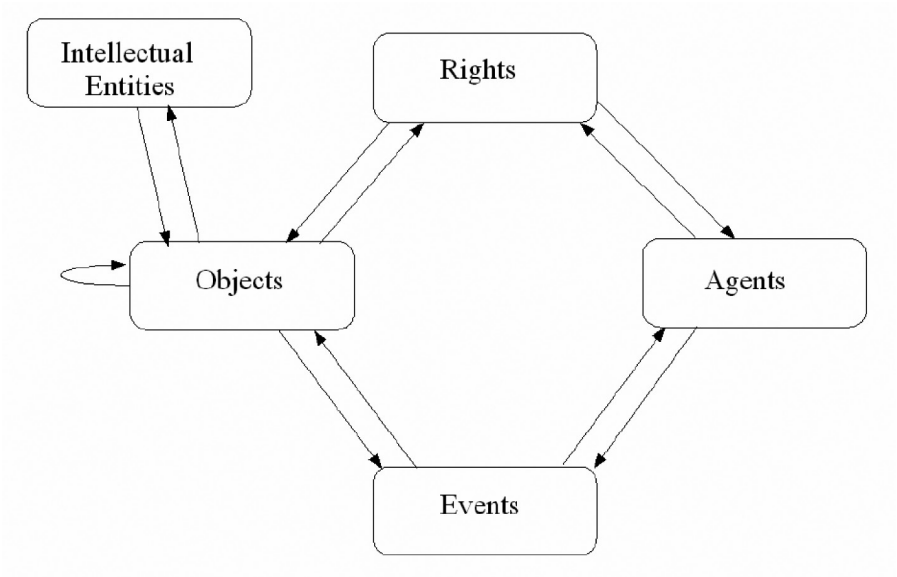
Ein *Event* ist in PREMIS eine identifizierbare Aktion oder ein Ereignis, in das mindestens ein Objekt und/oder ein Agent einbezogen sind. In der *Event Entitiy* werden Informationen über diese Aktionen oder Ereignisse und ihre Resultate sowie ihre Beziehungen zu Objekten und Agenten beschrieben. Mit der lückenlosen Aufzeichnung der Ereignisse im Archiv kann die Geschichte und die Verwendung der digitalen Objekte im Archivsystem nachgewiesen werden. Die Dokumentation der Ereignisse dient also dem Nachweis der Provenienz, als Beleg für die Einhaltung von Rechten oder kann für Statistikfunktionen und Billing herangezogen werden.

Agent Entity

Ein PREMIS *Agent* ist definiert als Person, Organisation oder Software, welche auf ein Ereignis im digitalen Archiv bezogen ist. Mit der *Agent Entity* werden spezifische Informationen von Agenten beschrieben, die im Zusammenhang mit Langzeitarchivierungsereignissen und Rechtemanagement im Leben eines Datenobjektes auftreten. Informationen über Agenten dienen v.a. der eindeutigen Identifizierung eines Agents.

Rights-Entity

Für den Betrieb eines Langzeitarchivs ist es wichtig, von den mit den Objekten verbundenen Rechten mit Relevanz für die Planung und Durchführung von Aktionen für die digitale Langzeitarchivierung zu wissen. Das betrifft z.B. das Kopieren von Daten, die Umwandlung in andere Formate etc. Aussagen über diese Rechte und Erlaubnisse werden in der *Rights Entity* beschrieben. Seit PREMIS 2.0 können tiefer gehende Rechtekonstellationen und deren Kontexte beschrieben werden, wie z.B. spezifische Urheberrechte in einem bestimmten Rechtsraum.
Um das Zusammenspiel der einzelnen Entitäten besser veranschaulichen zu können, folgt eine grafische Darstellung des Datenmodells.



PREMIS Datenmodell in Version 2.0 [10]

10 http://www.loc.gov/premis/v2/premis-2-0.pdf

6.4 LMER

Tobias Steinke

Die Langzeitarchivierungsmetadaten für elektronische Ressourcen (LMER) wurden von der Deutschen Bibliothek entwickelt. Das Objektmodell basiert auf dem "Preservation Metadata: Metadata Implementation Schema" der Nationalbibliothek von Neuseeland (2003).
Ziele von LMER sind:

- Ergänzung zu existierenden bibliographischen Metadaten, deshalb nur Beschreibung der technischen Informationen zu einem Objekt und der technischen Veränderungshistorie
- Praxisrelevante Beschränkung auf Angaben, die größtenteils automatisch generiert werden können
- Identifizierung der Kernelemente, die für alle Dateikategorien und jedes Dateiformat gültig sind, sowie ein flexibler Teil für spezifische Metadaten
- Abzubilden als XML-Schema
- Dateiformatidentifikation über Referenz zu einer zu schaffenden File-Format-Registry
- Modularer Aufbau zur Integration in Containerformate wie METS

Historie

LMER entstand 2003 aus dem Bedarf für technische Metadaten im Vorhaben LZA-RegBib. Die erste Version 1.0 wurde 2004 als Referenzbeschreibung und XML-Schema veröffentlicht. 2005 erschien eine überarbeitete Version 1.2, die auch Grundlage für die Verwendung im Projekt kopal ist. Die Version 1.2 führte eine starke Modularisierung und damit einhergehende Aufteilung in mehrere XML-Schemas ein, die eine bessere Einbindung in METS ermöglichte. Als Resultat entstand das METS-Profile-Universelles-Objektformat (UOF), das auf METS 1.4 und LMER 1.2 basiert.

Objektmodell

In LMER meint ein Objekt eine logische Einheit, die aus beliebig vielen Dateien bestehen kann. Es gibt einen Metadatenabschnitt zum Objekt und je einen Metadatenabschnitt zu jeder zugehörigen Datei. Zum Objekt einer jeden Datei kann es Prozess-Abschnitte geben. Diese beschreiben die technische Veränderungshistorie, also vor allem die Anwendung der Langzeiterhaltungsstrategie Migration. Schließlich gibt es noch den Abschnitt Metadatenmodifikation, der

Änderungen an den Metadaten selbst dokumentiert und sich auf alle anderen Abschnitte bezieht. Dabei wird davon ausgegangen, dass sich alle relevanten Metadatenabschnitte in derselben XML-Datei befinden.

Die vier möglichen Abschnittsarten LMER-Objekt, LMER-Datei, LMER-Prozess und LMER-Modifikation werden jeweils durch ein eigenes XML-Schema beschrieben. Dadurch kann jeder Abschnitt eigenständig in anderen XML-Schemas wie METS eingesetzt werden. Es gibt jedoch auch ein zusammenfassendes XML-Schema für LMER, das anders als die einzelnen Schemas Abhängigkeiten und Muss-Felder definiert.

LMER-Objekt

Die Metadaten zum Objekt stellen über einen Persistent Identifier den Bezug zu bibliographischen Metadaten her. Zugleich finden sich dort u.a. Informationen zur Objektversion und zur Anzahl der zugehörigen Dateien.

LMER-Datei

Zu jeder Datei werden die technischen Informationen erfasst, wie sie auch von einem Dateisystem angezeigt werden (Name, Pfad, Größe, Erstellungsdatum), aber auch eine Referenz zu exakten Formatbestimmung. Zudem wird jede Datei einer Kategorie zugeordnet (Bild, Video, Audio etc.), die insbesondere für die spezifischen Metadaten relevant ist. Denn in einem speziellen Platzhalterelement des Datei-Abschnitts können dank des flexiblen Mechanismus von XML-Schemata beliebige XML-Metadaten zur spezifischen Bestimmung bestimmter Dateicharakteristiken hinterlegt werden. Ein Beispiel dafür ist die Ausgabe des Dateianalysewerkzeugs JHOVE.

LMER-Prozess

Die Metadaten in einem Prozess-Abschnitt beschreiben die Schritte und Resultate von technischen Veränderungen und Konvertierungen (Migrationen) an einem Objekt oder einzelnen Dateien eines Objekts. Gehört ein Prozess-Abschnitt zu einem Objekt, so bezeichnet er auch die Versionsnummer und die Kennung des Objekts, von dem die vorliegende Version abgeleitet wurde.

LMER-Modifikation

Die LMER-Daten werden in der Regel in einer oder mehreren XML-Dateien gespeichert. Veränderungen (Ergänzungen oder Korrekturen) der XML-Daten darin können im Modifikationsabschnitt aufgeführt werden.

Literatur

Referenzbeschreibung zu LMER 1.2:
http://nbn-resolving.de/?urn=urn:nbn:de:1111-2005041102
Referenzbeschreibung zum Universellen Objektformat (UOF):
http://kopal.langzeitarchivierung.de/downloads/kopal_Universelles_Objektformat.pdf

6.5 MIX

Tobias Steinke

MIX steht für „NISO Metadata for Images in XML“ und ist ein XML-Schema für technische Metadaten zur Verwaltung digitaler Bildsammlungen. Die Metadatenelemente dieses XML-Schemas werden durch den Standard ANSI/NISO Z39.87-2006 („Technical Metadata for Digital Still Images“) beschrieben. MIX wurde von der Library of Congress und dem MARC Standards Office entwickelt. Neben allgemeinen Informationen zu einer Datei werden insbesondere komplexe Informationen zu Bildeigenschaften wie Farbinformationen aufgenommen, sowie detaillierte Beschreibungen der technischen Werte der Erzeugungsgeräte wie Scanner oder Digitalkamera. Zusätzlich kann eine Veränderungshistorie in den Metadaten aufgeführt werden, wobei dies ausdrücklich als einfacher Ersatz für Institutionen gedacht ist, welche keine eigenen Langzeitarchivierungsmetadaten wie PREMIS nutzen. Es gibt keine Strukturinformationen in MIX, denn hierfür wird das ebenfalls von der Library of Congress stammende METS vorgesehen. Die aktuelle Version von MIX ist 1.0 von 2006. Ein öffentlicher Entwurf für MIX 2.0 liegt vor.

Offizielle Webseite: http://www.loc.gov/standards/mix/

7 Formate

7.1 Einführung

Jens Ludwig

Bereits in der alltäglichen Nutzung elektronischer Daten und Medien sind sich die meisten Nutzer der Existenz von Formaten und ihrer Schwierigkeiten bewusst. Es gehört zum digitalen Alltag, dass nicht jedes Videoformat mit jeder Software abspielbar ist, dass dasselbe Textverarbeitungsdokument manchmal von verschiedenen Programmen verschieden dargestellt wird und dass Programme im Speicherdialog eine Vielzahl von Formaten anbieten, von deren Vor- und Nachteilen man keine Ahnung hat. Für die langfristige Erhaltung von Informationen stellen sich diese Probleme in verschärfter Form. Formate sind ein wesentlicher Faktor für die Gefahr des technologischens Veraltens digitaler Informationen.

Dieses Kapitel soll dabei helfen, die wesentlichen Aspekte für den Umgang mit Formaten für die Langzeitarchivierung zu verstehen. In „Digitale Objekte und Formate“ werden dafür zuerst die begrifflichen Grundlagen gelegt: Was sind die digitalen Objekte, mit denen wir alltäglich umgehen, und welche Rol-

le spielen Formate? Der Abschnitt „Auswahlkriterien“ bietet Hilfestellung für eine der meist gestellten Fragen bezüglich der Langzeitarchivierung: Welches Format soll ich verwenden? Leider gibt es hier weder eine allgemeingültige Lösung, nicht ein Format, das alle anderen überflüssig macht, noch sind mit der sinnvollen Wahl eines Formates alle Aufgaben gelöst, die im Zusammenhang mit Formaten anfallen. „Formatcharakterisierung“ beschreibt zusammen mit den Aufgaben der Identifizierung von Formaten, der Validierung und der Extraktion von technischen Metadaten einige technische Werkzeuge, die dafür genutzt werden können. Den Abschluss bildet „File Format Registries“, das einige zentrale Verzeichnisse beschriebt, in denen Referenzinformationen über Formate gesammelt werden.

7.2 Digitale Objekte und Formate

Stefan E. Funk

Digitale Objekte

Die erste Frage, die im Zusammenhang mit der digitalen Langzeitarchivierung gestellt werden muss, ist sicherlich die nach den zu archivierenden Objekten. Welche Objekte möchte ich archivieren? Eine einfache Antwort lautet hier zunächst: digitale Objekte!

Eine Antwort auf die naheliegende Frage, was denn digitale Objekte eigentlich sind, gibt die Definition zum Begriff „digitales Objekt" aus dem Open Archival Information System (OAIS). Dieser Standard beschreibt ganz allgemein ein Archivsystem mit dessen benötigten Komponenten und deren Kommunikation untereinander, wie auch die Kommunikation vom und zum Nutzer. Ein digitales Objekt wird dort definiert als

An object composed of a set of bit sequences

(CCSDS 2001), also als ein aus einer Reihe von Bit-Sequenzen zusammengesetztes Objekt. Somit kann all das als ein digitales Objekt bezeichnet werden, das mit Hilfe eines Computers gespeichert und verarbeitet werden kann. Und dies entspricht tatsächlich der Menge der Materialien, die langzeitarchiviert werden sollen, vom einfachen Textdokument im .txt-Format über umfangreiche PDF-Dateien mit eingebetteten Multimedia-Dateien bis hin zu kompletten Betriebssystemen. Ein digitales Objekt kann beispielsweise eine Datei in einem spezifischen Dateiformat sein, zum Beispiel eine einzelne Grafik, ein Word-Dokument oder eine PDF-Datei. Als ein digitales Objekt können allerdings auch komplexere Objekte bezeichnet werden wie Anwendungsprogramme (beispipelsweise Microsoft Word und Mozilla Firefox), eine komplette Internet-Seite mit all ihren Texten, Grafiken und Videos, eine durchsuchbare Datenbank auf CD inklusive einer Suchoberfläche oder gar ein Betriebssystem wie Linux, Mac OS oder Windows.

Ein digitales Objektes kann auf drei Ebenen beschrieben werden, als *physisches Objekt*, als *logisches Objekt* und schließlich als *konzeptuelles Objekt*.

Als *physisches Objekt* sieht man die Menge der Zeichen an, die auf einem Informationsträger gespeichert sind – die rohe Manifestation der Daten auf dem Speichermedium. Die Art und Weise der physischen Beschaffenheit dieser Zeichen kann aufgrund der unterschiedlichen Beschaffenheit des Trägers

sehr unterschiedlich sein. Auf einer CD-ROM sind es die sogenannten „Pits“ und „Lands“ auf der Trägeroberfläche, bei magnetischen Datenträgern sind es Übergänge zwischen magnetisierten und nicht magnetisierten Teilchen. Auf der physischen Ebene haben die Bits keine weitere Bedeutung außer eben der, dass sie binär codierte Information enthalten, also entweder die „0“ oder die „1“. Auf dieser Ebene unterscheiden sich beispielsweise Bits, die zu einem Text gehören, in keiner Weise von Bits, die Teil eines Computerprogramms oder Teil einer Grafik sind.

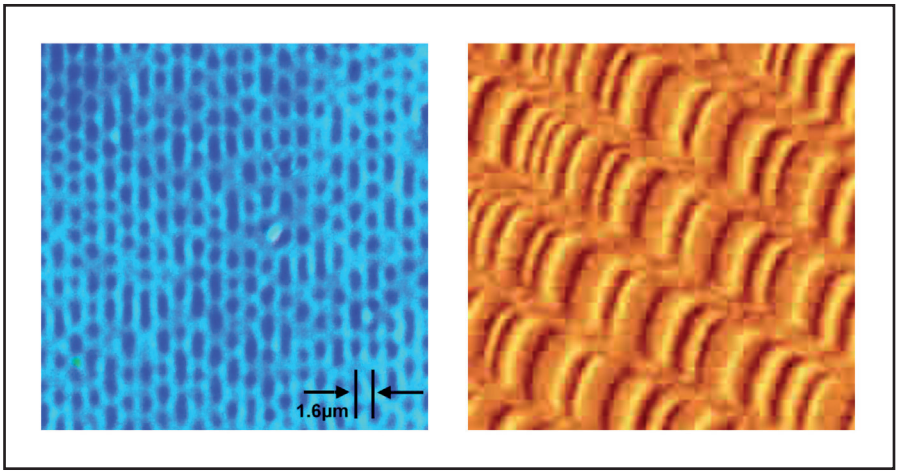


Abbildung 1: Das physische Objekt: „Nullen“ und „Einsen“ auf der Oberfläche einer CD-Rom (blau) und einer Festplatte (gelb) [1].

Die Erhaltung dieses Bitstreams (auch Bitstreamerhaltung) ist der erste Schritt zur Konservierung des gesamten digitalen Objekts, er bildet sozusagen die Grundlage aller weiteren Erhaltungs-Strategien.
Unter einem *logischen Objekt* versteht man eine Folge von Bits, die von einem Informationsträger gelesen und als eine Einheit angesehen werden kann. Diese können von einer entsprechenden Software als Format erkannt und verarbeitet werden. In dieser Ebene existiert das Objekt nicht nur als Bitstream, es hat bereits ein definiertes Format. Die Bitstreams sind auf dieser Ebene schon sehr viel spezieller als die Bits auf dem physischen Speichermedium. So müssen diese zunächst von dem Programm, das einen solchen Bitstream zum Beispiel

1 Bildquelle CD-Rom-Oberfläche: http://de.wikipedia.org/wiki/Datei:Compactdiscar.jpg, Bildquelle Festplatten-Oberfläche: http://leifi.physik.uni-muenchen.de/web_ph10/umwelt-technik/11festplatte/festplatte.htm
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

als eine Textdatei erkennen soll, als eine solche identifizieren. Erst wenn der Bitstream als korrekte Textdatei erkannt worden ist, kann er vom Programm als Dateiformat interpretiert werden.

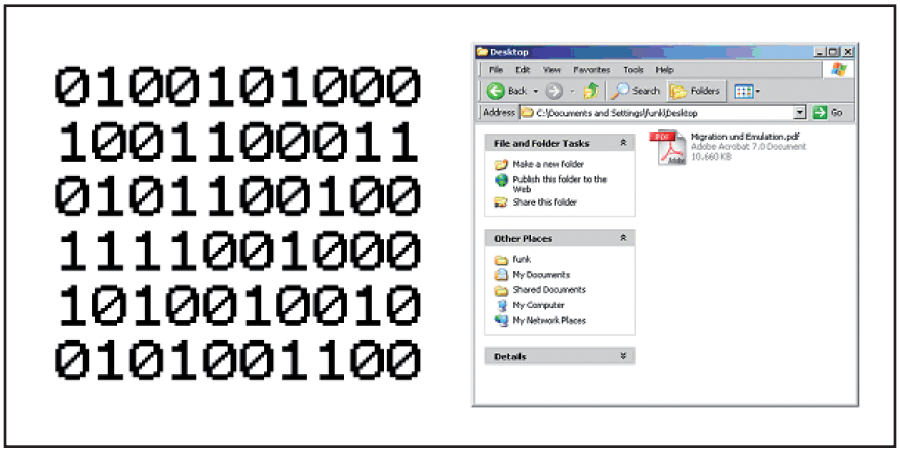


Abbildung 2: Das logische Objekt: Eine Bit-Folge als Repräsentation eines PDF-Dokuments

Will man diesen logischen Einheiten ihren Inhalt entlocken, muss das Format dieser Einheit genau bekannt sein. Ist ein Format nicht hinreichend bekannt oder existiert die zu dem Format gehörige Software nicht mehr, so wird die ursprüngliche Information des logischen Objektes sehr wahrscheinlich nicht mehr vollständig zu rekonstruieren sein. Um solche Verluste zu vermeiden, gibt es verschiedene Lösungsansätze, zwei davon sind Migration und Emulation.
Das *konzeptuelle Objekt* beschreibt zu guter Letzt die gesamte Funktionalität, die dem Benutzer des digitalen Objekts mit Hilfe von dazu passender Soft- und Hardware zur Verfügung steht – es ist das Objekt „zum Begreifen". Dies sind zunächst die Objekte, Zeichen und Töne, die der Mensch über seine Sinne wahrnimmt. Auch interaktive Dinge wie das Spielen eines Computerspiels oder eine durchsuchbare Datenbank zählen dazu, denn die Funktion eines Computerspiels ist es, gespielt werden zu können. Ein weiteres Beispiel ist eine komplexe Textdatei mit all ihren Editierungsmöglichkeiten, Tabellen und enthaltenen Bildern, die das verarbeitende Programm bietet.

Dieses konzeptuelle Objekt ist also die eigentliche, für den Betrachter bedeutungsvolle Einheit, sei es ein Buch, ein Musikstück, ein Film, ein Computerprogramm oder ein Videospiel. Diese Einheit ist es, die der Nachwelt erhalten bleiben soll und die es mit Hilfe der digitalen Langzeitarchivierung zu schützen gilt.

Das Ziel eines Langzeitarchivs ist es also, das konzeptuelle Objekt zu archivieren und dem Nutzer auch in ferner Zukunft Zugriff auf dessen Inhalte zu gewähren. Die Darstellung bzw. Nutzung des digitalen Objekts soll so nahe wie möglich den Originalzustand des Objekts zur Zeit der Archivierung wiederspiegeln. Dies ist nicht möglich, wenn sich bereits Probleme bei der Archivierung auf den unteren Ebenen, der logischen und der physischen Ebene, ergeben. Gibt es eine unbeabsichtigte Veränderung des Bitstreams durch fehlerhafte Datenträger oder existiert eine bestimmte Software nicht mehr, die den Bitstream als Datei erkennt, ist auch eine Nutzung des Objekts auf konzeptueller Ebene nicht mehr möglich.

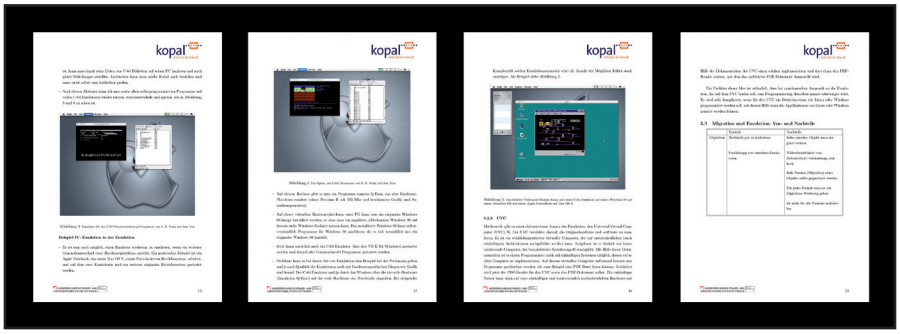


Abbildung 3: Das konzeptuelle Objekt: Die PDF-Datei mit allen ihren Anzeige- und Bearbeitungsmöglich-keiten

Formate

Ein Computer-Programm muss die Daten, die es verwaltet, als Bit-Folge auf einen dauerhaften Datenspeicher (zum Beispiel auf eine CD oder eine Festplatte) ablegen, damit sie auch nach Ausschalten des Computers sicher verwahrt sind. Sie können so später erneut in den Rechner geladen werden. Damit die gespeicherten Daten wieder genutzt werden können, ist es erforderlich, dass das ladende Programm die Bit-Folge exakt in der Weise interpretiert, wie es beim Speichern beabsichtigt war.

Um dies zu erreichen, müssen die Daten in einer Form vorliegen, die sowohl das speichernde als auch das ladende Programm gleichfalls „verstehen“ und interpretieren können. Ein Programm muss die Daten, die es verwaltet, in einem definierten *Dateiformat* speichern können. Dies bedeutet, alle zu speichernden Daten in eine genau definierte Ordnung zu bringen, um diese dann als eine Folge von Bits zu speichern, als sogenannten *Bitstream*. Die Bits, mit denen beispielsweise der Titel eines Dokuments gespeichert ist, müssen später auch wie-

der exakt von derselben Stelle und semantisch als Titel in das Programm geladen werden, damit das Dokument seine ursprüngliche Bedeutung behält. Somit muss das Programm das Format genau kennen und muss wissen, welche Bits des Bitstreams welche Bedeutung haben, um diese korrekt zu interpretieren und verarbeiten zu können.

Formate sind also wichtig, damit eine Bit-Folge semantisch korrekt ausgewertet werden kann. Sind zwei voneinander unabhängige Programme fähig, ihre Daten im selben Format zu speichern und wieder zu laden, ist ein gegenseitiger Datenaustausch möglich. Für die digitale Langzeitarchivierung sind Formate sehr relevant, weil hier zwischen dem Schreiben der Daten und dem Lesen eine lange Zeit vergehen kann. Die Gefahr von (semantischen) Datenverlusten ist daher sehr groß, denn ein Lesen der Daten ist nicht mehr möglich, wenn das Format nicht mehr interpretiert werden kann.

Eine *Format-Spezifikation* ist eine Beschreibung der Anordnung der Bits, das heißt eine Beschreibung, wie die Daten abgelegt und später interpretiert werden müssen, um das ursprüngliche Dokument zu erhalten. Grob kann zwischen proprietären und offenen *Dateiformaten* unterschieden werden. Bei proprietären Dateiformaten ist die Spezifikation oft nicht oder nicht hinreichend bekannt, bei offenen Formaten hingegen ist die Spezifikation frei zugänglich und oft gut dokumentiert. Aus einer Datei, deren Format und Spezifikation bekannt ist, kann die gespeicherte Information auch ohne das vielleicht nicht mehr verfügbare lesende Programm extrahiert werden, da mit der Spezifikation eine Anleitung zur semantischen Interpretation vorhanden ist.

Zum *Format-Standard* kann eine Format-Spezifikation dann werden, wenn sich das durch sie beschriebene Format weithin als einheitlich für eine bestimmte Nutzung durchgesetzt hat – auch und gerade gegenüber anderen Formaten – und es von vielen beachtet und genutzt wird. Ein solcher Vorgang kann entweder stillschweigend geschehen oder aber gezielt durch einen Normungsprozess herbeigeführt werden, indem eine möglichst breite Anwendergruppe solange an einer Spezifikation arbeitet, bis diese von allen Beteiligten akzeptiert wird und anwendbar erscheint. Als Ergebnis eines solchen Normungsprozesses wird die erarbeitete Format-Spezifikation als Norm von einer Behörde oder Organisation veröffentlicht und dokumentiert. Als Beispiel ist hier auf nationaler Ebene das Deutsches Institut für Normung e.V. (DIN) zu nennen, auf europäischer und internationaler Ebene das Europäisches Komitee für Normung (CEN) und die Internationale Organisation für Normung (ISO).

Literatur

Consultative Committee for Space Data Systems (2001): *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, BLUE BOOK, http://public.ccsds.org/publications/archive/650x0b1.pdf

Huth, Karsten, Andreas Lange (2004): *Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive*, http://www.archimuse.com/publishing/ichim04/2758_HuthLange.pdf

Thibodeau, Kenneth (2002): Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, In: *Council on Library and Information Resources: The State of Digital Preservation: An International Perspective,* http://www.clir.org/PUBS/reports/pub107/thibodeau.html

Abrams, Steffen, Sheila Morrissey, Tom Cramer (2008): *What? So what? The Next-Generation JHOVE2 Architecture for Format-Aware Characterization*, http://confluence.ucop.edu/download/attachments/3932229/Abrams_a70_pdf.pdf?version=1

7.3 Auswahlkriterien

Jens Ludwig

Formate sind in unterschiedlichem Maße dem Risiko zu veralten ausgesetzt. Daher ist es naheliegend die langfristige Nutzbarkeit der digitalen Objekte durch die Verwendung eines geeigneten Formates zu unterstützen. Bevor man aber versucht zu beantworten, welches Format theoretisch am besten für die Langzeitarchivierung geeignet ist, muss man sich klarmachen, was die begrenzenden Faktoren der Formatwahl sind.

Die wichtigste und in gewissem Sinne triviale Einschränkung der Formatauswahl ist, dass ein Format auch die benötigte Funktionalität aufweisen muss. Es gibt Formate mit identischen Funktionen, die leicht durcheinander ersetzt werden können, aber genauso Formate für Spezialzwecke, die man dann leider nicht mit für die Langzeitarchivierung besser geeigneten austauschen kann, weil diese Spezialfunktionen eben benötigt werden. Um ein Format auswählen zu können, muss man sich also bewusst sein, was für Funktionalitäten benötigt werden.

In diesem Zusammenhang gilt es auch die Position des „Langzeitarchivierers" in der Verarbeitungskette zu berücksichtigen: Muss schon bei der Bearbeitung und Erstellung des digitalen Objekts das richtige Format ausgewählt werden, weil z.B. ein Dokument genauso wiederverwendet und bearbeitet werden soll? Dann muss man selbst der Ersteller sein oder hinreichenden Einfluss auf die Erstellung haben, sonst muss man hinnehmen, was man bekommt. Oder reicht ggf. eine statische Version, die nur den visuellen Eindruck erhält, und es ist deshalb möglich, das Objekt in ein neues, selbst ausgewähltes Format zu überführen?

Und selbst wenn die digitalen Objekte in den nach bestem Wissen und Gewissen ausgesuchten Formaten vorliegen, heißt das nicht, dass das Problem gelöst ist. Quasi jedes Format kann veralten, auch wenn es sich einmal als die beste Wahl dargestellt hat, Anforderungen können sich ändern und der technische Fortschritt kann neue Formate ermöglichen und erfordern. Aus all diesen Gründen kann man keine dauerhafte Festlegung auf ein bestimmtes Format treffen.

Kriterien

Trotz dieser Einschränkungen und Absicherungen lassen sich aber eine Reihe von allgemeinen Faktoren aufführen, was für Formate für digitale Objekte sinnvoll sind, die langfristig genutzt werden sollen. Und es gibt eine Vielzahl von Katalogen, die solche Faktoren aufführen: z.B. eher klassische Aufstellungen wie Lormant et al. (2005), Stanescu (2004) oder Arms, Fleischhauer (2005), deren Autoren auch die informative Seite der Library of Congress zum Thema Formate pflegen (Arms, Fleischhauer 2007), aber auch spezialisierte wie Barkstrom, Folk (2002), die besonders Erwägungen für naturwissenschaftliche Forschungsdaten berücksichtigen, oder Christensen et al. (2004), die Kriterien für Kontainerformate zur Internetarchivierung aufstellen. So interessant die unterschiedlichen Perspektiven und Kriterien im Detail sein mögen, auf einer abstrakten Ebenen lassen sich die Kriterien zusammenfassen. Angelehnt an Rog, van Wijk (2008) sind zu nennen:

- Offenheit: Ist die Spezifikation des Formates frei verfügbar oder ist sie ein Betriebsgeheimnis eines Herstellers? Ist das Format durch Normungsinstitutionen standardisiert worden? Mit der Spezifikation besteht die Möglichkeit, dass Format zu verstehen und ggf. selbst Nutzungssoftware zu entwickeln, auch wenn es keinen Anbieter mehr gibt.
- Verbreitung: Wie verbreitet ist das Format? Wie häufig wird es genutzt, wieviel unabhängige Anbieter von Nutzungssoftware gibt es? Eine hohe Verbreitung ist ein Indiz dafür, dass das Format noch lange und von vieler Software unterstützt wird, da ein großer Markt dafür vorhanden ist.
- Komplexität: Wie kompliziert ist das Format? Technische Komplexität erschwert die fehlerfreie Entschlüsselung bzw. Nutzung. Je mehr Wissen zum Verständnis eines Formates notwendig ist, desto eher kann ein Teil des notwendigen Wissens verloren gehen.
- Schutzmechanismen: Kopierschütze und Verschlüsselungen mögen für bestimmte Anwendungen sinnvoll sein, für die Langzeitarchivierung sind sie es nicht. Die langfristige Erhaltung setzt das Kopieren der digitalen Objekte voraus und eine Verschlüsselung erfordert als Minimum die zusätzliche Kenntnis des Schlüssels und Verschlüsselungsverfahrens.
- Selbstdokumentation: Wenn ein Format die Integration von Metadaten ermöglicht, dann erleichtert das voraussichtlich das Verständnis des digitalen Objekts und verringert die Abhängigkeit von externen Metadatenquellen.
- Robustheit: Je robuster ein Format ist, desto weniger wirken sich Veränderungen aus. Wie stark wirken sich Fehler einzelner Bits auf die Nutz-

barkeit des gesamten Objekts aus? Gibt es nur einen kleinen, vernachlässigbaren Darstellungsfehler oder lässt es sich ggf. überhaupt nicht mehr nutzen? Wie kompatibel sind verschiedene Versionen bzw. Weiterentwicklungen des Formats untereinander?

- Abhängigkeiten: Formate, die weniger von spezieller Hard- oder Software oder anderen Ressourcen (z.B. Internetzugang) abhängig sind als andere, sind zu bevorzugen.

Wie bereits erwähnt wurde, sind über diese Kritierien hinaus die spezifisch benötigten Funktionalitäten zu erwägen. Diese selbst nur für bestimmte Medientypen auszuführen, würden den Umfang dieses Kapitels sprengen. Gute weiterführende Quellen für bestimmte Medientypen sind neben dem Kapitel „Vorgehensweise für ausgewählte Objekttypen“ dieses Handbuchs auch Arms, Fleischhauer (2007) und AHDS (2006).

Literatur

AHDS (arts and humanities data service): Preservation Handbooks. 2006. http://ahds.ac.uk/preservation/ahds-preservation-documents.htm

Arms, Caroline/ Fleischhauer, Carl: *Digital Formats: Factors for Sustainability, Functionality, and Quality.* 2005. Paper for IS&T Archiving 2005 Conference, Washington, D.C. http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pd

Arms, Caroline/ Fleischhauer, Carl (Hrsg.): Sustainability of Digital Formats. Planning for Library of Congress Collections. 2007. http://www.digitalpreservation.gov/formats/index.shtml

Barkstrom, Bruce R./ Folk, Mike: *Attributes of File Formats for Long Term Preservation of Scientific and Engineering Data in Digital Libraries.* 2002. http://www.ncsa.uiuc.edu/NARA/Sci_Formats_and_Archiving.doc

Christensen, Steen S. et al.: *Archival Data Format Requirements.* 2004. http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf

Lormant, Nicolas et al.: *How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?* 2005. Paper for PV 2005, The Royal Society, Edinburgh. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf

Rog, Judith/ van Wijk, Caroline: *Evaluating File Formats for Long-term Preservation.* 2008. http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf

Stanescu, Andreas: *Assessing the Durability of Formats in a Digital Preservation Environment.* In: D-Lib Magazine, November 2004, Volume 10 Number 11. doi:10.1045/november2004-stanescu

7.4 Formatcharakterisierung

Stefan E. Funk und Matthias Neubauer

Die Archivierung von digitalen Objekten steht und fällt mit der Charakterisierung und Validierung der verwendeten Dateiformate. Ohne die Information, wie die Nullen und Einsen des Bitstreams einer Datei zu interpretieren sind, ist der binäre Datenstrom schlicht unbrauchbar. Vergleichbar ist dies beispielsweise mit der Entzifferung alter Schriften und Sprachen, deren Syntax und Grammatik nicht mehr bekannt sind. Daher ist es für die digitale Langzeitarchivierung essentiell, die Dateien eines digitalen Objektes vor der Archivierung genauestens zu betrachten und zu kategorisieren.

Eine nach oben genannten Kriterien erfolgte Auswahl geeigneter Formate ist ein erster Schritt zu einer erfolgreichen Langzeitarchivierung. Eine automatisierte Charakterisierung der vorliegenden Formate ist ein weiterer Schritt. Die Speicherung der digitalen Objekte und deren Archivierung sollte unabhängig voneinander geschehen können, daher muss davon ausgegangen werden, dass außer dem zu archivierenden Objekt selbst keinerlei Daten zu dessen Format vorliegen.

Ziel einer Charakterisierung ist es, möglichst automatisiert das Format einer Datei zu identifizieren und durch Validierung zu kontrollieren, ob diese Datei auch deren Spezifikationen entspricht – bei einer sorgfältigen Auswahl des Formats ist diese ja bekannt. Eine einer Spezifikation entsprechende Datei kann später, beispielsweise für eine Format-Migration, nach dieser Spezifikation interpretiert werden und die Daten in ein aktuelleres Format umgewandelt werden. Außerdem sollen möglichst viele technische Daten über das Objekt (technische Metadaten) aus dem vorliegenden Objekt extrahiert werden, so dass eine Weiterverwendung auch in ferner Zukunft hoffentlich wahrscheinlich ist.

7.4.1 Identifizierung

Bei der *Identifizierung* eines digitalen Objekts geht es in erster Linie um die Frage, welches Format nun eigentlich vorliegt. Als Anhaltspunkte können zunächst interne oder externe Merkmale einer Datei herangezogen werden, zum Beispiel ein *HTTP Content-Type Header* oder ein *Mimetype* – zum Beispiel „text/xml“ für eine XML-Datei oder „application/pdf“ für eine PDF-Datei, die *Magic Number* oder als externes Merkmal eine *File Extension* (Dateiendung).

Die Dateiendung oder File Extension bezeichnet den Teil des Dateinamens, welcher rechts neben dem letzten Vorkommen eines Punkt-Zeichens liegt (wie

beispielsweise in „Datei.ext“). Dieses Merkmal ist jedoch meist nicht in einer Formatspezifikation festgelegt, sondern wird lediglich zur vereinfachten, oberflächlichen Erkennung und Eingruppierung von Dateien in Programmen und manchen Betriebssystemen genutzt. Vor allem aber kann die Dateiendung jederzeit frei geändert werden, was jedoch keinerlei Einfluss auf den Inhalt und damit auf das eigentliche Format der Datei hat. Daher ist es nicht ratsam, sich bei der Formaterkennung allein auf die Dateiendung zu verlassen, sondern in jedem Fall noch weitere Erkennungsmerkmale zu überprüfen, sofern dies möglich ist.

Einige Dateiformat-Spezifikationen definieren eine so genannte Magic Number. Dies ist ein Wert, welcher in einer Datei des entsprechenden Formats immer an einer in der Spezifikation bestimmten Stelle[2] der Binärdaten gesetzt sein muss. Anhand dieses Wertes kann zumindest sehr sicher angenommen werden, dass die fragliche Datei in einem dazu passenden Format vorliegt. Definiert ein Format keine Magic Number, kann meist nur durch den Versuch der Anwendung oder der Validierung der Datei des vermuteten Formats Klarheit darüber verschafft werden, ob die fragliche Datei tatsächlich in diesem Format abgespeichert wurde.

7.4.2 Validierung

Die *Validierung* oder auch Gültigkeitsprüfung ist ein wichtiger und notwendiger Schritt vor der Archivierung von Dateien. Auch wenn das Format einer zu archivierenden Datei sicher bestimmt werden konnte, garantiert dies noch nicht, dass die fragliche Datei korrekt gemäß den Formatspezifikationen aufgebaut ist. Enthält die Datei Teile, die gegen die Spezifikation verstoßen, kann eine Verarbeitung oder Darstellung der Datei unmöglich werden. Besonders fragwürdig, speziell im Hinblick auf die digitale Langzeitarchivierung, sind dabei proprietäre und gegebenenfalls undokumentierte Abweichungen von einer Spezifikation oder auch zu starke Fehlertoleranz eines Darstellungsprogrammes.

Ein gutes Beispiel hierfür ist HTML, bei dem zwar syntaktische und grammatikalische Regeln definiert sind, die aktuellen Browser jedoch versuchen, fehlerhafte Stellen der Datei einfach dennoch darzustellen oder individuell zu interpretieren. Wagt man nun einmal einen Blick in die „fernere“ Zukunft – beim heutigen Technologiewandel etwa 20-30 Jahre – dann werden die proprietären Darstellungsprogramme wie beispielsweise die unterschiedlich interpre-

2 Eine bestimmte Stelle in einer Datei wird oft als „Offset“ bezeichnet und mit einem hexadezimalen Wert addressiert

tierenden Web-Browser Internet Explorer und Firefox wohl nicht mehr existieren. Der einzige Anhaltspunkt, den ein zukünftiges Bereitstellungssystem hat, ist also die Formatspezifikation der darzustellenden Datei. Wenn diese jedoch nicht valide zu den Spezifikationen vorliegt, ist es zu diesem Zeitpunkt wohl nahezu unmöglich, proprietäre und undokumentierte Abweichungen oder das Umgehen bzw. Korrigieren von fehlerhaften Stellen nachzuvollziehen. Daher sollte schon zum Zeitpunkt der ersten Archivierung sichergestellt sein, dass eine zu archivierende Datei vollkommen mit einer gegebenen Formatspezifikation in Übereinstimmung ist.

Weiterhin kann untersucht werden, zu welchem Grad eine Formatspezifikation eingehalten wird – dies setzt eine erfolgreiche Identifizierung voraus. Als weiteres Beispiel kann eine XML-Datei beispielsweise in einem ersten Schritt *well-formed* (wohlgeformt) sein, so dass sie syntaktisch der XML-Spezifikation entspricht. In einem zweiten Schritt kann eine XML-Datei aber auch noch *valid* (valide) sein, wenn sie zum Beispiel einem XML-Schema entspricht, das wiederum feinere Angaben macht, wie die XML-Datei aufgebaut zu sein hat.

Da Format-Spezifikationen selbst nicht immer eindeutig zu interpretieren sind, sollte eine Validierung von Dateien gegen eine Spezifikation für die digitale Langzeitarchivierung möglichst konfigurierbar sein, so dass sie an lokale Bedürfnisse angepasst werden kann.

7.4.3 Extraktion, technische Metadaten und Tools

Mathias Neubauer

Wie bei jedem Vorhaben, das den Einsatz von Software beinhaltet, stellt sich auch bei der Langzeitarchivierung von digitalen Objekten die Frage nach den geeigneten Auswahlkriterien für die einzusetzenden Software-Tools.

Besonders im Bereich der Migrations- und Manipulationstools kann es von Vorteil sein, wenn neben dem eigentlichen Programm auch der dazugehörige Source-Code[3] der Software vorliegt. Auf diese Weise können die während der Ausführung des Programms durchgeführten Prozesse auch nach Jahren noch nachvollzogen werden, indem die genaue Abfolge der Aktionen im Source-

3 Der Source- oder auch Quellcode eines Programmes ist die les- und kompilierbare, aber nicht ausführbare Form eines Programmes. Er offenbahrt die Funktionsweise der Software und kann je nach Lizenzierung frei erweiter- oder veränderbar sein (Open Source Software).

Code verfolgt wird. Voraussetzung dafür ist natürlich, dass der Source-Code seinerseits ebenfalls langzeitarchiviert wird.

Nachfolgend werden nun einige Tool-Kategorien kurz vorgestellt, welche für die digitale Langzeitarchivierung relevant und hilfreich sein können.

Formaterkennung

Diese Kategorie bezeichnet Software, die zur Identifikation des Formats von Dateien eingesetzt wird. Die Ergebnisse, welche von diesen Tools geliefert werden, können sehr unterschiedlich sein, da es noch keine global gültige und einheitliche Format Registry gibt, auf die sich die Hersteller der Tools berufen können. Manche Tools nutzen jedoch schon die Identifier von Format Registry Prototypen wie PRONOM (beispielsweise „DROID", eine Java Applikation der National Archives von Großbritannien, ebenfalls Urheber von PRONOM (http://droid.sourceforge.net). Viele Tools werden als Ergebnis einen so genannten „Mime-Typ" zurückliefern. Dies ist jedoch eine sehr grobe Kategorisierung von Formattypen und für die Langzeitarchivierung ungeeignet, da zu ungenau.

Metadatengewinnung

Da es für die Langzeitarchivierung, insbesondere für die Migrationsbemühungen, von großem Vorteil ist, möglichst viele Details über das verwendete Format und die Eigenschaften einer Datei zu kennen, spielen Tools zur Metadatengewinnung eine sehr große Rolle. Prinzipiell kann man nie genug über eine archivierte Datei wissen, jedoch kann es durchaus sinnvoll sein, extrahierte Metadaten einmal auf ihre Qualität zu überprüfen und gegebenenfalls für die Langzeitarchivierung nur indirekt relevante Daten herauszufiltern, um das Archivierungssystem nicht mit unnötigen Daten zu belasten. Beispiel für ein solches Tool ist „JHOVE" (das JSTOR/Harvard Object Validation Environment der Harvard University Library, http://hul.harvard.edu/jhove/), mit dem sich auch Formaterkennung und Validierung durchführen lassen. Das Tool ist in Java geschrieben und lässt sich auch als Programmier-Bibliothek in eigene Anwendungen einbinden. Die generierten technischen Metadaten lassen sich sowohl in Standard-Textform, als auch in XML mit definiertem XML-Schema ausgeben.

Validierung

Validierungstools für Dateiformate stellen sicher, dass eine Datei, welche in einem fraglichen Format vorliegt, dessen Spezifikation auch vollkommen ent-

spricht. Dies ist eine wichtige Voraussetzung für die Archivierung und die spätere Verwertung, Anwendung und Migration beziehungsweise Emulation dieser Datei. Das bereits erwähnte Tool „JHOVE" kann in der aktuellen Version 1.1e die ihm bekannten Dateiformate validieren; verlässliche Validatoren existieren aber nicht für alle Dateiformate. Weit verbreitet und gut nutzbar sind beispielsweise XML Validatoren, die auch in XML Editoren wie „oXygen" (SyncRO Soft Ltd., http://www.oxygenxml.com) oder „XMLSpy" (Altova GmbH, http://www.altova.com/XMLSpy) integriert sein können.

Formatkorrektur

Auf dem Markt existiert eine mannigfaltige Auswahl an verschiedensten Korrekturprogrammen für fehlerbehaftete Dateien eines bestimmten Formats. Diese Tools versuchen selbstständig und automatisiert, Abweichungen gegenüber einer Formatspezifikation in einer Datei zu bereinigen, so dass diese beispielsweise von einem Validierungstool akzeptiert wird. Da diese Tools jedoch das ursprüngliche Originalobjekt verändern, ist hier besondere Vorsicht geboten! Dies hat sowohl rechtliche als auch programmatische Aspekte, die die Frage aufwerfen, ab wann eine Korrektur eines Originalobjektes als Veränderung gilt, und ob diese für die Archivierung gewünscht ist. Korrekturtools sind üblicherweise mit Validierungstools gekoppelt, da diese für ein sinnvolles Korrekturverfahren unerlässlich sind. Beispiel für ein solches Tool ist „PDF/A Live!" (intarsys consulting GmbH, (http://www.intarsys.de/produkte/pdf-a-live), welches zur Validierung und Korrektur von PDF/A konformen Dokumenten dient.

Konvertierungstools

Für Migrationsvorhaben sind Konvertierungstools, die eine Datei eines bestimmten Formats in ein mögliches Zielformat überführen, unerlässlich. Die Konvertierung sollte dabei idealerweise verlustfrei erfolgen, was jedoch in der Praxis leider nicht bei allen Formatkonvertierungen gewährleistet sein kann. Je nach Archivierungsstrategie kann es sinnvoll sein, proprietäre Dateiformate vor der Archivierung zunächst in ein Format mit offener Spezifikation zu konvertieren. Ein Beispiel hierfür wäre „Adobe Acrobat" (Adobe Systems GmbH, http://www.adobe.com/de/products/acrobat/), welches viele Formate in PDF[4] überführen kann.

4 Portable Document Format, Adobe Systems GmbH, Link: http://www.adobe.com/de/products/acrobat/adobepdf.html

Für Langzeitarchivierungsvorhaben empfiehlt sich eine individuelle Kombination der verschiedenen Kategorien, welche für das jeweilige Archivierungsvorhaben geeignet ist. Idealerweise sind verschiedene Kategorien in einem einzigen Open Source Tool vereint, beispielsweise was Formaterkennung, -konvertierung und -validierung betrifft. Formatbezogene Tools sind immer von aktuellen Entwicklungen abhängig, da auf diesem Sektor ständige Bewegung durch immer neue Formatdefinitionen herrscht. Tools, wie beispielsweise „JHOVE", die ein frei erweiterbares Modulsystem bieten, können hier klar im Vorteil sein. Dennoch sollte man sich im Klaren darüber sein, dass die Archivierung von digitalen Objekten nicht mittels eines einzigen universellen Tools erledigt werden kann, sondern dass diese mit fortwährenden Entwicklungsarbeiten verbunden ist. Die in diesem Kapitel genannten Tools können nur Beispiele für eine sehr große Palette an verfügbaren Tools sein, die beinahe täglich wächst.

7.5 File Format Registries

Andreas Aschenbrenner und Thomas Wollschläger

Zielsetzung und Stand der Dinge

Langzeitarchive für digitale Objekte benötigen aufgrund des ständigen Neuerscheinens und Veraltens von Dateiformaten aktuelle und inhaltlich präzise Informationen zu diesen Formaten. File Format Registries dienen dazu, den Nachweis und die Auffindung dieser Informationen in einer für Langzeitarchivierungsaktivitäten hinreichenden Präzision und Qualität zu gewährleisten. Da Aufbau und Pflege einer global gültigen File Format Registry für eine einzelne Institution so gut wie gar nicht zu leisten sind, müssen sinnvollerweise kooperativ erstellte und international abgestimmte Format Registries erstellt werden. Dies gewährleistet eine große Bandbreite, hohe Aktualität und kontrollierte Qualität solcher Unternehmungen.

File Format Registries können verschiedenen Zwecken dienen und dementsprechend unterschiedlich angelegt und folglich auch verschieden gut nachnutzbar sein. Hinter dem Aufbau solcher Registries stehen im Allgemeinen folgende Ziele:

- Formatidentifizierung
- Formatvalidierung
- Formatdeskription/-charakterisierung
- Formatlieferung/-ausgabe (zusammen mit einem Dokument)
- Formatumformung (z.B. Migration)
- Format-Risikomanagement (bei Wegfall von Formaten)

Für Langzeitarchivierungsvorhaben ist es zentral, nicht nur die Bewahrung, sondern auch den Zugriff auf Daten für künftige Generationen sicherzustellen. Es ist nötig, eine Registry anzulegen, die in ihrer Zielsetzung alle sechs genannten Zwecke kombiniert. Viele bereits existierende oder anvisierte Registries genügen nur einigen dieser Ziele, meistens den ersten drei.

Beispielhaft für derzeit existierende File Format Registries können angeführt werden:

(I) die File Format Encyclopedia,
http://pipin.tmd.ns.ac.yu/extra/fileformat/
(II) FILExt,
http://filext.com/
(III) Library of Congress Digital Formats,
http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml
(IV) C.E. Codere's File Format site,
http://magicdb.org/stdfiles.html
(V) PRONOM,
http://www.nationalarchives.gov.uk/pronom/
(VI) das Global Digital Format Registry,
http://hul.harvard.edu/gdfr/
(VIIa) Representation Information Registry Repository,
http://registry.dcc.ac.uk/omar
(VIIb) DCC RI RegRep,
http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCRegRepV04
(VIII) FCLA Data Formats,
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Bewertung von File Format Registries

Um zu beurteilen bzw. zu bewerten, ob sich spezielle File Format Registries für eine Referenzierung bzw. Einbindung in das eigene Archivsystem eignen, sollten sie sorgfältig analysiert werden. Sinnvoll können z.B. folgende Kriterien als Ausgangspunkt gewählt werden:

- Was ist der Inhalt der jeweiligen Registry? Wie umfassend ist sie aufgebaut?
- Ist der Inhalt vollständig im Hinblick auf die gewählte Archivierungsstrategie?
- Gibt es erkennbare Schwerpunkte?
- Wie werden Beschreibungen in die Registry aufgenommen? (Governance und Editorial Process)
- Ist die Registry langlebig? Welche Organisation und Finanzierung steckt dahinter?
- Wie kann auf die Registry zugegriffen werden? Wie können ihre Inhalte in eine lokale Archivierungsumgebung eingebunden werden?

Künftig werden File Format Registries eine Reihe von Anforderungen adressieren müssen, die von den im Aufbau bzw. Betrieb befindlichen Langzeit-Archivsystemen gestellt werden. Dazu gehören u.a. folgende Komplexe:

I) Vertrauenswürdigkeit von Formaten

Welche Rolle spielt die qualitative Bewertung eines Formats für die technische Prozessierung? Braucht man beispielsweise unterschiedliche Migrationsroutinen für Formate unterschiedlicher Vertrauenswürdigkeit? Wie kann dann ein Kriterienkatalog für die Skalierung der *confidence* (Vertrauenswürdigkeit) eines Formats aussehen und entwickelt werden? Unter Umständen müssen hier noch weitere Erfahrungen mit Migrationen und Emulationen gemacht werden, um im Einzelfall zu einem Urteil zu kommen. Es sollte jedoch eine Art von standardisiertem Vokabular und Kriteriengebrauch erreicht werden und transparent sein.

II) Persistent Identifier

Wie können *Persistent Identifier* (dauerhafte und eindeutige Adressierungen) von File Formats sinnvoll generiert werden? So kann es bestimmte Vorteile haben, Verwandtschafts- und Abstammungsverhältnisse von File Formats bereits am Identifier ablesen zu können. Die Identifizierung durch „Magic Numbers" scheint zu diesem Zweck ebenso wenig praktikabel wie die anhand eventueller ISO-Nummern. Die vermutlich bessere Art der Identifizierung ist die anhand von Persistent Identifiers wie URN oder DOI.

III) ID-Mapping

Wie kann ein Mapping verschiedener Identifikationssysteme (Persistent Identifier, interne Identifier der Archivsysteme, ISO-Nummer, PRONOM ID, etc.) durch Web Services erreicht werden, um in Zukunft die Möglichkeit des Datenaustausches mit anderen File Format Registries zu ermöglichen?

IV) Integration spezieller Lösungen

Wie kann in die bisherigen nachnutzbaren Überlegungen anderer Institutionen die Möglichkeit integriert werden, spezifische Lösungen für den Datenaustausch bereit zu halten? Dies betrifft beispielsweise die Möglichkeit, lokale Sichten zu erzeugen, lokale *Preservation Policies* zuzulassen oder aber mit bestimmten Kontrollstatus von eingespielten Records (z.B. „imported", „approved", „deleted") zu arbeiten.

Literatur

Abrams, Seaman: *Towards a global digital format registry.* 69th IFLA 2003. http://archive.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

Representation and Rendering Project: *File Format Report.* 2003. http://www.leeds.ac.uk/reprend/

Lars Clausen: *Handling file formats.* May 2004. http://netarchive.dk/publikationer/FileFormats-2004.pdf

8 Digitale Erhaltungsstrategien

8.1 Einführung

Stefan E. Funk

Wie lassen sich die Dinge bewahren, die uns wichtig sind, Objekte, die wir der Nachwelt am allerliebsten in genau dem Zustand, in dem sie uns vorliegen, erhalten wollen?

Handelt es sich bei diesen Objekten um Texte oder Schriften, wissen wir, dass Stein- und Tontafeln sowie Papyri bei geeigneter Behandlung mehrere tausend Jahre überdauern können. Auch bei Büchern haben wir in den letzten Jahrhunderten Kenntnisse darüber gesammelt, wie diese zu behandeln sind bzw. wie diese beschaffen sein müssen, um nicht der unfreiwilligen Zerstörung durch zum Beispiel Säurefraß oder Rost aus eisenhaltiger Tinte anheim zu fallen. Auch Mikrofilme aus Cellulose mit Silberfilm-Beschichtung sind bei richtiger Lagerung viele Jahrzehnte, vielleicht sogar Jahrhunderte, haltbar. Alle diese Medien haben den Vorteil, dass sie, wenn sie als die Objekte, die sie sind, erhalten werden können, von der Nachwelt ohne viele Hilfsmittel interpretiert

werden können. Texte können direkt von Tafeln oder aus Büchern gelesen und Mikrofilme mit Hilfe eines Vergrößerungsgerätes recht einfach lesbar gemacht werden.

Bei den digitalen Objekten gibt es zwei grundlegende Unterschiede zu den oben genannten analogen Medien: Zum einen werden die digitalen Informationen als Bits (auf Datenträgern) gespeichert. Ein Bit ist eine Informationseinheit und hat entweder den Wert „0“ oder den Wert „1“. Eine Menge dieser Nullen und Einsen wird als Bitstream bezeichnet. Die Lebensdauer der Bits auf diesen Datenträgern kennen wir entweder nur aus Laborversuchen oder wir haben noch nicht genug Erfahrungswerte für eine sichere Angabe der Lebensdauer über einen langen Zeitraum hinweg sammeln können. Schließlich existieren diese Datenträger erst seit einigen Jahren (bei DVDs) oder Jahrzehnten (bei CDs). Eine Reaktion auf die Unsicherheit über die Lebensdauer dieser Medien ist die Bitstreamerhaltung sowie die Mikroverfilmung. Zum anderen ist keines der digitalen Objekte ohne technische Hilfsmittel nutzbar. Selbst wenn wir die Nullen und Einsen ohne Hilfsmittel von den Medien lesen könnten, dann könnten wir wenig bis gar nichts mit diesen Informationen anfangen. Da diese konzeptuellen Objekte digital kodiert auf den Medien gespeichert sind, bedarf es spezieller Hilfsmittel, die diese Informationen interpretieren können. Als Hilfsmittel dieser Art ist einerseits die Hardware zu sehen, die die Daten von den Medien lesen kann (beispielsweise CD- bzw. DVD-Laufwerke) und natürlich die Computer, die diese Daten weiterverarbeiten. Andererseits wird die passende Software benötigt, die die Daten interpretiert und so die digitalen Objekte als konzeptuelle Objekte erst oder wieder nutzbar macht.

Kann der Bitstream nicht mehr interpretiert werden, weil das Wissen um eine korrekte Interpretation verloren ging, ist der Inhalt des konzeptuellen Objektes verloren, obwohl die eigentlichen Daten (der Bitstream) noch vorhanden sind. Lösungsansätze für dieses Problem sind die Migration und die Emulation. Eine weitere Idee ist es, in einem so genannten Computermuseum die originale Hard- und Software bereitzustellen und so die konzeptuellen Objekte zu erhalten.

8.2 Bitstream Preservation

Dagmar Ullrich

Grundlage aller Archivierungsaktivitäten ist der physische Erhalt der Datenobjekte, die Bitstream[1] Preservation. Es wird eine Speicherstrategie vorgeschlagen, die auf einer redundanten Datenhaltung auf mindestens zwei unterschiedlichen, marktüblichen und standardisierten Speichertechniken basiert. Die eingesetzten Speichermedien sollten regelmäßig durch aktuelle ersetzt werden, um sowohl dem physischen Verfall der Speichermedien als auch dem Veralten der eingesetzten Techniken vorzubeugen. Es werden vier Arten von Migrationsprozessen vorgestellt. Das sind: Refreshment, Replication, Repackaging und Transformation. Als Medienmigration im engeren Sinne werden nur die beiden ersten, Refreshment und Replication, angesehen. Sie bezeichnen das Auswechseln einzelner Datenträger (refreshing) oder eine Änderung eingesetzter Speicherverfahren (replication). Durch die kurzen Lebenszyklen digitaler Speichermedien erfolgt ein Erneuern der Trägermedien oft im Rahmen der Aktualisierung der eingesetzten Speichertechnik.

Physischer Erhalt der Datenobjekte

Um digitale Daten langfristig verfügbar zu halten, muss an zwei Stellen angesetzt werden. Zum einen muss der physische Erhalt des gespeicherten Datenobjekts (Bitstreams) auf einem entsprechenden Speichermedium gesichert werden. Zum anderen muss dafür Sorge getragen werden, dass dieser Bitstream auch interpretierbar bleibt, d.h. dass eine entsprechende Hard- und Software-Umgebung verfügbar ist, in der die Daten für einen menschlichen Betrachter lesbar gemacht werden können. Ohne den unbeschädigten Bitstream sind diese weiterführenden Archivierungsaktivitäten sinnlos. Der physische Erhalt der Datenobjekte wird auch als „Bitstream Preservation" bezeichnet. Für den physischen Erhalt des Bitstreams ist eine zuverlässige Speicherstrategie erforderlich.

---

1 Der Begriff „Bitstream" wird hier als selbsterklärend angesehen. Eine Erläuterung des Begriffs findet sich in: Rothenberg, Jeff (1999): *Ensuring the Longevity of Digital Information.* http://www.clir.org/pubs/archives/ensuring.pdf
Bei diesem Text handelt es sich um eine ausführlichere Fassung eines gleichnamigen Artikels, der 1995 in der Zeitschrift „Scientific American", Band 272, Nummer 1, Seiten 42-47 erschienen ist.
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Verfahrensvorschläge für eine Bitstream Preservation

Die nachstehenden vier Verfahrensvorschläge können als Grundlage für eine zuverlässige Speicherstrategie zur Sicherstellung des physischen Erhalts der archivierten Datenobjekte verwendet werden:[2]

1. *Redundante Datenhaltung*: Die Daten sollten in mehrfacher Kopie vorliegen. Zur Sicherung gegen äußere Einflüsse empfiehlt sich auch eine räumlich getrennte Aufbewahrung der unterschiedlichen Kopien.
2. *Diversität eingesetzter Speichertechnik*: Die Daten sollten auf mindestens zwei unterschiedlichen Datenträgertypen gesichert werden.
3. *Standards*: Die verwendeten Speichermedien sollten internationalen Standards entsprechen und auf dem Markt eine weite Verbreitung aufweisen.
4. *Regelmäßige Medienmigration*: Die verwendeten Speichertechniken bzw. Datenträger müssen regelmäßig durch neue ersetzt werden.

Redundanz, Speichertechniken und Standards

Eine mehrfach *redundante Datenhaltung* ist in vielen Bereichen der Datensicherung üblich. Bei wertvollen, insbesondere bei nicht reproduzierbaren Daten wird man sich nicht auf eine einzige Kopie verlassen wollen. Um das Risiko äußerer Einflüsse wie Wasser- oder Brandschäden zu verringern, bietet sich die räumlich getrennte Aufbewahrung der Kopien an. Um auch die Gefahr eines Datenverlusts durch menschliches Versagen oder Vorsatz einzuschränken, kann eine Aufbewahrung bei zwei unabhängigen organisatorischen Einheiten in das Redundanzszenario mit einbezogen werden. Zusätzliche Sicherheit lässt sich gewinnen, indem die jeweiligen Kopien *auf unterschiedlichen Speichertechniken* gehalten werden. Dies mindert das Risiko eines Datenverlusts durch Veralterung einer der eingesetzten Techniken. Sofern vorhanden, sollten Fehlererkennungs- und Korrekturmechanismen zur Sicherung der Datenintegrität eingesetzt werden. Weiter sollte die Funktionstüchtigkeit der Speichermedien und Lesegeräte anhand von Fehlerstatistiken überwacht werden. Die sachgerechte Handhabung von Datenträgern und Lesegeräten ist in jedem Fall vorauszusetzen. Alle

2 Die Auflistung erhebt keinen Anspruch auf Vollständigkeit. Ähnliche Aufstellungen finden sich z.B. in: Rathje, Ulf (2002): Technisches Konzept für die Datenarchivierung im Bundesarchiv. In: Der Archivar, H. 2, Jahrgang 55, S.117-120, http://www.bundesarchiv.de/imperia/md/content/abteilungen/abtb/1.pdf und: o.V. (o.J.) Digital preservation. Calimera Guidelines. S.3. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf

verwendeten Speichertechniken bzw. -medien sollten auf internationalen *Standards* basieren und über eine möglichst breite Nutzergruppe verfügen.

Regelmäßige Medienmigration

Als Medienmigration kann jeder Vorgang betrachtet werden, bei dem das physische Trägermedium eines Datenobjekts innerhalb eines Archivs geändert und der Vorgang mit der Absicht durchgeführt wird, das Datenobjekt zu erhalten, indem die alte Instanz durch die neue ersetzt wird. Eine entsprechende Definition von „Digital Migration" findet sich im OAIS-Referenzmodell[3]:

> *Digital Migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS. It is distinguished from transfers in general by three attributes:*
>
> *- a focus on the Preservation of the full information content*
>
> *- a perspective that the new archival implementation of the information is a replacement for the old; and*
>
> *- full control and responsibility over all aspects of the transfer resides with the OAIS.*

Im OAIS-Referenzmodell werden vier Arten der Migration genannt: Refreshment, Replication, Repackaging und Transformation.[4]

Refreshment: Als Refreshment werden Migrationsprozesse bezeichnet, bei denen einzelne Datenträger gegen neue, gleichartige Datenträger ausgetauscht werden. Die Daten auf einem Datenträger werden direkt auf einen neuen Datenträger gleichen Typs kopiert, der anschließend den Platz des alten in der Speicherinfrastrukur des Archivs einnimmt. Weder an den Daten noch an der Speicherinfrastruktur werden also Änderungen vorgenommen, es wird lediglich ein Datenträger gegen einen gleichartigen anderen ausgetauscht.

Replication: Eine Replication ist ein Migrationsprozess, bei dem ebenfalls Daten von einem Datenträger auf einen neuen kopiert werden. Bei der Replica-

---

3 Consultative Committee for Space Data Systems (CCSDS) (2002): *Reference Model for an Open Archival Information System (OAIS).* Blue Book. Washington DC. Seite 5-1. vgl. auch Kapitel 4.
http://public.ccsds.org/publications/archive/650x0b1.pdf

4 Consultative Committee for Space Data Systems (CCSDS) (2002): *Reference Model for an Open Archival Information System (OAIS).* A.a.O. Seite 5-4.

tion jedoch kann es sich bei dem neuen Datenträger auch um einen andersartigen, z.B. aktuelleren, handeln. Andersartige Datenträger erfordern eine entsprechende Anpassung der Speicherinfrastruktur. Der neue Datenträger kann in der Regel nicht unmittelbar den Platz des alten einnehmen. Der wesentliche Unterschied zum Refreshment liegt daher in den mit dem Kopierprozess einhergehenden Änderungen der verwendeten Speicherinfrastruktur.

Repackaging: Ein Repackaging ist ein Migrationsprozess, bei dem ein sogenanntes Archivpaket verändert wird. Diese Änderung betrifft nicht die eigentlichen Inhaltsdaten, sondern die Struktur des Archivpakets.

Transformation: eine Transformation ist ein Migrationsprozess, bei dem auch die Inhaltsdaten des Archivpakets verändert werden.

Refreshment und Replication können als Medienmigrationen im engeren Sinne angesehen werden. Der Umkopierprozess erfolgt in beiden Fällen mit der Absicht, das Trägermedium zu ersetzen, unabhängig davon, welche Inhalte auf ihm abgelegt sind. Die Replication wird im Folgenden im Sinne eines Technologiewechsels interpretiert.[5] Ein Refreshment beschränkt sich dagegen auf den Wechsel einzelner Datenträger innerhalb einer Speichertechnik, z.B. einer Magnetbandgeneration. Bei Repackaging und Transformation dagegen werden auch die Datenobjekte selbst umgeschrieben. Ein Beispiel für ein Repackaging ist die Änderung des Packformats von ZIP zu TAR. Eine Formatmigration, z.B. von JPG zu TIFF, ist dagegen eine Transformation, da die Inhalte des Archivpakets verändert werden. Die Unterscheidung dieser vier Arten von Migrationen erleichtert die begriffliche Abgrenzung einer Medienmigration von einer Formatmigration. Eine Formatmigration umfasst letztlich immer auch eine Medienmigration, da ein neues Datenobjekt erstellt und auf einem eigenen Trägermedium abgelegt wird. Die Formatmigration erfolgt aber mit Blick auf die künftige Interpretierbarkeit des Bitsreams, die Medienmigration im engeren Sinne hingegen dient dessen Erhalt. Für die Bitstream Preservation sind nur die beiden ersten, Refreshment und Replication, wesentlich, da die beiden anderen den Bitstream verändern. Ein Refreshment ist in der Regel weniger aufwändig als eine Replication, da nicht das Speicherverfahren, sondern nur einzelne Datenträger erneuert werden.

5 Eine Replication muss nach der zitierten Definition nicht notwendig von einem veralteten Medium auf ein aktuelleres erfolgen, sondern ggf. auch auf ein gleichartiges. In der Praxis wird das aber eher selten der Fall sein.

Refreshment und Replication

Ein Erneuern (refreshing) einzelner Datenträger kann aufgrund von Fehlerraten oder auf der Basis bestimmter Kriterien wie Zugriffshäufigkeit oder Alter erfolgen. Der Aufwand solcher Maßnahmen ist gegen die Wahrscheinlichkeit eines Datenverlusts durch einen fehlerhaften Datenträger abzuwägen. Auf der einen Seite können zusätzliche Kontrollverfahren eine sehr hohe Systemlast erzeugen, die den aktiven Zugriff auf die Daten beträchtlich einschränken kann. Zudem sind die Beurteilungskriterien wie Zugriffshäufigkeit, Alter und ggf. die tolerierbare Fehlerrate oft strittig und zum Teil nur mit teurer Spezialsoftware oder auch gar nicht feststellbar. Nicht selten können sie auch im Einzelfall durch Unterschiede in Produktionsablauf oder Handhabung zwischen Datenträgern desselben Typs stark variieren. Auf der anderen Seite wird die Haltbarkeit von Trägermedien aufgrund des raschen Technologiewandels meist gar nicht ausgereizt. Die Wahrscheinlichkeit schadhafter Datenträger durch altersbedingten Verfall ist daher eher gering. Um diesen Zusammenhang deutlich zu machen, kann die durchschnittliche Lebensdauer eines Datenträgers von seiner durchschnittlichen Verfallszeit unterschieden werden.[6]

„*Medium Expected Lifetime (MEL)*: The estimated amount of time the media will be supported and will be operational within the electronic deposit system."

"*Medium Decay Time (MDT)*: The estimated amount of time the medium should operate without substantial read and write errors."

Die Definition der durchschnittlichen Lebensdauer enthält zwei durch „und" verbundene Zeitangaben. Die eine bezieht sich auf die Dauer der Unterstützung eines Speichermediums durch den Hersteller, die andere auf die Dauer des Einsatzes eines Speichermediums im digitalen Archiv. Diese beiden Zeitspannen können durchaus differieren. Nicht selten zwingt die wegfallende Unterstützung durch den Hersteller zur Migration, auch wenn die vorhandenen Systeme voll funktionsfähig sind und noch weiter betrieben werden könnten. Für Speichertechniken, die vom Hersteller nicht mehr unterstützt werden, können Ersatzteile oder technische Betreuung nicht mehr garantiert werden. Ihr Weiterbetrieb ist daher nicht ratsam. Der Begriff der durchschnittlichen Lebensdauer wird aus diesen Gründen hier als die durchschnittlich zu erwartende Hersteller-Unterstützung interpretiert. Solange diese durchschnittliche Lebensdauer unter der durchschnittlichen Verfallszeit liegt, ist ein Ausfall einzelner

6 Van Diessen, Raymond J. und van Rijnsoever, Ben J. (2002): *Managing Media Migration in a Deposit System. IBM/KB Long-Term Preservation Study Report Series Nr. 5*. Amsterdam: IBM Niederlande. S.4.
http://www-5.ibm.com/nl/dias/resource/migration.pdf

Datenträgern selten zu erwarten. Statt aufwändiger Kontrollen der Datenträger kann es in diesem Fall einfacher sein, auf eine redundante Datenhaltung zu vertrauen, im konkreten Fehlerfall einzelne Datenträger oder Laufwerke zu ersetzen und den gesamten Bestand im Rahmen eines Technologiewechsels (Replication) komplett auszutauschen.

Eine Replication im Sinne eines Technologiewechsels umfasst Änderungen in der bestehenden Speicherinfrastrukur. Erforderliche Technologiewechsel können sehr unterschiedlich ausfallen. Sie können von einer Magnetbandgeneration zur nächsten reichen oder einen vollständigen Wechsel z.B. von Magnetbändern zu optischen Medien bedeuten. Im ersten Schritt muss die neue Speichertechnik in die bestehende Infrastruktur integriert werden. Anschließend müssen die betroffenen Datenbestände von der alten Technik auf die neue umkopiert werden. Bei großen Datenmengen mit ggf. hohen Sicherheits- oder Verfügbarkeitsansprüchen können diese Umkopierprozesse aufwändig und langwierig sein. Die Lesegeschwindigkeit der älteren Speichermedien wird in der Regel langsamer sein als die Schreibgeschwindigkeit der neuen. Beide müssen für einen Kopierprozess koordiniert werden, ggf. über Zwischenspeicher. Der Übertragungsvorgang muss abgeschlossen sein, bevor die alte Speichertechnik unbrauchbar wird. An diesem Punkt sei auf die oben ausgeführte Interpretation von „Medium Expected Lifetime“ hingewiesen. Dass der Migrationsprozess abgeschlossen sein muss, bevor eine Speichertechnik nicht mehr auf dem Markt ist, wäre ein sehr hoher Anspruch, da viele Speichermedien nur drei bis fünf Jahre lang angeboten werden. Unter Umständen kann ein solcher Anspruch je nach Wert der betroffenen Daten gerechtfertigt sein. Häufig bieten Hersteller die Unterstützung von Speichermedien einige Jahre länger an, als diese Technik aktiv vertrieben wird. Dies verlängert die zuverlässige Einsatzdauer von Speichertechniken. Eine zusätzliche Sicherheit kann in diesem Kontext auch der Verfahrensvorschlag unterschiedliche Speichertechniken einzusetzen bieten.

Zusammenfassung

Ein Langzeitarchiv muss über zuverlässige Speicherstrategien verfügen, die nicht nur ein „Refreshment“ eingesetzter Datenträger innerhalb einer Speichertechnik ermöglichen, sondern darüber hinaus auch die Erneuerung ganzer Speichertechniken. Solche Strategien müssen sicherstellen, dass zu keinem Zeitpunkt Datenbestände unzugänglich werden, weil ihre Trägermedien nicht mehr lesbar sind.

Literatur

Rothenberg, Jeff (1999), *Ensuring the Longevity of Digital Information*. http://www.clir.org/pubs/archives/ensuring.pdf
Bei diesem Text handelt es sich um eine ausführlichere Fassung eines gleichnamigen Artikels, der 1995 in der Zeitschrift „Scientific American“, Band 272, Nummer 1, Seiten 42-47 erschienen ist.

Rathje, Ulf (2002): *Technisches Konzept für die Datenarchivierung im Bundesarchiv*. In: Der Archivar, H. 2, Jahrgang 55, S.117-120. http://www.bundesarchiv.de/imperia/md/content/abteilungen/abtb/1.pdf

o.V. (o.J.) *Digital preservation*. Calimera Guidelines. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf

Consultative Committee for Space Data Systems (CCSDS) (2002): *Reference Model for an Open Archival Information System (OAIS)*. Blue Book. Washington DC. Seite 5-1. http://public.ccsds.org/publications/archive/650x0b1.pdf

Van Diessen, Raymond J. und van Rijnsoever, Ben J. (2002): *Managing Media Migration in a Deposit System*. IBM/KB Long-Term Preservation Study Report Series Nr. 5. Amsterdam: IBM Niederlande. http://www-5.ibm.com/nl/dias/resource/migration.pdf

8.3 Migration

Stefan E. Funk

Migration und Emulation

Wenn die Archivierung des Bitstreams sichergestellt ist (siehe Bitstreamerhaltung), kann man beginnen, sich über die Archivierung und vor allem über die Nutzung von digitalen Objekten Gedanken zu machen. Bei nicht digitalen Medien wie Büchern und Mikrofilmen hat man in den letzten Jahrzehnten und Jahrhunderten sehr viel Erfahrung mit deren Erhaltung gesammelt, das heißt, auf physikalischer Ebene konnten und können diese Medien sehr lange verfügbar gehalten werden. Ein Buch braucht als zu erhaltendes Objekt auch nur auf der physischen Ebene betrachtet zu werden, denn zum Benutzen eines Buches reicht es aus, das Buch selbst zu erhalten und so die Lesbarkeit zu gewährleisten.

Zwei Strategien, welche die Lesbarkeit der archivierten digitalen Dokumente über lange Zeit (Long Term) garantieren sollen, sind zum einen die Migration und zum anderen die Emulation. „Long term" wird vom Consultative Committee for Space Data Systems (CCSDS) definiert als:

> *„Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely."*

Die Migration passt die digitalen Objekte selbst einem neuen Umfeld an, die Dokumente werden zum Beispiel von einem veralteten Dateiformat in ein aktuelles konvertiert. Mit der Emulation wird das originäre Umfeld der digitalen Objekte simuliert, das neue Umfeld also an die digitalen Objekte angepasst. Diese Strategien können alternativ genutzt werden; sie sind unabhängig voneinander.

Um ein digitales Dokument archivieren und später wieder darauf zugreifen zu können, sind möglichst umfassende Metadaten nötig, also Daten, die das digitale Objekt möglichst genau beschreiben. Dazu gehören in erster Linie die technischen Metadaten. Für die Migration sind weiterhin die Provenance Metadaten wichtig, die wie erläutert die Herkunft des Objekts beschreiben. Deskriptive Metadaten sind aus technischer Sicht nicht so interessant. Sie werden benötigt, um später einen schnellen und komfortablen Zugriff auf die Objekte zu ermöglichen. Rechtliche Metadaten können schließlich genutzt werden, um Einschränkungen für die Migration, die Emulation und den Zugriff auf die digitalen Objekte festzulegen.

Migration

Mit dem Stichwort Migration werden innerhalb der Langzeitarchivierungs-Community unterschiedliche Prozesse bezeichnet. Dies sind sowohl die Datenträgermigration als auch die Daten- oder Formatmigration.

Bei der Datenträgermigration werden Daten von einem Träger auf einen anderen kopiert, z.B. von Festplatte auf CD, von DVD auf Band etc. Diese Art der Migration ist die Grundlage der physischen Erhaltung der Daten, der Bitstream Preservation.

Bei einer Datenmigration (auch Formatmigration genannt) werden Daten von einem Datenformat in ein aktuelleres, möglichst standardisiertes und offen gelegtes Format überführt. Dies sollte geschehen, wenn die Gefahr besteht, dass archivierte Objekte aufgrund ihres Formates nicht mehr benutzt werden können. Das Objekt selbst wird so verändert, dass seine Inhalte und Konzepte erhalten bleiben, es jedoch auf aktuellen Rechnern angezeigt und benutzt werden kann. Problematisch ist bei einer Datenmigration der möglicherweise damit einhergehende Verlust an Informationen. So ist es zum Beispiel möglich, dass sich das äußere Erscheinungsbild der Daten ändert oder - noch gravierender - Teile der Daten verloren gehen.

Eine verlustfreie Migration ist dann möglich, wenn sowohl das Original-Format wie auch das Ziel-Format eindeutig spezifiziert sind, diese Spezifikationen bekannt sind UND eine Übersetzung von dem einen in das andere Format ohne Probleme möglich ist. Hier gilt: Je einfacher und übersichtlicher die Formate, desto größer ist die Wahrscheinlichkeit einer verlustfreien Migration. Bei der Migration komplexer Datei-Formate ist ein Verlust an Informationen wahrscheinlicher, da der Umfang einer komplexen Migration nicht unbedingt absehbar ist. Eine Migration eines Commodore-64 Computerspiels in ein heute spielbares Format für einen PC ist sicherlich möglich, jedoch ist es (a) sehr aufwändig, (b) schlecht bzw. gar nicht automatisierbar und (c) das Ergebnis (sehr wahrscheinlich) weit vom Original entfernt.

Beispiel: Alte und neue PCs

- Sie haben einen recht alten PC, auf dem Sie seit langem Ihre Texte schreiben, zum Beispiel mit einer älteren Version von Word 95 (Betriebssystem: Windows 95). Sie speichern Ihre Daten auf Diskette.
- Ihr neuer Rechner, den Sie sich angeschafft haben, läuft unter Windows XP mit Word 2003 und hat kein Diskettenlaufwerk mehr.
- Nun stehen Sie zunächst vor dem Problem, wie Sie Ihre Daten auf den neuen Rechner übertragen. Wenn Sie Glück haben, hat Ihr alter Rechner

schon USB, sodass Sie Ihre Daten mit einem USB-Stick übertragen können. Vielleicht haben Sie auch noch ein Diskettenlaufwerk, auf das Sie zurückgreifen können. Oder aber Ihr alter Rechner kann sich ins Internet einwählen und Ihre Daten können von dort mit dem neuen Rechner heruntergeladen werden. Hier ist unter Umständen ein wenig zu tun. Es gibt jedoch noch genügend Möglichkeiten, Ihre Daten zu übertragen.

- Nehmen wir an, Ihre Daten sind sicher und korrekt übertragen worden. Wenn Sie Glück haben, meldet sich Word 2003 und sagt, Ihre Dateien seien in einem alten .doc-Format gespeichert und müssen in das aktuelle Format konvertiert werden. Diese Konvertierung ist dann eine Migration in ein neues, aktuelleres .doc-Format. Wenn die Migration erfolgreich abläuft, sieht Ihr Dokument aus wie auf dem alten Rechner unter Word 95. Es besteht jedoch die Möglichkeit, dass Ihr Dokument sich verändert hat (Formatierung, Schriftart, Schriftgröße etc.).
- Sollten Sie Pech haben, erkennt Word das alte Format nicht und eine Migration ist nicht automatisch möglich. Dann bleibt noch die Möglichkeit, die alten Dateien mit einem Zwischenschritt über ein anderes Textformat, das beide Textprogramme beherrschen, zu konvertieren. Sicherlich können beide Programme einfache Textdateien verarbeiten (.txt), vielleicht auch Dateien im Rich-Text-Format (.rtf). Sie müssen nun Ihre Dokumente mit dem alten Word alle als Text- oder RTF-Datei neu speichern, diese erneut (wie oben beschrieben) auf den neuen Rechner übertragen und dann mit dem neuen Word (als Text- oder RTF-Datei) wieder öffnen. Sehr wahrscheinlich sind dann sehr viele Formatierungen (Inhaltsverzeichnisse, Überschriften, Schriftdicken, Schriftarten, etc.) verloren gegangen, da eine .txt-Datei keinerlei solcher Dinge speichern kann. Nur der Text entspricht dem originalen Dokument. Mit einer RTF-Datei haben Sie sicherlich weniger Informationsverlust. Sie führen also praktisch zwei Migrationen durch: .doc (Word 95) – .txt (bzw. .rtf) – .doc (Word 2003), siehe hierzu die Abbildungen 1 und 2.

kopal
Daten für die Zukunft

Pressemitteilung

Test des für kopal entwickelten Archivsystems erfolgreich
Fazit der Testphase zur Abnahme durch die Projektpartner

Das kopal-System für die Langzeitarchivierung digitaler Objekte hat am 28.11.2005 eine erste 30-tägige Testphase erfolgreich durchlaufen. Damit wurde das von der IBM Deutschland für das Projekt kopal weiterentwickelte System DIAS (Digital Information Archiving System) durch die Projektpartner Die Deutsche Bibliothek, die Niedersächsische Staats- und Universitätsbibliothek (SUB) Göttingen und die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) abgenommen. Die Deutsche Bibliothek und die SUB Göttingen haben eine Funktionsprüfung des DIAS im Fernzugriff auf das bei der GWDG gehostete System durchgeführt. Beide Bibliotheken haben Testobjekte aus ihren elektronischen Beständen in die Tests eingebracht wie etwa Dissertationen, Zeitschriften und Monographien sowie Objekte des Göttinger Digitalisierungszentrums. Das wissenschaftliche Rechenzentrum GWDG stellte die Testumgebung zur Verfügung und führte Tests bezüglich betriebsüblichen Störungen und Systemausfällen durch.

Die von Mitte Oktober bis Ende November durchgeführten Tests schließen die beiden ersten Arbeitspakete der Entwicklungsphase des Projektes ab. Die Weiterentwicklung des DIAS zu einem kooperativ betriebenen Archivsystem wird unter Anwendung internationaler Standards von IBM übernommen. Die Projektmitarbeiter Der Deutschen Bibliothek und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen erstellen auf den weiterentwickelten DIAS-Core abgestimmte Open Source Softwareprodukte. Diese kopal-Tools ermöglichen eine einfache und skalierbare Einspeisung der Daten in das Archiv und einen komfortablen Zugriff auf die archivierten Materialien.

In der nun abgeschlossenen Testphase wurde die Funktionalität des DIAS im Zusammenspiel mit den entwickelten Tools zum automatisierten Einspeisen (Ingest) und Abfragen (Access) untersucht. Dokumentiert wurde außerdem das Verhalten des Systems bezüglich der Größe der Objekte sowie Vollständigkeit und Korrektheit der Lieferungen. Defekte oder unvollständige Datenpakete wurden eingespielt, um die im System integrierte Fehlererkennung zu testen. Es wurden mögliche Störungen und Systemausfälle simuliert, die erfahrungsgemäß im praktischen Betrieb vorkommen.

beiden Institutionen archiviert werden. Die Entwicklung von Komponenten für Administration und Error-Tracking befindet sich in der Vorbereitung.

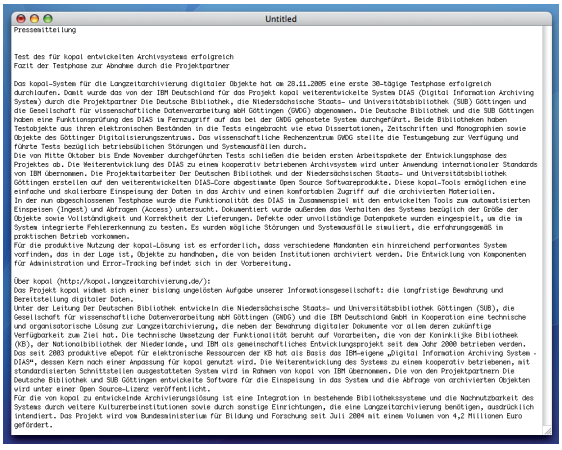
Über kopal (http://kopal.langzeitarchivierung.de/):

Das Projekt kopal widmet sich einer bislang ungelösten Aufgabe unserer Informationsgesellschaft: die langfristige Bewahrung und Bereitstellung digitaler Daten.

Unter der Leitung Der Deutschen Bibliothek entwickeln die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB), die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und die IBM Deutschland GmbH in Kooperation eine technische und organisatorische Lösung zur Langzeitarchivierung, die neben der Bewahrung digitaler Dokumente vor allem deren zukünftige Verfügbarkeit zum Ziel hat. Die technische Umsetzung der Funktionalität beruht auf Vorarbeiten, die von der Koninklijke Bibliotheek (KB), der Nationalbibliothek der Niederlande, und IBM als gemeinschaftliches Entwicklungsprojekt seit dem Jahr 2000 betrieben werden. Das seit 2003 produktive eDepot für elektronische Ressourcen der KB hat als Basis das IBM-eigene „Digital Information Archiving System – DIAS", dessen Kern nach einer Anpassung für kopal genutzt wird. Die Weiterentwicklung des Systems zu einem kooperativ betriebenen, mit standardisierten Schnittstellen ausgestatteten System wird im Rahmen von kopal von IBM übernommen. Die von den Projektpartnern Die Deutsche Bibliothek und SUB Göttingen entwickelte Software für die Einspeisung in das System und die Abfrage von archivierten Objekten wird unter einer Open Source-Lizenz veröffentlicht.

Für die von kopal zu entwickelnde Archivierungslösung ist eine Integration in bestehende Bibliothekssysteme und die Nachnutzbarkeit des Systems durch weitere Kulturerbeinstitutionen sowie durch sonstige Einrichtungen, die eine Langzeitarchivierung benötigen, ausdrücklich intendiert. Das Projekt wird vom Bundesministerium für Bildung und Forschung seit Juli 2004 mit einem Volumen von 4,2 Millionen Euro gefördert.

Abbildung 1: Ein Word-Dokument mit Grafiken, Formatierungen, Link, etc.



Untitled

Pressemitteilung

Test des für kopal entwickelten Archivsystems erfolgreich
Fazit der Testphase zur Abnahme durch die Projektpartner

Das kopal-System für die Langzeitarchivierung digitaler Objekte hat am 28.11.2005 eine erste 30-tägige Testphase erfolgreich durchlaufen. Damit wurde das von der IBM Deutschland für das Projekt kopal weiterentwickelte System DIAS (Digital Information Archiving System) durch die Projektpartner Die Deutsche Bibliothek, die Niedersächsische Staats- und Universitätsbibliothek (SUB) Göttingen und die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) abgenommen. Die Deutsche Bibliothek und die SUB Göttingen haben eine Funktionsprüfung des DIAS im Fernzugriff auf das bei der GWDG gehostete System durchgeführt. Beide Bibliotheken haben Testobjekte aus ihren elektronischen Beständen in die Tests eingebracht wie etwa Dissertationen, Zeitschriften und Monographien sowie Objekte des Göttinger Digitalisierungszentrums. Das wissenschaftliche Rechenzentrum GWDG stellte die Testumgebung zur Verfügung und führte Tests bezüglich betriebsüblichen Störungen und Systemausfällen durch.
Die von Mitte Oktober bis Ende November durchgeführten Tests schließen die beiden ersten Arbeitspakete der Entwicklungsphase des Projektes ab. Die Weiterentwicklung des DIAS zu einem kooperativ betriebenen Archivsystem wird unter Anwendung internationaler Standards von IBM übernommen. Die Projektmitarbeiter Der Deutschen Bibliothek und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen erstellen auf den weiterentwickelten DIAS-Core abgestimmte Open Source Softwareprodukte. Diese kopal-Tools ermöglichen eine einfache und skalierbare Einspeisung der Daten in das Archiv und einen komfortablen Zugriff auf die archivierten Materialien.
In der nun abgeschlossenen Testphase wurde die Funktionalität des DIAS im Zusammenspiel mit den entwickelten Tools zum automatisierten Einspeisen (Ingest) und Abfragen (Access) untersucht. Dokumentiert wurde außerdem das Verhalten des Systems bezüglich der Größe der Objekte sowie Vollständigkeit und Korrektheit der Lieferungen. Defekte oder unvollständige Datenpakete wurden eingespielt, um die im System integrierte Fehlererkennung zu testen. Es wurden mögliche Störungen und Systemausfälle simuliert, die erfahrungsgemäß im praktischen Betrieb vorkommen.
Für die produktive Nutzung der kopal-Lösung ist es erforderlich, dass verschiedene Mandanten ein hinreichend performantes System vorfinden, das in der Lage ist, Objekte zu handhaben, die von beiden Institutionen archiviert werden. Die Entwicklung von Komponenten für Administration und Error-Tracking befindet sich in der Vorbereitung.

Über kopal (http://kopal.langzeitarchivierung.de/):
Das Projekt kopal widmet sich einer bislang ungelösten Aufgabe unserer Informationsgesellschaft: die langfristige Bewahrung und Bereitstellung digitaler Daten.
Unter der Leitung Der Deutschen Bibliothek entwickeln die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB), die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und die IBM Deutschland GmbH in Kooperation eine technische und organisatorische Lösung zur Langzeitarchivierung, die neben der Bewahrung digitaler Dokumente vor allem deren zukünftige Verfügbarkeit zum Ziel hat. Die technische Umsetzung der Funktionalität beruht auf Vorarbeiten, die von der Koninklijke Bibliotheek (KB), der Nationalbibliothek der Niederlande, und IBM als gemeinschaftliches Entwicklungsprojekt seit dem Jahr 2000 betrieben werden. Das seit 2003 produktive eDepot für elektronische Ressourcen der KB hat als Basis das IBM-eigene „Digital Information Archiving System - DIAS", dessen Kern nach einer Anpassung für kopal genutzt wird. Die Weiterentwicklung des Systems zu einem kooperativ betriebenen, mit standardisierten Schnittstellen ausgestatteten System wird im Rahmen von kopal von IBM übernommen. Die von den Projektpartnern Die Deutsche Bibliothek und SUB Göttingen entwickelte Software für die Einspeisung in das System und die Abfrage von archivierten Objekten wird unter einer Open Source-Lizenz veröffentlicht.
Für die von kopal zu entwickelnde Archivierungslösung ist eine Integration in bestehende Bibliothekssysteme und die Nachnutzbarkeit des Systems durch weitere Kulturerbeinstitutionen sowie durch sonstige Einrichtungen, die eine Langzeitarchivierung benötigen, ausdrücklich intendiert. Das Projekt wird vom Bundesministerium für Bildung und Forschung seit Juli 2004 mit einem Volumen von 4,2 Millionen Euro gefördert.

Abbildung 2: Das selbe Dokument im .txt-Format ohne Formatierungen

Beispiel: Zeichenkodierungen

- Eine Organisation, die in den 80er Jahren ihre Daten mit IBM Mainframes bearbeitet hat, möchte diese Daten auch auf späteren Systemen nutzen können. Die IBM Mainframes nutzten einen Zeichenstandard namens EBCDIC.[7]
- In den 90er Jahren installierte Rechner nutzten den ASCII Zeichencode (American National Standard Code for Information Interchange), welcher nicht alle Zeichen des EBCDIC abdeckte. Die Organisation musste sich nun entscheiden, ob sie alle Dokumente nach ASCII konvertierten (und einen permanenten Verlust von Daten hinnahmen) oder sie nur bei Bedarf in ASCII umwandelten und die Originaldaten in EBCDIC beließen. So hatte man den gleichen Verlust beim Umwandeln, jedoch für spätere Zeit die Originaldaten erhalten.
- Zum Jahrtausendwechsel begann UNICODE[8] die Welt zu erobern und tatsächlich enthält UNICODE alle Zeichen des EBCDIC, sodass nun alle Dokumente 1:1 von EBCDIC in UNICODE konvertiert werden konnten (sofern die Originaldateien noch existierten!). Bei einer sofortigen Konvertierung in ASCII wären tatsächlich Daten verloren gegangen.

Zusammenfassung: Vor- und Nachteile von Migration

Vorteile von Migration

- Migration ist technisch (verglichen mit Emulation) gut zu realisieren.
- Migration kann in vielen Fällen automatisiert werden.
- Die migrierten Dokumente sind unabhängig von weiteren Komponenten (abgesehen von der aktuellen Darstellungssoftware).
- Die originalen Objekte können aufbewahrt werden, um evtl. später darauf zurückgreifen zu können.

Nachteile von Migration

- Jedes Objekt muss einzeln migriert werden.
- Die Wahrscheinlichkeit von Datenverlust bzw. Datenveränderung ist (besonders über mehrere Migrationsschritte) sehr hoch.
- Jede Version (Migration) eines Objekts inklusive des Original-Dokuments sollte gespeichert werden. Damit ist unter Umständen ein hoher Speicherplatzbedarf verbunden.

---

7 Extended Binary Coded Decimal Interchange Code: http://www.natural-innovations.com/computing/asciiebcdic.html

8 http://www.unicode.org

- Für jedes Format und für jeden Migrations-Schritt muss es ein Migrations-Werkzeug geben.
- Migration ist nicht für alle Formate realisierbar.

Literatur

Consultative Committee for Space Data Systems (2001): *Reference Model for an Open Archi-val Information System (OAIS),* CCSDS 650.0-B-1, BLUE BOOK, http://public.ccsds.org/publications/archive/650x0b1.pdf

Jenkins, Clare (2002): *Cedars Guide to: Digital Preservation Strategies*, http://www.webarchive.org.uk/wayback/archive/20050409230000/http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html

8.4 Emulation

Stefan E. Funk

Mit Emulation (Nachbildung, Nachahmung, von lat. aemulator = Nacheiferer) versucht man die auftretenden Verluste einer Datenformatmigration zu umgehen, indem man die originale Umgebung der archivierten digitalen Objekte nachbildet. Emulation kann auf verschiedenen Ebenen stattfinden:

- zum einen auf der Ebene von Anwendungssoftware,
- zum anderen auf der Ebene von Betriebssystemen und zu guter Letzt
- auf der Ebene von Hardware-Plattformen.

So kann zum Beispiel die originale Hardware des digitalen Objekts als Software mit einem Programm nachgebildet werden, welches das archivierte Betriebssystem und die darauf aufbauenden Softwarekomponenten laden kann (Emulation von Hardware-Plattformen). Ein Beispiel für die Emulation von Betriebssystemen wäre ein MS-DOS-Emulator[9], der die Programme für dieses schon etwas ältere Betriebssystem auf aktuellen Rechnern ausführen kann. Ein Beispiel für den ersten Fall wäre etwa ein Programm zum Anzeigen und Bearbeiten von sehr alten Microsoft Word-Dateien (.doc), die das aktuelle Word nicht mehr lesen kann. Auf diese Weise wird die Funktionalität dieser alten und nicht mehr verfügbaren Soft- oder Hardware emuliert und die Inhalte bzw. die Funktionalität der damit erstellten Dokumente erhalten. Im Gegensatz zur Migration, bei der jeweils eine neue und aktuellere Version des digitalen Objektes selbst erzeugt wird, werden die originalen Objekte bei der Emulation nicht verändert. Stattdessen muss man für jede neue Hardwarearchitektur die Emulationssoftware anpassen, im schlechtesten Fall muss diese jedes Mal neu entwickelt werden. Wenn das aus irgendeinem Grund nicht geschieht, ist der komplette Datenbestand der betroffenen Objekte unter Umständen nicht mehr nutzbar und damit für die Nachwelt verloren.

Emulation von Anwendungssoftware

Da es um die Darstellung der digitalen Dokumente geht, die wir vorhin beschrieben haben, ist die Emulation der Software, die mit diesen Dokumenten arbeitet, eine erste Möglichkeit. So kann auf einem aktuellen System ein Pro-

9 DOS – Disc Operating System, näheres unter: http://www.operating-system.org/betriebssystem/_german/bs-msdos.htm

gramm entwickelt werden, das archivierte digitale Objekte in einem bestimmten Format öffnen, anzeigen oder bearbeiten kann, auf die mit aktueller Software auf diesem System nicht mehr zugegriffen werden kann, weil vielleicht die Originalsoftware nicht mehr existiert oder auf aktuellen Systemen nicht mehr lauffähig ist.

Wenn wir zum Beispiel eine PDF-Datei aus dem Jahr 1998, Version 1.2, darstellen möchten, und der aktuelle Acrobat Reader 7.0 stellt das Dokument nicht mehr richtig dar, müssen wir einen PDF-Reader für diese PDF-Version auf einem aktuellen Betriebssystem programmieren, sprich: einen alten PDF-Reader emulieren. Dieser sollte dann alle PDF-Dateien der Version 1.2 darstellen können. Für jeden Generationswechsel von Hardware oder Betriebssystem würde so ein PDF-Reader benötigt, um den Zugriff auf die PDF-Dokumente in Version 1.2 auch in Zukunft zu gewährleisten. Die genaue Kenntnis des PDF-Formats ist hierzu zwingend erforderlich.

Emulation von Betriebssystemen und Hardware-Plattformen

Bei einigen Anwendungen kann es sinnvoll sein, eine komplette Hardware-Plattform zu emulieren, zum Beispiel wenn es kein einheitliches Format für bestimmte Anwendungen gibt. Hier ist der Commodore-64 ein gutes Beispiel. Die Spiele für den C-64 waren eigenständige Programme, die direkt auf dem Rechner liefen, soll heissen, es wird direkt die Hardware inklusive des Betriebssystems[10] benötigt und nicht ein Programm, das diese Spiele ausführt (wie ein PDF-Viewer).

Es muss also ein Commodore-64 in Software implementiert werden, der sich genau so verhält wie die Hardware und das Betriebssystem des originalen Commodore-64 und auf einem aktuellen Computersystem lauffähig ist. Diese C-64-Emulatoren gibt es für nahezu alle aktuellen Computersysteme und auch weitere Emulatoren für andere ältere Systeme sind erhältlich.[11]

Die Emulation eines Betriebssystems oder einer Hardware-Plattform ist eine sehr komplexe Sache, die schon für einen C-64-Emulator sehr viel Arbeit bedeutet. Man kann jedoch auch die Hardware eines PC in Software nachbilden, um dann auf einem solchen virtuellen PC beliebige Betriebssysteme und die auf ihnen laufenden Anwendungsprogramme oder auch Spiele zu starten (die Be-

10 Eine Trennung von Hardware und Betriebssystem ist beim Commodore-64 nicht nötig, da diese beiden Komponenten sehr eng zusammenhängen. Auch andere „Betriebssysteme" des C-64, wie zum Beispiel GEOS, setzen direkt auf das Betriebssystem des C-64 auf.

11 Hier einige Adressen im Internet zum Thema Emulatoren: http://www.aep-emu.de/, http://www.homecomputermuseum.de/

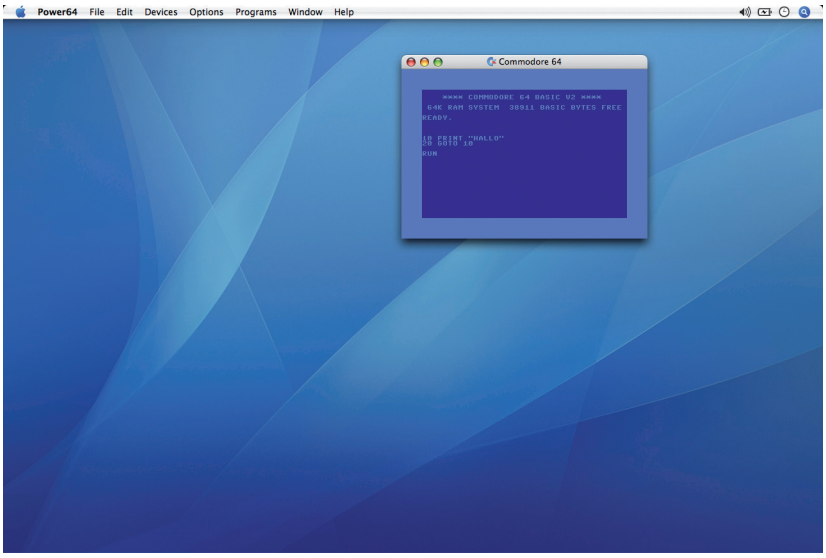


Abbildung 1: Power 64, ein Commodore-64 Emulator für Mac OS X

triebssysteme wie auch die Programme bleiben dann im Originalzustand). Dies bedeutet im Allgemeinen, dass eine gute Performanz auf der aktuellen Hardware vorhanden sein muss. Eine Emulation eines Commodore-64 auf einem aktuellen PC ist jedoch keine performanzkritische Anwendung. Für zukünftige Computersysteme, die unsere heutigen emulieren sollen, wird im Allgemeinen davon ausgegangen, dass deren Performanz weitaus höher ist als heute, sodass auch hier die Performanz für eine erfolgreiche Emulation ausreichen dürfte.

Beispiel: Migration und Emulation alter C-64 Programme

- Da der Commodore 64 ein sehr beliebter und weit verbreiteter Homecomputer war, gibt es sehr viele Emulatoren für nahezu alle aktuellen Computersysteme. Viele Videospiele, die es für den C-64 gab, sind im Internet als C-64 Disk-Image zu finden. Die darin enthaltenen Programme können dann mit den Emulatoren geladen und genutzt werden. Als alter C-64 Nutzer stand ich also nicht vor dem Problem, meine Spiele von alten 5,25-Zoll Disketten auf neuere Datenträger migrieren zu müssen. Ein Emulator für den Apple unter Mac OS X ist Power64[12], siehe Abbildung 1.
- Anders sah es hingegen für die Programme aus, die ich vor mehr als 20 Jahren auf dem C-64 selbst programmiert habe. Es handelt sich hier

12 http://www.infinite-loop.at/Power64/index.html

um viele Programme in Commodore-64 BASIC. Die Frage, die sich mir stellte, war nun die, ob und wie ich diese Daten von meinen alten (auf dem Original C-64 noch laufenden) 5,25 Zoll-Disketten von 1982 bis 1987 auf die Festplatte meines PC kopieren und ich diese Daten auch für den C-64-Emulator nutzen kann.

- Der erste Versuch, einfach ein vor einigen Jahren noch gebräuchliches 5,25 Zoll-Laufwerk[13] an den PC anzuschließen und die C-64 Daten am PC auszulesen, schlug zunächst einmal fehl. Grund hierfür waren die unterschiedlichen Dichten und die unterschiedlichen Dateisysteme der 5,25 Zoll-Disketten. Auf eine Diskette des C-64 war Platz für 170 KB, damals einfache Dichte (single density). Die Disketten für den PC hatten jedoch doppelte Dichte (double density) oder gar hohe Dichte (high density), sodass das mit zur Verfügung stehende Diskettenlaufwerk die C-64 Disketten nicht lesen konnte.
- Nach kurzer Recherche entdeckte ich eine Seite im Internet (die Community für den C-64 ist immer noch enorm groß), die Schaltpläne für einige Kabel abbildete, mit denen man seinen PC mit den Diskettenlaufwerken seines C-64 verbinden konnte. Mit Hilfe des Programmes Star Commander[14]**,** das unter DOS läuft, kann man damit seine Daten von C-64 Disketten auf seinen PC kopieren und auch gleich Disk-Images erstellen. Inzwischen kann man solche Kabel auch bestellen und muss nicht selbst zum Lötkolben greifen (Für die Nutzung dieses Programms muss natürlich eine lauffähige DOS-Version zur Verfügung stehen, ist keine verfügbar, kann evtl. eine emuliert werden :-)
- Nach diesen Aktionen kann ich nun meine alten selbst erstellten Programme auf vielen C-64 Emulatoren wieder nutzen, weiterentwickeln und spielen, wie in Abbildung 2 und 3 zu sehen ist (und das sogar auf mehreren virtuellen Commodore-64 gleichzeitig).

13 Den ersten Versuch unternahm ich vor etwa vier Jahren, 5,25-Zoll-Diskettenlaufwerke waren nicht mehr wirklich gebräuchlich, aber noch erhältlich. Heute werden selbst die 3,5-Zoll-Laufwerke schon nicht mehr mit einem neuen Rechner verkauft. Neue Medien zum Datenaustausch und zur Speicherung sind heute USB-Stick, DVD, CD-ROM und Festplatte.

14 http://sta.c64.org/sc.html

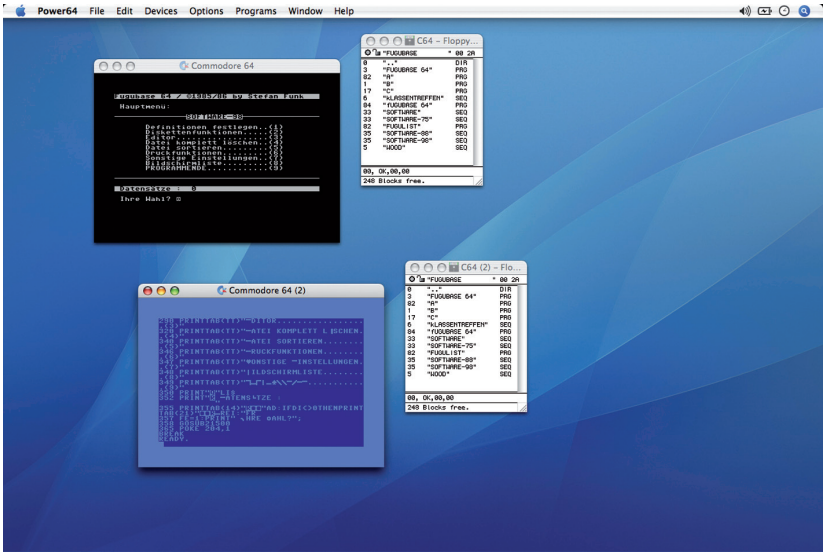


Abbildung 2: Fugubase 64, ein Datenverwaltungs-Programm in Basic für den C-64, emuliert unter Mac OS X (S. E. Funk, 1985/86)

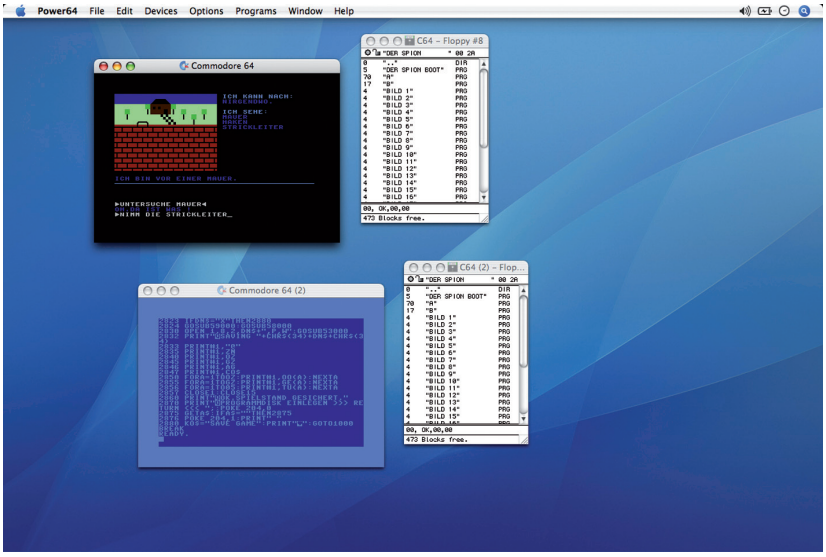


Abbildung 3: Der Spion, ein Adventure in Basic für den C-64, emuliert unter Max OS X (S. E. Funk, 1987)

Beispiel: Eine Emulation in der Emulation

- Es ist nun auch möglich, einen Emulator wiederum zu emulieren, wenn ein weiterer Generationswechsel einer Hardwareplattform ansteht. Ein praktisches Beispiel ist ein Apple Notebook, das unter Mac OS X, einem Unix-basierten Betriebssystem, arbeitet. Auf diesem werden zwei Emulatoren und ein weiteres originales Betriebssystem gestartet.
- Auf diesem Rechner wird das Programm Q gestartet[15], das eine Hardware-Plattform emuliert (einen Pentium x86 mit diversen Grafik-, Sound- und weiteren Hardwarekomponenten). Es basiert auf dem CPU-Emulator QEMU.[16]
- Auf dieser virtuellen Hardwareplattform kann nun ein originales Windows 98 installiert werden, so dass man ein reguläres, altbekanntes Windows 98 auf diesem nicht-Windows-Rechner nutzen kann. Das installierte Windows 98 kann selbstverständlich alle Programme für Windows 98 ausführen, da es sich tatsächlich um ein originales Windows 98 handelt. Sogar ein Windows-Update über das Internet ist möglich.
- Jetzt kann natürlich auch ein C-64 Emulator für Windows, hier der VICE[17], gestartet werden. Darauf laufen nun alle altbekannten und beliebten Commodore-64 Programme.
- Probleme kann es bei dieser Art von Emulation zum Beispiel bei der Performanz geben und je nach Qualität der Emulatoren auch mit hardwarespezifischen Dingen wie Grafik, Sound und angeschlossener Peripherie (Mäuse, Joysticks, etc.). Der C-64 Emulator muss schließlich durch Windows über die virtuelle Hardware (Emulation QEMU) auf die reale Hardware des Notebooks zugreifen. Bei steigender Komplexität solcher Emulationsszenarien wird die Anzahl der möglichen Fehler stark ansteigen. Als Beispiel siehe Abbildung 4.

Der Universal Virtual Computer (UVC)

Mittlerweile gibt es einen elaborierteren Ansatz der Emulation, den Universal Virtual Computer (UVC) von IBM. Der UVC ist ein wohldokumentierter virtueller Computer**,** der auf unterschiedlichen (auch zukünftigen) Architekturen nachgebildet werden kann. Aufgebaut ist er ähnlich wie ein heute existierender Computer, der beispielsweise Speicherzugriff ermöglicht. Mit Hilfe dieser Dokumentation ist es einem Programmierer auch auf zukünftigen Systemen

15 http://www.kju-app.org/
16 http://www.nongnu.org/qemu/
17 http://www.viceteam.org/

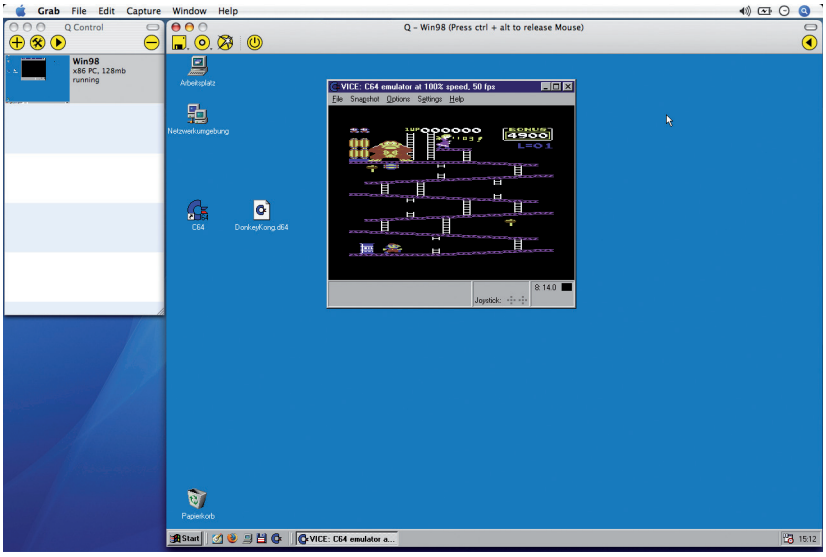


Abbildung 4: Das Videospiel Donkey Kong auf einem C-64 Emulator auf einem Windows 98 auf einem virtuellen Pentium auf einem Apple PowerBook unter Mac OS X

möglich, diesen virtuellen Computer zu implementieren. Auf diesem virtuellen Computer aufbauend können nun Programme geschrieben werden, die zum Beispiel eine PDF-Datei lesen oder Grafiken darstellen können.

Archiviert wird jetzt der PDF-Reader (der Bildbetrachter), der für den UVC programmiert wurde, sowie das originale PDF-Dokument (oder die originale Grafik) selbst. Ein zukünftiger Nutzer kann dann auf einer zukünftigen und wahrscheinlich hoch entwickelten Hardware auch in ferner Zukunft noch mit Hilfe der Dokumentation des UVC einen solchen implementieren und mit Hilfe dieses virtuellen Computers den PDF-Reader starten, mit dem das archivierte PDF-Dokument dargestellt wird. Die Dokumentation muss selbstverständlich erhalten bleiben und lesbar sein.

Ein Problem dieser Idee ist sicherlich, dass bei zunehmendem Anspruch an die Emulation, die auf dem UVC laufen soll, eine Programmierung derselben immer schwieriger wird. Es wird sehr kompliziert, wenn für den UVC ein Betriebssystem wie Linux oder Windows programmiert werden soll, mit dessen Hilfe dann die Applikationen von Linux oder Windows genutzt werden können. Schon eine nachprogrammierte Version eines Textverarbeitungsprogrammes wie zum Beispiel Word, mit dem später alte Word-Dokumente (.doc) auf dem UVC gelesen und bearbeitet werden können, ist ein höchst umfangreiches Un-

ternehmen. Zumal hier nicht nur die Formatbeschreibung, sondern auch alle Programmfunktionen bekannt sein müssen.

Zusammenfassung: Vor- und Nachteile von Emulation

Vorteile von Emulation

- Bei der Emulation bleiben die Originalobjekte unverändert.
- Eine Konvertierung der Objekte ist nicht nötig.
- Für die Emulation wird weniger Speicherplatz benötigt, da keine Migrationen gespeichert werden müssen**.**

Nachteile von Emulation

- Für komplizierte Objekte/Systeme (wie Betriebssysteme oder Anwendungsprogramme) sind Emulatoren technisch schwer zu implementieren.
- Es entsteht ein hoher Aufwand pro Hardware-Generationswechsel. Es müssen für jede Plattform neue Emulatoren entwickelt werden.
- Die Spezifikationen für die zu emulierenden Objekte/Systeme sind nicht immer hinreichend bekannt.

Literatur

Lorie, Raymond (2002): *the UVC: a method for preserving digital documents – proof of con-cept*, http://www-5.ibm.com/nl/dias/resource/uvc.pdf

Nationaal Archief (2005): *Technical Description of the Universal Virtual Computer (UVC) - Data preservation process for spreadsheets,* http://www.digitaleduurzaamheid.nl/bibliotheek/docs/TDUVCv1.pdf

Oltmans, Eric; Nanda Kol (2005): *A Comparison Between Migration and Emulation in Terms of Costs*, http://www.rlg.org/en/page.php?Page_ID=20571#article0

8.5 Computermuseum

Karsten Huth

Die Erhaltung von Hardware ist auf lange Sicht keine vielversprechende Strategie zur Bewahrung von digitalen Objekten. Solange aber keine technischen Möglichkeiten zum Transfer von Daten aus den alten Systemen auf aktuelle Plattformen sowie zur Migration oder Emulation zur Verfügung stehen, ist der Erhalt des originalen Systems ein notwendiger erster Schritt zur Erhaltung eines digitalen Bestandes. Zudem gibt es auch Museen, die ihren Besuchern den Eindruck der historischen Software mitsamt der historischen Hardware vermitteln wollen. Dieser Artikel gibt eine kurze Einführung in die Probleme, die sich vor allem durch die Auflösungserscheinungen der Werkstoffe ergeben.

Definition

Auch wenn man die Strategie der „Hardware Preservation", also der Erhaltung von Hardware, als Methode zur Langzeitarchivierung auf keinen Fall empfehlen sollte, so ist es leider alltägliche Praxis, dass digitale Langzeitarchive auch obsolete Hardware vorhalten müssen, zumindest bis sie in der Lage sind, besser geeignete Strategien durchzuführen. Aber gerade in den Anfängen eines digitalen Archivs, wenn es noch über keinen geregelten Workflow verfügt, werden digitale Objekte oft auf ihren originalen Datenträgern oder mitsamt ihrer originalen Hardware/Software-Umgebung abgeliefert. Dies betrifft vor allem digitale Objekte, die technologisch obsolet geworden sind. Deshalb sind in der Praxis, wenn auch ungewollt, Computermuseen eher die Regel als eine Ausnahme.

Leider hat sich der Begriff „Computermuseum" im deutschen Sprachraum verfestigt. Passender wäre der Begriff „Hardware-/Software-Konservierung", denn die konservierten Computer müssen nicht unbedingt nur im Rahmen eines Museums erhalten werden. Man muss vielmehr differenzieren zwischen:

1. Hardware Preservation als Strategie zur Archivierung von digitalen Objekten:
   Eigentliches Ziel ist die Erhaltung der digitalen Objekte. Zu diesem Zweck versucht man die ursprüngliche Hardware/Software-Plattform so lange wie möglich am Laufen zu halten.

2. Hardware Preservation im Rahmen eines Technikmuseums:
   Wird im ersten Fall die Hardware/Software-Plattform nur erhalten, um den Zugriff auf die digitalen Objekte zu ermöglichen, so ist hier die

ursprüngliche Hardware/Software Plattform das zentrale Objekt der konservatorischen Bemühungen. Während im ersten Fall Reparaturen an der Hardware einzig der Lauffähigkeit der Rechner dienen, so fallen im Rahmen eines Technikmuseums auch ethische Gesichtspunkte bei der Restaurierung ins Gewicht. Die Erhaltung der Funktion ist bei einer Reparatur nicht mehr das einzige Kriterium, es sollten auch möglichst die historisch adäquaten Bauteile verwendet werden. Diese Auflage erschwert die beinahe unmögliche Aufgabe der Hardware-Konservierung noch zusätzlich.

Bei einem technischen Museum liegt die Motivation zur Konservierung von Hardware auf der Hand. Die historische Hardware zusammen mit der originalen Software sind die Sammelobjekte und Exponate des Museums. Deswegen müssen sie solange wie möglich in einem präsentablen Zustand erhalten werden. Daneben gibt es aber auch noch weitere Gründe, die für die Hardware Preservation als Archivierungsstrategie sprechen.

Gründe zur Aufrechterhaltung eines Computermuseums

- Keine andere Strategie erhält soviel vom intrinsischen Wert der digitalen Objekte (Look and Feel). An Authentizität ist dieser Ansatz nicht zu übertreffen.[18]
- Bei komplexen digitalen Objekten, für die Migration nicht in Frage kommt, und eine Emulation der Hardware/Software Umgebung noch nicht möglich ist, ist die Hardware Preservation die einzige Möglichkeit, um das Objekt zumindest für einen Übergangszeitraum zu erhalten.[19]
- Zur Unterstützung von anderen Archivierungsstrategien kann die zeitweise Erhaltung der originalen Plattformen notwendig sein. Man kann z.B. nur durch einen Vergleich mit der ursprünglichen Hardware/Software-Plattform überprüfen, ob ein Emulatorprogramm korrekt arbeitet oder nicht.[20]

---

18 Borghoff, Uwe M. et al. (2003): *Methoden zur Erhaltung digitaler Dokumente.* 1. Aufl. Heidelberg: dpunkt-Verl., 2003: S. 16-18

19 Jones, Maggie/ Beagrie, Neil (o.J.): *Preservation Management of Digital Materials: A Handbook.* Digital Preservation Coalition.

20 Rothenberg, Jeff (1998): *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources.* Washington D.C.: Council on Library and Information Resources: S. 12-13 http://www.clir.org/pubs/reports/rothenberg/inadequacy.html

Probleme der Hardware Preservation

Ob man ein Hardware-Museum aus dem ersten oder dem zweiten Grund führt, in beiden Fällen hat man mit den gleichen Problemen zu kämpfen. Zum einen ergeben sich auf lange Sicht gesehen große organisatorische und zum anderen rein technische Probleme der Konservierung von Hardware und Datenträgern.

1. Organisatorische Probleme:

- Die Menge an zu lagerndem und zu verwaltendem Material wird stetig wachsen. Da nicht nur die Rechner, sondern auch Peripheriegeräte und Datenträger gelagert werden müssen, steigen Platzbedarf und Lagerungsaufwand enorm an. „Selbst heute schon erscheint es unrealistisch, sämtliche bisher entwickelten Computertypen in einem Museum zu versammeln, geschweige denn dies für die Zukunft sicher zu stellen."[21]
- Techniker und Experten, die historische Computer bedienen und gegebenenfalls reparieren können, werden über kurz oder lang nicht mehr zur Verfügung stehen. Mit wachsendem Bestand müssten die Mitarbeiter des Museums ihr Fachwissen ständig erweitern, oder der Bedarf an Technikexperten und neuen Mitarbeitern würde ständig wachsen.[22]
- Die Nutzung der digitalen Objekte ist nur sehr eingeschränkt möglich. Da die obsoleten Computersysteme von der aktuellen Technologie abgeschnitten sind, könnte der Nutzer nur im Computermuseum auf die Objekte zugreifen.[23]

2. Technische Probleme:

- Die technischen Geräte und Bausteine haben nur eine begrenzte Lebenserwartung. Da für obsolete Systeme keine Ersatzteile mehr produziert werden, ist die Restaurierung eines Systems irgendwann nicht mehr möglich.[24]
- Neben der Hardware muss auch die originale Softwareumgebung erhalten und archiviert werden. Diese muss natürlich auf den entsprechenden Datenträgern vorgehalten werden. Da Datenträger ebenso wie die Hardware nur eine begrenzte Lebensdauer haben, müssen die Software und die Daten von Zeit zu Zeit auf neue, frischere Datenträger des gleichen

21 s. Borghoff (2003); a.a.O.

22 Dooijes, Edo Hans (200): *Old computers, now and in the future.* Department of Computerscience/University of Amsterdam. http://www.science.uva.nl/museum/pdfs/oldcomputers_dec2000.pdf

23 s. Rothenberg (1998), a.a.O.

24 s. Borghoff (2003), a.a.O.

Typs, oder zumindest auf passende Datenträger des gleichen Computersystems umkopiert werden. Da jedoch Datenträger eines obsoleten Systems nicht mehr hergestellt werden, stößt diese Praxis zwangsläufig an ihre Grenze und Software und Daten gehen verloren.[25]

Auftretende Schäden bei der Lagerung

Es gibt wenig Literatur über die tatsächlich in der Praxis auftretenden Schäden. Der folgende Abschnitt bezieht sich auf eine Umfrage in Computermuseen. Diese Umfrage war Teil einer Abschlussarbeit an der San Francisco State University im Fach Museum Studies. Die folgende Aufzählung ist eine vorläufige Rangliste der auftretenden Probleme.[26]

- Zerfall von Gummiteilen: Gummi wird für viele Bauteile der Hardware verwendet. Riemen in Motoren, Rollen in Magnetbänderlaufwerken, Lochkartenleser und Drucker, um nur einige Beispiele zu nennen. Gummi ist anfällig für Oxidation. Harte Oberflächen werden durch Oxidation weich und klebrig. Mit fortschreitendem Zerfall kann der Gummi wieder verhärten und dabei brüchig werden.
- Zerfall von Schaumstoffisolierungen: Schaumstoff wird hauptsächlich zur Lärmisolierung und Luftfilterung in Computern verwendet. Vor allem Schaumstoff aus Polyurethan ist sehr anfällig für eine ungewollte Oxidation. Das Material verfärbt sich zunächst und zerfällt dann in einzelne Krümel.
- Verfärbung von Plastikteilen: UV-Licht verändert die chemische Zusammensetzung der Plastikgehäuse. Die Funktion des Geräts wird dadurch zwar nicht beeinträchtigt, aber die Farbe des Gehäuses verändert sich merklich ins Gelb-bräunliche.
- Schäden durch Staub: Staub greift sowohl das Äußere der Hardware als auch ihr Innenleben an. Staub ist nur eine grobe Umschreibung für eine Vielzahl an Schadstoffen, wie z.B. Ruß, Ammoniumnitrat, Ammoniumsulfat und Schwefelsäure. Mit dem Staub lagert sich Salz und Feuchtigkeit an den Bauteilen ab. Dadurch wird die Anfälligkeit für Rost oder Schimmel erhöht. Lüfter mit Ventilatoren zur Kühlung von Prozessoren ziehen den Staub in das Gehäuse des Rechners.
- Zerfall der Batterien: Leckende Batterien können das Innenleben eines Rechners zerstören. Batterien sind Behälter bestehend aus Metall und

25 s. Rothenberg (1998), a.a.O.

26 Gibson, Mark A. (2006): *The conservation of computers and other high-tech artifacts . Unique problemes and long-term solutions*: Thesis M.A. San Francisco : San Francisco State University

Metalloxid, eingetaucht in eine Flüssigkeit oder ein Gel aus Elektrolyten. Batterien sind sehr anfällig für Rost. Bei extrem unsachgemäßer Behandlung können sie sogar explodieren. Austretende Elektrolyte können Schaltkreise zersetzen.

- Korrosion: Metall ist ein häufiger Werkstoff in elektronischen Geräten. Metall wird vor allem für das Gehäuse sowie für Klammern, Schrauben und Federn verwendet.
- Beschädigte Kondensatoren: Ähnlich wie bei einer Batterie ist ein Elektrolyt wesentlicher Bestandteil eines Kondensators. Der Elektrolyt kann eine Flüssigkeit, eine Paste oder ein Gel sein. Problematisch wird es, wenn der Elektrolyt austrocknet, da dann der Kondensator nicht mehr arbeitet. Trocknet der Elektrolyt nicht aus, kann der Kondensator lecken, so dass der Elektrolyt austritt und ähnlichen Schaden anrichtet, wie eine kaputte Batterie. Kondensatoren, die lange ungenutzt bleiben, können explodieren.
- Zerfall des Plastiks: Plastik löst sich über einen längeren Zeitraum hinweg auf. Der sogenannte Weichmacher, ein chemischer Stoff, der bei der Produktion beigemengt wird, tritt in milchartigen Tropfen aus dem Material aus. Bei bestimmten Plastiksorten riecht die austretende Feuchtigkeit nach Essig. Der Prozess beeinträchtigt auch die Haltbarkeit von anderen Materialien, die mit dem zerfallenden Plastik verbunden sind.
- Schimmel: Bei einigen Monitoren aus den siebziger und achtziger Jahren kann Schimmel an der Innenseite der Mattscheibe auftreten.

Stark gefährdete Geräte und Bauteile

Von den oben genannten möglichen Schäden sind die folgenden Bauteile am häufigsten betroffen:

- Schaltkreise, die auf Dauer ausfallen.
- Kondensatoren, die ausfallen oder explodieren.
- Ausfall von batteriebetriebenen Speicherkarten und EPROMs, sowie damit einhergehender Datenverlust.
- Durch kaputte Gummirollen zerstörte Kartenleser und Magnetbandlaufwerke.
- Verstaubte und verschmutzte Kontakte.
- Gebrochene oder verloren gegangene Kabel.[27]

27 s. Dooijes (2000), a.a.O.

Gesundheitsschädliche Stoffe und Risiken

Zu beachten ist, dass Restauratoren mit gesundheitsgefährdenden Stoffen am Arbeitsplatz in Kontakt kommen können. Welche Stoffe in Frage kommen, hängt vom Alter und der Bauart der Hardware ab. Dokumentiert ist das Auftreten von:

- Quecksilber
- Blei (auch bleihaltige Farbe)
- Polychloriertem Biphenyl (PCB)
- Thorium und anderen radioaktiven Substanzen
- Asbest
- Cadmium

Besondere Vorsicht ist beim Umgang mit Batterien (vor allem defekten, leckenden Batterien) und Kondensatoren geboten. Abgesehen davon, dass Kondensatoren oft gesundheitsgefährdende Stoffe enthalten, können sie auch in stillgelegtem Zustand über Jahre hin eine hohe elektrische Spannung aufrechterhalten. Wenn Kondensatoren nach längerer Zeit wieder unter Strom gesetzt werden, können sie explodieren.[28]

Empfehlung zur Lagerung und Restaurierung:

Die Hardware sollte bei der Lagerung möglichst vor Licht geschützt werden. Ideal ist ein Helligkeitswert um 50 Lux. Fensterscheiben sollten die ultraviolette Strahlung herausfiltern. Dadurch wird der Zerfall von Plastik und Gummi verlangsamt. Ebenso ist eine möglichst niedrige Raumtemperatur, unter 20°C, sowie eine relative Luftfeuchtigkeit von unter 50% ratsam. Beides verlangsamt den Zerfall von Gummi und Plastik. Die niedrige Luftfeuchtigkeit verringert die Wahrscheinlichkeit von Korrosion. Vor der Inbetriebnahme eines Rechners sollte abgelagerter Staub durch vorsichtiges Absaugen entfernt werden. Dabei ist erhöhte Sorgfalt geboten, damit keine elektrostatische Energie die Schaltkreise beschädigt und keine wichtigen Teile mit eingesaugt werden. Mit einer zuvor geerdeten Pinzette können gröbere Staubknäuel beseitigt werden. Batterien sollten während der Lagerung möglichst aus der Hardware entfernt werden. Weitverbreitete Batterietypen sollten nicht gelagert werden. Wenn die Hardware in Betrieb genommen wird, werden frische Batterien des betreffenden Typs eingesetzt. Seltene, obsolete Batterietypen sollten separat gelagert werden. Alle genannten Maßnahmen können den Zerfall der Hardware jedoch

28 s. Gibson (2006), a.a.O.

nur verlangsamen. Aufzuhalten ist er nicht. Defekte Bauteile werden oft durch das Ausschlachten von Hardware gleicher Bauart ersetzt. Dabei werden alle intakten Teile zu einer funktionierenden Hardwareeinheit zusammengefügt. Natürlich stößt dieses Verfahren irgendwann an seine Grenzen.

Bereits eingetretene Schäden sollten durch Restaurierungsarbeiten abgemildert werden. Auslaufende Flüssigkeiten aus Kondensatoren oder Batterien sollte man umgehend mit Isopropanol-Lösung entfernen.

Dokumentation

Ein Computermuseum kommt natürlich um die korrekte Verzeichnung seiner Artefakte (Hardware und Software) nicht herum. Zusätzlich werden Informationen über den Betrieb, die Bedienung und die verwendete Technik der Hardware und Software benötigt. Des Weiteren sollten Informationen über den Erhaltungszustand und potentiell anfällige Bauteile der Hardware erhoben und gesammelt werden. Wie bei anderen Erhaltungsstrategien fallen auch hier Metadaten an, die gespeichert und erschlossen werden wollen. Schon bei der Aufnahme eines obsoleten Systems in das Archiv sollte darauf geachtet werden, dass die notwendigen Zusatzinformationen verfügbar sind (z.B. Betriebshandbücher über die Hardware/Software, technische Beschreibungen und Zeichnungen usw.). Da diese Informationen bei älteren Systemen meistens nur in gedruckter Form vorliegen, sollte auch hier Raum für die Lagerung mit einkalkuliert oder eine Digitalisierung der Informationen erwogen werden.[29]

Beispieldaten des Computerspiele Museums Berlin

Die Softwaresammlung umfasst zurzeit 12.000 Titel über eine Zeitspanne von 1972 bis heute. Die Software wird getrennt von der Hardware in normalen Büroräumen gelagert und hat einen Platzbedarf von ca. 70 qm.

In der Hardwaresammlung des Computerspiele Museums befinden sich augenblicklich 2180 Sammlungsstücke. Sie sind in einer Datenbank inklusive Foto erfasst und inventarisiert. Die Sammlung besteht aus Videospielautomaten, Videospiele Konsolen, Heimcomputern, Handhelds, technischen Zusatzteilen (Laufwerke, Controller, Monitore etc.). Des Weiteren besitzt das Museum eine umfangreiche Sammlung gedruckter Informationen wie Computerspiele, Magazine und Handbücher. Diese sind in einer gesonderten Datenbank erfasst. Die Hardwaresammlung ist auf ca. 200 qm an der Peripherie Berlins untergebracht. Der Hauptgrund dafür ist die günstigere

29 s. Dooijes (2000), a.a.O.

Miete für die Räume, als das in zentralerer Lage möglich wäre. Die Räume sind beheizbar und entsprechen größtenteils ebenfalls Bürostandard.[30]

30 Daten stammen von Herrn Andreas Lange, Kurator des Computerspielemuseums Berlin (2006)

8.6 Mikroverfilmung

Christian Keitel

Die Eignung des Mikrofilms als analoger oder digitaler Datenträger für digital vorliegende Bildinformation wird diskutiert, notwendige Rahmenbedingungen werden benannt.

Ein ungelöstes Problem bei der langfristigen Archivierung digitaler Informationen ist die begrenzte Haltbarkeit digitaler Datenträger. Künstliche Alterungstests sagen CDs, DVDs und Magnetbändern nur eine wenige Jahre währende Haltbarkeit voraus, während herkömmliche Trägermedien wie z.B. Pergament oder Papier mehrere Jahrhunderte als Datenspeicher dienen können. Hervorragende Ergebnisse erzielt bei diesen Tests insbesondere der Mikrofilm. Bei geeigneter (kühler) Lagerung wird ihm eine Haltbarkeit von über 500 Jahren vorausgesagt. Verschiedene Projekte versuchen daher, diese Eigenschaften auch für die Archivierung genuin digitaler Objekte einzusetzen. Neben der Haltbarkeit des Datenträgers sind dabei auch Aspekte wie Formate, Metadaten und Kosten zu bedenken.

In Anlehnung an die Sicherungs- und Ersatzverfilmung herkömmlicher Archivalien wurden zunächst digitale Informationen auf Mikrofilm als Bild ausbelichtet und eine spätere Benutzung in einem geeigneten Lesegerät (Mikrofilmreader) geplant. Das menschliche Auge benötigt zur Ansicht dieser Bilder nur eine Lupe als optische Vergrößerungshilfe. Erinnert sei in diesem Zusammenhang an das in den Anfängen des EDV-Einsatzes in Bibliotheken übliche COM-Verfahren (Computer Output on Microfilm/-fiche) zur Produktion von Katalog-Kopien. In letzter Zeit wird zunehmend von einer Benutzung im Computer gesprochen, was eine vorangehende Redigitalisierung voraussetzt. Dieses Szenario entwickelt die herkömmliche Verwendung des Mikrofilms weiter, sie mündet in einer gegenseitigen Verschränkung digitaler und analoger Techniken. Genuin digitale Daten werden dabei ebenso wie digitalisierte Daten von ursprünglich analogen Objekten/Archivalien auf Mikrofilm ausbelichtet und bei Bedarf zu einem späteren Zeitpunkt über einen Scanner redigitalisiert, um dann erneut digital im Computer benutzt zu werden. Eine derartige Konversionsstrategie erfordert im Vergleich mit der Verwendung des Mikrofilms als Benutzungsmedium einen wesentlich höheren Technikeinsatz.

Neben der Haltbarkeit des Datenträgers liegt ein zweiter Vorteil darin, dass die auf dem Mikrofilm als Bilder abgelegten Informationen nicht regelmäßig wie bei der Migrationsstrategie in neue Formate überführt werden müssen. Völ-

lig unabhängig von Formaterwägungen ist der Mikrofilm jedoch nicht, da er über die Ablagestruktur von Primär- und v.a. Metadaten gewisse Ansprüche an das Zielformat bei der Redigitalisierung stellt, z.B. die bei den Metadaten angewandte Form der Strukturierung. Die Vorteile im Bereich der Formate verlieren sich, wenn der Mikrofilm als digitales Speichermedium begriffen wird, auf dem die Informationen nicht mehr als Bild, sondern als eine endlose Abfolge von Nullen und Einsen binär, d.h. als *Bitstream*, abgelegt werden. Es bleibt dann allein die Haltbarkeit des Datenträgers bestehen, die in den meisten Fällen die Zeit, in der das verwendete Dateiformat noch von künftigen Computern verstanden wird, um ein Vielfaches übersteigen dürfte. Auf der anderen Seite entstehen für diese Zeit verglichen mit anderen digitalen Speichermedien nur sehr geringe Erhaltungskosten.

Bei der Ausbelichtung der digitalen Objekte ist darauf zu achten, dass neben den Primärdaten auch die zugehörigen Metadaten auf dem Film abgelegt werden. Verglichen mit rein digitalen Erhaltungsstrategien kann dabei zum einen die für ein Verständnis unabdingbare Einheit von Meta- und Primärdaten leichter bewahrt werden. Das von OAIS definierte archivische Informationspaket (AIP) wird hier physisch konstruiert. Zum anderen verspricht die Ablage auf Mikrofilm auch Vorteile beim Nachweis von Authentizität und Integrität, da die Daten selbst nur schwer manipuliert werden können (die Möglichkeit ergibt sich nur durch die erneute Herstellung eines Films).

Vor einer Abwägung der unterschiedlichen Erhaltungsstrategien sollten sowohl die Benutzungsbedingungen als auch die Kosten beachtet werden, die bei der Ausbelichtung, Lagerung und erneuten Redigitalisierung entstehen. Schließlich ist zu überlegen, in welcher Form die Informationen künftig verwendet werden sollen. Während der Einsatz des Mikrofilms bei Rasterbildern (nichtkodierten Informationen) naheliegt, müssen kodierte Informationen nach erfolgter Redigitalisierung erneut in Zeichen umgewandelt werden. Die Fehlerhäufigkeit der eingesetzten Software muss dabei gegen die zu erwartenden Vorteile aufgewogen werden.

Literatur

Projekt ARCHE, s. http://www.landesarchiv-bw.de >>> Aktuelles >>> Projekte

Eine Bibliographie findet sich beim Forum Bestandserhaltung unter http://www.uni-muenster.de/Forum-Bestandserhaltung/konversion/digi.html

9 Access

9.1 Einführung

Karsten Huth

Der Titel dieses Kapitels ist ein Begriff aus dem grundlegenden ISO Standard OAIS. Access steht dort für ein abstraktes Funktionsmodul (bestehend aus einer Menge von Einzelfunktionalitäten), welches im Wesentlichen den Zugriff auf die im Archiv vorgehaltenen Informationen regelt. Das Modul Access ist die Schnittstelle zwischen den OAIS-Modulen „Data Management", „Administration" und „Archival Storage".[1] Zudem ist das Access-Modul die Visitenkarte eines OAIS für die Außenwelt. Nutzer eines Langzeitarchivs treten ausschließlich über dieses Modul mit dem Archiv in Kontakt und erhalten gegebenenfalls Zugriff auf die Archivinformationen. In der digital vernetzten Welt kann man davon ausgehen, dass der Nutzer von zu Hause aus über ein Netzwerk in den Beständen eines Archivs recherchiert. Entsprechende technische Funktionali-

1 Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information System: Blue Book*. Washington, DC. Page 4-14ff

täten wie Datenbankanfragen an Online-Kataloge oder elektronische Findmittel werden bei vielen Langzeitarchiven zum Service gehören. Die Möglichkeit von Fernanfragen an Datenbanken ist jedoch keine besondere Eigenart eines Langzeitarchivs. Wesentlich sind folgende Fragen:

- Wie können die Informationsobjekte (z. T. auch als konzeptuelle Objekte bezeichnet) dauerhaft korrekt adressiert und nachgewiesen werden, wenn die logischen Objekte (z.B. Dateien, Datenobjekte) im Zuge von Migrationen technisch verändert werden und im Archiv in verschiedenen technischen Repräsentationen vorliegen?[2]
- Wie kann der Nutzer erkennen, dass die an ihn gelieferte Archivinformation auch integer und authentisch ist?[3]
- Wie kann das Archiv bei fortwährendem technologischem Wandel gewährleisten, dass die Nutzer die erhaltenen Informationen mit ihren verfügbaren technischen und intellektuellen Mitteln auch interpretieren können?

Erst wenn sich ein Archiv in Bezug auf den Zugriff mit den oben genannten Fragen befasst, handelt es strategisch im Sinne der Langzeitarchivierung. Die entsprechenden Maßnahmen bestehen natürlich zum Teil aus der Einführung und Implementierung von geeigneten technischen Infrastrukturen und Lösungen. Da die technischen Lösungen aber mit der Zeit auch veralten und ersetzt werden müssen, sind die organisatorisch-strategischen Maßnahmen eines Archivs von entscheidender Bedeutung. Unter diesem Gesichtspunkt sind Standardisierungen von globalen dauerhaften Identifikatoren, Zugriffsschnittstellen, Qualitätsmanagement und Zusammenschlüsse von Archiven unter gemeinsamen Zugriffsportalen eine wichtige Aufgabe für die nationale und internationale Gemeinde der Gedächtnisorganisationen.

---

2 vgl. Funk, Stefan: *Kap 7.2 Digitale Objekte und Formate*

3 nestor - Materialien 8: nestor - Kompetenznetzwerk Langzeitarchivierung / Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Version 1 (Entwurf zur Öffentlichen Kommentierung), Juni 2006, Frankfurt am Main : nestor c/o Die Deutsche Bibliothek, urn:nbn:de:0008-2006060710; Punkt 6.3 S. 16. Als Version 2 unter http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_08.pdf
abrufbar Es ist davon auszugehen, dass in den nächsten Jahren eine Klärung erfolgt, welche weiteren Persistenten Identifikatoren – außer den in diesem Kapitel vertieft Behandelten URN und DOI – in der Anwendung Bedeutung erhalten werden.

9.2 Workflows für den Objektzugriff

Dirk von Suchodoletz

Es genügt nicht, lediglich ein digitales Objekt bit-getreu zu bewahren, sondern es sind Vorkehrungen zu treffen, um dieses Objekt zu einem späteren Zeitpunkt wieder darstellen zu können. Hierzu dienen bestimmte Workflows, die ein Langzeitarchiv implementieren sollte. Deshalb beschäftigt sich dieser Abschnitt mit der theoretischen Fundierung und Formalisierung technischer Abläufe, wie sie beispielsweise mit dem DIAS-Projekt zu Beginn des Jahrtausends an der Königlichen Bibliothek der Niederlande eingeführt wurden[4].

Der zentrale Ausgangspunkt der Überlegungen liegt darin begründet, dass digitale Objekte nicht allein aus sich heraus genutzt oder betrachtet werden können. Stattdessen benötigen sie einen geeigneten Kontext, damit auf sie zugegriffen werden kann. Dieser Kontext, im Folgenden Erstellungs- oder Nutzungsumgebung genannt, muss geeignete Hardware- und Softwarekomponenten so zusammenfügen, dass je nach Typ des digitalen Objekts dessen Erstellungsumgebung oder ein geeignetes Äquivalent erzeugt wird. Für diese Schritte der Wiederherstellung sind seitens des Archivbetreibers geeignete Workflows (Arbeitsabläufe) vorzusehen. Um diese beschreiben zu können, sind sogenannte „View-Paths" ein zentrales Konzept. Diese Darstellungspfade liefern die grundlegenden Anweisungen zur Konstruktion geeigneter technischer Workflows für das Sichtbarmachen oder Ablaufenlassen verschiedener digitaler Objekte.

Den Betreibern eines digitalen Langzeitarchivs wachsen aus diesen Überlegungen verschiedene Aufgaben zu. Hierzu zählen die Bestimmung des Typs eines Objekts bei der Einstellung ins Archiv (dem sog. Ingest in der OAIS-Terminologie) und die Beschaffung und Ablage der notwendigen Metadaten, auf die an anderer Stelle in diesem Handbuch ausführlich eingegangen wird.

Für den späteren Objektzugriff spielt die Überprüfung, inwieweit im Langzeitarchivierungssystem eine für diesen Objekttyp passende Nutzungsumgebung vorhanden ist, eine besondere Rolle. Deren technischer Workflow wird nachfolgend näher ausgeführt. Dabei können View-Path und Nutzungsumgebung je nach Art der betreibenden Organisation und ihrer spezifischen Anforderungen, die typischerweise durch „Significant Properties"[5] beschrieben

4 Fußnotentext:Vgl. van Diessen; Steenbakkers 2002 und van Diessen 2002, S. 16f

5 Vgl. http://www.significantproperties.org.uk/about.html sowie http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.aspx

werden, unterschiedlich aussehen. Dies resultiert je nach Benutzergruppe oder Einrichtung in verschiedene Kriterien, nach denen Darstellungspfade bestimmt werden. Es lassen sich drei wesentliche Phasen voneinander unterscheiden (Abbildung 1):

- Notwendige Arbeitsschritte und Handlungen bei der Objektaufnahme in das OAIS-konforme Archiv.
- Workflows, die im Laufe des Archivbetriebs umzusetzen sind.
- Abläufe für den Objektzugriff nach der Objektausgabe an den Endbenutzer.

Der View-Path zum Zugriff auf die unterschiedlichen, im Langzeitarchiv abgelegten Objekttypen ist im Moment der Archivausgabe festzulegen. An dieser Stelle müssen diverse Workflows implementiert sein, die es erst erlauben, dass ein späterer Archivnutzer tatsächlich auf das gewünschte Objekt zugreifen kann. Hierbei spielt die auf das Objekt angewendete Langzeitstrategie, ob Migration oder Emulation, keine Rolle (Abbildung 2). In jedem Fall muss das Archivmanagement dafür sorgen, dass eine passende Nutzungsumgebung bereitgestellt wird.

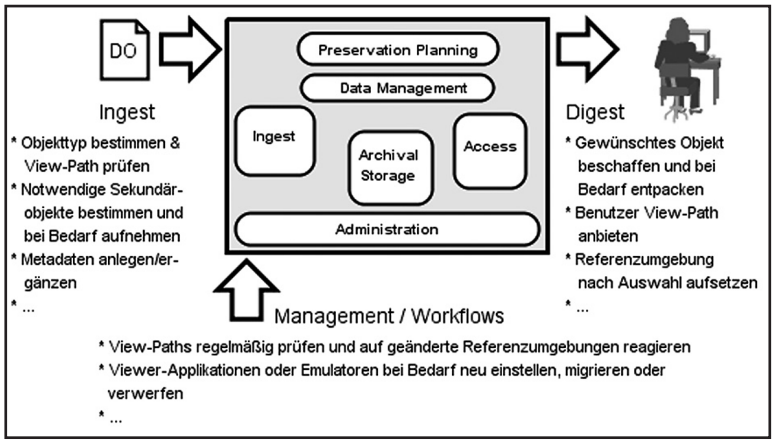


Abbildung 1: Bei der Unterhaltung eines digitalen Langzeitarchivs sind eine Reihe verschiedener technischer Workflows festzulegen und umzusetzen.

Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

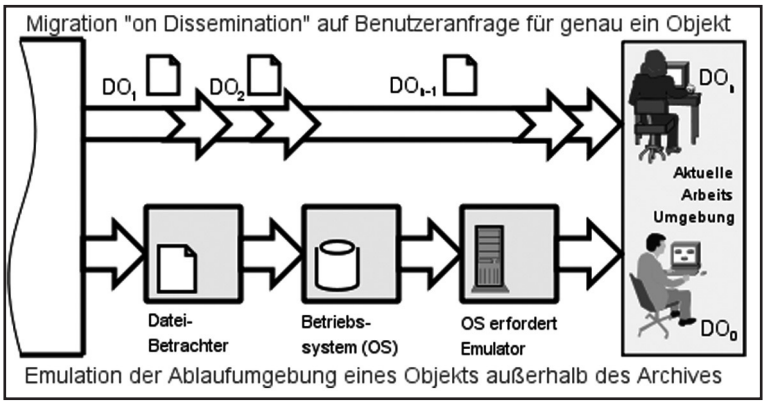


Abbildung 2: Unabhängig von der eingesetzten Archivierungsstrategie eines Objekts muss für einen geeigneten Zugriff gesorgt werden.

Formalisierter Zugriff auf archivierte Objekte

Die Wiederherstellung von Nutzungsumgebungen oder entsprechender Äquivalente lässt sich durch View-Paths formalisieren. Diese Wege starten vom darzustellenden oder auszuführenden digitalen Objekt. Sie reichen je nach angewendeter Archivierungsstrategie über verschiedene Zwischenschritte bis in die tatsächliche Arbeitsumgebung des Archivnutzers. Da digitale Archivalien, im Zusammenhang mit View-Path auch Primärobjekte genannt, nicht aus sich allein heraus genutzt werden können, werden weitere digitale Objekte benötigt. Hierzu zählen Viewer, Hilfsapplikationen, Betriebssysteme oder Emulatoren. Sie sind als Hilfsmittel, im Folgenden als Sekundärobjekte bezeichnet, nicht von primärem Archivierungsinteresse, jedoch zwingend ebenfalls zu berücksichtigen.

Das Konzept der Darstellungspfade wurde ursprünglich im Zuge des eingangs genannten DIAS-Projektes an der Königlichen Bibliothek der Niederlande entwickelt[6]. Die Abbildung 3 zeigt einen typischen View-Path-Verlauf ausgehend vom digitalen Objekt. Im Beispiel wurde es mittels einer bestimmten Software erzeugt, die ihrerseits wiederum auf einem Betriebssystem ausgeführt werden kann, das seinerseits wegen des Nicht-Vorhandenseins des originalen Rech-

6 Vgl. http://www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html sowie van Diessen; Steenbakkers (2002).

ners einen Hardwareemulator erfordert. Dieser läuft als Applikation in der Arbeitsumgebung des Archivnutzers.

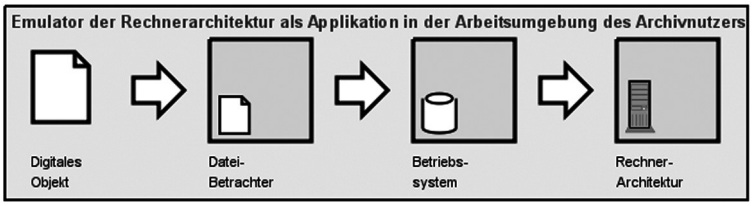


Abbildung 3: Das Beispiel zeigt einen generellen View-Path zum Rendering/Ablaufenlassen eines digitalen Objekts eines bestimmten Typs unter Einsatz der Emulationsstrategie.

Weitere View-Path-Beispiele

Ein View-Path eines migrierten Objekts wäre entsprechend kürzer: Da eine in der aktuellen Arbeitsumgebung ablaufende Applikation genutzt werden kann, muss lediglich diese passend zum jeweiligen Objekttyp bereitgestellt werden. Umgekehrt kann sich ein View-Path bei geschachtelter Emulation verlängern: Steht der im obigen Beispiel genannte Hardwareemulator nur für ein älteres Betriebssystem statt der aktuellen Umgebung zur Verfügung, würde ein Zwischenschritt aus diesem obsoleten Betriebssystem mit passendem Emulator für die aktuelle Umgebung angehängt werden (Abbildung 4 unten).

Während der Ausgangspunkt des View-Paths durch das Primärobjekt fixiert ist, wird sich, erzwungen durch den technologischen Fortschritt und die sukzessive Obsoleszenz vorhandener Rechnerplattformen, der Endpunkt des View-Path im Zeitablauf verschieben. Weiterhin hängt die Länge eines Darstellungspfades vom Typ des Objekts ab: Ist eine Applikation wie z.B. eine alte Datenbank von primärem Interesse, so entfällt beispielsweise der Zwischenschritt der Erstellungs- oder Anzeigeapplikation, da sie direkt auf einem Betriebssystem ausgeführt werden kann.

View-Paths müssen nicht automatisch eindeutig bestimmt sein. Aus Sicht des Archivmanagements bieten sich generell folgende Szenarien für Darstellungspfade an:

- Es existiert zum gegebenen Zeitpunkt ein Weg vom Primärobjekt zu seiner Darstellung oder Ausführung.

- Es existieren mehrere verschiedene View-Paths. Diese sind mit geeigneten Metriken zu versehen. Diese erlauben die Bewertung zur Verfügung stehender Alternativen und werden weiter hinten besprochen.
- Es kann Archivobjekte geben, zu denen zu bestimmten Zeitpunkten keine View-Paths konstruierbar sind.

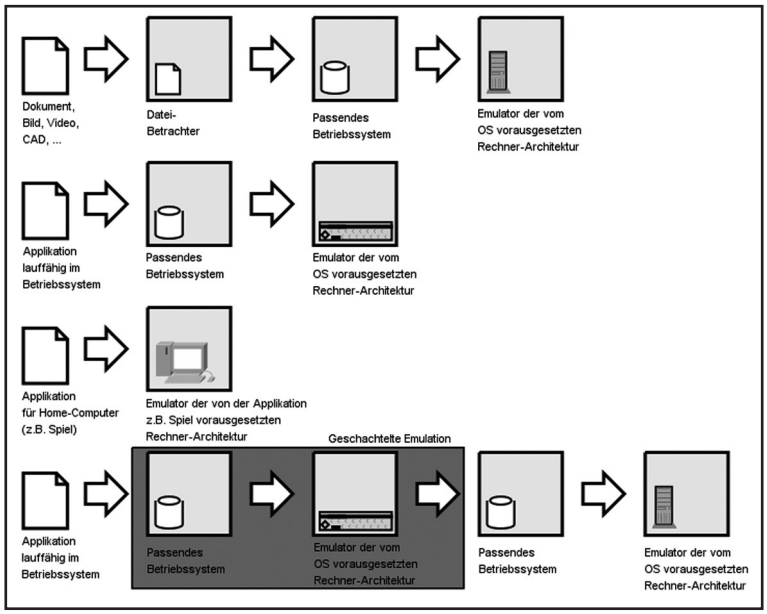


Abbildung 4: Die Länge eines View-Path hängt sowohl vom Typ des Objekts als auch von der eingesetzten Archivierungsstrategie ab.

Referenzumgebung - Endpunkt von View-Paths

Referenzumgebungen haben jeweils festgelegte Eigenschaften, die sich aus der Definition ihrer Hard- und Softwareumgebung bestimmen. Dieses sind die aktuellen Arbeitsumgebungen, in denen sich Archivbenutzer bewegen. Sie ändern sich zwangsläufig im Laufe der Zeit und sind unter anderem durch die Beschaffungspolitik der jeweiligen Gedächtnisorganisation determiniert. View-Paths sollten sich auf wohldefinierte Referenzumgebungen[7] beziehen.[8] Geeig-

7 Gedächtnisorganisationen haben typischerweise bestimmte Arbeitsumgebungen an ihren Arbeitsplätzen für ihre eigenen Archivare oder externe Nutzer (Abbildung 1). Diese Umgebungen können sie in einem gewissen Umfang selbst mitbestimmen. Vgl. a. van Diessen 2002.

8 van Diessen (2002b)

nete Referenzumgebungen versuchen deshalb, in möglichst kompakter und gut bedienbarer Form ein ganzes Spektrum von Nutzungsumgebungen zur Verfügung zu stellen. Dabei sollte die Basisplattform möglichst der jeweils aktuellen Hardware mit jeweils üblichen Betriebssystemen entsprechen. Das verhindert einerseits das Entstehen eines Hardwaremuseums mit hohen Betriebskosten und andererseits findet sich der Benutzer zumindest für das Basissystem in gewohnter Umgebung wieder.

Eine Referenzumgebung sollte in der Lage sein, neben der jeweiligen Nutzungsumgebung zusätzlich die notwendigen Hinweise zum Einrichten und zur Bedienung bereitzustellen. Dies schließt den geeigneten Zugriff auf die Objektmetadaten mit ein. Weitere Kriterien liegen in der Güte der Darstellung der Nutzungsumgebung, was durch den jeweils eingesetzten Emulator und seine Benutzer-Interfaces in der jeweiligen Referenzumgebung mitbestimmt wird.

Andererseits können Referenzumgebungen als Endpunkte eines View-Paths diesen umgekehrt mitbestimmen (Abbildung 5). Wenn technische, finanzielle oder personelle Restriktionen bestimmte Referenzumgebungen, die potenziell nachgefragt werden, nicht erlauben, kann dies die Wahl eines Emulators wesentlich beeinflussen. Die Rekursion kann sich unter Umständen noch weiter auf frühere Entscheidungsknoten auswirken.

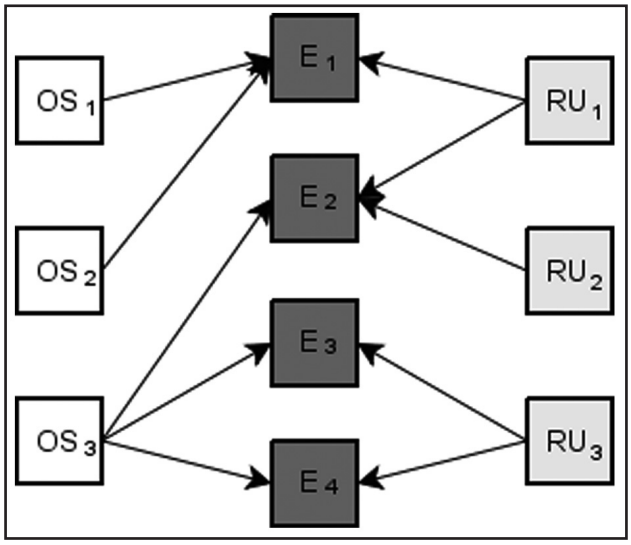


Abbildung 5: Die Auswahl der Emulatoren (E) wird in vielen Fällen nicht nur durch das Betriebssystem (OS) im View-Path, sondern auch von der anderen Seite durch die verfügbaren Referenzumgebungen (RU) beeinflusst.

Wegen ihrer spezielleren Anforderungen, die durch die eingesetzten Emulatoren und Viewer beeinflusst werden, ist es für die Betreiber von Langzeitarchiven vielfach sinnvoll, eine für die Objekte ihres Archivs relevante Referenzplattform selbst zu definieren und bereitzustellen. Diese unterscheidet sich je nach Anwendung und Gedächtnisorganisation: Bibliotheken und Archive brauchen in erster Linie Viewer für ihre migrierten statischen Objekte, die sie ihren Nutzern innerhalb ihrer Recherchesysteme anbieten wollen. Darüber hinaus können Emulatoren für nicht-migrierbare Archivinhalte und als Kontrollwerkzeug für mehrfach migrierte Objekte benötigt werden.[9]

Technische Museen oder Ausstellungen leben eher von interaktiven Objekten. Die Referenz-Workstation ist entsprechend den zu zeigenden Exponaten zu bestücken. Ähnliches gilt für multimediale Kunstobjekte. Hier könnten die Significant Properties jedoch sehr spezielle Anforderungen an eine möglichst authentische Präsentation stellen.

Für Firmen oder Institutionen werden in den meisten Fällen lediglich View-Paths erforderlich sein, die sich bereits mittels X86-Virtualisierern[10] komplett erstellen lassen. Die erwarteten Objekte sind eher statischer Natur und wurden typischerweise auf PCs verschiedener Generationen erstellt. Generell muss es sich bei den eingesetzten Referenz-Workstations nicht um die jeweils allerneueste Hardwaregeneration handeln. Stattdessen sollte jene Technologie angestrebt werden, die einen optimalen Austausch erlaubt und den Anforderungen der jeweiligen Nutzer gerecht wird.

Je nach Art des Archivs kann ein Datenaustausch mit der Außenwelt erforderlich werden: Nicht alle Objekte müssen bereits von Anfang an im Archiv abgelegt sein. Es können später immer wieder Objekte, beispielsweise aus Nachlässen von Wissenschaftlern, Künstlern oder Politikern auftauchen. In solchen Fällen wird es zu Zwecken der Datenarchäologie[11] von Interesse sein, externe Objekte in bestehende Workflows einspeisen zu können. Umgekehrt sollen vielleicht Objekte für externe Nutzung speicher- oder ausdruckbar sein.

---

9 Die Darstellung eines Objekts via Emulation wird typischerweise deutlich aufwändiger sein. Gerade für häufig nachgefragte, statische Objekte bietet sich deshalb die Migration an. Bei Zweifeln an der Authentizität kann mittels Emulation das Ergebnis des n-ten Migrationsschritts mit dem unveränderten Originalobjekt verglichen werden.

10 Softwareprodukte wie VMware oder VirtualBox erlauben das Nachbilden eines X86er PCs auf einer X86er 32 oder 64-bit Maschine. Solange die Treiberunterstützung besteht, können ältere Betriebssysteme, wie Windows95 oder 2000 bequem innerhalb einer Applikation (dem Virtualisierer) auf dem Standard-Desktop des Benutzers ausgeführt werden.

11 Im Sinne des Umgangs (Bestimmung) mit sehr alten, lange schon nicht mehr zeit-typischen digitalen Objekten.

Für alle Schritte muss ein ausreichendes Bedienungswissen vorhanden sein. Hierzu werden sich neue Berufsfelder, wie das eines digitalen Archivars herausbilden müssen, um ausgebildetes Fachpersonal auch für recht alte Nutzungsumgebungen vorhalten zu können. Selbst wenn diese nicht mehr direkt auf der Originalhardware ablaufen, so müssen sie innerhalb der Emulatoren bedient werden können.

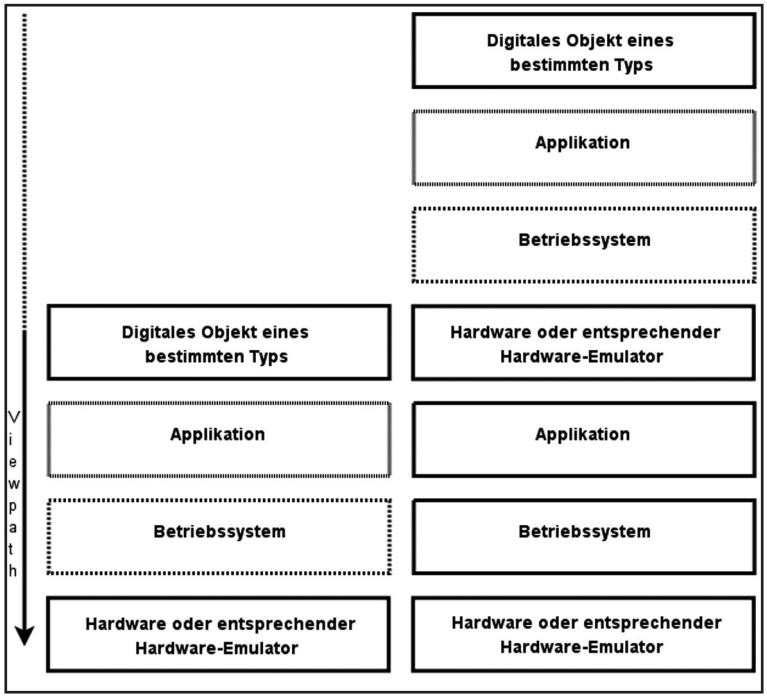


Abbildung 6: Ein View-Path in einer Darstellung als Schichtenmodell. Je nach Typ des Objekts muss nicht jede Schicht durch eine Softwarekomponente repräsentiert sein.

View-Path-Varianten und -Auswahl

Einen View-Path kann man als Entscheidungsbaum interpretieren, an dessen Wurzel das interessierende Primärobjekt steht. Ein Blatt ohne weitere Verzweigungen stellt das Ende des Pfades in Form einer gültigen Referenzumgebung dar. Zur Veranschaulichung des Aufbaus der geforderten Nutzungsumgebung sollte man sich am besten ein Schichtenmodell vorstellen, wie es in Abbildung 6 präsentiert wird.

Eine ganze Reihe digitaler Objekte können durch mehr als eine Applikation dargestellt oder ablauffähig gemacht werden. Dabei können die Ergebnisse in Authentizität, Komplexität oder Qualität differieren. Aus diesen Überlegungen

folgen Pfadverzweigungen und auf der Schicht der Applikation eine Auswahlmöglichkeit. Ähnliches wiederholt sich für die Anforderung der Applikation nach einem Betriebssystem. Auf dieser Ebene kann erneut eine Verzweigung auftreten. Die Rekursion setzt sich mit dem Betriebssystem und einer Menge an geeigneten Hardwareemulatoren fort.

Da der technologische Fortschritt wesentlichen Einfluss auf die Referenzumgebung hat und nur bedingt durch den Archivbetreiber beeinflusst werden kann, bestimmen sich View-Path und Referenzumgebung gegenseitig. Auf den Zwischenschichten stehen Betriebssysteme und Hardwareemulatoren über Gerätetreiber in Abhängigkeit zueinander.

Die Modellierung des View-Paths in Schichten erfolgt nicht eng fixiert: So reduziert sich beispielsweise bei einem digitalen Primärobjekt in Form eines Programms die Zahl der Schichten. Ähnliches gilt für einfache Plattformen wie bei Home-Computern, wo keine Trennung zwischen Betriebssystem und Applikation vorliegt (Abbildung 4 Mitte). Darüber hinaus können Schichten wiederum gestapelt sein, wenn es für einen bestimmten Emulator erforderlich wird, seinerseits eine geeignete Nutzungsumgebung herzustellen, was im rechten Teil von Abbildung 6 gezeigt wird.

Metriken als Entscheidungskriterien

Eine sinnvolle Erweiterung des etwas starren Ansatzes im ursprünglichen DIAS-Preservation-Modell (van Diessen; Steenbakkers 2002) könnte in der Gewichtung der einzelnen View-Path-Varianten liegen. Dies ließe sich durch eine beschreibende Metrik formalisieren. Gerade wenn an einem Knoten mehr als eine Option zur Auswahl steht (Abbildung 7), erscheint es sinnvoll:

- Präferenzen der Archivnutzer beispielsweise in Form der Auswahl der Applikation, des Betriebssystems oder der Referenzplattform zuzulassen.
- Gewichtungen vorzunehmen, ob beispielsweise besonderer Wert auf die Authentizität der Darstellung (van Diessen; van der Werf-Davelaar 2002) oder eine besonders einfache Nutzung gelegt wird.
- Vergleiche zwischen verschiedenen Wegen zuzulassen, um die Sicherheit und Qualität der Darstellung der Primärobjekte besser abzusichern.
- Den Aufwand abzuschätzen, der mit den verschiedenen Darstellungspfaden verbunden ist, um bei Bedarf eine zusätzliche ökonomische Bewertung zu erlauben.

Ein Ergebnis könnten mehrdimensionale Metriken sein, die mit den Objektmetadaten gespeichert und durch das Archivmanagement regelmäßig aktualisiert werden. Hierzu sollte eine Rückkopplung mit den Archivbenutzern erfolgen. So könnten in die Aktualisierungen Bewertungen seitens der Nutzer einfließen, die auf dem praktischen Einsatz bestimmter View-Paths sowie ihrer Handhabbarkeit und Vollständigkeit beruhen.

Aggregation von View-Paths

Wie erläutert, kann für bestimmte Objekttypen mehr als ein Darstellungspfad existieren. Dieses kann die Wahrscheinlichkeit des langfristig erfolgreichen Zugriffs verbessern – jedoch zum Preis potenziell höherer Kosten. Ausgehend vom Objekttyp und der eventuell notwendigen Applikation erhält man weitere View-Paths für andere Objekte, die automatisch anfallen: Ein einfaches Beispiel demonstrieren die sogenannten Office-Pakete, Zusammenstellungen verschiedener Applikationen. Sie können mit einer Vielfalt von Formaten umgehen – nicht nur mit denen der enthaltenen Teilkomponenten, sondern über Importfilter hinaus mit einer Reihe weiterer Dateiformate.[12]

Diese Betrachtungen können dazu dienen, eine potenziell überflüssige Referenzplattform zu identifizieren, die nur für ein bestimmtes Objekt vorgehalten wird, das aber auf alternativen Wegen ebenfalls darstellbar ist. So muss beispielsweise zur Betrachtung eines PDFs nicht eine besonders seltene Plattform genutzt werden, wenn ein gleichwertiger Viewer auf einer mehrfach genutzten anderen Plattform ebenfalls ablauffähig ist (Abbildung 7).

Schwieriger wird jedoch eine Zusammenlegung, wenn von den alternativen View-Paths nicht bekannt ist, ob sie ebenfalls zu 100% das gefragte Objekt rekonstruieren. Ein typisches Beispiel ist der Import von Word-Dokumenten in einer anderen als der Erstellungsapplikation. An diesem Punkt könnten ebenfalls die im vorangegangenen Abschnitt vorgenommenen Überlegungen zu Benutzerrückmeldungen in Betracht gezogen werden.

Mit jedem View-Path sind bestimmte Kosten verbunden, die vom Archivmanagement beachtet werden müssen. Diese lassen sich für jede Schicht eines View-Paths abschätzen. Übersteigen sie eine gewisse Schwelle, könnten ökonomische Erwägungen an die Aufnahme von Primärobjekten in das Archiv und das Archivmanagement geknüpft werden. In solchen Fällen könnten bestimmte

12 Ein Paket wie OpenOffice kann bezogen auf seine Importfilter für die verschiedenen Teilanwendungen wie Textverarbeitung, Tabellenkalkulation oder Präsentation inklusive der Unterstützung verschiedener älterer Versionen schnell über 100 verschiedene Formate statischer Objekte lesen.

Dateiformate abgelehnt werden, weil deren spätere Rekonstruktion zu hohe Aufwendungen erwarten lässt.

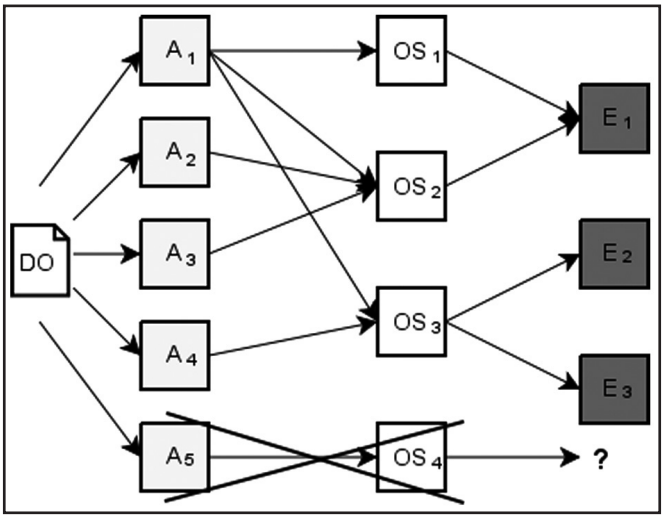


Abbildung 7: Redundante View-Paths zur Darstellung oder zum Abspielen eines digitalen Objekts (DO). Dieses kann im Beispiel von fünf verschiedenen Applikationen (A) geöffnet werden, die auf vier verschiedenen Betriebssystemen (OS) laufen können, wobei für ein OS kein geeigneter Emulator (E) bereitsteht.

So sollte je nach Situation oder Archiv hinterfragt werden, einen Darstellungspfad aufrecht zu erhalten, wenn sinnvolle Alternativen existieren. Liegt beispielsweise ein Primärobjekt nach dem PDF 1.0 Standard vor, welches mit einem Werkzeug in einer Windows 3.11 Nutzungsumgebung erzeugt wurde, muss deshalb diese Umgebung nicht zwingend erhalten werden. Unter bestimmten Bedingungen kann auf einen View-Path verzichtet werden:

- Es existiert eine ausreichende Anzahl weiterer, gleichwertiger und dabei einfacherer Darstellungspfade.
- Wegen der guten und vollständigen Beschreibung des Formats ist es deutlich einfacher, einen Viewer für die jeweils aktuellen Arbeitsumgebungen zu migrieren, als alte Nutzungsumgebungen durch Emulatoren zu sichern.
- Dieser Objekttyp ist der einzige, der eine Windows 3.11 Umgebung potenziell verlangt.

- Es gibt kein spezielles Interesse, diese Nutzungsumgebung aufrecht zu erhalten, da sie nicht im Fokus der Institution liegt.

Ein solches Vorgehen ließe sich auf andere View-Paths ausdehnen, die für eine Reihe von Dateiformaten und Applikationen Apple- oder alternativ Microsoft-Betriebssysteme voraussetzen. Wenn beispielsweise kein gesondertes Bedürfnis speziell nach dedizierten Apple-Nutzungsumgebungen besteht, weil beispielsweise die Art der Benutzerinteraktion, das Aussehen der grafischen Oberfläche und der Applikation in dieser von eigenständigem Interesse sind, könnte ein solcher Zweig im Archiv geschlossen werden. Besonders gut lässt sich dies am OpenOffice veranschaulichen, welches für etliche kommerzielle und freie UNIX-Varianten, wie BSD, Solaris oder Linux, Mac-OS X und selbstredend für die Windows-Betriebssysteme angeboten wird.

Ähnlich liegt der Fall in der trivialen Vervielfachung der View-Paths durch unterschiedliche Anpassungen von Programmen und Betriebssystemen an nationale und regionale Besonderheiten. Hierzu zählen natürliche Sprachen, spezielle Zeichen oder auch Währungen und Einheiten. Während Open-Source-Betriebssysteme und Applikationen schon lange mehrere Sprachpakete in einer Installation erlauben, setzte dies sich bei kommerziellen Systemen erst recht spät durch. Das Bedürfnis von einem Betriebssystem wie Windows 2000 alle verschiedenen Landesversionen aufzubewahren, lässt sich vermutlich besser in Kooperationen ähnlicher Gedächtnisinstitutionen über Landesgrenzen hinweg erreichen. Dies würde zudem Kosten und Aufwand für redundante Lizensierungen reduzieren helfen.

Einfacher und damit oft günstiger zu pflegenden View-Paths kann der Vorrang vor anderen eingeräumt werden. Jedoch sind auch hier Voraussagen schwierig und es kann die Gefahr bestehen, dass sich mit dem Wandel der Referenzumgebungen die Kostenstrukturen erneut ändern. Andererseits lassen sich Vorkehrungen treffen, dass seltenere View-Paths zumindest an spezialisierten Institutionen mit besonderem Sammelauftrag und entsprechender Finanzierung weiter betreut werden.

Die verschiedenen Strategien der Langzeitarchivierung der Emulatoren – Migration oder Schachtelung generieren unterschiedliche Aufwendungen im Archivbetrieb:

- Die geschachtelte Emulation sorgt für eher längere Darstellungspfade bei geringem Migrationsbedarf. Der Aufwand entsteht beim zunehmend komplexer werdenden Zugriff auf das Primärobjekt.

- Die Migration von Emulatoren, Universal Virtual Machines und modulare Ansätze[13] generieren eher kurze View-Paths bei einfacherem Zugriff auf das Primärobjekt. Jedoch liegt der Aufwand im regelmäßigen Update aller benötigter Emulatoren oder ihrer Teilkomponenten.

Der erste Ansatz ist aus diesem Grund eher für Objekte mit seltenem Zugriff oder Institutionen mit kleinen, speziell ausgebildeten Nutzerkreisen wie Archive, geeignet. Die zweite Strategie passt sicherlich besser auf viel genutzte Objekttypen mit größeren Anwenderkreisen. Eine Reihe von Kostenfaktoren kann durch gemeinschaftliche Anstrengungen und verteilte Aufgaben reduziert werden.

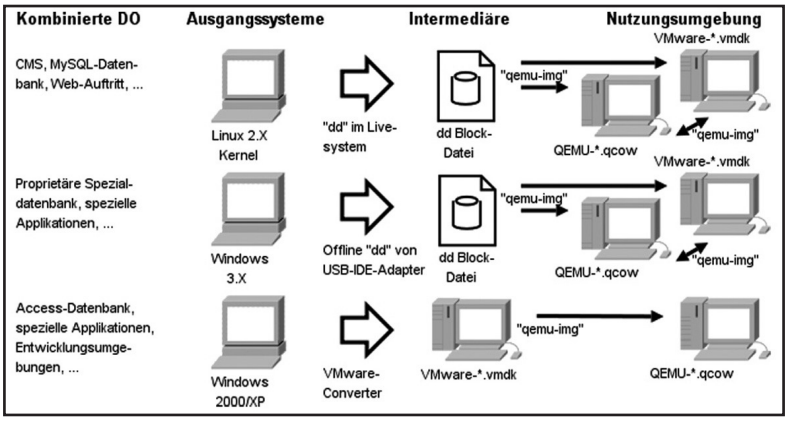


Abbildung 8: In Abhängigkeit vom zu archivierenden Gesamtsystem kombinierter digitaler Objekte existieren verschiedene Wege zur Archivierung.

Kombinierte Archivobjekte

Primärobjekte mit bereits im Paket enthaltenen notwendigen Sekundärobjekten wie bestimmte Metadaten zur Beschreibung, Viewer und Hilfsprogramme, zusammengeführt in einem Archival Information Package (AIP), können bereits aufbereitet vorgehalten werden.

Solche kombinierten Objekte könnten aus der Überführung eines physischen Systems in ein virtuelles entstehen. Dieses Vorgehen lässt sich beispielsweise auf sehr spezielle dynamische Objekte, wie Datenbanken von Content Management Systemen (CMS) für WWW-Auftritte, Firmendaten aus Produk-

13 Vgl. van der Hoeven; van Diessel; van der Meer (2005).

tionsanlagen oder Entwicklungsabteilungen vorstellen (Abbildung 8): Statt Daten, Applikation und notwendiges Betriebssystem im Zuge der Archivierung voneinander zu lösen, könnte eine komplette Maschine in eine emulierte Umgebung überführt werden. Damit bleibt die Serverfunktionalität mit ihren gesamten Einstellungen als eine Einheit erhalten. Das Verfahren kann auf diese Weise sicherstellen, dass ein AIP über Archive hinweg ohne einschränkende Nebenbedingungen ausgetauscht werden kann.

Das Zusammenfassen bestimmter View-Paths in einer gemeinsamen Umgebung könnte einerseits für bestimmte Forschungsthemen und andererseits für eine Vereinfachung der View-Path-Erstellung für den Benutzerbetrieb sinnvoll sein.

Eine etwas anders gelagerte Zusammenfassung schlagen Reichherzer[14] und Brown (2006) vor, um auf Datensätze der öffentlichen Administration zu späteren Zeitpunkten zugreifen zu können. Solche typischerweise in einem komprimierten Archiv zusammengefassten Daten von Erhebungen, Statistiken oder Auswertungen umfassen eine Reihe typischer Objektformate einer gewissen Epoche. Diese könnten in einem gemeinsamen Container untergebracht sein und Hilfsprogramme, wie den Dokumentenausdruck nach Postscript oder PDF enthalten. Passende virtuelle Drucker, die durch Postscript- oder PDF-Generatoren repräsentiert werden, können fehlende Exportfilter alter Applikationen ersetzen.

View-Paths als OAIS-Workflows

Im OAIS-Referenzmodell übernimmt das Management des digitalen Langzeitarchivs eine Reihe von Aufgaben, die sich mit dem Lifecycle-Management und damit verbundenen Workflows von Primärobjekten befassen. Hierfür wurden besonders die Arbeitsprozesse für die Objektausgabe betrachtet. Für das langfristige Management wird insbesondere die Aufgabe des Preservation Planning interessant.[15]

Ein zentrales Moment ist die regelmäßige Kontrolle der View-Paths bei einem Wechsel der Referenzumgebung als Bezugsgröße. Jeder Plattformwechsel stellt neue Anforderungen für die Wiederherstellung von Nutzungsumgebungen. Bei dieser Überprüfung handelt es sich um einen iterativen Prozess, der über alle registrierten Objekttypen des Archivs abläuft. Hierfür ist jeweils eine passende Strategie für den Übergang von einer Referenzumgebung auf eine

14 Die Idee hierzu findet sich bereits bei Rothenberg (1998) sowie Reichherzer und Brown (2006).

15 van Diessen (2002a)

neue zu suchen. Generell gilt: Neue Objekttypen und Dateiformate erfordern neue Darstellungspfade.

Es ergeben sich verschiedene Anforderungen an den Archivbetrieb:

- Erstellung eines Hintergrundarchivs - In diesem werden die einzelnen Elemente des View-Path dauerhaft abgelegt. Sie werden dann genauso behandelt wie Primärobjekte. An dieser Stelle kann überlegt werden, ob bestimmte Einzelobjekte, wie Emulatoren, spezifische Hilfsprogramme und Beschreibungen in einem AIP gebündelt oder einzeln abgelegt werden.
- Betrieb eines Online-Archivs für den Direktzugriff - Für häufig nachgefragte Sekundärobjekte kann es sinnvoll sein, diese zusätzlich zum Langzeitarchiv in einem speziellen Archiv, wie einem aktuellen Dateisystem einer Referenzumgebung, vorzuhalten. Das kann einerseits das Langzeitarchiv entlasten und andererseits zu einem beschleunigten Ablauf der View-Path-Erstellung führen.
- Anlage eines View-Path-Caches - Für häufiger nachgefragte und aufwändiger zu generierende Darstellungspfade kann die Vorhaltung vorbereiteter Nutzungsumgebungen den Aufwand für Nutzer und Archivbetreiber reduzieren. Diese Caches könnten als Teil des Online-Archivs oder direkt auf der Referenzplattform abgelegt sein.

Die aufgezeigten Überlegungen haben klare ökonomische Implikationen für die mit digitalen Objekten befassten Gedächtnisorganisationen. Sie werden sich im Zuge der technischen Workflows, die sich mit der Wiederherstellung von Nutzungsumgebungen befassen, einen Teil der notwendigen Entwicklungen der Emulatoren und Viewer selbst leisten oder diese Leistung am Markt einkaufen müssen. Entstehen bei den angestrebten Entwicklungen offene Standards und Werkzeuge, wie PRONOM[16] oder Dioscuri,[17] können sich einerseits die Belastungen der Einzelinstitution in Grenzen halten und andererseits verbindliche Verfahren entwickeln, die von einer breiten Anwendergemeinschaft unterstützt werden.

16 Vgl. http://dioscuri.sourceforge.net/

17 Vgl. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Literatur

The Preservation Manager for the e-Depot: http://www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html

van Diessen, Raymond; Steenbakkers, Johan F. (2002): *The Long-Term Preservation Study of the DNEP project - an overview of the results,* The Hague, The Netherlands; http://www.kb.nl/hrd/dd/dd_onderzoek/reports/1-overview.pdf

van Diessen, Raymond (2002): *Preservation Requirements in a Deposit System*, The Hague, The Netherlands: http://www.kb.nl/hrd/dd/dd_onderzoek/reports/3-preservation.pdf

van der Hoeven, Jeffrey; van Diessen, Raymond; van der Meer; K. (2005):*Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects*, Journal of Information Science, Vol. 31, No. 3, 196-208; DOI: 10.1177/0165551505052347

van Diessen, Raymond; van der Werf-Davelaar, Titia (2002): *Authenticity in a Digital Environment,* The Hague, The Netherlands, http://www.kb.nl/hrd/dd/dd_onderzoek/reports/2-authenticity.pdf

Rothenberg, Jeff (1998): Avoiding Technological Quicksand: *Finding a Viable Technical Foundation for Digital Preservation in „The State of Digital Preservation: An International Perspective"*, Washington, http://www.clir.org/pubs/reports/rothenberg/contents.html

Reichherzer, Thomas; Geoffrey, Geoffrey (2006): *Quantifying software requirements for supporting archived office documents using emulation*, International Conference on Digital Libraries 06: Proceedings of the 6th ACM/IEEE-CS joint Conference on Digital Libraries; S. 86-94

9.3 Retrieval

Matthias Neubauer

Genauso wichtig wie die sichere Archivierung der digitalen Objekte ist auch die Möglichkeit, diese Objekte wieder aus dem Archiv herauszuholen und zu nutzen. Dabei muss gewährleistet sein, dass die Objekte den Zustand und den Informationsgehalt zum Zeitpunkt des Einspielens in das Archivsystem widerspiegeln. Im Idealfall sollte das Objekt noch exakt so abrufbar sein, wie es einmal in das Archiv eingespielt wurde. Je nach Verwendungszweck kann es jedoch auch sinnvoll sein, eher eine migrierte Form eines Objektes abzurufen. Einige wichtige Punkte, die es beim Zugriff von archivierten Objekten zu beachten gilt, sollen im Folgenden kurz erläutert werden.

Objektidentifikation

Zunächst ist eine eindeutige Identifikation des abzurufenden Objektes wichtig. Zu dieser Thematik existieren vielerlei Lösungen und Philosophien. Einige werden in den folgenden Kapiteln zum Thema „Persistent Identifier" vorgestellt. Grundsätzlich muss es anhand der verwendeten Identifizierungen möglich sein, jedwede Form und Version eines digitalen Objektes aus dem Langzeitarchiv abzurufen. Dies kann gegebenenfalls auch durch eine Kombination von externen und internen Identifikatoren realisiert werden.

Datenkonsistenz

Die Unversehrtheit der Daten hat höchste Priorität. Innerhalb des Archivs sollte durch geeignete Routinen zwar sichergestellt sein, dass der originale digitale Datenstrom erhalten bleibt. Jedoch können auch - und vor allem - bei der Übertragung der Daten aus dem Archiv heraus Inkonsistenzen durch Übertragungsfehler oder andere Störeinflüsse entstehen. Idealerweise sollte daher bei jedem Zugriff auf ein Archivobjekt über Checksummenvergleiche die Unversehrtheit der Daten sichergestellt werden. Je nach Art und Status der Daten kann diese Überprüfung auch nur stichprobenartig erfolgen.

Versionsmanagement

Je nach Verwendungszweck der Daten kann es entweder sinnvoll sein, das ursprüngliche Originalobjekt aus dem Archiv herauszuholen oder aber auch eine migrierte Form zu nutzen. Die höchste Authentizität wird man sicherlich mit

dem ursprünglichen Objekt erreichen, jedoch kann es sich auf zukünftigen Systemen sehr schwierig gestalten, die erhaltenen Daten aufzubereiten und zu nutzen (mehr darüber im Kapitel über Emulation und Migration). Ein gutes Langzeitarchivierungssystem sollte nach Möglichkeit sowohl Originalversion und letzte Migrationsform, als auch alle dazwischen liegenden Objektversionen zugreifbar halten, um eine vollkommene Transparenz und Rekonstruierbarkeit zu gewährleisten.

Interpretation und Aufbereitung der Daten

Sofern das digitale Objekt zum Zweck einer Präsentation oder Weiternutzung abgerufen wurde, muss es durch geeignete Methoden aufbereitet und verfügbar gemacht werden. Schon beim Einspielen der Daten in das Archivsystem ist daher darauf zu achten, dass man die Struktur des Objektes in den beiliegenden Metadaten dokumentiert. Zudem kann es notwendig sein, die innerhalb eines Archivsystems verwendeten Schlüsselnummern zur eindeutigen Identifikation von Dateiformaten zu entschlüsseln und auf ein anderes System einzustellen.

Caching

Unter dem Begriff „Caching" versteht man die Pufferung oft genutzter Daten in einem schnell verfügbaren und hochperformanten Zwischenspeicher. Im Falle des Retrieval aus einem Langzeitarchivierungssystem ist dies dann sinnvoll, wenn die Archivobjekte auch als Basis für Präsentationssysteme und den täglichen Zugriff dienen sollen. Um das Archivsystem nicht mit unnötigen Anfragen nach häufig genutzten Objekten zu belasten, wird ein lokaler Zwischenspeicher angelegt, der stark frequentierte Objekte vorhält und gegebenenfalls mit einer neuen Version innerhalb des Archivsystems synchronisiert beziehungsweise aktualisiert. Bei einem Zugriff auf das Objekt wird also nicht direkt das Archivsystem angesprochen, sondern zuerst geprüft, ob das Objekt bereits in der gewünschten Version lokal vorliegt. Eine kurze Kommunikation mit dem Archivsystem findet lediglich statt, um den Status und die Konsistenz des lokal vorliegenden Objektes zu validieren.

Sichere Übertragungswege

Um die Datensicherheit und den Datenschutz zu gewährleisten, sind sichere Übertragungswege zwischen dem Langzeitarchivierungssystem und dem zugreifenden System unerlässlich. Zwar kann eine etwaige Manipulation der Daten und Objekte durch die bereits angesprochene Checksummenüberprüfung erkannt werden, jedoch schützt dies nicht vor dem unerlaubten Zugriff Dritter

auf die Objekte des Archivsystems. Dies kann sowohl über sogenanntes Abhören der Datenleitung geschehen, als auch dadurch, dass unbefugte Dritte an Zugangsdaten und Netzwerkadressen des Archivsystems gelangen. Hier ist es daher sinnvoll, mit eindeutigen Befugnissen, sicheren Übertragungsprotokollen (wie HTTPS oder SFTP) und idealerweise Signaturschlüsseln und restriktiven IP-Freigaben zu arbeiten.

Datenübernahme in ein neues Archivsystem

Ein digitales Langzeitarchivsystem sollte die Möglichkeit bieten, alle Objekte zum Zwecke einer Migration auf ein neues oder anderes Archivsystem als Gesamtpaket oder als einzelne Objekte abzurufen. Verbunden mit dem einzelnen Objekt oder dem Gesamtpaket sollten auch alle gesammelten Metadaten sein. Sie sollten nach Möglichkeit komplett in das neue Archivsystem übernommen werden.

Diese Punkte sollten bei der Planung und Umsetzung von Zugriffsstrategien auf ein Archivsystem beachtet und mit einbezogen werden. Für individuelle Lösungen werden sicherlich auch noch weitere Faktoren eine Rolle spielen. Die jeweiligen Implementierungen sind natürlich auch stark von dem verwendeten Archivsystem abhängig.

9.4 Persistent Identifier (PI) – ein Überblick

Kathrin Schroeder

Warum Persistent Identifier?

Wer eine Printpublikation bestellt, kennt i.d.R. die ISBN – eine weltweit als eindeutig angesehene Nummer. Damit kann die Bestellung sicher ausgeführt werden. Eine ähnliche Nummerierung bieten Persistent Identifier für elektronische Publikationen, die im Internet veröffentlicht werden. Damit können sehr unterschiedliche digitale Objekte wie z.B. PDF-Dokumente, Bilder, Tonaufnahmen oder Animationen dauerhaft identifiziert und aufgefunden werden.

Als „ISBN für digitale Objekte" sind die gängigen Internetadressen, die Uniform Resource Locators (URL) nicht geeignet, da diese sich zu häufig ändern.[18] Stabile, weltweit eindeutige Identifier sind für ein digitales Langzeitarchiv unumgänglich, wie dies z.B. auch aus dem OAIS-Referenzmodell hervorgeht. Ein von außen sichtbarer stabiler Identifier ist für die zuverlässige Referenzierung sowie für die sichere Verknüpfung von Metadaten mit dem Objekt wichtig.

Kriterien

Kriterien an PI-Systeme können sehr unterschiedlich sein. Exemplarisch sind Kriterien, die in Der Deutschen Nationalbibliothek für die Entscheidung für ein PI-System zugrunde gelegt wurden, aufgeführt.

Standardisierung

- Verankerung in internationalen Standards

Funktionale Anforderungen

- Standortunabhängigkeit des Identifiers
- Persistenz
- weltweite Eindeutigkeit
- Der Identifier ist adressierbar und anklickbar (Resolving).

18 Weiterführende Informationen zu „Adressierung im Internet und Leistungsgrenzen standortgebundener Verweise" vgl. http://www.persistent-identifier.de/?link=202

- Es kann von einem PI gleichzeitig auf mehrere Kopien des Dokumentes (1:n-Beziehung) verwiesen werden.

Flexibilität, Skalierbarkeit

- Das PI-System ist skalierbar und
- flexibel in der PI-Anwendung selbst, d.h. es können neue Funktionalitäten hinzukommen, ohne die Konformität zum Standard zu gefährden.

Technologieunabhängigkeit und Kompatibilität

- Das PI-System ist generisch sowie protokoll- und technologieunabhängig als auch
- kompatibel mit existierenden Anwendungen und Diensten wie z.B. OpenURL, SFX, Z39.50, SRU/SRW.

Anwendung, Referenzen

- Wie verbreitet und international akzeptiert ist das PI-System?

Businessmodell und nachhaltiger Bestand

- Folgekosten (Businessmodell), Nachhaltigkeit des technischen Systems

PI-Beispiele

Nachfolgend werden die gegenwärtig als Persistent Identifier bekannten und publizierten Systeme, Spezifikationen und Standards tabellarisch vorgestellt. Zu Beginn wird das einzelne PI-System optisch hervorgehoben („Kürzel – vollständiger Name"). Die PI-Systeme sind alphabetisch geordnet.

Jede Tabelle beinhaltet die nachfolgenden Elemente:

Kurzbezeichnung	*allgemein verwendete oder bekannte Abkürzung des PI-Systems*
Erläuterung	*kurze, allgemeine inhaltliche Erläuterungen über das Ziel sowie die Funktionalitäten des PI-Systems*
Syntax	*Darstellung der allgemeinen Syntax des PIs Zusätzlich wird der jeweilige PI als URN dargestellt.*
Beispiel	*ein oder mehrere Beispiele für einen PI*
Identifizierung / Registry	*kurze Angaben, was mit dem PI identifiziert wird und ob ein Registry gepflegt wird*

Resolving	*Wird ein Resolving unterstützt, d.h. kann der Identifier in einer klickbaren Form dem Nutzer angeboten werden*
Anwender	*Anwendergruppen, Institutionen, Organisationen, die das PI-System unterstützen, z.T. erfolgt dies in Auswahl*
Tool-Adaption	*Vorhandene Tools, Adaption in Digital Library Tools oder anderen Content Provider Systemen*
Referenz	*Internetquellen, Die Angabe erfolgt in Form von URLs*

ARK - Archival Resource Key

Kurzbezeichnung	*ARK*
Erläuterung	*ARK (Archival Resource Key) ist ein Identifizierungsschema für den dauerhaften Zugriff auf digitale Objekte. Der Identifier kann unterschiedlich verwendet werden: Als Link* *•von einem Objekt zur zuständigen Institution,* *•von einem Objekt zu Metadaten und* *•zu einem Objekt oder dessen adäquater Kopie.*
Syntax	*[http://NMAH/]ark:/NAAN/Name[Qualifier]* *NMAH: Name Mapping Authority Hostport* *ark: ARK-Label* *NAAN: Name Assigning Authority Number* *Name: NAA-assigned* *Qualifier: NMA-supported*
Beispiel	*http://foobar.zaf.org/ark:/12025/654xz321/s3/f8.05v.tiff* *Als URN:* *urn:ark:/12025/654xz321/s3/f8.05v.tiff*
Identifizierung / Registry	*- ARK-Vergabe für alle Objekte* *- zentrales Registry für Namensräume*

Resolving	*Ja, ein zentrales Register der ARK-Resolving-Dienste soll in einer „globalen Datenbank“ erfolgen, die gegenwärtig nicht von einer internationalen Agentur wie z.B. der IANA betreut wird.*
Anwender	*15 angemeldete Institutionen:* *(Eigenauskunft) Darunter:* *California Digital Library,* *LoC,* *National Library of Medicine,* *WIPO,* *University Libraries* *Internet Archive,* *DCC,* *National Library of France*
Tool-Adaption	*Entwicklung der California Digital Library: Noid (Nice Opaque Identifier) Minting and Binding Tool*
Referenz	*http://www.cdlib.org/inside/diglib/ark/*
Bemerkungen	*Allerdings muss bei Kopien der spezif. Resolving-Service angegeben werden.*

DOI – Digital Object Identifier

Kurzbezeichnung	*DOI*
Erläuterung	*Anwendungen von Digital Object Identifiers (DOI) werden seit 1998 durch die International DOI Foundation (IDF) koordiniert. Dem DOI liegt ein System zur Identifizierung und dem Austausch von jeder Entität geistigen Eigentums zugrunde. Gleichzeitig werden mit dem DOI technische und organisatorische Rahmenbedingungen bereitgestellt, die eine Verwaltung digitaler Objekte sowie die Verknüpfung der Produzenten oder Informationsdienstleistern mit den Kunden erlauben. Dadurch wird die Möglichkeit geschaffen, Dienste für elektronische Ressourcen, die eingeschränkt zugänglich sind, auf Basis von DOIs zu entwickeln und zu automatisieren.* *Das DOI-System besteht aus den folgenden drei Komponenten:* *• Metadaten,* *• dem DOI als Persistent Identifier und* *• der technischen Implementation des Handle-Systems.* *Institutionen, die einen Dienst mit einem individuellen Profil aufbauen wollen, können dies in Form von Registration Agencies umsetzen. Das bekannteste Beispiel ist CrossRef, in dem die Metadaten und Speicherorte von Referenzen verwaltet und durch externe Institutionen weiterverarbeitet werden können.* *Die DOI-Foundation ist eine Non-Profit-Organisation, deren Kosten durch Mitgliedsbeiträge, den Verkauf von DOI-Präfixen und den vergebenen DOI-Nummern kompensiert werden.* *Die Struktur von DOIs wurde seit 2001 in Form eines ANSI/NISO-Standards (Z39.84) standardisiert, welche die Komponenten der Handles widerspiegelt:*

Syntax	*Präfix / Suffix*
Beispiel	*10.1045/march99-bunker* *Der Zahlencode "10" bezeichnet die Strings als DOIs, die unmittelbar an den Punkt grenzende Zahlenfolge "1045" steht für die vergebende Institution z.B. eine Registration Agency. Der alphanumerische String im Anschluss an den Schrägstrich identifiziert das Objekt z.B. einen Zeitschriftenartikel.* *Als URN:* *urn:doi:10.1045/march99-bunker*
Identifizierung / Registry	*- DOI-Vergabe für alle Objekte* *- zentrale Registrierung von Diensten,* *- Nutzer müssen sich bei den Serviceagenturen registrieren*
Resolving	*- Ja, Handle-System als technische Basis* *- Zentraler Resolving-Service* *- verschiedene, nicht kommunizierte dezentrale Dienste*
Anwender	*- 7 Registration Agencies (RA)* *Copyright Agency, CrossRef, mEDRA, Nielson BookData, OPOCE, Bowker, TIB Hannover* *- CrossRef-Beteiligte: 338* *CrossRef-Nutzer* *- Bibliotheken (970, auch LoC)* *- Verlage (1528)*
Tool-Adaption	*Tools, welche die Nutzung von DOIs vereinfachen und die Funktionalität erweitern: http://www.doi.org/tools.html* *Digital Library Tools von ExLibris*
Referenz	*http://www.doi.org*
Bemerkungen	*- DOIs sind URN-konform.* *- kostenpflichtiger Service* *- gestaffelte Servicegebühren*

ERRoL - Extensible Repository Resource Locator

Kurzbezeichnung	*ERRoL*
Erläuterung	*Ein ERRoL ist eine URL, die sich nicht ändert und kann Metadaten, Content oder andere Ressourcen eines OAI-Repositories identifizieren.*
Syntax	*„http://errol.oclc.org/“ + oai-identifier*
Beispiel	*http://errol.oclc.org/oai:xmlregistry.oclc.org:demo/ISBN/0521555132.ListERRoLs* *http://errol.oclc.org/oai:xmlregistry.oclc.org:demo/ISBN/0521555132.html* *http://errol.oclc.org/ep.eur.nl/hdl:1765/9*
Identifizierung / Registry	*OAI Registry at UIUC (Grainger Engineering Library Information Center at University of Illinois at Urbana-Champaign) http://gita.grainger.uiuc.edu/registry/ListRepoIds.asp?self=1*
Resolving	*http-Redirect*
Anwender	*Nicht zu ermitteln*
Tool-Adaption	*DSpace*
Referenz	*http://errol.oclc.org/* *http://www.oclc.org/research/projects/oairesolver/*
Bemerkungen	*Erscheint experimentell. Kein echter Persistent Identifier, da URLs aktualisiert werden müssen .*

GRI – Grid Resource Identifier

Kurzbezeichnung	*GRI*
Erläuterung	*Die Spezifikationen definieren GRI für eindeutige, dauerhafte Identifier für verteilte Ressourcen sowie deren Metadaten.*
Syntax	*s. URN-Syntax*
Beispiel	*urn:dais:dataset:b4136aa4-2d11-42bd-aa61-8e8aa5223211* *urn:instruments:telescope:nasa:hubble* *urn:physics:colliders:cern* *urn:lsid:pdb.org:1AFT:1*
Identifizierung / Registry	*s. URN*

Resolving	*Im Rahmen von applikationsabhängigen Diensten wie z.B. Web-Services.*
Anwender	*School of Computing Science, University of Newcastle upon Tyne, Arjuna Technologies http://www.neresc.ac.uk/projects/gaf/*
Tool-Adaption	*http://www.neresc.ac.uk/projects/CoreGRID/*
Referenz	*http://www.neresc.ac.uk/ws-gaf/grid-resource/*
Bemerkungen	*GRI sind URN-konform.*

GRid - Global Release Identifier

Kurzbezeichnung	*GRid*
Erläuterung	*GRid ist ein System, um Releases of Tonaufnahmen für die elektronische Distribution eindeutig zu identifizieren. Das System kann Identifizierungssysteme in der Musikindustrie integrieren. Dazu gehören ein Minimalset an Metadaten, um Rechte (DRM) eindeutig zuordnen zu können.*
Syntax	*A Release Identifier consists of 18 characters, and is alphanumeric, using the Arabic numerals 0 to 9 and letters of the Roman alphabet (with the exception of I and O). It is divided into its five elements in the following order:* • *Identifier Scheme* • *Issuer Code* • *IP Bundle Number* • *Check Digit*

Beispiel	*A1-2425G-ABC1234002-M* *A1 - Identifier Scheme (i.e. Release Identifier for the recording industry)* *2425G - Issuer Code – (for example ABC Records)* *ABC1234002 - IP Bundle Number (for example an electronic release composed of a sound and music video recording, screensaver, biography and another associated video asset)* *M - Check Digit*
Identifizierung / Registry	*RITCO, an associated company of IFPI Secretariat, has been appointed as the Registration Agency.*
Resolving	*Resource Discovery Service*
Anwender	*Unklar*
Tool-Adaption	*unklar*
Referenz	*ISO 7064: 1983, Data Processing – Check Character Systems* *ISO 646: 1991, Information Technology – ISO 7-bit Coded Character Set for Information Exchange.*
Bemerkungen	*Kostenpflichtige Registrierung (150 GBP) für einen Issuer Code für 1 Jahr.*

GUUID / UUID

Kurzbezeichnung	*GUUID / UUID*
Erläuterung	*GUIDs (Globally Unique IDentifier) sind unter der Bezeichnung „UUID“ als URN-Namespace bereits bei der IANA registriert. Aufgrund des Bekanntheitsgrades werden diese erwähnt.* *Ein UUID (Universal Unique Identifier) ist eine 128-bit Nummer zur eindeutigen Identifizierung von Objekten oder anderen Entities im Internet.* *UUIDs wurden ursprünglich in dem Apollo Computer-Netzwerk, später im Rahmen der Open Software Foundation's (OSF), Distributed Computing Environment (DCE) und anschließend innerhalb der Microsoft Windows Platforms verwendet.*
Syntax	*s. URN-Syntax*
Beispiel	*urn:aps:node:0fe46720-7d30-11da-a72b-0800200c9a66*
Identifizierung / Registry	*URN-Namespace-Registry*
Resolving	*Kein*
Anwender	*Softwareprojekte*
Tool-Adaption	*UUID-Generatoren: http://kruithof.xs4all.nl/uuid/uuidgen* *http://www.uuidgenerator.com/* *http://sporkmonger.com/*
Referenz	*http://www.ietf.org/rfc/rfc4122.txt*
Bemerkungen	*In der Spezifikation wird ein Algorithmus zur Generierung von UUIDs beschrieben. Wichtig ist der Ansatz, dass weltweit eindeutige Identifiers ohne (zentrale) Registrierung generiert und in unterschiedlichen Applikationen sowie verschiedene Objekttypen verwendet werden können. Wobei deutlich gesagt wird, dass UUIDs *nicht* auflösbar sind.*

Handle

Kurzbezeichnung	*Handle*
Erläuterung	*Das Handle-System ist die technische Grundlage für DOI-Anwendungen. Es ist eine technische Entwicklung der Corporation for National Research Initiatives. Mit dem Handle-System werden Funktionen, welche die Vergabe, Administration und Auflösung von PIs in Form von Handles erlauben, bereitgestellt. Die technische Basis bildet ein Protokoll-Set mit Referenz-Implementationen wie z.B. DOI, LoC.*
Syntax	*Handle ::= Handle Naming Authority "/" Handle Local Name* *Das Präfix ist ein numerischer Code, der die Institution bezeichnet. Das Suffix kann sich aus einer beliebigen Zeichenkette zusammensetzen.*
Beispiel	*Als URN: urn:handle:10.1045/january99-bearman*
Identifizierung / Registry	*Zentrales Handle-Registry für die Präfixe.*
Resolving	*Handle-Service*
Anwender	*DOI-Anwender, LoC, DSpace-Anwender*
Tool-Adaption	*DSpace*
Referenz	*http://www.handle.net*
Bemerkungen	*Handles sind URN-konform.*

InfoURI

Kurzbezeichnung	*InfoURI*
Erläuterung	*InfoURI ist ein Identifier für Ressourcen, die über kein Äquivalent innerhalb des URI-Raumes verfügen wie z.B. LCCN. Sie sind nur für die Identifizierung gedacht, nicht für die Auflösung. Es ist ein NISO-Standard.*
Syntax	*„info:“ namespace „/“ identifier [„#“ fragment]* *info-scheme = “info”* *info-identifier = namespace “/” identifier* *namespace = scheme* *identifier = path-segments*
Beispiel	*info:lccn/n78089035* *Als URN:* *urn:info:lccn/n78089035*
Identifizierung / Registry	*Zentrales Registry für Namespaces*
Resolving	*nein*
Anwender	*18 Anwender:* *LoC, OCLC, DOI etc.*
Tool-Adaption	*Entwicklung für die Adaption von OpenURL-Services*
Referenz	*http://info-uri.info/*
Bemerkungen	*Zusammenarbeit mit OpenURL.*

NLA - Australische Nationalbibliothek

Kurzbezeichnung	*Keine vorhanden, aber die Identifier beginnen mit NLA*
Erläuterung	
Syntax	*Abhängig von den einzelnen Typen elektronischen Materiales werden die Identifier nach verschiedenen Algorithmen gebildet.* *Beispiel* *Collection Identifier* *nla.pic, nla.ms, nla.map, nla.gen, nla.mus, nla.aus, nla.arc*
Beispiel	*Manuscript Material* *collection id-collection no.-series no.-item no.-sequence no.- role code-generation code* *nla.ms-ms8822-001-0001-001-m*
Identifizierung / Registry	*Objekte, die archiviert werden. Es existiert ein lokales Registry.*
Resolving	*Ja, für die lokalen Identifier*
Anwender	*ANL, Zweigstellen, Kooperationspartner*
Tool-Adaption	
Referenz	*http://www.nla.gov.au/initiatives/persistence.html*
Bemerkungen	*Dies ist eine Eigenentwicklung. Es werden keine internationalen Standards berücksichtigt.*

LSID - Life Science Identifier

Kurzbezeichnung	*LSID*
Erläuterung	*Die OMG (Object Management Group) spezifiziert LSID als Standard für ein Benennungsschema für biologische Entitäten innerhalb der "Life Science Domains" und die Notwendigkeit eines Resolving-Dienstes, der spezifiziert, wie auf die Entitäten zugegriffen werden kann.*

Syntax	*The LSID declaration consists of the following parts, separated by double colons:* • *"URN"* • *"LSID"* • *authority identification* • *namespace identification* • *object identification* • *optionally: revision identification.* *If revision field is omitted then the trailing colon is also omitted.*
Beispiel	*URN:LSID:ebi.ac.uk:SWISS-PROT. accession:P34355:3 URN:LSID:rcsb. org:PDB:1D4X:22 URN:LSID:ncbi.nlm.nih. gov:GenBank.accession:NT_001063:2*
Identifizierung / Registry	*s. URN*
Resolving	*DDDS/DNS, Web-Service*
Anwender	*undurchsichtig*
Tool-Adaption	
Referenz	*http://www.omg.org/docs/dtc/04-05-01.pdf* • *"OMG Life Sciences Identifiers Specification." - Main reference page.* • *Interoperable Informatics Infrastructure Consortium (I3C)* • *Life Sciences Identifiers. An OMG Final Adopted Specification which has been approved by the OMG board and technical plenaries. Document Reference: dtc/04-05-01. 40 pages.* • *LSID Resolution Protocol Project. Info from IBM.* • *"Identity and Interoperability in Bioinformatics." By Tim Clark (I3C Editorial Board Member). In Briefings in Bioinformatics (March 2003).* *"Build an LSID authority on Linux." By Stefan Atev (IBM)*
Bemerkungen	

POI - PURL-Based Object Identifier

Kurzbezeichnung	*POI*
Erläuterung	*POI ist eine einfache Spezifikation als Resource-Identifier auf Grundlage des PURL-Systems und ist als „oai-identifier“ für das OAI-PMH entwickelt worden. POIs dienen als Identifier für Ressourcen, die in den Metadaten von OAI-konformen Repositories beschrieben sind. POIs können auch explizit für Ressourcen verwendet werden.*
Syntax	*“http://purl.org/poi/”namespace-identifier “/” local-identifier* *namespace-identifier = domainname-word “.” domainname* *domainname = domainname-word [“.”domainname]* *domainname-word = alpha *(alphanum | “-”)* *local-identifier = 1*uric*
Beispiel	*http://www.ukoln.ac.uk/distributed-systems/poi/*
Identifizierung / Registry	*kein*
Resolving	*Ja, wenn dieser über das OAI-Repository bereitgestellt wird, wobei der PURL-Resolver empfohlen wird.*
Anwender	*unklar*
Tool-Adaption	*POI-Lookup-Tools* *http://www.rdn.ac.uk/poi/*
Referenz	*POI Resolver Guidelines* *http://www.ukoln.ac.uk/distributed-systems/poi/resolver-guidelines/* *“The PURL-based Object Identifier (POI).” By Andy Powell (UKOLN, University of Bath), Jeff Young (OCLC), and Thom Hickey (OCLC). 2003/05/03. http://www.ukoln.ac.uk/distributed-systems/poi/*
Bemerkungen	

PURL – Persistent URL

Kurzbezeichnung	*PURL*
Erläuterung	*PURL (Persistent URL) wurde vom „Online Computer Library Center“ (OCLC) 1995 im Rahmen des „Internet Cataloging Projects“, das durch das U.S. Department of Education finanziert wurde, eingeführt, um die Adressdarstellung für die Katalogisierung von Internetressourcen zu verbessern. PURLs sind keine Persistent-Identifier, können jedoch in bestehende Standards wie URN überführt werden. Technisch betrachtet wird bei PURL der existierende Internet-Standard „HTTP-redirect“ angewendet, um PURLs in die URLs aufzulösen.*
Syntax	*http://purl.oclc.org/OCLC/PURL/FAQ* *- protocol* *- resolver address* *- name*
Beispiel	*http://purl.oclc.org/keith/home* *Als URN:* *urn:/org/oclc/purl/keith/home*
Identifizierung / Registry	*Kein Registry*
Resolving	*ja, jedoch wird nur ein lolaker Resolver installiert.*
Anwender	*Keine Auskunft möglich (lt. Stuart Weibel)* *- OCLC* *- United States Government Printing Office (GPO)* *- LoC*
Tool-Adaption	*PURL-Software*
Referenz	*http://purl.org*

Bemerkungen	*- kein zentrales Registry* *- Die genaue Anzahl von vergebenen PURLs ist unbekannt. ??* *- Ein Test der DOI-Foundation ergab, dass nur 57% der getesteten PURLs auflösbar waren.* *- Experimentell von OCLC eingeführt.* *- Es ist keine Weiterentwicklung vorgesehen.*

URN – Uniform Resource Name

Kurzbezeichnung	*URN*
Erläuterung	*Der Uniform Resource Name (URN) existiert seit 1992 und ist ein Standard zur Adressierung von Objekten, für die eine institutionelle Verpflichtung zur persistenten, standortunabhängigen Identifizierung der Ressourcen besteht. URNs wurden mit dem Ziel konzipiert, die Kosten für die Bereitstellung von Gateways sowie die Nutzung von URNs so gering wie möglich zu halten - vergleichbar mit existierenden Namensräumen wie z.B. URLs. Aus diesem Grund wurde in Standards festgelegt, wie bereits existierende oder angewendete Namensräume bzw. Nummernsysteme einfach in das URN-Schema sowie die gängigen Protokolle wie z.B. HTTP (Hypertext Transfer Protocol) oder Schemas wie z.B. URLs integriert werden können.* *Der URN als Standard wird von der Internet Engineering Task Force (IETF) kontrolliert, die organisatorisch in die Internet Assigned Numbering Authority (IANA) eingegliedert ist. Sie ist für die Erarbeitung und Veröffentlichung der entsprechenden Standards in Form von "Request for Comments" (RFCs) zuständig. Diese umfassen die folgenden Bereiche:* • *URN-Syntax (RFC 2141),* • *funktionale Anforderungen an URNs (RFC 1737),* • *Registrierung von URN-Namensräumen (z.B. RFCs 3406, 2288, 3187, NBN: 3188),* • *URN-Auflösungsverfahren (RFCs 3401, 3402, 3403, 3404).*

Syntax	*URN:NID:NISS* *URNs bestehen aus mehreren hierarchisch aufgebauten Teilbereichen. Dazu zählen der Namensraum (Namespace, NID), der sich aus mehreren untergeordneten Unternamensräumen (Subnamespaces, SNID) zusammensetzen kann, sowie der Namensraumbezeichner (Namespace Specific String, NISS).*
Beispiel	*urn:nbn:de:bsz:93-opus-59* *Als URL / URI:* *http://nbn-resolving.de/* *urn:nbn:de:bsz:93-opus-59* *Als OpenURL:* *http.//[openURL-service]?identifier=* *urn:nbn:de:bsz:93-opus-59* *Als InfoURI:* *info:urn/urn:nbn:de:bsz:93-opus-59* *Als ARK:* *http://[NMAH]ark:/NAAM/* *urn:nbn:de:bsz:93-opus-59* *Als DOI:* *10.1111/urn:nbn:de:bsz:93-opus-59*
Identifizierung / Registry	*Überblick über den Status registrierter URN-Namensräume (unvollständig)* *http://www.iana.org/assignments/* *urn-namespaces/*
Resolving	*Es gibt mehrere Möglichkeiten:* *- http-Redirect (Umleitung der URN zur URL)* *- DNS (Domain Name System)*

Anwender	*CLEI Code* *IETF* *IPTC* *ISAN* *ISBN* *ISSN* *NewsML* *OASIS* *OMA* *Resources* *XML.org* *Web3D* *MACE* *MPEG* *Universal Content Identifier* *TV-Anytime Forum* *Federated Content* *Government (NZ)* *Empfehlung: OAI 2.0: oai-identifier als URNs verwenden* *<u>NBN:</u>* *Finnland,* *Niederlande,* *Norwegen,* *Österreich,* *Portugal,* *Slovenien,* *Schweden,* *Schweiz,* *Tschechien,* *Ungarn,* *UK*
Tool-Adaption	*OPUS, DigiTool (ExLibris), Miless*

Referenzen	*Internetstandards:* *http://www.ietf.org/rfc/rfc1737.txt* *http://www.ietf.org/rfc/rfc2141.txt* *http://www.ietf.org/rfc/rfc3406.txt* *http://www.ietf.org/rfc/rfc288.txt* *http://www.ietf.org/rfc/rfc3187.txt* *http://www.ietf.org/rfc/rfc3188.txt* *http://www.ietf.org/rfc/rfc3401.txt* *http://www.ietf.org/rfc/rfc3402.txt* *http://www.ietf.org/rfc/rfc3403.txt* *http://www.ietf.org/rfc/rfc3404.txt* *URN-Prüfziffer Der Deutschen Bibliothek:* *http://www.pruefziffernberechnung.de/U/URN.shtml*
Bemerkungen	*Innerhalb der URNs sind sowohl die Integration bereits bestehender Nummernsysteme (z.B. ISBN) als auch institutionsgebundene Nummernsysteme auf regionaler oder internationaler Ebene als Namensräume möglich. Dazu zählt auch die „National Bibliography Number“ (NBN, RFC 3188), ein international verwalteter Namensraum der Nationalbibliotheken, an dem Die Deutsche Bibliothek beteiligt ist.*

XRI - Extensible Resource Identifier

Kurzbezeichnung	*XRI*
Erläuterung	*XRI wurde vom TC OASIS entwickelt. XRI erweitert die generische URI-Syntax, um "extensible, location-, application-, and transport-independent identification that provides addressability not just of resources, but also of their attributes and versions." zu gewährleisten. Segmente oder Ressourcen können persistent identifiziert und/oder zu adressiert werden. Die Persistenz des Identifiers wird mit den Zielen der URNs gleichgestellt.*
Syntax	*xri: authority / path ? query # fragment*
Beispiel	*xri://@example.org*agency*department/ docs/govdoc.pdf* *XRI mit URN: xri://@example.bookstore/!(ur n:ISBN:0-395-36341-1)*
Identifizierung / Registry	*nein*
Resolving	*OpenXRI.org server*
Anwender	*12 Förderer http://www.openxri.org/*
Tool-Adaption	

Referenz	*http://www.openxri.org/* • *"OASIS Releases Extensible Resource Identifier (XRI) Specification for Review." News story 2005-04-07.* • *XRI Generic Syntax and Resolution Specification 1.0 Approved Committee Draft. PDF source posted by Drummond Reed (Cordance), Tuesday, 20 January 2004, 03:00pm.* • *XRI Requirements and Glossary Version 1.0. 12-June-2003. 28 pages. [source .DOC, cache]* • *OASIS Extensible Resource Identifier TC web site* • *XRI TC Charter* • *"OASIS TC Promotes Extensible Resource Identifier (XRI) Specification." News story 2004-01-19. See also "OASIS Members Form XRI Data Interchange (XDI) Technical Committee."*
Bemerkungen	

Referenzen

Beschreibung	Referenz
Überblicksdarstellung von PI-Systemen des EPICUR-Projektes	http://www.persistent-identifier.de/?link=204
PADI – Preserving Access to Digital Information	http://www.nla.gov.au/padi/topics/36.html
nestor-Informationsdatenbank, Themenschwerpunkt: Persistente Identifikatoren	http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?show=21
ERPANET Workshop „Persistent Identifier“, 2004	http://www.erpanet.org/events/2004/cork/index.php

9.4.1 Der Uniform Resource Name (URN)

Christa Schöning-Walter

Damit digitale Objekte auf Dauer zitierfähig sind, müssen stabile Referenzen vorhanden sein, die auch über technische und organisatorische Veränderungen hinweg eine verlässliche Adressierung ermöglichen. Grundlegend für den dauerhaften Zugang ist die Langzeitverfügbarkeit der digitalen Objekte an sich. Die Speicherung in vertrauenswürdigen Archiven ist dafür eine unabdingbare Voraussetzung. Persistent Identifier *haben in diesem Kontext die zentrale Funktion, die Objekte digitaler Sammlungen langfristig und weltweit eindeutig zu identifizieren.*

Sammlung und Langzeitarchivierung von Netzpublikationen bei der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek (DNB) hat den Auftrag, das kulturelle und wissenschaftliche Erbe Deutschlands in seiner seit 1913 veröffentlichten Form zu sammeln, dauerhaft zu bewahren und für die Nutzung zugänglich zu machen. Seit dem Inkrafttreten des *Gesetzes über die Deutsche Nationalbibliothek* vom 22. Juni 2006 gehören auch Netzpublikationen zum Sammelauftrag.[19] Als Netzpublikationen gelten jegliche Darstellungen in Schrift, Bild oder Ton, die in öffentlichen Netzen bereitgestellt werden. Elektronische Zeitschriften, E-Books, Hochschulprüfungsarbeiten, Forschungsberichte, Kongressschriften und Lehrmaterialien gehören ebenso dazu wie Digitalisate alter Drucke, Musikdateien oder Webseiten. Am 17. Oktober 2008 ist zudem die Pflichtablieferungsverordnung neu gefasst worden. Sie konkretisiert den gesetzlichen Sammelauftrag der DNB.

Um die Benutzbarkeit ihrer digitalen Sammlungen auch in Zukunft gewährleisten zu können, engagiert sich die DNB auf dem Gebiet der Langzeitarchivierung. Grundlagen für die Langzeiterhaltung und Langzeitbenutzbarkeit digitaler Objekte sind in dem vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekt „kopal – Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen“ entwickelt worden (*http://kopal.langzeitarchivierung.de*). Fortlaufende Anpassungen der Datenbestände an den Stand der Technik (Migrationen, Emulationen) sollen dafür sorgen, dass digitale Samm-

19 http://www.d-nb.de/netzpub/index.htm

lungen trotz Weiterentwicklungen bei den Hard- und Softwaresystemen verfügbar bleiben.

Im Zuge der Langzeitarchivierung muss neben dem Erhalt der digitalen Daten an sich und ihrer Interpretierbarkeit auch die Identifizierbarkeit der Objekte sichergestellt werden. Eindeutige Bezeichner, die über den gesamten Lebenszyklus hinweg mit den Netzpublikationen und ihren Metadaten verbunden bleiben, ermöglichen es, die Objekte in digitalen Sammlungen persistent (dauerhaft) zu identifizieren – d.h. sie dem Nutzer auch über Systemgrenzen und Systemwechsel hinweg verlässlich zur Verfügung zu stellen.

Die DNB verwendet als Schema für die Identifizierung digitaler Ressourcen den *Uniform Resource Name* (URN). Das Konzept zur Langzeitarchivierung beinhaltet, dass alle Netzpublikationen, die bei der DNB gesammelt, erschlossen und archiviert werden, zwingend einen Persistent Identifier benötigen.[20] Die Zuordnung erfolgt spätestens bei der Erschließung einer Netzpublikation durch die DNB. Der Service zur Registrierung und Auflösung von URNs ist ein kooperativer Dienst, der von den verlegenden Institutionen mit genutzt werden kann. Idealerweise erfolgt die URN-Vergabe schon im Zuge der Veröffentlichung einer Netzpublikation. Dann ist es möglich, alle Speicherorte von vornherein mit in das System aufzunehmen.

Das URN-Schema

Die Wurzeln des URN reichen zurück bis in die frühen 1990er Jahre. Das Funktionsschema gehört zu den Basiskonzepten, die im Zusammenhang mit dem Entwurf einer Architektur für das *World Wide Web* (WWW) entstanden sind. Der URN ist ein *Uniform Resource Identifier* (URI). URIs werden im globalen Informations- und Kommunikationsnetz für die Identifizierung jeglicher zu adressierender physikalischer oder abstrakter Ressourcen benutzt (z.B. für den Zugriff auf Objekte, den Aufruf von Webservices, die Zustellung von Nachrichten etc.).[21] Das *World Wide Web Consortium* (W3C) betont die besondere Bedeutung dieser Technologie:

> *The Web is an information space. [...] URIs are the points in that space. Unlike web data formats [...] and web protocols [...] there is only one Web naming/addressing technology: URIs.*[22]

---

20 Schöning-Walter (2002)

21 Tim Berners-Lee (2005): Uniform Resource Identifier (URI) – Generic Syntax. http://www.ietf.org/rfc/rfc3986.txt

22 W3C: Web Naming and Addressing Overview. http://www.w3.org/Addressing/.

Das URN-Schema hat den Status eines durch die *Internet Assigned Numbers Authority* (IANA) legitimierten de facto-Internetstandards.[23] Beschrieben ist das Schema in verschiedenen so genannten *Requests for Comments* (RFCs), wie sie üblicherweise von den Arbeitsgruppen der *Internet Engineering Task Force* (IETF) veröffentlicht werden.

URNs sind als weltweit gültige, eindeutige Namen für Informationsressourcen im WWW konzipiert worden. Die Entwicklung zielte darauf, Unabhängigkeit vom Ort der Speicherung und vom Zugriffsprotokoll zu erreichen (*RFC 2141*, URN Syntax).[24]

> *Uniform Resource Names (URNs) are intended to serve as persistent, location-independent resource identifiers and are designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space. Therefore, the URN syntax provides a means to encode character data in a form that can be sent in existing protocols, transcribed on most keyboards, etc.*

Das Schema fächert sich in sogenannte Namensräume auf, die ebenfalls bei IANA registriert werden müssen.[25] Mit der Verzeichnung eines Namensraums werden sowohl der Geltungsbereich (welche Art von digitalen Objekten soll identifiziert werden) als auch spezifische Regeln festgelegt (*RFC 3406*, URN Namespace Definition Mechanisms).[26] Zu den bisher bei IANA angemeldeten URN-Namensräumen gehören u.a. (Stand: 15. Februar 2009):

- urn:issn – International Serials Number (*RFC 3044*),
- urn:isbn – International Standards Books Number (*RFC 3187*),
- urn:nbn – National Bibliography Number (*RFC 3188*),
- urn:uuid – Universally Unique Identifiers (*RFC 4122;* für verteilte Softwaresysteme),
- urn:isan – International Standard Audiovisual Number (*RFC 4246*).

RFC 1737 (Functional Requirements for URNs) beschreibt die Anforderungen, die grundsätzlich jeder URN-Namensraum erfüllen muss.[27] Das sind:

---

23 http://www.ietf.org/rfc.html

24 Ryan Moats (1997): URN Syntax. http://www.ietf.org/rfc/rfc2141.txt

25 Die bei IANA registrierten URN-Namensräume sind verzeichnet unter http//www.iana.org/assignments/urn-namespaces.

26 Leslie L. Daigle et. al. (2002): URN Namespace Definition Mechanisms. http://www.ietf.org/rfc/rfc3406.txt.

27 Larry Mainter, Karen Sollins (1994): Functional Requirements for Uniform Resource Names. http://www.ietf.org/rfc/rfc1737.txt.

- *Global Scope*: Gültigkeit der Namen weltweit,
- *Global Uniqueness*: Eindeutigkeit der Namen weltweit,
- *Persistence*: Gültigkeit der Namen auf Dauer,
- *Scalability*: das Namensschema muss beliebig viele Objekte bezeichnen können,
- *Legacy Support*: schon vorhandene Bezeichnungs- und Nummerierungssysteme, Zugriffsschemata oder Protokolle müssen eingebettet werden können, sofern sie regelkonform sind,
- *Extensibility*: das Namensschema muss bei Bedarf weiterentwickelt werden können,
- *Indepencence*: die beteiligten Institutionen selbst legen die Regeln für das Namensschema fest,
- *Resolution*: die Auflösung von URNs in Zugriffsadressen erfolgt über einen Resolvingdienst.

URNs verweisen nicht selbst auf die Informationsressourcen. Ein zwischengeschalteter Resolvingmechanismus führt die Auflösung durch (*RFC 2276*, Architectural Principles of URN Resolution).[28] In der Regel werden URNs über ein Register in URLs umgewandelt. Dieses Prinzip ermöglicht es, den Aufwand für die Pflege von Zugriffsadressen relativ gering zu halten.

Die National Bibliography Number

Die *National Bibliography Number* (NBN) ist ein registrierter URN-Namensraum mit maßgeblicher Bedeutung für den Bibliotheks- und Archivbereich.[29] Das Konzept beruht auf einer Initiative der *Conference of European National Librarians* (CENL) und wurde im Jahr 2001 unter Federführung der Finnischen Nationalbibliothek entwickelt, um digitale Publikationen in den Nationalbibliografien verzeichnen zu können (*RFC 3188*, Using National Bibliography Numbers as URNs).[30]

Die NBN ist international gültig. Auf nationaler Ebene sind üblicherweise die Nationalbibliotheken für den Namensraum verantwortlich. In Deutsch-

28 Karen Sollins (1998): Architectural Principles of Uniform Resource Name Resolution. http://www.ietf.org/rfc/rfc2276.txt.

29 Hans-Werner Hilse, Jochen Kothe (2006): Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. London: Consortium of European Research Libraries. Amsterdam: European Commission on Preservation and Access. urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

30 Juha Hakala (1998): Using National Bibliography Numbers as Uniform Resource Names. http://www.ietf.org/rfc/rfc3188.txt.

land hat die DNB die Koordinierungsfunktion übernommen. In den Jahren 2002 bis 2005 wurde im Rahmen des BMBF-Projekts „EPICUR – Enhancements of Persistent Identifier Services" eine Infrastruktur aufgebaut (*http://www.persistent-identifier.de*).[31] Die DNB betreibt seither einen URN-Resolver für Deutschland, Österreich und die Schweiz. Beim Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) existiert dazu ein Spiegelsystem, das bei Bedarf den Resolvingservice übernehmen kann.

Die Entwicklungen in EPICUR waren vorrangig auf die Verzeichnung von Hochschulschriften ausgerichtet. In Zusammenarbeit mit Hochschulbibliotheken, Verbundzentralen und Forschungseinrichtungen wurden URN-Registrierungsverfahren entwickelt, die mittlerweile stark vereinheitlicht und in der Praxis recht gut etabliert sind. Die Regeln zur Vergabe von URNs für Hochschulschriften sind in der *URN-Strategie der Deutschen Nationalbibliothek* beschrieben.[32] Auch die Schweizerische Nationalbibliothek (NB) nimmt seit mehreren Jahren aktiv am URN-Verfahren teil. Seit August 2008 gibt es ein Handbuch zur Anwendung der NBN in der Schweiz.[33]

Die DNB hat die Verwendung des NBN-Schemas mittlerweile auf alle Netzpublikationen ausgedehnt, die im Rahmen ihres erweiterten Sammelauftrags erschlossen werden. Dies hat zur Folge, dass auch fortlaufende Sammelwerke (z.B. Zeitschriften oder Schriftenreihen), granulare Erscheinungsformen (z.B. die Beiträge in einer Zeitschrift oder die Einzelseiten eines Digitalisats) sowie dynamische Publikationsformen (z.B. veränderliche Webseiten) schrittweise mit in die URN-Strategie eingebunden werden müssen. Jedes digitale Objekt, das einzeln identifizierbar und adressierbar sein soll, muss als inhaltlich stabile, eigenständige Einheit in das Langzeitarchiv aufgenommen werden und benötigt für den Zugang einen eigenen Persistent Identifier.

Die Erschließung elektronischer Zeitschriften bei der DNB wird gegenwärtig dahingehend neu ausgerichtet, in der Zukunft auch Fachartikel in der Nationalbibliografie zu verzeichnen und mit einem URN zu kennzeichnen. Im Förderprogramm *Kulturelle Überlieferung* der Deutschen Forschungsgemeinschaft (DFG) wird sogar eine persistente Identifizierung bis auf die Ebene der Einzelseiten digitalisierter Drucke gefordert.[34] Vor diesem Hintergrund erprobt die

---

31 Kathrin Schroeder (2005): EPICUR. In: Dialog mit Bibliotheken. 2005/1. S. 58-61.

32 EPICUR: Uniform Resource Names (URN) – Strategie der Deutschen Nationalbibliothek (2006). http://www.persistent-identifier.de/?link=3352. urn:nbn:de:1111-200606299.

33 e-Helvetica URN-Handbuch. August 2008. Version 1.0. http://www.nb.admin.ch/slb/slb_professionnel/01693/01695/01706/index.html?lang=de.

34 Deutsche Forschungsgemeinschaft: Praxisregeln im Förderprogramm *Kulturelle Überlieferung*. Kap. 5.7. Zitieren, persistente Adressierung. http://www.dfg.de/forschungsfoerderung/

ULB Sachsen-Anhalt in einem Modellprojekt die Vergabe von NBNs für die digitalisierten Drucke aus der Sammlung Ponickau (es handelt sich um ca. 10.000 Drucke mit insgesamt ca. 600.000 Seiten).[35]

Die Struktur der NBN

Das URN-Schema ist streng hierarchisch aufgebaut und gliedert sich in ein Präfix und ein Suffix. Wie andere URI-Schemata auch (z.B. *http:*, *ftp:*, *mailto:*) wird der URN durch seine Bezeichnung gekennzeichnet, gefolgt von einem Doppelpunkt. Der grundsätzliche Aufbau eines URN lautet:

urn:[NID]:[SNID]-[NISS]

Präfix:

- NID *Namespace Identifier (hier: nbn)*
- SNID *Subnamespace Identifier*

Suffix:

- NISS *Namespace Specific String*

Durch Gliederung in Unternamensräume (Subnamespaces) kann die auf internationaler Ebene eingeleitete hierarchische Strukturierung auf nationaler Ebene weiter fortgesetzt werden. Ein zentrales Strukturelement ist das Länderkennzeichen. Ein URN, der mit urn:nbn:de beginnt, drückt aus, dass es sich um eine NBN für eine in Deutschland veröffentlichte Publikation handelt (urn:nbn:ch gilt für die Schweiz, urn:nbn:at für Österreich), die über den URN-Resolver bei der DNB aufgelöst werden kann.

Die Option zur Gliederung der NBN in Unternamensräume wird genutzt, um interessierten Institutionen die Möglichkeit einzuräumen, selbst die Persistent Identifier für ihre Netzpublikationen zu vergeben. Für Deutschland, Österreich und die Schweiz erfolgt die Registrierung von Unternamensräumen bei der DNB. Die Bezeichnung muss eindeutig sein. Als Kennzeichen für Unternamensräume (Subnamespace Identifier) können verwendet werden:

---

formulare/download/12_151.pdf.

35 Dorothea Sommer, Christa Schöning-Walter, Kay Heiligenhaus (2008): URN Granular: Persistente Identifizierung und Adressierung von Einzelseiten digitalisierter Drucke. Ein Projekt der Deutschen Nationalbibliothek und der Universitäts- und Landesbibliothek Sachsen-Anhalt. In: ABI-Technik. Heft 2/2008. S. 106-114.

- das Bibliothekssigel oder die ISIL (*International Standard Identifier for Libraries and Related Organizations*),[36] ggf. kombiniert mit dem Kürzel des Bibliotheksverbundes,
- eine 4-stellige (fortlaufende) Nummer oder
- eine alphanumerische Zeichenkette.

Die Möglichkeiten der Differenzierung eines Namensraums sind vielfältig. Das Präfix in seiner Gesamtheit hat letztlich die Funktion, den Anwendungsbereich zu spezifizieren (Wofür wird der Name benutzt? Wer ist verantwortlich?) und ist definiert als derjenige Teil des URN, der im Resolver als fester Bestandteil des Namens verzeichnet ist. Beispiele zulässiger Namensräume sind:

- urn:nbn:de:101,[37]
- urn:nbn:de:0008,[38]
- urn:nbn:de:0100,[39]
- urn:nbn:de:tu-darmstadt,[40]
- urn:nbn:de:gbv:3,[41]
- urn:nbn:de:gbv:3:1.[42]

Demgegenüber kennzeichnet das Suffix das digitale Objekt an sich (NISS, Namespace Specific String). Die Regeln für die Bildung des Suffix werden von der Institution festgelegt, die den Unternamensraum besitzt. Bereits existierende Nummerierungssysteme können eingebettet werden, wenn sie mit den Konventionen des URN-Schemas übereinstimmen.

Erlaubte Zeichen für die Bildung einer NBN sind alle alphanumerischen Zeichen und zusätzlich einige Sonderzeichen. Nach den in EPICUR festgelegten Regeln ist die letzte Ziffer immer eine automatisch berechnete Prüfziffer.[43] Beispiele zulässiger Namen sind:

36 http://sigel.staatsbibliothek-berlin.de/isil.html
37 Namensraum der DNB
38 Namensraum des nestor-Projekts
39 Namensraum des Suhrkamp-Verlags
40 Namensraum der TU Darmstadt
41 Namensraum der ULB Sachsen-Anhalt
42 Namensraum der ULB Sachsen-Anhalt für die Digitalisate der Sammlung Ponickau
43 EPICUR: Beschreibung des Algorithmus zur Berechnung der Prüfziffer. http://www.persistent-identifier.de/?link=316

- urn:nbn:ch:bel-110142[44]
- urn:nbn:de:bsz:16-opus-88271[45]
- urn:nbn:de:gbv:7-isbn-90-6984-508-3-8[46]
- urn:nbn:de:tib-10.1594/WDCC/EH4106_6H_1CO2IS92A_U306[47]

Registrierung und Bekanntmachung der NBN

Die NBN wird aktiv, sobald der URN und mindestens ein gültiger URL im Resolver verzeichnet sind. Idealerweise ist der URN einer Netzpublikation bereits auflösbar, sobald er erstmals bekannt gemacht wird (z.B. durch Verzeichnung auf einer Webseite oder in einem Online-Katalog).
Für die Übermittlung von URNs und URLs an das Resolvingsystem stehen verschiedene Transferschnittstellen und ein standardisiertes Datenaustauschformat zur Verfügung. Mögliche Registrierungsverfahren sind:

- OAI-Harvesting: Das Verfahren eignet sich besonders für Massendaten. Grundlage bildet das im Projekt EPICUR entwickelte Datenaustauschformat xepicur.[48] Neue Daten werden täglich eingesammelt, teilweise sogar mehrmals täglich.
- Mailverfahren: Die automatische Registrierung kann auch in Form einer elektronischen Nachricht erfolgen, indem eine xepicur-Datei als Dateianhang an den URN-Resolver gesendet wird.
- EPICUR-Frontend: Das Webinterface ermöglicht die manuelle Erfassung einzelner URNs.[49]

Mit Ablieferung einer Netzpublikation bei der DNB wird der URN in die Nationalbibliografie und den Online-Katalog übernommen und anschließend über die Datendienste weiter verbreitet. Die verlegende Stelle ihrerseits kann den URN durch Einbettung in die Publikation oder Verzeichnung auf einer vorgeschalteten Webseite bekannt machen – und damit die Nutzung des Persistent Identifier unterstützen.

---

44 Heft einer Verlagszeitschrift, URN vergeben von der Schweizerischen Nationalbibliothek
45 Dissertation, URN vergeben von der Universität Heidelberg
46 Forschungsbericht, URN vergeben von der SUB Göttingen
47 Forschungsdatensatz des World Data Center for Climate (WDCC), URN vergeben von der DOI-Registrierungsagentur bei der TIB Hannover
48 xepicur - XML-Datentransferformat zur Verwaltung von Persistent Identifiers. http://www.persistent-identifier.de/?link=210.
49 Persistent Identifier – Management der Deutschen Nationalbibliothek. https://ssl.nbn-resolving.de/frontend/

Für sammelpflichtige Netzpublikationen erfüllt die DNB die Aufgabe der Langzeitarchivierung und ergänzt im URN-Resolver die Archivadresse, sobald die Publikation in ihren Geschäftsgang gelangt. Noch nicht registrierte URNs werden bei der Erschließung nachträglich verzeichnet. Besitzt eine Netzpublikation keinen URN, dann übernimmt die DNB die Zuordnung eines eindeutigen Namens, um die Identifizierbarkeit aller digitalen Objekte im Langzeitarchiv gewährleisten zu können.

Auflösung der NBN

Der URN-Resolver sorgt für die Auflösung der Namen in Zugriffsadressen. Ein URN muss dafür mit der Basisadresse des Resolvers (*http://nbn-resolving.de*) verknüpft werden:

- http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8 oder
- http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

Diese Funktionalität ist in der Regel direkt eingebettet in die Systeme, die den URN entweder anzeigen oder ihn für den Zugriff auf ein digitales Objekt benutzen (z.B. Dokumentenserver, Online-Kataloge etc.). Für einen individuellen Zugriff kann die EPICUR-Webseite benutzt werden.[50] Die handelsüblichen Browser unterstützen die Auflösung des URN bisher nicht.

Abhängig von den benutzten Parametern wird entweder der direkte Zugang zum digitalen Objekt hergestellt oder der Resolver gibt die Liste aller registrierten URLs zurück. Bei Vorhandensein mehrerer URLs existiert ein Standardverhalten des Resolvers. Vorrangig wird der URL mit der höchsten Priorität aufgelöst. Falls dieser URL nicht erreichbar ist, wird der URL mit der nächsten Priorität benutzt. Durch Pflege des URN (Aktualisierung der Zugriffsadressen bei Veränderungen) lässt sich erreichen, dass die Verknüpfung mit dem Original erhalten bleibt, bis die Netzpublikation vor Ort (auf dem Hochschulschriftenserver, auf dem Verlagsserver, im Institutional Repository etc.) nicht mehr verfügbar ist. Die Archivversion bei der DNB erlangt erst dann Bedeutung für den Zugang, wenn andere Zugriffsmöglichkeiten nicht mehr funktionieren.

50 http://www.persistent-identifier.de/?link=610

Abb.: URN-Auflösung über die EPICUR-Webseite

Zusammenfassung

Durch Nutzung von Persistent Identifiern lassen sich die Nachteile einer standortbezogenen Identifizierung und Adressierung digitaler Objekte weitgehend überwinden. Während die genaue Speicheradresse (URL) meistens nicht auf Dauer benutzbar ist, behalten URN-basierte Referenzen in Online-Katalogen, Bibliografien, Portalen oder Publikationen auch dann ihre Gültigkeit, wenn sich der Speicherort einer Netzpublikation verändert. Das hat auch Vorteile für die Zitierfähigkeit digitaler Quellen in der Praxis des wissenschaftlichen Arbeitens.
Persistenz ist keine Eigenschaft der URNs an sich. Sie erfordert abgestimmte Regeln sowie eine Pflege der Daten im Resolvingsystem. Die Infrastruktur muss in der Lage sein, die URNs solange nachzuweisen und aufzulösen, wie die Netzpublikationen selbst oder Referenzen darauf irgendwo existieren. Sind diese Voraussetzungen erfüllt, dann kann im Zusammenspiel mit der Langzeitarchivierung auch für die Objekte digitaler Sammlungen über lange Zeiträume hinweg der Zugang gewährleistet werden.

Quellenangaben

Berners-Lee, Tim (2005): *Uniform Resource Identifier (URI) – Generic Syntax*. http://www.ietf.org/rfc/rfc3986.txt

Daigle, Leslie L et. al. (2002): *URN Namespace Definition Mechanisms*. http://www.ietf.org/rfc/rfc3406.txt

e-Helvetica *URN-Handbuch*. (2008). Version 1.0. http://www.nb.admin.ch/slb/slb_professionnel/01693/01695/01706/index.html?lang=de

EPICUR: *Uniform Resource Names (URN) – Strategie der Deutschen Nationalbibliothek* (2006). http://www.persistent-identifier.de/?link=3352. urn:nbn:de:1111-200606299.

Hakala, Juha (1998): *Using National Bibliography Numbers as Uniform Resource Names*. http://www.ietf.org/rfc/rfc3188.txt.

Hilse, Hans-Werner; Kothe, Jochen (2006): *Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. London: Consortium of European Research Libraries*. Amsterdam: European Commission on Preservation and Access. urn:nbn:de:gbv:7-isbn-90-6984-508-3-8.

Mainter, Larry ; Sollins, Karen (1994): *Functional Requirements for Uniform Resource Names*. http://www.ietf.org/rfc/rfc1737.txt

Moats, Ryan (1997): URN Syntax. http://www.ietf.org/rfc/rfc2141.txt

Schöning-Walter; Christa (2008): *Persistent Identifier für Netzpublikationen*. In: Dialog mit Bibliotheken. 2008/1. S. 32-38.

Schroeder, Kathrin (2005): *EPICUR*. In: Dialog mit Bibliotheken. 2005/1. S. 58-61.

Sollins, Karen (1998): *Architectural Principles of Uniform Resource Name Resolution*. http://www.ietf.org/rfc/rfc2276.txt

Sommer, Dorothea; Schöning-Walter, Christa; Heiligenhaus, Kay (2008): *URN Granular*: Persistente Identifizierung und Adressierung von Einzelseiten digitalisierter Drucke. Ein Projekt der Deutschen Nationalbibliothek und der Universitäts- und Landesbibliothek Sachsen-Anhalt. In: ABI-Technik. Heft 2/2008. S. 106-114.

9.4.2 Der Digital Objekt Identifier (DOI)

Jan Brase

Der Digital Object Identifier (DOI)

Der Digital Object Identifier (DOI) wurde 1997 eingeführt, um Einheiten geistigen Eigentums in einer interoperativen digitalen Umgebung eindeutig zu identifizieren, zu beschreiben und zu verwalten. Verwaltet wird das DOI-System durch die 1998 gegründete International DOI Foundation (IDF).[51]

Der DOI-Name ist ein dauerhafter persistenter Identifier, der zur Zitierung und Verlinkung von elektronischen Ressourcen (Texte, aber Primärdaten oder andere Inhalte) verwendet wird. Über den DOI-Namen sind einer Ressource aktuelle und strukturierte Metadaten zugeordnet.

Ein DOI-Name unterscheidet sich von anderen, gewöhnlich im Internet verwendeten Verweissystemen wie der URL, weil er dauerhaft mit der Ressource als Entität verknüpft ist und nicht lediglich mit dem Ort, an dem die Ressource platziert ist.

Der DOI-Name identifiziert eine Entität direkt und unmittelbar, also nicht eine Eigenschaft des Objekts (eine Adresse ist lediglich eine Eigenschaft des Objekts, die verändert werden und dann ggf. nicht mehr zur Identifikation des Objekts herangezogen werden kann).

Das IDF-System besteht aus der „International DOI Foundation“ selbst, der eine Reihe von Registrierungsagenturen („Registration Agencies“; RA) zugeordnet sind. Für die Aufgaben einer RA können sich beliebige kommerzielle oder nicht kommerzielle Organisationen bewerben, die ein definiertes Interesse einer Gemeinschaft vorweisen können, digitale Objekte zu referenzieren.

Technik

Das DOI-System baut technisch auf dem Handle-System auf. Das Handle System wurde seit 1994 von der US-amerikanischen Corporation for National Research Initiatives (CNRI)[52] als verteiltes System für den Informationsaustausch entwickelt. Handles setzen direkt auf das IP-Protokoll auf und sind eingebettet in ein vollständiges technisches Verwaltungsprotokoll mit festgelegter Prüfung der Authentizität der Benutzer und ihrer Autorisierung. Durch das Handle-Sy-

51 http://www.doi.org/

52 http://www.cnri.reston.va.us/ bzw. http://www.handle.net

stem wird ein Protokoll zur Datenpflege und zur Abfrage der mit dem Handle verknüpften Informationen definiert. Diese Informationen können beliebige Metadaten sein, der Regelfall ist aber, dass die URL des Objektes abgefragt wird, zu dem das Handle registriert wurde. Weiterhin stellt CNRI auch kostenlos Software zur Verfügung, die dieses definierte Protokoll auf einem Server implementiert (und der damit zum sog. Handle-Server wird).

Ein DOI-Name besteht genau wie ein Handle immer aus einem Präfix und einem Suffix, wobei beide durch einen Schrägstrich getrennt sind und das Präfix eines DOI-Namens immer mit „10." Beginnt. Beispiele für DOI-Namen sind:

doi:10.1038/35057062
doi:10.1594/WDCC/CCSRNIES_SRES_B2

Die Auflösung eines DOI-Namens erfolgt nun über einen der oben erwähnten Handle-Server. Dabei sind in jedem Handle-Server weltweit sämtliche DOI-Namen auflösbar. Dieser große Vorteil gegenüber anderen PI-Systemen ergibt sich einerseits durch die eindeutige Zuordnung eines DOI-Präfix an den Handle-Server, mit dem dieser DOI-Name registriert wird und andererseits durch die Existenz eines zentralen Servers bei der CNRI, der zu jedem DOI-Präfix die IP des passenden Handle-Servers registriert hat. Erhält nun ein Handle-Server irgendwo im Netz den Auftrag einen DOI-Namen aufzulösen, fragt er den zentralen Server bei der CNRI nach der IP-Adresse des Handle-Servers, der den DOI-Namen registriert hat und erhält von diesem die geforderte URL.

DOI-Modell

Die Vergabe von DOI-Namen erfolgt wie oben erwähnt nur durch die DOI-Registrierungsagenturen, die eine Lizenz von der IDF erwerben. Dadurch wird sichergestellt, dass jeder registrierte DOI-Name sich an die von der IDF vorgegebenen Standards hält. Diese Standards sind als Committee Draft der ISO Working Group TC46 SC9 WG7 (Project 26324 Digital Object Identifier system) veröffentlicht und sollen ein anerkannter ISO Standard werden. Zum Stand 12/07 gibt es 8 DOI-Registrierungsagenturen, die teilweise kommerzielle, teilweise nicht-kommerzielle Ziele verfolgen. Bei den Agenturen handelt es sich um

• Copyright Agency Ltd[53], CrossRef[54], mEDRA[55], Nielsen BookData[56] und R.R. Bowker[57] als Vertreter des Verlagswesens,

• Wanfang Data Co., Ltd[58] als Agentur für den Chinesischen Markt,

• OPOCE (Office des publications EU)[59], dem Verlag der EU, der alle offiziellen Dokumente der EU registriert

• Technische Informationsbibliothek (TIB) als nicht-kommerzielle Agentur für Primärdaten und wissenschaftliche Information

Dieses Lizenz-Modell wird häufig gleichgesetzt mit einer kommerziellen Ausrichtung des DOI-Systems, doch steht es jeder Registrierungsagentur frei, in welcher Höhe sie Geld für die Vergabe von DOI-Namen verlangt. Auch muss berücksichtigt werden, dass – anders als bei allen anderen PI-Systemen – nach der Vergabe von DOI-Namen durch die Verwendung des Handle-Systems für das Resolving- bzw. für die Registrierungs-Infrastruktur keine weiteren Kosten entstehen.

Die TIB als DOI Registrierungsagentur für Primärdaten

Der Zugang zu wissenschaftlichen Primärdaten ist eine grundlegende Voraussetzung für die Forschungsarbeit vor allem in den Naturwissenschaften. Deshalb ist es notwendig, bestehende und zum Teil auch neu aufkommende Einschränkungen bei der Datenverfügbarkeit zu vermindern.
Traditionell sind Primärdaten eingebettet in einen singulären Forschungsprozess, ausgeführt von einer definierten Gruppe von Forschern, geprägt von einer linearen Wertschöpfungskette:

Experiment	⇒	Primärdaten	⇒	Sekundärdaten	⇒	Publikation
	Akkumulation		Datenanalyse		Peer-Review	

53 http://www.copyright.com.au/
54 http://www.crossref.org/
55 http://www.medra.org/
56 http://www.nielsenbookdata.co.uk/
57 http://www.bowker.com/
58 http://www.wanfangdata.com/
59 http://www.publications.eu.int/

Durch die Möglichkeiten der neuen Technologien und des Internets können einzelne Bestandteile des Forschungszyklus in separate Aktivitäten aufgeteilt werden (Daten-Sammlung, Daten-Auswertung, Daten-Speicherung, usw.) die von verschiedenen Einrichtungen oder Forschungsgruppen durchgeführt werden können. Die Einführung eines begleitenden Archivs und die Referenzierung einzelner Wissenschaftlicher Inhalte durch persistente Identifier wie einen DOI-Namen schafft die Möglichkeit anstelle eines linearen Forschungsansatzes, den Wissenschaftlerarbeitsplatz einzubinden in einen idealen Zyklus der Information und des Wissens (siehe Abbildung 1), in dem durch Zentrale Datenarchive als Datenmanager Mehrwerte geschaffen werden können und so für alle Datennutzer, aber auch für die Datenautoren selber ein neuer Zugang zu Wissen gestaltet wird.

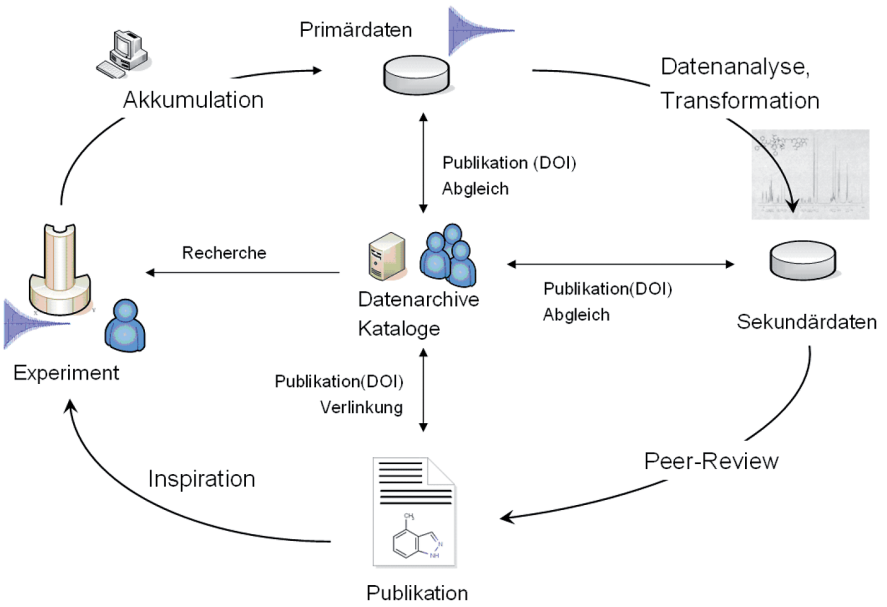


Abbildung 1: Ein idealer Zyklus der Information und des Wissens

Der DFG-Ausschuss „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ hat 2004 ein Projekt[60] gestartet, um den Zugang zu wissenschaftlichen Primärdaten zu verbessern. Aus diesem Projekt heraus ist die TIB seit Mai 2005 weltweit erste DOI-Registrierungsagentur für wissenschaftliche Daten. Beispielhaft im Bereich der Geowissenschaften werden Primärdatensätze

60 http://www.std-doi.de

registriert. Die Datensätze selber verbleiben bei den lokalen Datenzentren und die TIB vergibt für jeden Datensatz einen DOI-Namen.

Der Datensatz wird somit eine eigene zitierfähige Einheit. Mittlerweile wurden über dieses System über 500.000 Datensätze mit einer DOI versehen und zitierfähig gemacht. Die Metadatenbeschreibungen der Datensätze werden zentral an der TIB gespeichert. Diese Beschreibungen enthalten alle Angaben, die nach ISO 690-2 (ISO 1997) zur Zitierung elektronischer Medien verlangt werden.

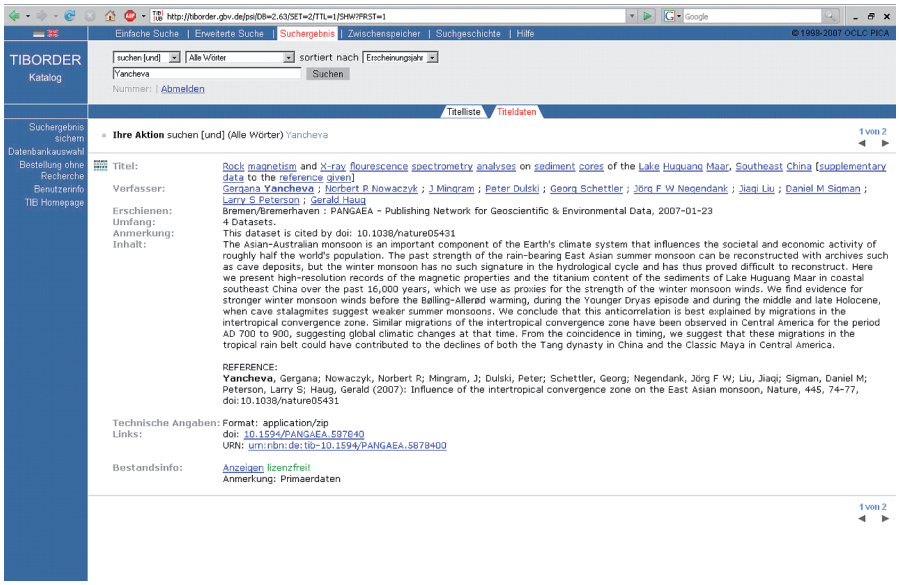


Abbildung.2: Anzeige eines Primärdatensatzes im Online-Katalog der TIB Hannover

Zusätzlich werden Sammlungen oder Auswertungen von Primärdatensätzen auch in den Katalog der TIB aufgenommen. Die Anzeige eines Primärdatensatzes im Katalog der TIB sehen sie in Abbildung 2.

Die DOI Registrierung erfolgt bei der TIB immer in Kooperation mit lokalen Datenspeichern als sog. Publikationsagenten, also jenen Einrichtungen, die weiterhin für Qualitätssicherung und die Pflege und Speicherung der Inhalte, sowie die Metadatenerzeugung zuständig sind. Die Datensätze selber verbleiben bei diesen lokalen Datenzentren, die TIB speichert die Metadaten und macht alle registrierten Inhalte über eine Datenbank suchbar.[61]

61 Brase (2004); Lautenschlager et al. (2005)

Für die Registrierung von Datensätzen wurde an der TIB ein Webservice eingerichtet. Komplementär wurden bei den Publikationsagenten entsprechende Klienten eingerichtet, die sowohl eine automatisierte als auch manuelle Registrierung ermöglichen. In allen Datenzentren sind die SOAP[62]-Klienten vollständig in die Archivierungsumgebung integriert, so dass zusätzlicher Arbeitsaufwand für die Registrierung entfällt. Mithilfe dieser Infrastruktur sind bisher problemlos mehrere hunderttausend DOI Namen registriert worden. Das System baut seitens der TIB auf dem XML-basierten Publishing-Framework COCOON von Apache auf. Dazu wurde COCOON um eine integrierte Webservice-Schnittstelle erweitert, wodurch die Anbindung von weiterer Software überflüssig wird. Die modulare Struktur des Systems erlaubt es, dieses auf einfache Weise auf alle weiteren Inhalte, die mit DOI Namen registriert werden, anzupassen.

Status

Die DOI-Registrierung von Primärdaten ermöglicht eine elegante Verlinkung zwischen einem Wissenschaftlichen Artikel und den im Artikel analysierten Primärdaten. Artikel und Datensatz sind durch die DOI in gleicher Weise eigenständig zitierbar.
So wird beispielsweise der Datensatz:
G.Yancheva, . R Nowaczyk et al (2007)
Rock magnetism and X-ray flourescence spectrometry analyses on sediment cores of the Lake Huguang Maar, Southeast China, PANGAEA
doi:10.1594/PANGAEA.587840
in folgendem Artikel zitiert.
G. Yancheva, N. R. Nowaczyk et al (2007)
Influence of the intertropical convergence zone on the East Asian monsoon
Nature 445, 74-77
doi:10.1038/nature05431
Mittlerweile hat die TIB ihr Angebot auch auf andere Inhaltsformen ausgeweitet.[63] Als Beispiele seien hier genannt:

---

62 SOAP steht für Simple Object Access Protocol, ein Netzwerkprotokoll, mit dessen Hilfe Daten zwischen Systemen ausgetauscht werden können

63 Weitere Informationen zu den Aufgaben der TIB als DOI-Registrierungsagentur und dem Nachweis von Primärdaten durch DOI-Namen sind auf den Internetseiten der TIB zu finden
http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/ und
http://www.tib-hannover.de/de/spezialsammlungen/forschungsdaten/

• doi:10.1594/EURORAD/CASE.1113 in Kooperation mit dem European Congress for Radiology (ECR) wurden über 6.500 medizinische Fallstudien registriert.

• doi:10.2312/EGPGV/EGPGV06/027-034 in Kooperation mit der European Association for Computer Graphics (Eurographics) wurden über 300 Artikel (Graue Literatur) registriert.

• doi:10.1594/ecrystals.chem.soton.ac.uk/145 Gemeinsam mit dem Projekt eBank des UK Office for Library Networking wurden erstmals DOI Namen für Kristallstrukturen vergeben.

• doi:10.2314/CERN-THESIS-2007-001 in Kooperation mit dem CERN werden DOI Namen für Berichte und Dissertationen vergeben

• doi:10.2314/511535090 Seit Sommer 2007 vergibt die TIB auch DOI Namen für BMBF Forschungsberichte.

DOI-Namen und Langzeitarchivierung

Die Referenzierung von Ressourcen mit persistenten Identifiern ist ein wichtiger Bestandteil jedes Langzeitarchivierungskonzeptes. Der Identifier selber kann natürlich keine dauerhafte Verfügbarkeit sicherstellen, sondern stellt nur eine Technik dar, die in ein Gesamtkonzept eingebunden werden muss. Ein Vorteil der DOI ist hier sicherlich einerseits der zentrale Ansatz durch die überwachende Einrichtung der IDF, der die Einhaltung von Standards garantiert und andererseits die breite Verwendung der DOI im Verlagswesen, das an einer dauerhaften Verfügbarkeit naturgemäß interessiert ist. In sehr großen Zeiträumen gerechnet gibt es natürlich weder für die dauerhafte Existenz der IDF noch der CNRI eine Garantie. Allerdings ist die Technik des Handle Systems so ausgelegt, dass eine Registrierungsagentur jederzeit komplett selbstständig die Auflösbarkeit ihrer DOI-Namen sicherstellen kann.

Literatur

Brase, Jan (2004): *Using Digital Library Techniques - Registration of Scientific Primary Data.* Lecture Notes in Computer Science, 3232: 488-494.

International Organisation for Standardisation (ISO) (1997): ISO 690-2:1997 Information and documentation, TC 46/SC 9

Lautenschlager, M., Diepenbroek, M., Grobe, H., Klump, J. and Paliouras, E. (2005): *World Data Center Cluster „Earth System Research" - An Approach for a Common Data Infrastructure in Geosciences.* EOS, Transactions, American Geophysical Union, 86(52, Fall Meeting Suppl.): Abstract IN43C-02.

Uhlir, Paul F. (2003): *The Role of Scientific and Technical Data and Information in the Public Domain,* National Academic Press, Washington DC

10 Hardware

10.1 Einführung

Stefan Strathmann

Einer der entscheidenden Gründe, warum eine digitale Langzeitarchivierung notwendig ist und warum sie sich wesentlich von der analogen Bestandserhaltung unterscheidet, ist die rasch voranschreitende Entwicklung im Bereich der Hardware. Mit dieser Entwicklung geht einher, dass heute noch aktuelle Hardware schon in sehr kurzer Zeit veraltet ist. Die Hardware ist aber eine Grundvoraussetzung zur Nutzung digitaler Objekte. Es müssen also Maßnahmen getroffen werden, der Obsoleszenz von Hardware Umgebungen entgegen zu wirken.

Die Veralterung von Hardware ist – anders als viele andere Aspekte der LZA – auch für Laien und nicht in die Materie eingearbeitete Interessenten sehr leicht nachvollziehbar: wer erinnert sich nicht noch vage an verschiedene Diskettentypen, auf denen vor wenigen Jahren noch wichtige Daten gespeichert wurden. Doch heute verfügen die meisten von uns nicht mehr über entspre-

chende Lesegeräte oder die Daten sind nicht mehr lesbar, weil die Speichermedien durch den Alterungsprozess zerstört wurden.

Nicht alle Speichermedien sind für alle Zwecke (der digitalen Langzeitarchivierung) gleich gut geeignet und es Bedarf einer sorgfältigen Auswahl der Hardware-Umgebung wenn man digitale Objekte langfristig zur Nutzung bereitstellen möchte. Insbesondere die Lebensdauer von verschiedenen Speichermedien kann sehr unterschiedlich sein und muß bei allen Planungen zur LZA berücksichtigt werden.

Das folgende Kapitel untersucht die Anforderungen an eine für LZA Zwekke geeignete Hardware Umgebung und an Speichermedien, bevor die Funktionsweisen und Besonderheiten von Magnetbändern und Festplatten erläutert werden.

Die Herausgeber dieses Handbuches sind bestrebt in künftigen Überarbeitungen dieses Hardware-Kapitel noch deutlich zu erweitern und bspw. auch ein Unterkapitel zu optischen Speichermedien aufzunehmen.

10.2 Hardware-Environment

Dagmar Ullrich

Digitale Datenobjekte benötigen eine Interpretationsumgebung, um ihren Inhalt für Menschen zugänglich zu machen. Diese Umgebung kann in unterschiedliche Schichten gegliedert werden, deren unterste die Hardware-Umgebung bildet. Diese Einteilung wird anhand eines Schichtenmodells, dem „Preservation Layer Model" veranschaulicht. Die Hardware-Umgebung umfasst nicht nur eine geeignete Rechnerarchitektur zur Darstellung der Inhalte, sondern auch eine funktionsfähige Speicherumgebung für den physischen Erhalt und die Bereitstellung des digitalen Datenobjektes.

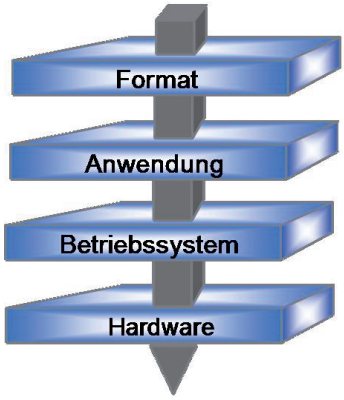
Interpretationsumgebung digitaler Objekte und „Preservation Layer Model"

Um ein digitales Datenobjekt lesbar zu halten, muss eine entsprechende Interpretationsumgebung verfügbar sein. Diese umfasst Hardware, Betriebssystem und Anwendungssoftware. Um z.B. eine Word-Datei anzuzeigen wird eine passende Version von MS-Word benötigt. Für die Installation der Anwendungssoftware muss ein geeignetes Betriebssystem verfügbar sein, das seinerseits auf eine entsprechende Rechnerarchitektur angewiesen ist. In der Regel gibt es mehrere mögliche Kombinationen. Die Lesbarkeit digitaler Daten ist nur so lange sichergestellt, wie mindestens eine solche gültige Kombination einsatzfähig ist. Dieser Zusammenhang wird im Konzept des „Preservation Layer Models" herausgearbeitet. Die nachstehende Grafik veranschaulicht dieses Konzept.[1]

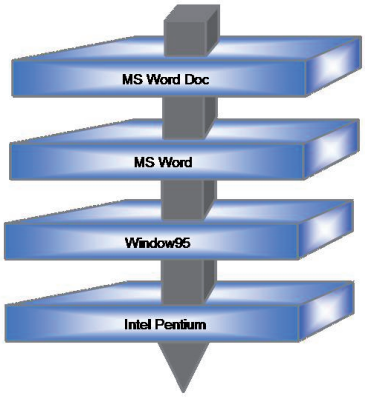
Eine funktionsfähige Kombination der verschiedenen Ebenen wird als gültiger „View Path" eines digitalen Datenobjektes bezeichnet und kann dem entsprechenden Objekt zugeordnet werden. Das Preservation Layer Model wurde an der Nationalbibliothek der Niederlande gemeinsam mit IBM entwickelt, um rechtzeitig zu erkennen, wann ein Datenobjekt Gefahr läuft, ohne gültigen View Path und damit nicht mehr lesbar zu sein. Zeichnet sich der Wegfall einer Komponente ab, lässt sich automatisch feststellen, welche View Paths und somit welche Datenobjekte betroffen sind. Auf dieser Grundlage kann dann entweder eine Emulationsstrategie entwickelt oder eine Migration betroffener

1 Eine ausführliche Beschreibung des Preservation Layer Models findet sich in:
Van Diessen, Raymond J. (2002): *preservation requirements in a deposit system.*
Amsterdam: IBM Netherlands. S. 7-15.
http://www-05.ibm.com/nl/dias/resource/preservation.pdf
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Datenobjekte durchgeführt werden. Im Falle einer Formatmigration werden alle darunter liegenden Ebenen automatisch mit aktualisiert. Die Hard- und Softwareumgebung des alten Formats wird nicht mehr benötigt. Will man jedoch das Originalformat erhalten, müssen auch Betriebssystem und Rechnerarchitektur als Laufzeitumgebung der Interpretationssoftware vorhanden sein. Nicht immer hat man die Wahl zwischen diesen beiden Möglichkeiten. Es gibt eine Reihe digitaler Objekte, die sich nicht oder nur mit unverhältnismäßig hohem Aufwand in ein aktuelles Format migrieren lassen. Hierzu gehören vor allem solche Objekte, die selbst ausführbare Software enthalten, z.B. Informationsdatenbanken oder Computerspiele. Hier ist die Verfügbarkeit eines geeigneten Betriebssystems und einer Hardwareplattform (nahezu) unumgänglich. Um eine Laufzeitumgebung verfügbar zu halten, gibt es zwei Möglichkeiten. Zum einen kann die Originalhardware aufbewahrt werden (vgl. hierzu Kapitel 12.4 Computermuseum). Zum anderen kann die ursprüngliche Laufzeitumgebung emuliert werden (vgl. hierzu Kapitel 12.3 Emulation). Es existieren bereits unterschiedliche Emulatoren für Hardwareplattformen[2] und Betriebssysteme.



Preservation Layer Model



Beispiel: View Path

2 Als Beispiel für die Emulation einer Rechnerarchitektur kann „Dioscuri“ genannt werden. Dioscuri ist eine Java-basierte Emulationssoftware für x86-Systeme. http://dioscuri.sourceforge.net/

Speicherung und Bereitstellung des digitalen Objekts

Aber nicht nur die Interpretierbarkeit der Informationsobjekte erfordert eine passende Umgebung. Bereits auf der Ebene des Bitstream-Erhalts wird neben dem Speichermedium auch eine Umgebung vorausgesetzt, die das Medium ausliest und die Datenströme an die Darstellungsschicht weitergibt. So brauchen Magnetbänder, CD-ROMs oder DVDs entsprechende Laufwerke und zugehörige Treiber- und Verwaltungssoftware. Bei einer Festplatte sind passende Speicherbusse und ein Betriebssystem, das die Formatierung des eingesetzten Dateisystems verwalten kann, erforderlich.

Literatur

Van Diessen, Raymond J. (2002): *preservation requirements in a deposit system*. Amsterdam: IBM Netherlands. S. 7-15. http://www-05.ibm.com/nl/dias/resource/preservation.pdf

10.3 Digitale Speichermedien

Dagmar Ullrich

Datenträger, egal ob analog oder digital, sind nur begrenzt haltbar und müssen früher oder später ausgewechselt werden, um Informationsverlust zu verhindern. Digitale Datenträger veralten in der Regel wesentlich schneller als übliche analoge Medien. Zudem hängt ihre Lesbarkeit von der Verfügbarkeit funktionstüchtiger Lesegeräte ab. Zu den gängigen digitalen Speichermedien zählen Festplatten, Magnetbänder und optische Medien wie CD-ROM oder DVD. Die Unterschiede in Haltbarkeit und Speichereigenschaften entscheiden darüber, inwieweit und in welcher Kombination sie für die Langzeitarchivierung eingesetzt werden können.

Lebensdauer von Trägermedien

Um Informationen über die Zeit verfügbar zu halten, müssen sie auf einem zuverlässigen Trägermedium vorliegen. Die Haltbarkeit des Trägermediums ist von wesentlicher Bedeutung für die Verfügbarkeit der Information. Seine begrenzte Lebensdauer erfordert ein rechtzeitiges Übertragen auf ein neues Medium. Mündlich tradierte Gedächtnisinhalte werden durch Auswendiglernen von einer Generation an die nächste weitergereicht. Schriftstücke wie Urkunden, Bücher oder Verträge werden bei Bedarf durch Kopieren vor dem Verfall des Trägermediums geschützt. Auch digitale Daten benötigen Trägermedien, die erhalten und ggf. erneuert werden müssen.[3] Im Vergleich zu herkömmlichen analogen Datenträgern sind digitale Datenträger jedoch in der Regel deutlich kurzlebiger. Neben ihrer Kurzlebigkeit spielt für digitale Datenträger noch ein weiterer Aspekt eine Rolle: Es wird eine Nutzungsumgebung benötigt, um die Datenobjekte zugänglich zu machen. Um ein digitales Trägermedium, z.B. ein Magnetband oder eine CD-ROM lesen zu können, ist ein entsprechendes Laufwerk und die zugehörige Treibersoftware nötig. Wenn man von der Lebensdauer eines digitalen Datenträgers spricht, muss dabei stets auch die Verfügbarkeit der entsprechenden Nutzungsumgebung (Lesegerät und Betriebssystem mit Treibersoftware) im Auge behalten werden. Eine CD-ROM ohne Laufwerk enthält verlorene Daten, selbst wenn die CD-ROM völlig intakt ist.

3 Der Nachweis der Authentizität ist bei analogem Material wesentlich stärker als bei digitalen Daten an das Trägermedium gebunden. Bei Kopiervorgängen muss dies berücksichtigt werden. Vgl. hierzu Kapitel 8.1.

Die wichtigsten digitalen Speichermedien

In den folgenden Kapiteln werden die drei wichtigsten digitalen Speichermedien, nämlich Festplatte, Magnetbänder und optische Medien vorgestellt. Die genannten Trägermedien lassen sich in zwei Gruppen einteilen: magnetische Medien wie Festplatten und Magnetbänder und optische Medien wie CD-ROM oder DVD. Eine andere mögliche Gruppierung unterscheidet nach Online- und Offline-Speicher. Festplatten werden als Online-Speicher bezeichnet, da sie in der Regel konstant eingeschaltet und für den Zugriff verfügbar sind, Offline-Speichermedien (Magnetbänder, CD-ROM, DVD) dagegen werden nur im Bedarfsfall in ein Laufwerk eingelegt und ausgelesen. Eine dritte mögliche Einteilung der drei Medientypen trennt Medien mit Direktzugriff von sogenannten sequentiellen Medien. Beim Direktzugriff kann ein Schreib-/Lesekopf direkt über der gesuchten Stelle positioniert werden. Beim sequentiellen Zugriff muss einer Schreib-/Lesespur gefolgt werden, bis der relevante Abschnitt erreicht wurde. Festplatten arbeiten mit Direktzugriff. Magnetbänder sind dagegen sequentielle Medien. Durch die Online-Verfügbarkeit und den Direktzugriff ist die Festplatte nach wie vor das schnellste der drei gängigen Speichermedien. Dafür ist sie derzeit noch das verschleißanfälligste und teuerste Speichermedium.[4] Die genannten Medientypen werden oft in Kombination eingesetzt. Dabei werden die Medien so angeordnet, dass teure und performante Medien, zumeist Festplatten, Daten mit hoher Zugriffshäufigkeit vorhalten, weniger oft angeforderte Daten dagegen auf preiswerte Offline-Medien ausgelagert werden. Eine solche Anordnung von Speichermedien wird auch als „Hierarchisches Speichermanagement“ (HSM) bezeichnet. Eine entsprechende Empfehlung findet sich in Calimera Guidelines for Digital Preservation:

> *Strategies for both online and offline storage will be needed. Delivery files in continual use will need to be stored online, on servers. Master files are best stored offline since they are less frequently accessed.*[5]

Bei größeren Unternehmen und Rechenzentren werden die unterschiedlichen Speichermedien zu umfangreichen Speichernetzwerken zusammengeschlossen.

---

4 Ob Festplatten immer noch teurer sind als Bandspeicher ist eine derzeit viel diskutierte Frage. Eine interessante Untersuchung findet sich in:
McAdam, Dianne (2005): *Is Tape Really Cheaper Than Disk?*. White Paper. Nashua: Data Mobility Group.
http://www-03.ibm.com/industries/media/doc/content/bin/DMG_tape_disk.pdf?g_type=pspot

5 o.V. (o.J.) *Digital preservation.* Calimera Guidelines. S.6.
http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf

Die verschiedenen Arten von Speichernetzwerken ermöglichen eine gut skalierbare, redundante Speicherung auf unterschiedlichen Medien. In den meisten Fällen kommen hierfür gängige Backup- oder Spiegelungsmechanismen in lokalen Speichernetzwerken zum Einsatz. Andere Konzepte sehen das Zusammenwirken räumlich weit voneinander entfernter Speicherkomponenten vor. Hierzu gehören auch Peer-to-Peer-Netzwerke, wie sie z.B. von der Open Source Software „Lots of Copies Keep Stuff Safe" (LOCKSS)[6] eingesetzt werden.

Speichermedien in der Langzeitarchivierung

Die nachstehende Tabelle vergleicht Festplatte, Bandspeicher und Optische Medien hinsichtlich ihrer Eignung für unterschiedliche Archivierungszeiträume.[7] Die Eignung eines Speichermediums hängt von den Nutzungsanforderungen und ggf. seiner Kombination mit anderen Speichermedien ab. In diesem Sinne gibt es kein für die Langzeitarchivierung in besonderer Weise geeignetes Speichermedium. Vielmehr empfiehlt es sich, eine Speicherstrategie aufzustellen, die den unterschiedlichen Anforderungen der Archivdaten und der durchschnittlichen Lebensdauer der eingesetzten Speichertechniken gerecht werden kann.

Anforderung	Disk	Bandspeicher	Optische Medien
Häufiger Zugriff	Y	N	N
Schnelle Zugriffszeit	Y	N	Vielleicht
Kurze Archivierung (< 1 Jahr)	Y	Y	Y
Mittlere Archivierung (< 10 Jahre)	N	Y	Y
Lange Archivierung (< 20 Jahre)	N	Y	Y
Auslagerung	N	Y	Y
Unveränderbar	N	mit WORM Tape	mit WORM Disc

6 http://www.lockss.org/lockss/Home

7 Arbeitgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag. S. 45.

Literatur

McAdam, Dianne (2005): *Is Tape Really Cheaper Than Disk*?. White Paper. Nashua: Data Mobility Group.

http://www-03.ibm.com/industries/media/doc/content/bin/DMG_tape_disk.pdf?g_type=pspot

o.V. (o.J.) Digital preservation. Calimera Guidelines. http://www.calimera.org/Lists/Guidelines%20PDF/Digital_preservation.pdf

Arbeitgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag.

10.3.1 Magnetbänder

Dagmar Ullrich

Magnetbänder speichern Daten auf einem entsprechend beschichteten Kunststoffband. Dabei können zwei unterschiedliche Verfahren eingesetzt werden, das Linear-Verfahren oder das Schrägspur-Verfahren. Gängige Bandtechnologien verfügen über Funktionen zur Datenkompression und Kontrollverfahren zur Sicherung der Datenintegrität. Die wichtigsten aktuellen Bandtechnologien werden im Überblick vorgestellt. Als Lesegeräte können Einzellaufwerke, automatische Bandwechsler oder umfangreiche Magnetband-Bibliotheken dienen. Verschleiß der Magnetbänder und damit ihrer Lebensdauer hängen von der Nutzungsweise und Laufwerksbeschaffenheit ab und fallen daher unterschiedlich aus. Die Haltbarkeit hängt darüber hinaus von der sachgerechten Lagerung ab. Regelmäßige Fehlerkontrollen und -korrekturen sind für einen zuverlässigen Betrieb erforderlich. Magnetbänder eignen sich für die langfristige Speicherung von Datenobjekten, auf die kein schneller oder häufiger Zugriff erfolgt, oder für zusätzliche Sicherungskopien.

Funktionsweise von Magnetbändern

Die Datenspeicherung erfolgt durch Magnetisierung eines entsprechend beschichteten Kunststoffbandes. Dabei können zwei unterschiedliche Verfahren eingesetzt werden: das Linear-Verfahren und das Schrägspur-Verfahren. Beim Linear-Verfahren wird auf parallel über die gesamte Bandlänge verlaufende Spuren nacheinander geschrieben. Dabei wird das Band bis zum Ende einer Spur in eine Richtung unter dem Magnetkopf vorbeibewegt. Ist das Ende des Bandes erreicht, ändert sich die Richtung, und die nächste Spur wird bearbeitet. Dieses Verfahren wird auch lineare Serpentinenaufzeichnung genannt. Beim Schrägspur-Verfahren (Helical Scan) dagegen verlaufen die Spuren nicht parallel zum Band, sondern schräg von einer Kante zur anderen. Der rotierende Magnetkopf steht bei diesem Verfahren schräg zum Band. Die wichtigsten Bandtechnologien, die auf dem Linear-Verfahren beruhen, sind „**L**inear **T**ape **O**pen“ (LTO), „**D**igital **L**inear **T**ape (DLT), die Nachfolgetechnologie Super-DLT und „**A**dvanced **D**igital **R**ecording“ (ADR). Für das Schrägspurverfahren können als wichtigste Vertreter „Advanced Intelligent Tape“ (AIT), Mammoth-Tapes, „**D**igital **A**udio **T**apes“ (DAT) und „**D**igital **T**ape **F**ormat“ (DTF) genannt werden. Die jeweiligen Technologien nutzen verschiedene Bandbreiten. Gängige Bandformate sind 4 mm, 8 mm, ¼ Zoll (6,2 mm) und ½ Zoll (12,5 mm). Die Kapazitäten liegen im Gigabyte-Bereich mit aktuellen Maximalwerten bei bis zu 1,6 Terabyte (LTO4, mit Datenkompression). Ebenso wie die Bandkapazität

hat sich auch die erreichbare Transferrate in den letzten Jahren stark erhöht. Die meisten Bandtechnologien nutzen Datenkompressionsverfahren, um die Kapazität und die Geschwindigkeit zusätzlich zu steigern. Diese Entwicklung wird durch den Konkurrenzdruck immer preiswerteren Festplattenspeichers gefördert. Zur Sicherung der Datenintegrität verfügen die meisten Bandtechnologien über Kontrollverfahren, die sowohl beim Schreiben als auch bei jedem Lesezugriff eingesetzt werden.

Übersicht der wichtigsten Bandtechnologien

Die nachstehende Tabelle listet die oben genannten Technologien im Überblick.[8] Es wurden bewusst auch auslaufende Technologien in die Tabelle aufgenommen (ADR, DTF). Das hat drei Gründe: Erstens werden diese Technologien noch vielerorts eingesetzt, zweitens erlauben die älteren Angaben eine anschauliche Darstellung des Kapazitäts- und Performance-Wachstums in den letzten Jahren und drittens zeigt sich hier, wie schnell Bandtechnologien veraltern und vom Markt verschwinden, auch wenn die Medien selbst eine wesentlich längere Lebensdauer haben.

Einzellaufwerke und Bandbibliotheken

Magnetbänder werden für Schreib- und Lesevorgänge in ihre zugehörigen Bandlaufwerke eingelegt. Bei kleineren Unternehmen werden in der Regel Einzellaufwerke eingesetzt. Sie werden im Bedarfsfall direkt an einen Rechner angeschlossen und das Einlegen des Bandes erfolgt manuell. Bei steigender Datenmenge und Rechnerzahl kommen automatische Bandwechsler zum Einsatz. Diese Erweiterungen können beliebig skalierbar zu umfangreichen Bandroboter-Systemen (Bandbibliotheken) ausgebaut werden, die über eine Vielzahl von Laufwerken und Bandstellplätzen verfügen. Solche Bandbibliotheken erreichen Ausbaustufen im Petabyte-Bereich.

---

8 Die Tabelle wurde entnommen und modifiziert aus:
Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag. S. 71.
Wo erforderlich, sind die Angaben über die Webseiten der Hersteller aktualisiert worden.

Technologie	Aktuelle Version	Kapazität ohne Kompression	Transferrate (MB/sec)	Verfahren	Band-format	Weiterführende Informationen [18.08.2007]
ADR[1]	ADR 2	60 GB	4	Linear	8 mm	www.speicherguide.de
AIT	AIT-4	200 GB	24	Helical Scan	8 mm	www.aittape.com
DAT	DAT-72	36 GB	6	Helical Scan	4 mm	www.datmgm.com
DLT	DLT-V4	160 GB	10	Linear	½ Zoll	www.dlttape.com
DTF[2]	DTF-2	200 GB	24	Helical Scan	½ Zoll	www.speicherguide.de
LTO-Ultrium	LTO-4	8400 GB	160	Linear	½ Zoll	www.lto.org
Mammoth[3]	M2	40 GB	12	Helical Scan	8 mm	www.speicherguide.de
S-DLT	SDLT 600A	300 GB	36	Linear	½ Zoll	www.dlttape.com

1 Die Herstellerfirma OnStream hat 2003 Konkurs anmelden müssen, sodass die Fortführung dieser Technologie unklar ist.
2 Die DTF-Technologie wird seit 2004 nicht fortgeführt.
3 Die Herstellerfirma Exabyte wurde 2006 von Tandberg Data übernommen. Seitdem wird das Mammoth-Format nicht weiterentwickelt.

Verschleiß und Lebensdauer von Magnetbändern und Laufwerken

Die Lebensdauer von Magnetbändern wird üblicherweise mit 2 - 30 Jahre angegeben. Die Autoren von „Speichern, Sichern und Archivieren auf Bandtechnologie“ geben sogar eine geschätzte Lebensdauer von mindestens 30 Jahren an:

> *Für die magnetische Datenspeicherung mit einer 50-jährigen Erfahrung im Einsatz als Massenspeicher kann man sicherlich heute mit Rückblick auf die Vergangenheit unter kontrollierten Bedingungen eine Lebensdauerschätzung von mindestens 30 Jahren gewährleisten.*[9]

9 Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit.* Eschborn: AWV-Eigenverlag. S.85

Die große Spannbreite der Schätzungen erklärt sich durch die unterschiedlichen Bandtechnologien. Auch äußere Faktoren wie Lagerbedingungen und Nutzungszyklen spielen eine wesentliche Rolle für die Haltbarkeit. Da Magnetbänder stets ein passendes Laufwerk benötigen, hängt ihre Lebensdauer auch von der Verfügbarkeit eines funktionstüchtigen Laufwerks ab. Ein schadhaftes Laufwerk kann ein völlig intaktes Band komplett zerstören und somit zu einem Totalverlust der gespeicherten Daten führen. Magnetbänder sollten kühl, trocken und staubfrei gelagert werden. Nach einem Transport oder anderweitiger Zwischenlagerung sollten sie vor Einsatz mind. 24 Stunden akklimatisiert werden. Neben der Lagerung spielt der Einsatzbereich eines Magnetbandes mit der daraus resultierenden Anzahl an Schreib- und Lesevorgängen eine Rolle. Je nach Bandtechnologie und Materialqualität ist der Verschleiß beim Lesen oder Beschreiben eines Tapes unterschiedlich hoch. Auch der Verlauf von Lese- oder Schreibvorgängen beeinflusst die Haltbarkeit der Bänder und Laufwerke. Werden kleine Dateneinheiten im Start-Stopp-Verfahren auf das Magnetband geschrieben, mindert das nicht nur Speicherkapazität und Geschwindigkeit, sondern stellt auch eine wesentlich höhere mechanische Beanspruchung von Bändern und Laufwerken dar. Aus diesem Grund bieten neuere Technologien eine anpassbare Bandgeschwindigkeit (ADR) oder den Einsatz von Zwischenpuffern. Laufwerke, die einen ununterbrochenen Datenfluss ermöglichen, werden auch Streamer, die Zugriffsart als Streaming Mode bezeichnet.

Da den Lebensdauerangaben von Herstellern bestimmte Lagerungs- und Nutzungsvoraussetzungen zugrunde liegen, sollte man sich auf diese Angaben nicht ohne weiteres verlassen. Eine regelmäßige Überprüfung der Funktionstüchtigkeit von Bändern und Laufwerken ist in jedem Fall ratsam. Einige Bandtechnologien bringen Funktionen zur Ermittlung von Fehlerraten bei Lesevorgängen und interne Korrekturmechanismen mit. Aus diesen Angaben können Fehlerstatistiken erstellt werden, die ein rechtzeitiges Auswechseln von Medien und Hardware ermöglichen.

Trotz der verhältnismäßig langen Lebensdauer von Magnetbändern und deren Laufwerken sollte nicht übersehen werden, dass die eingesetzten Technologien oft wesentlich kürzere Lebenszyklen haben. Wie bereits oben aus der Tabelle hervorgeht, verschwinden Hersteller vom Markt oder die Weiterentwicklung einer Produktfamile wird aus anderen Gründen eingestellt. Zwar wird üblicherweise die Wartung vorhandener Systeme angeboten, oft aber mit zeitlicher Begrenzung. Aber auch bei der Weiterentwicklung einer Produktfamilie ist die Kompatibilität von einer Generation zur nächsten nicht selbstverständlich. Nicht selten können z.B. Laufwerke einer neuen Generation ältere Bänder

zwar lesen, aber nicht mehr beschreiben. Das technische Konzept für die Datenarchivierung des Bundesarchivs sieht daher folgendes vor:

> *Es sollen nur Datenträger verwendet werden, für die internationale Standards gelten, die am Markt eine ausgesprochen weite Verbreitung haben, als haltbar gelten und daher auch in anderen Nationalarchiven und Forschungseinrichtungen eingesetzt werden. Mit diesen Grundsätzen soll das Risiko minimiert werden, dass der gewählte Archiv-Datenträger vom Markt verschwindet bzw. überraschend von einem Hersteller nicht mehr produziert wird und nicht mehr gelesen werden kann, weil die Laufwerke nicht mehr verfügbar sind.*[10]

Magnetbänder in der Langzeitarchivierung

Magnetbänder sind durch ihre vergleichsweise lange Haltbarkeit für die Langzeitarchivierung digitaler Datenbestände gut geeignet. Dies gilt allerdings nur dann, wenn die Daten in dem gespeicherten Format lange unverändert aufbewahrt werden sollen und die Zugriffzahlen eher gering ausfallen. Sind hohe Zugriffszahlen zu erwarten oder ein kurzer Formatmigrationszyklus sollten Bänder in Kombination mit schnellen Medien wie Festplatten zum Speichern von Sicherungskopien eingesetzt werden.

Literatur

Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV) (2003): *Speichern, Sichern und Archivieren auf Bandtechnologien. Eine aktuelle Übersicht zu Sicherheit, Haltbarkeit und Beschaffenheit*. Eschborn: AWV-Eigenverlag.

Rathje, Ulf (2002): *Technisches Konzept für die Datenarchivierung im Bundesarchiv*. In: Der Archivar, H. 2, Jahrgang 55, S.117-120.

10 Rathje, Ulf (2002): *Technisches Konzept für die Datenarchivierung im Bundesarchiv*. In: Der Archivar, H. 2, Jahrgang 55, S.117-120. (Zitat S. 119).

10.3.2 Festplatten

Dagmar Ullrich

Festplatten sind magnetische Speichermedien. Sie speichern Daten mittels eines Schreib-/Lesekopfes, der über drehenden Platten direkt positioniert wird. Die wichtigsten Speicherbusse (S)-ATA, SCSI, SAS und Fibre Channel werden vorgestellt. Festplatten können einzeln oder im Verbund als Speichersubsysteme genutzt werden. Unterschiedliche Speicherkomponenten können komplexe Speichernetzwerke bilden. Die Lebensdauer von Festplatten wird üblicherweise zwischen 3 und 10 Jahren geschätzt. Umgebungseinflüsse wie magnetische Felder, Stöße oder Vibrationen, aber auch Betriebstemperatur und Nutzungszyklen beeinflussen die Haltbarkeit von Festplatten. Festplatten eignen sich für Kurzzeitarchivierung bzw. in Kombination mit anderen Medien zur Verbesserung von Zugriffszeiten. Für eine revisionssichere Archivierung kommen sie in „Content Addressed Storage-Systemen" zum Einsatz, die über Inhalts-Hashes die Datenauthentizität sicherstellen.

Funktionsweise und Speicherbusse

Festplatten speichern Daten durch ein magnetisches Aufzeichnungsverfahren. Die Daten werden im direkten Zugriff (random access) von einem positionierbaren Schreib-/Lesekopf auf die rotierenden Plattenoberflächen geschrieben bzw. von dort gelesen. Festplatten können beliebig oft beschrieben und gelesen werden. Die aktuelle Maximalkapazität einer einzelnen Festplatte liegt bei einem Terabyte. Festplatten zeichnen sich gegenüber sequentiellen Medien wie Magnetbändern durch schnellen Zugriff auf die benötigten Informationsblöcke aus. Die Zugriffsgeschwindigkeit einer Festplatte hängt vor allem von der Positionierzeit des Schreib-/Lesekopfes, der Umdrehungsgeschwindigkeit der Platten und der Übertragungsrate, mit der die Daten von/zur Platte übertragen werden, ab. Die Übertragungsrate wird wesentlich von der Wahl des Speicherbusses, der Anbindung der Festplatte an den Systembus, bestimmt. Die Speicherbusse lassen sich in parallele und serielle Busse unterscheiden. Die Entwicklung paralleler Busse ist rückläufig, da bei zunehmender Übertragungsrate die Synchronisation der Datenflüsse immer schwieriger wird. Die wichtigsten Standards für Speicherbusse sind: „Advanced Technology-Attachment" (ATA). Dieser ursprünglich parallele Bus wird heute fast ausschließlich seriell als S-ATA eingesetzt. „Small Computer Systems Interface" (SCSI) wurde ebenfalls ursprünglich als paralleler Bus entwickelt und wird heute vorwiegend seriell als Serial-Attached-SCSI (SAS) betrieben. Dieses Bussystem zeichnet sich durch

hohe Übertragungsraten und einfache Konfiguration aus. Fibre Channel[11] (FC) ist ein originär serieller Bus. Er ermöglicht die Hochgeschwindigkeitsübertragung großer Datenmengen und die Verbindung von Speicherkomponenten mit unterschiedlichen Schnittstellen. Er kommt daher hauptsächlich bei größeren Speichersubsystemen oder komplexen Speichernetzwerken zum Einsatz.
Festplatten werden häufig nach ihren Schnittstellen als (S-)ATA-, SCSI- oder SAS-Platten bezeichnet. SCSI- oder SAS-Platten bieten schnelle Zugriffszeiten, sind jedoch im Vergleich zu S-ATA-Platten teuer. S-ATA-Platten dienen vorwiegend dem Speichern großer Datenmengen mit weniger hohen Zugriffsanforderungen. Die ursprünglich aus dem Notebook-Umfeld stammende, heute zunehmend aber auch als mobiles Speichermedium z.B. für Backup-Zwecke eingesetzte USB-Platte, basiert derzeit intern meist auf einer Platte mit (S)-ATA-Schnittstelle.

Einzelfestplatten und Festplattensubsysteme

Festplatten können intern in PCs oder Servern eingebaut oder auch als extern angeschlossener Datenspeicher eingesetzt werden. Die Kapazität einzelner Platten kann durch ihren Zusammenschluss zu Speichersubsystemen (Disk-Arrays) bis in den Petabyte-Bereich[12] erweitert werden. Solche Speichersubsysteme werden meist als RAID-Systeme bezeichnet. RAID steht für „Redundant Array of Independent[13] Disks. "Redundant" weist hier auf den wichtigsten Einssatzzweck dieser Systeme hin: Der Zusammenschluss von Einzelplatten dient nicht nur der Kapazitätserweiterung, sondern vorwiegend der verbesserten Ausfallsicherheit und Verfügbarkeit. Die Platten in RAID-Systemen können so konfiguriert werden, dass bei Ausfall einzelner Platten die betroffenen Daten über die verbliebenen Platten im laufenden Betrieb rekonstruiert werden können. In RAID-Systemen kommen üblicherweise SCSI-Platten zum Einsatz. Zunehmend werden aus Kostengründen auch (S-)ATA-Platten eingesetzt, wobei das Subsystem selbst über SCSI oder FC mit dem Speichernetzwerk verbunden wird. Interessant mit Blick auf ihre Langlebigkeit sind die verhältnismäßig neuen MAID-Systeme. MAID steht für „Massive Array of Idle Disks". Im Unterschied zu herkömmlichen Festplatten-RAIDs sind die Platten dieser

11 Die Bezeichnung Fibre Channel kann insofern irreführend sein, als dass dieser serielle Speicherbus sowohl mit Glasfaser als auch mittels herkömmlicher Kupferkabel umgesetzt werden kann.

12 Werden Speichersubsysteme in dieser Größenordnung ausgebaut, können derzeit noch Schwierigkeiten bei der Speicherverwaltung durch das Betriebssystem auftreten.

13 Da RAID-Systeme die Möglichkeit bieten, auch preiswerte Festplatten mit hoher Ausfallsicherheit zu betreiben, wird das „I" in RAID auch mit „inexpensive" übersetzt.

Speicher-Arrays nicht konstant drehend, sondern werden nur im Bedarfsfall aktiviert. Dies mindert den Verschleiß ebenso wie Stromverbrauch und Wärmeentwicklung, kann aber zu Einbußen in der Zugriffsgeschwindigkeit führen.

Ausfallursachen und Lebensdauer von Festplatten

Die Lebensdauer von Festplatten wird sehr unterschiedlich eingeschätzt. Zumeist wird eine Lebensdauer zwischen 3 und 10 Jahren angenommen. Es finden sich jedoch auch wesentlich höhere Angaben von bis zu 30 Jahren. In der Regel werden als Haupteinflüsse die Betriebstemperatur und der mechanische Verschleiß angesehen. Die übliche Betriebstemperatur sollte bei 30°-45°C liegen, zu hohe, aber auch sehr niedrige Temperaturen können der Festplatte schaden. Ein mechanischer Verschleiß ist bei allen beweglichen Teilen möglich. So sind die Lager der drehenden Platten und der bewegliche Schreib-/Lesekopf bei hohen Zugriffszahlen verschleißgefährdet. Die Gefahr, dass Platten durch lange Ruhezeiten beschädigt werden („sticky disk"), ist bei modernen Platten deutlich verringert worden. Zwei Risiken sind bei Festplatten besonders ernst zu nehmen, da sie einen Totalverlust der Daten bedeuten können: zum einen der so genannte Head-Crash. Ein Head-Crash bedeutet, dass der Schreib-/Lesekopf die drehenden Platten berührt und dabei die Plattenbeschichtung zerstört. Zum anderen können umgebende Magnetfelder die magnetischen Aufzeichnungen schädigen. Festplatten sollten daher in einer Umgebung aufbewahrt werden, die keine magnetischen Felder aufweist, gleichmäßig temperiert ist und die Platte keinen unnötigen Stößen oder sonstigen physischen Beeinträchtigungen aussetzt. In welchem Maße die unterschiedlichen Einflüsse die Lebensdauer von Festplatten beeinträchtigen, wird üblicherweise durch Extrapolation von Labortests festgelegt. Hieraus resultieren die Herstellerangaben zu Lebensdauer und Garantiezeiten. Die Lebensdauer einer Festplatte wird üblicherweise mit „mean time before failure" (MTBF) angegeben. Diese Angabe legt die Stunden fest, die eine Platte betrieben werden kann, bevor Fehler zu erwarten sind. Die Betriebsdauer sollte sich jedoch nicht nur an der MTBF ausrichten, da im Produktivbetrieb oft deutliche Abweichungen von diesen Werten feststellbar sind. Es empfiehlt sich stets auch der Einsatz und die Weiterentwicklung von Überwachungssoftware.

Festplatten in der Langzeitarchivierung

Welche Rolle kann ein Medium, dem eine durchschnittliche Lebensdauer von 5 Jahren zugesprochen wird, für die Langzeitarchivierung von digitalen Da-

tenbeständen spielen? Als Trägermedium zur langfristigen Speicherung von Daten sind langlebigere Medien wie Magnetbänder nicht nur aufgrund ihrer Lebensdauer, sondern auch aus Kostengründen in der Regel besser geeignet. Festplatten können aber in zwei möglichen Szenarien auch für Langzeitarchivierungszwecke sinnvoll sein. Zum einen können sie die Zugriffszeiten auf Archivinhalte deutlich verbessern, wenn sie in Kombination mit anderen Medien in einem hierarchischen Speichermanagement eingesetzt werden. Zum anderen können beispielsweise Formatmigrationen schon nach kurzer Zeit für einen Teil der Archivobjekte erforderlich werden. In diesem Fall ist eine langfristige Speicherung der Dateien gar nicht erforderlich, sondern viel eher deren zeitnahes Auslesen und Wiedereinstellen nach erfolgter Formataktualisierung. Die veralteten Originalversionen können dann auf ein langlebiges Medium ausgelagert werden. Für die jeweils aktuellen Versionen jedoch, die möglicherweise einen kurzen Formatmigrationszyklus haben, kann eine Festplatte ein durchaus geeignetes Trägermedium sein.

Revisionssichere Archivierung mit Content Addressed Storage-Systemen (CAS)

In Wirtschaftsunternehmen und im Gesundheitswesen sind die Anforderungen an Archivierungsverfahren oft an die Erfüllung gesetzlicher Auflagen gebunden. Zu diesen Auflagen gehört oft der Nachweis der Datenauthentizität. Eine Möglichkeit, diese geforderte Revisionssicherheit herzustellen, liegt in der Verwendung von Speichermedien, die nicht überschrieben werden können. Hierfür wurde in der Vergangenheit auf WORM-Medien (Write Once Read Many) zurückgegriffen. Heute werden CD-ROM oder DVD bevorzugt. Eine Alternative hierzu stellen so genannte CAS-Systeme auf Festplattenbasis dar. CAS-Systeme nutzen gut skalierbare Festplattenspeicher in Kombination mit internen Servern und einer eigenen Verwaltungssoftware. Das Grundprinzip beruht auf der Erstellung von Checksummen bzw. Hashes zu jedem eingestellten Inhalt. Über diese Inhalts-Hashes werden die Objekte adressiert. Der Hash-Wert sichert dabei die Authentizität des über ihn adressierten Inhalts. Dieses Verfahren ist an die Verfügbarkeit des CAS-Systems und der Funktionstüchtigkeit der eingesetzten Hardware gebunden. In der Regel können einzelne Komponenten im laufenden Betrieb ausgetauscht und aktualisiert werden.

11 Speichersysteme mit Langzeitarchivierungsanspruch

11.1 Einführung

Heike Neuroth

Dieses Kapitel gibt eine generelle Einführung in die technischen Systeme für die Langzeitarchivierung digitaler Objekte. Dabei werden internationale Entwicklungen ebenso berücksichtigt wie nationale praktische Beispiele gegeben.

Insgesamt bleibt festzuhalten, dass es nicht DIE technische Lösung gibt, sondern je nach Art der digitalen Sammlung, vorhandenem technischen Know-How, Bedarf, potentiellem Nutzungsszenarium und Budget verschiedene Möglichkeiten in Frage kommen. Es kann auch durchaus sein, dass an einer Institution zwei oder gar mehrere Archivsysteme parallel implementiert werden müssen.

Festzuhalten bleibt auch, dass mit der Diskussion um die Publikation und Nachnutzung von Forschungsdaten das Thema „technisches Archivsystem" mehr und mehr auch in die (wissenschaftliche) Breite getragen wird. So hat sich zum Beispiel im Frühjahr 2009 zum ersten Mal eine Arbeitsgruppe im Rahmen

des Open Grid Forums (OGF[1]) gegründet, die sich mit Grid und Repositorien beschäftigt. Dies ist eine Notwendigkeit, die sich aus den „data-driven sciences" mehr und mehr ergibt, da die zum Teil sehr teuer produzierten Datenmengen im TeraByte bzw. PetaByte Bereich, denen meist eine Grid Anwendung zugrunde liegt (z.B. CERN, Teilchenphysik), nicht reproduziert werden können und für die Nachnutzung interpretierbar bereit gestellt werden sollen.

1 http://www.ogf.org/

11.2 Repository Systeme – Archivsoftware zum Herunterladen

Andreas Aschenbrenner

Organisationen und Institutionen arbeiten bereits länger an technischen Systemen zur Verwaltung und Publikation ihrer digitalen Informationsobjekte (z.B. Dokumente, Multimedia, Rohdaten). Diese Dienste sind zum Beispiel die Publikationsserver, die verschiedene Hochschulen in Deutschland mittlerweile bereitstellen.

Während solche Aktivitäten früher eher abteilungs- und verwendungsspezifisch und häufig ad-hoc angegangen wurden, hat sich inzwischen die Situation deutlich geändert. Eine breite Community teilt ähnliche Anforderungen an solche Systeme, tauscht ihre Erfahrungen hierzu aus und entwickelt häufig gemeinsam auf der Basis des Open Source-Konzeptes – entsprechende Softwaresysteme. Inzwischen zeichnet sich eine gewisse Konvergenz der Technologien ab und ein Trend zu Offenheit und Interoperabilität. Viele Nutzer von Publikationsservern kennen - ohne es zu wissen - die typischen Web-Präsentationen von den drei bekanntesten Open Source Repositories; vielleicht sogar von den Webseiten ihrer eigenen Universität.

Dieses Kapitel des nestor Handbuchs präsentiert einen kurzen Überblick über existierende technische Systeme und damit verbundene Aktivitäten. Im Weiteren wird nicht auf kontextspezifische Anforderungen und organisatorische Strukturen oder Konzepte wie „institutional repositories", „trusted repositories", „open access repositories" eingegangen.

Diese sogenannten Repository Systeme decken je nach Fokus und Zielgruppe unterschiedliche Funktionen ab[2]:

- **Verwaltung von Informationsobjekten** (wo sind die Objekte wie abgespeichert, redundante Speicherung)
- **Metadatenverwaltung**, zur Identifikation, Administration und langfristigen Erhaltung von Informationsobjekten

2 Diese kurze Auflistung kann nicht vollständig sein und listet nur einige Kern-Funktionalitäten unterschiedlicher Fokusgruppen und Ziele. Für weitere technische Funktionen siehe z.B. den ISO Standard zu einem „Open Archival Information System" (OAIS) (http://public.ccsds.org/publications/archive/650x0b1.pdf), das DELOS Reference Model (http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid=) und andere.
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

- **Workflow**-Unterstützung zur Registrierung von Informationsobjekten (Ingest)
- **Zugang** durch Identifikation, Suchmechanismen etc
- **Präsentation**, **Einbettung** in die Nutzungsumgebung, Unterstützung von **Kollaboration**
- **Analyse** der Nutzung (Nutzungsstatistiken) und Archivinhalte
- **Vernetzung** der Objekte untereinander und mit Kontextdaten
- Unterstützung von Mechanismen zur **Langzeitarchivierung**

Entsprechende Gesamtpakete, die einige dieser Funktionen für ein bestimmtes Anwendungsgebiet umsetzen, sind bereits in den 90er Jahren aufgekommen, darunter der CERN Document Server[3] oder der Hochschulschriftenserver der Universität Stuttgart OPUS[4]. Andere Institutionen haben eigene Systeme entwickelt oder bestehende Systeme aufgegriffen und für ihre Bedürfnisse angepasst, wo dies sinnvoll möglich war. Inzwischen gibt es z.B. über 200 Installationen der EPrints[5] Software für die institutionelle Verwaltung von eprints (Dissertationen, Journale, Berichte, etc) und mehr als 300 in dem Verzeichnis OpenDOAR[6] nachgewiesene Installationen von DSpace. DSpace[7] - vormals eine institutionelle Repository Software des MIT[8] - hat substantielle Interessensgruppen in China, Indien und anderen Ländern, eine entsprechend große Community („DSpace Federation“) und eine Unterstützung durch die Wirtschaft.

Heute gibt es eine Vielzahl von Repository Systemen, wie z.B. die Überblicksarbeiten von OSI und nestor zeigen.[9] Besonders gefragt sind zurzeit vor allem folgende drei Repository Systeme, die auch auf der internationalen Repository Konferenz, der OpenRepositories[10], jeweils eigene Sessions haben:

- DSpace. http://www.dspace.org/
- Fedora. http://www.fedora-commons.org
- ePrints. http://www.eprints.org/

---

3 CERN Document Server (CDS). http://cds.cern.ch/, CERN - European Organization for Nuclear Research, http://www.cern.ch

4 OPUS Hochschulschriftenserver der Universität Stuttgart. http://elib.uni-stuttgart.de/opus/

5 EPrints. http://www.eprints.org/

6 OpenDOAR - Directory of Open Access Repositories. http://www.opendoar.org/

7 DSpace. http://www.dspace.org/

8 Massachusetts Institute of Technology. http://web.mit.edu/

9 siehe die entsprechenden Literaturverweise unten

10 OpenRepositories. http://www.openrepositories.org/

Trotz der voneinander unabhängigen Entwicklung einzelner Repository Systeme und deren spezifischen Anwendungsgebiete ist eine klare Tendenz zu Offenheit und Interoperabilität in der Community zu erkennen. Der Austausch wird allein schon dadurch gefördert, dass manche Institutionen mehrere Installationen von unterschiedlichen Systemen bei sich führen, um unterschiedlichen Anforderungen in ihrer Organisation gerecht zu werden. Aber auch die Sichtbarkeit der Open Access Bewegung[11] und aufkommende e-Science Mechanismen zur Vernetzung unterschiedlichster Daten und Dienste untereinander[12] fördern die Offenheit und Interoperabilität von Repository Systemen. Projekte wie Driver[13] und OAI-ORE[14] arbeiten auf eine internationale Föderation von Repositories hin. CRIG[15] widmet sich vor allem der Standardisierung von Schnittstellen. Trotz ihres kurzen Bestehens hat die internationale Repository Community bereits eine bedeutende Entwicklung hinter sich und die aktuellen Trends und Potenziale deuten auf eine Ausweitung und die verstärkte Relevanz des Themas.

Literatur

Semple, Najla (2006). *Digital Repositories*. Digital Curation Centre. 5 April 2006. http://www.dcc.ac.uk/resource/briefing-papers/digital-repositories/

Heery, Rachel & Anderson, Sheila, *Digital Repositories Review*. 2005. http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf

Vergleich bestehender Archivierungssysteme / Uwe M. Borghoff u. Mitarb. Univ. d. Bundeswehr München, Fak. f. Informatik, Inst. f. Softwaretechnologie. - Frankfurt am Main : nestor c/o Die Deutsche Bibliothek, 2005. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20050117016

reUSE *White Paper on Digital Repositories*. March 2005. http://www2.uibk.ac.at/reuse/docs/reuse-d11_whitepaper_10.pdf

JISC *Digital Repositories Programme*. http://www.jisc.ac.uk/index.cfm?name=programme_digital_repositories

---

11 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

12 Zum Beispiel die Verknüpfung von Publikationen mit den zugrunde liegenden wissenschaftlichen Rohdaten und Diensten zur Analyse, siehe auch Kapitel 15.5.

13 Driver - Digital Repository Infrastructure Vision for European Research. European Project IST-2.5.6.3. http://www.driver-repository.eu/

14 Open Archives: Object Reuse and Exchange (ORE). http://www.openarchives.org/ore/

15 JISC Common Repository Interfaces Group (CRIG). http://www.ukoln.ac.uk/repositories/digirep/index/CRIG

Open Repositories 2007 : 2nd International Conference on Open Repositories, 23-26 January 2007. Marriott Rivercenter, San Antonio, Texas, US. http://openrepositories.org/

OSI Guide to Institutional Repository Software. http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf

11.3 Speichersysteme mit Langzeitarchivierungsanspruch

Karsten Huth, Kathrin Schroeder und Natascha Schumann

Einführung

Dieser Beitrag gibt einen Überblick über Speichersysteme mit Archivierungsanspruch. Dabei stehen weniger die technischen Ausprägungen im Mittelpunkt, als vielmehr die allgemeinen Bedingungen, beispielsweise die Entstehungsgeschichte, denn oftmals sind diese Systeme aus Projekten zu konkreten Anwendungsszenarien heraus entstanden. Außerdem soll die generelle Strategie der Langzeitarchivierung dargestellt werden. Die Auswahl ist nicht vollständig, es wurde versucht, die gängigsten Systeme zu berücksichtigen.

Als Beispiele für Lösungen aus Projekten bzw. für konkrete Anwendungsfelder werden DIMAG (Digitales Magazin des Landesarchivs Baden-Württemberg), BABS (Bibliothekarisches Archivierungs- und Bereitstellungssystem der Bayerischen Staatsbibliothek), das Digitale Archiv des Bundesarchiv und PANDORA (Preserving and Accessing Networked Documentary Resources in Australia) vorgestellt. Aus dem Bereich der Institutional Repositories Software werden DigiTool von Ex Libris und Fedora (Flexible Extensible Digital Object and Repository Architecture) erläutert und abschließend Portico, kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen) und LOCKSS (Lots of Copies Keep Stuff Save) dargestellt.

DIMAG

DIMAG steht für das Digitale Magazin des Landesarchivs Baden Württemberg[16]. Es wurde konzipiert für verschiedene Formen von digitalen Archivalien, seien es elektronische Akten aus Behördensystemen, Statistiken aus Behörden oder Datenbanken. Die Software des DIMAG wurde vom Landesarchiv in Eigenregie entwickelt. Das System setzt auf offene Softwareprodukte (LINUX, PHP, MySQL und Apache), so dass die Architektur weitestgehend unabhän-

16 Keitel; Lang; Naumann: Konzeption und Aufbau eines digitalen Archivs: Von der Skizze zum Prototypen In: Erfahrungen mit der Übernahme digitaler Daten. Bewertung, Übernahme, Aufbereitung, Speicherung, Datenmanagement – Veröffentlichungen des Archivs der Stadt Stuttgart Bd. 99 Im Internet unter http://www.landesarchiv-bw.de/sixcms/media.php/25/aufsatz_labw_aufbau.pdf

gig von kommerziellen Anbietern ist. Gesichert werden die Daten auf einem RAID-Festplattensystem. Durch die Offenheit des RAID für einen Datentransfer auf andere Medien erhält sich das Archiv die Möglichkeit, den Speicher um eine Tapelibrary zu erweitern. Auch eine Konversion ausgewählter Datenobjekte für eine Belichtung auf Mikrofilm ist denkbar.

Das Produktivsystem steht in Ludwigsburg, Sicherheitskopien gehen an das Hauptstaatsarchiv in Stuttgart und das Generallandesarchiv in Karlsruhe. Das Speichersystem prüft stetig die Integrität und Authentizität der Daten anhand von gespeicherten Hashwertdateien. Abgelegt werden die Daten innerhalb des DIMAG in einem speziell geordneten Filesystem. Dieses Filesystem ist auch dann verfügbar, wenn das Archiv die Kontrolle über die laufende DIMAG-Software verlieren sollte. In dem Filesystem werden sowohl alle Metadaten als auch alle Inhaltsdaten gespeichert. Damit sind die Metadaten für den Fall eines Datenbankverlustes gesichert. Natürlich werden die für eine Recherche relevanten Teile der Metadatensätze in eine Datenbank importiert.

Das Filesystem des DIMAG baut sich aus festgelegten Knoten auf. Unter der Tektonik (= hierarchische Ordnungssystematik der Bestände eines Archivs) des Landesarchivs befindet sich der Knoten „digitales Objekt“, der wiederum mehrere Unterknoten enhalten kann. Diese Unterknoten werden Repräsentationen genannt. Jede Repräsentation enthält dieselbe Information, ist aber technisch verschieden (z.B. eine Repräsentation als Microsoft Office Format und die zweite Repräsentation als PDF/A Format). Repräsentation Nummer eins ist immer das Format, in dem das digitale Objekt an das Archiv übergeben wurde. Auf der Ebene „digitales Objekt“ protokolliert eine XML-Datei die technische Übernahme und die weitere Bearbeitung im Archiv. Unter einem Knoten „Repräsentation“ werden die primären Dateien abgelegt. Die Metadaten zu jedem Knoten und jeder Primärdatei werden jeweils in einer eigenen XML-Datei abgelegt. Alle Metadaten- und Primärdateien werden durch errechnete Hashwerte in eigenen MD5-Dateien gesichert.

Alle Rechte an der DIMAG-Software liegen beim Landesarchiv Baden-Württemberg. Bislang wird das System nur vom Landesarchiv betrieben.

BABS

Das Akronym BABS steht für das Bibliothekarische Archivierungs- und Bereitstellungssystem der Bayerischen Staatsbibliothek (BSB). Unter dem Namen wurde 2005 ein kooperatives Projekt zwischen der Bayerischen Staatsbibliothek und dem Leibniz-Rechenzentrum (LRZ) begonnen, das zum Ziel hatte,

eine organisatorisch-technische Infrastruktur für die Langzeitarchivierung von Netzpublikationen aufzubauen[17]. In BABS werden Retrodigitalisate aus der Produktion des Münchner Digitalisierungszentrums (MDZ) und seit 2008 auch die Bibliothekskopien aus der Public-Private-Partnership der BSB mit Google archiviert sowie auch elektronische Publikationen weiterer Produzenten – amtliche Veröffentlichungen, wissenschaftlich relevante Websites, freiwillige Ablieferungen kommerzieller Verlage etc.

Die Funktionalitäten Ingest, Data Management und Access werden einerseits von dem am MDZ entwickelten Electronic Publishing System ZEND (Zentrale Erfassungs- und Nachweisdatenbank) für die Retrodigitalisate, andererseits von dem Digital Asset Managementsystem DigiTool (siehe auch weiter unten) der Firma Ex Libris für elektronische Publikationen bereitgestellt.

Die Aufgabe des Archival Storage übernimmt das robotergesteuerte Archiv- und Backupsystem mit dem Softwarepaket Tivoli Storage Manager der Firma IBM am Leibniz-Rechenzentrum.

Derzeit (Stand: Januar 2009) wird in BABS ein Datenvolumen von 99,2 TB archiviert.

In einem weiteren Projekt (BABS2) soll die bestehende Infrastruktur nun zu einem vertrauenswürdigen und skalierbaren digitalen Langzeitarchiv ausgebaut werden, um den Herausforderungen rasch wachsender Datenmengen sowie gesetzlicher Verpflichtungen (Erlass über die Abgabe Amtlicher Veröffentlichungen an Bibliotheken, Pflichtstückegesetz) gewachsen zu sein.

Digitales Archiv

Das Digitale Archiv[18] ist die Archivierungslösung des Bundesarchivs. Potenzielle Nachnutzer sind alle Bundesbehörden.

Mit dem Digitalen Archiv können Daten und Metadaten aus disparaten Systemen der Behörden kontrolliert, fehlerfrei und effizient archivtauglich aufbereitet sowie in das Bundesarchiv überführt werden. Eine Pilotanwendung ist erfolgreich getestet worden, der Produktivbetrieb wurde im Oktober 2008 aufgenommen. Die Lösung wurde mit Hewlett Packard (HP) als Generalunternehmer und dem Partner SER geschaffen.

Der Gesamtprozess von der abgebenden Stelle bis in das Storage-System orientiert sich strikt an dem Standard DIN ISO 14721:2003 (Open Archival Information System - OAIS[19]). Technisch nutzt der Prozess zwei Komponenten,

17 BABS-Website: www.babs-muenchen.de

18 http://www.bundesarchiv.de/aktuelles/fachinformation/00054/index.html

19 http://public.ccsds.org/publications/archive/650x0b1.pdf

eine Workflowkomponente für die weitgehend automatisierte Eingangsbearbeitung (Standard-Archivierungsmodul - SAM) und eine Archivierungskomponente mit einer skalierbaren Storage-Lösung für die revisionssichere Speicherung des elektronischen Archivguts.
Kosten und Nutzen:

- Entlastung der Behörden von nicht mehr laufend benötigten Unterlagen
- Aufbau einer zentralen IT-Infrastrukur für die langfristige Speicherung
- komfortable Rückgriffsmöglichkeiten auf archivierte Unterlagen

PANDORA

PANDORA[20], das australische Web-Archiv, wurde 1996 von der Australischen Nationalbibliothek ins Leben gerufen und wird inzwischen von neun weiteren Bibliotheken bzw. Gedächtniseinrichtungen getragen. Es beinhaltet eine Sammlung von Kopien von Online Publikationen, die in Bezug zu Australien stehen. Dabei stehen v.a. Regierungsdokumente, wissenschaftliche Zeitschriften sowie Proceeding-Bände im Fokus. In der Regel sind die archivierten Publikationen frei zugänglich. Allen Ressourcen wird automatisch ein Persistent Identifier zugewiesen.

Archiviert wird nicht nur der Inhalt, sondern auch das „Look and Feel", sofern das möglich ist.

Die Architektur von PANDORA besteht aus dem Archivierungssystem PANDAS, dem Speichersystem DOSS, einem Bereitstellungssystem sowie einer Suchmaschine. Die Strategien zur Langzeitarchivierung beinhalten sowohl die technische Erhaltung durch Hardware und Software als auch, je nach Format, Migration und Emulation.

DigiTool

DigiTool[21] von Ex Libris ist ein Digital Asset Management Tool zur Verwaltung von digitalen Inhalten. Es wird von etlichen Institutional Repositories genutzt. Neben der Verwaltung von digitalen Objekten kann es auch zur Archivierung genutzt werden. Grundlage bildet das OAIS-Referenzmodell. Unterstützt werden Persistent Identifier und die Erstellung von Metadaten unter Verwendung

20 http://pandora.nla.gov.au/
21 http://www.exlibrisgroup.com/category/DigiToolOverview

des Metadatenstandards METS[22]. Mit DigiTool können unterschiedliche Dokumentenarten und Formate verwaltet werden sowie der Ingest-Prozess nach OAIS durchgeführt werden. DigiTool ermöglicht die Integration unterschiedlicher Sammlungen und bietet verschiedene Suchmöglichkeiten.

Im Januar 2009 wurde mit Rosetta[23] von Ex Libris ein eigenes Archivierungssystem gelauncht. Dieses ist direkt als Angebot für Nationalbibliotheken, Museen und weitere Gedächtnisorganisationen als Archivierungssystem gedacht. Das System wurde zusammen mit der Nationalbibliothek von Neuseeland entwickelt. Es hat eine verteilte Architektur und ist skalierbar. Kopien zum Gebrauch und die Dokumente für die Langzeitarchivierung werden getrennt gehalten. Es ist OAIS konform und orientiert sich an Richtlinien für vertrauenswürdige Archive.

FEDORA

Fedora[24] steht für Flexible Extensible Digital Object and Repository Architecture. Entwickelt wurde es an der Cornell University und an der University of Virginia Library. Zunächst als Projekt gefördert, wird Fedora seit 2007 als Non-Profit-Organisation geführt und ist als Open Source Software lizenziert. In erster Linie ist Fedora eine Repository Anwendung, die auch für Archivierungszwecke genutzt werden kann. Es bietet eine Metadatenbasierte Verwaltung der Daten und Unterstützung beim Ingestprozess.

Neben beschreibenden Metadaten werden auch technische Metadaten erfasst, die mittels JHOVE[25] und aus der Formatregistry PRONOM[26] gewonnen werden. PREMIS[27] und weitere LZA relevanten Metadaten können integriert werden. Fedora ist OAIS-konform und unterstützt die Migration. Alle Objekte erhalten Persistent Identifier und es erfolgt eine automatische Versionierung.

22 http://www.loc.gov/standards/mets/
23 http://www.exlibrisgroup.com/category/ExLibrisRosettaOverview
24 http://www.fedora.info/
25 http://hul.harvard.edu/jhove/
26 http://www.nationalarchives.gov.uk/pronom/
27 http://www.oclc.org/research/projects/pmwg/

PORTICO

PORTICO[28] kommt ursprünglich aus dem wissenschaftlichen Bereich und hat sich zum Ziel gesetzt, wissenschaftliche e-Journale in Zusammenarbeit mit Verlagen und Bibliotheken dauerhaft zu archivieren. Gespeichert wird der Inhalt in der Form, in der er veröffentlicht wurde, nicht aber veränderte oder korrigierte Fassungen. Ebenso wenig werden Kontextinformationen, z.B. das „Look and Feel" gespeichert. In den Quelldateien können Grafiken, Text oder andere Ressourcen enthalten sein, die den Artikel ausmachen.
Nach Lieferung der Originaldatei wird diese in ein eigenes Format migriert. Dieses Format basiert auf dem „Journal Publishing Tag Set". Die Archivierungsmethode von PORTICO basiert in erster Linie auf Migration, das heisst, die Dateien werden, wenn nötig, in ein aktuelleres Format umkopiert. Zusätzliche Dienste werden nicht angeboten.
Portico dient als Sicherheitsnetz, das heißt, die Ressourcen werden nur im Notfall herausgegeben und sind nicht für den täglichen Gebrauch gedacht. Die Kosten werden einerseits von den Autoren und andererseits von den Bibliotheken getragen.

kopal

Im Rahmen des Projekts kopal[29] (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen), an dem die Deutsche Nationalbibliothek, die SUB Göttingen, die GWDG Göttingen und IBM Deutschland beteiligt waren, wurde ein digitales Langzeitarchiv auf Basis des DIAS-Systems von IBM entwikkelt. Die im kopal-Projekt entwickelte Open Source Software koLibRI[30] (kopal Library for Retrieval and Ingest) ermöglicht das Erstellen, Einspielen und Abfragen von Archivpaketen (Objekt und zusätzliche Metadaten). Da die Arbeitsabläufe je nach Einrichtung variieren, erlaubt koLibRI, diese je nach Bedarf zu konfigurieren. Das Modell ist flexibel und bietet unterschiedliche Nutzungsmodelle. Da das kopal-System mandantenfähig ist, bietet es sich als zentrale Lösung für unterschiedliche Institutionen an. Der Kern, das DIAS-System, ist beim Dienstleister GWDG gehostet und wird mit Hilfe der koLibRI-Software von den Mandanten im Fernzugriff angesprochen. Zur Gewährung der Langzeitverfügbarkeit unterstützt das kopal-System durch entsprechende Metadaten

28 http://www.portico.org/
29 http://kopal.langzeitarchivierung.de/index.php.de
30 http://kopal.langzeitarchivierung.de/index_koLibRI.php.de

und einen Migrationsmanager die Dateiformatmigration.

Für das kopal-System bestehen drei verschiedene Nutzungsoptionen. 1. Als „kopal-Mandant" erhält eine Einrichtung einen eigenen Bereich des Archivsystems, den sie selbstständig verwaltet. Der Serverbetrieb bleibt allerding ausgelagert. 2. Eine Institution lässt ihre digitalen Daten durch einen „kopal-Mandanten" archivieren. 3. Eine Einrichtung installiert und konfiguriert ihr eigenes kopal-basiertes Archivsystem.

LOCKSS

LOCKSS[31] steht für Lots of Copies Keep Stuff Save. LOCKSS ist eine Kooperation mehrerer Bibliotheken. Initiiert wurde das Projekt von der Stanford University. Inzwischen sind mehr als 150 Bibliotheken beteiligt, das heisst, sie haben eine LOCKSS-Box in Gebrauch. Das ist ein Rechner, der mit der (Open Source) LOCKSS- Archivierungssoftware ausgestattet wird. Die zu archivierenden Ressourcen werden über einen Webcrawler geharvestet. Die Inhalte werden regelmäßig mit denen der anderen Boxen abgeglichen. LOCKSS bietet Zugang zu den Daten und auch zu den Metadaten. Außerdem bietet es eine Verwaltungsebene, die die Mitarbeiter zur Erfassung und zum Abgleich nutzen können. Nachdem der Herausgeber dem Harvesten zugestimmt hat, gibt er die exakte Harvesting-Adresse an.

Die Boxen kommunizieren miteinander und im Falle eines Datenverlustes bei einer Bibliothek springen die anderen ein, um ein nutzbares Exemplar zur Verfügung zu stellen.

Der Zugriff auf die Ressourcen kann auf zwei Arten erfolgen: Entweder wird im Falle der Nichterreichbarkeit auf der Ursprungsseite auf eine archivierte Kopie weitergeleitet oder es wird eine Infrastruktur implementiert, die einen Zugang via SFX erlaubt.

LOCKSS ist format-unabhängig und für alle Arten von Webinhalten nutzbar. Neben dem Inhalt wird ebenso das „Look and Feel" gespeichert. Als Strategie zur Sicherung der Verfügbarkeit der Objekte wird Formatmigration genutzt.

Eine Erweiterung gibt es mit dem Projekt CLOCKSS[32] (Controlled LOCKSS), das als „Dark Archive" nur im Notfall Zugriff auf die archivierten Objekte erlaubt.

---

31 http://www.lockss.org/lockss/Home

32 http://www.clockss.org/clockss/Home

12 Technischer Workflow

12.1 Einführende Bemerkungen und Begriffsklärungen

Reinhard Altenhöner

Immer dann, wenn Termini und Methoden in den Kontext eines spezifischen Themas oder spezialisierter Abläufe eingeführt werden, entsteht Bedarf für einen der eigentlichen Beschäftigung mit dem Gegenstand vorhergehenden Definitions- und Klärungsschritt. So auch in diesem Fall, wenn generelle Methoden zur Beschreibung und zur Modellierung von Abläufen auf die Langzeitarchivierung übertragen werden. Konkret: Die Langzeitarchivierung als relativ neues Arbeitsgebiet, in dem bislang der Schwerpunkt stark auf forschungsnahen oder gar experimentellen Ansätzen lag, wird beim Übergang zu produktiven Systemen und operativen Ablaufproblemen mit neuen Aufgabenstellungen konfrontiert. Und bislang steht für diesen Übergang keine spezifische Methodologie zur Verfügung, die im Sinne eines Vorgehensmodells konkrete Schritte für die Workflowentwicklung benennt.

Beim Übergang in den Langzeitarchivierungsbetrieb geht es um umfassende Arbeitsabläufe, um die massenhafte Prozessierung von (automatisierten) Arbeitsschritten und es wird sinnvollerweise auf das Erfahrungswissen und die

Methodik aus anderen Arbeitsbereichen und Geschäftsfeldern zurückgegriffen, um spezifische Antworten für eine Umsetzung im Umfeld der Langzeitarchivierung zu entwickeln. Günstig ist in diesem Zusammenhang, dass der Bewusstheitsgrad, mit dem Arbeitsprozesse im kommerziellen Kontext – oft über aufwändige Beratungsdienste durch einschlägige Anbieter - organisatorisch und technisch modelliert bzw. erneuert werden, hoch ist. Das gilt sicher generell für das Thema (technische) Prozessorganisation, um so mehr aber für das Arbeitsfeld der Langzeitarchivierung, das insbesondere in Bibliotheken, Archiven und Museen zunehmend wichtiger wird, das aber bislang bis auf wenige Ausnahmen noch nicht in größerem Umfang etabliert und in die allgemeinen Arbeitsabläufe generell integriert ist. Es folgen daher hier zunächst einige einführende Begriffsklärungen, die dann im nächsten Schritt für die konkrete Thematik Langzeitarchivierung methodisch-konzeptionell aufgegriffen werden, um schließlich in einem weiteren Schritt den bislang erreichten Praxisstand an einigen Beispielen etwas eingehender zu betrachten. Ergänzend noch der Hinweis, dass in diesem Handbuch zwischen dem organisatorischen[1] und dem technischen Workflow differenziert wird.

Der Begriff des Workflow wird im Deutschen im Allgemeinen mit dem Begriff des Geschäftsprozesses gleichgesetzt. Aus der abstrahierenden Beschreibung von Einzelfällen entsteht die Basis dafür, Abläufe systematisch als Arbeits- oder Geschäftsprozess zu beschreiben, um zum Beispiel daraus Schulungsmaterial für MitarbeiterInnen zu generieren, aber auch um Schwachstellen zu identifizieren oder neue Fallgruppen zu integrieren. Für die Etablierung neuer Geschäftsprozesse, für die bislang keine Vorlagen oder Matrizen existieren, wird auf die Ergebnisse aus einem Anforderungserhebungsprozess zurückgegriffen; das Requirements Engineering bildet einen eigenen methodischen Ansatz zur systematischen Aufarbeitung der zu lösenden Ausgangssituation. Mit der unterhalb der Ebene des Geschäftsprozesses liegenden Ebene der Arbeitsschritte – der Arbeits/Geschäftsprozess (work process) ist als eine geordnete Folge von Arbeitsschritten definiert - wird ein relativ hoher Detaillierungsgrad angestrebt, der es erlaubt, auf feingranularer Stufe Abläufe differenziert zu verstehen.

Hat man regelbasiert Abläufe beschrieben, tut sich erst die Möglichkeit auf, Geschäftsprozesse zu planen, bewusst systematischer einzugreifen, Teile oder ganze Abläufe neu zunächst abstrakt zu modellieren und dann zum Beispiel in Form von Arbeitsanweisungen praktisch umzusetzen. Auf diese Weise werden Abläufe steuerbar, sie können „gemanaged" werden. In diesen Prozessen wer-

1 Vgl. hierzu auch den von den Herausgebern dieses Handbuchs vorgesehenen Artikel zu organisatorischen Aspekten des Workflow.

den dann Dokumente, Informationen oder auch Aufgaben und Objekte von einem Teilnehmer zum anderen gereicht, die dann nach prozessorientierten Regeln bearbeitet werden. In klassischer Definition wird der Workflow übrigens häufig mit der teilweisen oder vollständigen Automatisierung eines Geschäftsprozesses gleichgesetzt.[2] Dahinter steht die Ansicht, den Reorganisationsbedarf in Institutionen mit der Einführung von IT-gestützten Verfahren bedienen zu können mit der manchmal fatalen Folge, dass anstelle einer eingehenden Analyse der Ausgangssituation die gegebene Organisation an ein gegebenes IT-Verfahren angeglichen wird.

Enger auf den Bereich der öffentlichen Verwaltung bezogen und so auch in Bibliotheken gebraucht ist der Begriff des „Geschäftsgangs", in diesen Einrichtungen häufig festgemacht am Bearbeitungsobjekt, in der Regel Büchern oder auch Akten und dem Weg dieser Objekte durch die einzelnen Phasen seiner Bearbeitung. Gemeint ist hier letztlich – trotz der verwaltungstypischen Fokussierung auf die bearbeiteten Objekte – der Arbeitsablauf/Geschäftsprozess als die Gesamtheit aller Tätigkeiten zur Erzeugung eines Produktes bzw. zur Erstellung einer Dienstleistung.[3]

Ein „Workflow-System" bezeichnet dagegen explizit die IT-gestützte integrierte Vorgangsbearbeitung, in der Datenbank, Dokumentenmanagement und Prozessorganisation in einem Gesamtkonzept abgebildet werden.[4] Abläufe werden also technisch unterstützt, wenn nicht sogar überhaupt mit Hilfe technischer Werkzeuge und Methoden betrieben.

Aber auch die Modellierung / Aufnahme von Geschäftsprozessen selbst kann toolunterstützt erfolgen; solche Geschäftsprozeßmanagement-Tools dienen der Modellierung, Analyse, Simulation und Optimierung von Prozessen. Die entsprechenden Applikationen unterstützen in der Regel eine oder mehrere Methodiken, ihr Funktionsspektrum reicht von der Ist-Aufnahme bis zur Weitergabe der Daten an ein Workflow-Management-System. Im Mittelpunkt stehen dabei die Organisation, Aufgaben bzw. Ablauf der Aufgaben und die zugrundeliegenden Datenmodelle. Mit der Schnittstelle solcher Tools zum Beispiel zu Workflow-Management-Systemen beschäftigt sich die Workflow-Management-Coalition[5], die sich insbesondere die Austauschbarkeit der Daten und damit die Interoperabilität zwischen unterschiedlichen, zum Teil spezialisierten

2 Martin (1999), S. 2.

3 Verwaltungslexikon (2008), Eintrag Workflow. Damit der englischen Ausgangsbedeutung des Begriffs folgend.

4 Verwaltungslexikon (2008), aaO.

5 http://www.wfmc.org/
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Tools durch entsprechende Standardisierungsanstrengungen auf die Fahnen geschrieben hat.

Der Begriff des „technischen Workflows" schließlich wird im Allgemeinen primär für die Abläufe verwandt, die einen hohen Automatisierungsgrad bereits haben oder wenigstens das Potential dazu. Entsprechend bezeichnet man mit dem Begriff des „Technischen Workflow-Management" die Systeme, die durch eine geringe Involviertheit von Menschen und eine hohe Wiederholbarkeit bei geringen Fehlerquoten gekennzeichnet sind.

Damit ist klar, dass der Begriff des technischen Workflow im Kontext der Langzeitarchivierung geradezu programmatischen Charakter hat, da angesichts der großen Objektmengen und ihrer prinzipiell gegebenen Eigenschaften als digitale Publikation ein hoher Automatisierungsgrad besonders bedeutsam ist. Und gleichzeitig liegt es nahe, sich bewusst auf Methoden und Werkzeuge aus dem Bereich des (technischen) Workflowmanagement zu beziehen.

12.2 Workflow in der Langzeitarchivierung: Methode und Herangehensweise

Reinhard Altenhöner

Die allmähliche Einführung der Langzeitarchivierung in das reguläre Auftragsportfolio von Bibliotheken und anderen Kulturerbeeinrichtungen mit immer höheren Bindungsquoten von Personal und anderen Ressourcen erzeugt(e) zunächst neue, häufig isolierte und händisch durchgeführte Abläufe, verändert aber auch in einer ganzheitlichen Betrachtung Arbeitsabläufe und die sie modellierenden Geschäftsprozesse. So ist schon für sich die Einspielung von Daten in ein Langzeitarchiv ein komplexer Vorgang, in dem eine ganze Reihe von aufeinander bezogenen bzw. voneinander abhängenden Aktivitäten ablaufen. Vor allem aber die zunehmende Relevanz der technischen und operativen Bewältigung der Aufgabe verlangt nach einer systematischen Modellierung der Geschäftsprozesse, also dem Einstieg in ein systematisches (technisches) Workflowmanagement. Es gilt allerdings festzustellen, dass selbst in Einrichtungen, die bereits seit einigen Jahren Erfahrungen mit dem Betrieb von Langzeitarchiven und ihrer Integration in die jeweilige Systemlandschaft gesammelt haben, häufig noch isolierte Bearbeitungsketten ablaufen, die zudem keinesfalls wirklichen Vollständigkeitsgrad haben, also alle Anforderungs- /arbeitsfelder abdecken und außerdem vielfach noch manuelle Eingriffe erfordern, insbesondere auf dem Gebiet des Fehlermanagements.

Diese Feststellung bedeutet aber auch, dass der Erfahrungshorizont zum technischen Workflow und insbesondere zum Management insgesamt noch gering ist, also hier noch konkrete Erfahrungen vor allem im Umgang mit großen Mengen und insbesondere auch im automatisierten Qualitätsmanagement gewonnen werden müssen. Insofern hat die Beschäftigung mit dem technischen Workflow derzeit noch viele theoretische Elemente und ‚propädeutischen' Charakter.

In einer Situation, in der verschiedene (bereits existente und neu entwickelte) Arbeitsprozesse ineinander greifen und auch verschiedene Organisationseinheiten an ein und demselben Vorgang beteiligt sind, ist die Modellbildung im Sinne der Geschäftsprozessmodellierung ein Beitrag zu einer umfassenden Optimierung. Damit befinden sich Bibliotheken, Archive und Museen in einer Situation, die man mit den Anstrengungen der Privatwirtschaft Anfang der 1990er Jahre vergleichen kann, als dort die Modellierung von Geschäftsprozessen unter verschärften Wettbewerbs- und Kostendruckbedingungen systematischer als zuvor angegangen wurde. Auch wenn im öffentlich finanzierten Um-

feld in besonderem Maße historisch geprägte Organisationsformen gegeben sind, die eine vorgangsbezogene Sicht erschweren, führt an der grundsätzlichen Anforderung der Neu-Modellierung aus systematischer Sicht kein Weg vorbei. Diese wird im Umfeld des technischen Workflow immer stark auch von der informationstechnischen Entwicklungsseite getrieben sein, denn Ziel der Geschäftsprozessmodellierung ist letztlich ihre technische Abbildung.

Übergeordnete Ziele dieses Herangehens, also der systematischen Modellierung und eines methodenbewussten Workflowmanagements sind[6]:

- Verbesserung der Prozessqualität
- Vereinheitlichung der Prozesse
- Schnellere und zuverlässigere Bearbeitung von Aufträgen (extern und intern)
- Reduzierung der Durchlaufzeiten
- Kostenreduktion
- Verbesserte Verfügbarkeit von Information / Dokumentation
- Erhöhte Prozessflexibilität
- Erhöhung der Transparenz der Prozesse (Statusermittlung, Dokumentation von Entscheidungen), Qualitätssicherung
- Automatische Eingriffsmöglichkeiten: Dokumentation, Eskalation bei Zeitüberschreitungen, Verteilung von Aufgaben und Verantwortlichkeiten
- Vermeidung von Redundanz, mangelnder Aktualität und Inkonsistenz durch Mehrfachschritte

Natürlich lassen sich kleine isolierte Prozesse oder Prozesselemente durch individuelle Programmierung jeweils neu umsetzen. Dies geschah in der Vergangenheit vielfach für einzelne Objektklassen oder auch einzelne Datenübergabe- oder –tauschprozesse. Aber schon beim Zusammenführen bzw. Hintereinandersetzen der einzelnen Teilschritte bedarf es einer Gesamtlogik für das Management des Ablaufs dieser Schritte. Fehlt diese Logik, entstehen letztlich viele immer wieder manuelle neu anzustoßende Teilkonstrukte mit dazu häufig proprietären „Konstruktions“elementen. Schon insofern ist die systematische Analyse verschiedener wiederkehrender Arbeitsabläufe ein sinnvoller Ansatz, um so zur Modellierung auch komplexer Vorgänge im Bereich der Langzeitarchivierung zu kommen.

6 Die folgende summarische Zusammenstellung betrifft sowohl organisatorische wie technische Aspekte des Workflowmanagements. Eine Trennung ist theoretisch zwar möglich, praktisch aber nicht sinnvoll.

Ziel dieses systematischen Ansatzes ist es, Services zu definieren, die auch in anderen Kontexten (wieder) verwendbar sind. Sie bilden Arbeitsschritte granular ab, die so in verschiedenen Umfeldern vorkommen (können), beispielsweise das Aufmachen eines Bearbeitungsfalls für ein Objekt und die IT-gestützte Verwaltung verschiedener Be-/Verarbeitungsschritte dieses Objekts. In dieser Perspektive entsteht der Geschäftsprozess für eine Klasse von Objekten aus der Zusammenfügung verschiedener Basisservices, die miteinander interoperabel sind. Dass diese Herangehensweise sehr stark mit dem Modell der Serviceorientierten Architektur (SOA) bei der Entwicklung IT-basierter Lösungen korrespondiert, ist dabei kein Zufall. Voraussetzung dafür ist aber die systematische Modellierung der Arbeits- oder Geschäftsprozesse, die vorgeben, welche Services wann und wie gebraucht werden. Die Prozessmodellierung bildet also die Basis für die Implementierung, die Prozesse selbst dienen der Orchestrierung, dem Zusammenspiel und der Aufeinanderabstimmung der Services. In einem optimalen (Infrastruktur)-Umfeld können so die Arbeitsschritte als kleinere Einheit eines Geschäftsprozesses verschiedene Services lose zusammenbringen.

Der Ansatz, Services nachnutzbar zu gestalten, bezieht sich in der Regel auf eine Organisation. Zwar wird immer wieder versucht, Geschäftsprozesse aus einem institutionellen Umfeld auf ein anderes zu übertragen, allerdings erweist sich dies in der Praxis als außerordentlich schwierig: Zu stark sind die Abweichungen der einzelnen Arbeitsschritte voneinander und zu unterschiedlich die jeweiligen Prioritäten und Schwerpunktsetzungen in den einzelnen Institutionen. Hinzu kommt außerdem noch, dass der Prozess der Modellierung und Ausgestaltung von Geschäftsprozessen selbst erhebliche Erkenntnisgewinne in der jeweiligen Organisation mit sich bringt, die für eine erfolgreiche Einführung neuer oder veränderter Geschäftsprozesse unverzichtbar sind.

Die Informatik hat für die Modellierung und Notation von Geschäftsprozessen verschiedene methodische Herangehensweisen entwickelt, zum Beispiel die Ereignisgesteuerten Prozessketen (EPK), eine von Scheer und Mitarbeitern entwickelte Sprache zur Modellierung von Geschäftsprozessen[7] und vor allem die Unified Modeling Language (UML) der Object Management Group (OMG), die in der Praxis heute dominierende (technische) „Sprache“ für die Modellierung von Daten, Verhalten, Interaktion und Aktivitäten.[8]

Legt man also UML als Syntax fest, sind noch methodische Festlegungen für die Herangehensweise zu treffen und es liegt nahe, sich für die vorbereitende Modellierung von technischen Abläufen in der Langzeitarchivierung am OAIS-

7 Keller (1992)

8 OMG Infrastructure (2007) und OMG Superstructure (2007)

Modell zu orientieren, das die prinzipiellen Aufgaben im Umfeld der Langzeitarchivierung in funktionaler Perspektive beschreibt und an anderer Stelle dieser Enzyklopädie ausführlich beschrieben wird.[9] Für den Bereich des Ingest einzubeziehen ist der Producer-Archive Interface Methodology Abstract Standard" (CCSDS 651.0-B-1), der insbesondere Validierungsmechanismen betrachtet.[10]

Einzelne Funktionen lassen sich so vor der Folie bisher bereits gemachter Erfahrungen allgemein beschreiben. Beispiele für diese übergreifenden Basisprozesse sind (ich nenne nur Beispiele für unmittelbar aus dem Kontext der Langzeitarchivierung heraus relevante Prozesse):

- Plattform- und Systemübergreifendes Taskmanagement
- Daten- und Objekttransfer-Mimik (z.B. OAI, ORE)
- Extraktion und Generierung von Metadaten (METS, LMER)
- Validierung von Dokumentformaten (z.B. JHOVE)
- Persistente Adressierung und Zugriffsmanagement auf Objektebene
- Speicherprozesse
- ID-Management
- Inhaltsauswahl / Basisrecherche
- Migrationsprozesse / Formatkonvertierungen
- On-the-fly-Generierung einer Bereitstellungsumgebung

9 Vgl. hierzu den entsprechenden Artikel von Nils Brübach / Manuela Queitsch / Hans Liegmann (†) in dieser Enzyklopädie als Kapitel 4: „Das Referenzmodell OAIS - Open Archival Information System"

10 Vgl. hierzu http://public.ccsds.org/publications/archive/651x0b1.pdf

12.3 Technisches Workflowmanagement in der Praxis: Erfahrungen und Ergebnisse

Reinhard Altenhöner

Massenprozesse in der Langzeitarchivierung sind noch wenig etabliert; daher ist der Umfang praktischer Erfahrungen noch begrenzt. Wichtige Erkenntnisse konnte sowohl in der technischen Workflowentwicklung als auch in der praktischen Umsetzung die niederländische Nationalbibliothek sammeln. Auch in der Deutschen Nationalbibliothek liegen erste Erfahrungen vor[11]: Nach einer Gesetzesnovelle Mitte des Jahres 2006 hat sie die Zuständigkeit für die Erhaltung der Langzeitverfügbarkeit deutscher Online – oder Netzpublikationen erhalten und steht nun vor sehr konkreten Herausforderungen, die derzeit zu einer umfassenden Reorganisation des technischen Workflow führen.[12] Mit dem Inkrafttreten des neuen Gesetzes und der damit verbundenen deutlich erweiterten Verpflichtung, die Aufgabe der Langzeitarchivierung zu erfüllen, stellt sich hier die Frage in einer neuen Dimension: Wie wird die Bibliothek die neuen Abläufe organisieren, welche technischen Methoden und Anwendungen werden im Massenverfahren eingesetzt? Da gleichzeitig die alten Arbeitsabläufe und –verfahren weiterlaufen, stellt sich die Frage der Integration in ganz anderer Weise. Zwar ist die Bibliothek in der glücklichen Situation, für die neuen Aufgaben zusätzliche Ressourcen erhalten zu haben, doch würden diese nicht eine nahtlose Imitation des organisatorisch-operativen Workflows auf Basis der existierenden Systeme abdecken – das ergibt sich schon aus den Mengen, um die es geht.

Die Königliche Bibliothek der Niederlande (KB) betreibt seit dem Jahr 2003 das OAIS-kompatible Archivierungssystem DIAS der Firma IBM operativ und hat im Laufe der gewonnenen Erfahrungen insbesondere organisatorisch eine ganze Reihe von Anpassungen vorgenommen.[13] Technisch gesehen wurde

---

11 Vgl. hierzu den einführenden Artikel von Maren Brodersen / Sabine Schrimpf im 18. Kapitel „Praxisbeispiele" dieser Enzyklopädie unter dem Titel „Langzeitarchivierung von elektronischen Publikationen durch die Deutsche Nationalbibliothek":

12 Es sei angemerkt, dass es eine ganze Reihe von weiteren Publikationen zum Thema gibt. So stellte etwa Clifton (2005) Worksflows der australischen Nationalbibliothek vor; diese beziehen sich allerdings auf die manuelle Behandlung von Objekten mittels einzelner Tools. Seit 2007 läuft in der australischen Nationalbibliothek ein Projekt zur Etablierung und IT-basierten Unterstützung der Datenmigration von physischen Datenträgern; noch ist es zu früh, um die Übertragbarkeit bzw. Nachnutzbarkeit des Ansatzes beurteilen zu können, vgl. http://prometheus-digi.sourceforge.net/

13 KB (2008)

eine auch in der KB weitgehend isolierte eigene Entwicklung aufgesetzt, die nur in geringem Maße an die sonstigen Abläufe der Bibliothek angebunden ist. Schwerpunkt liegt auf dem Ingest-Prozess, also dem Einspielen des in der Regel von Verlagen bereitgestellten publizierten Materials in das Archiv. Dieses erfolgt weitgehend automatisiert und es ist der Niederländischen Nationalbibliothek sehr schnell gelungen, die Fehlerquoten auf niedrige Promillebereiche zu drücken. Inzwischen sind mehr als zehn Millionen Objekte eingespielt, darunter auch (allerdings wenige) komplexe Objekte wie historische CD-ROMs. Für alle Objekte – es handelt sich in der weit überwiegenden Zahl um PDF-Dateien – gilt, dass in der eigentlichen Langzeitarchivumgebung nur rudimentäre Metadateninformationen gespeichert werden; die bibliographischen Informationen werden über das Recherchesystem der KB zur Verfügung gestellt.

Insgesamt ist es der KB gelungen, den technischen Workflow relativ unkompliziert und damit effizient und für hohe Durchsatzmengen geeignet zu gestalten. Dies war auch deswegen möglich, weil die Zahl der Lieferanten in das System in den Niederlanden zumindest in der Startsituation klein war, da wenige große Verlage bereits einen überwiegenden Anteil am Publikationsvolumen der Niederlande haben.

In Deutschland stellt sich die Situation anders dar: Hier bestimmen in einer zum Teil noch sehr traditionell geprägten Veröffentlichungslandschaft viele Verleger das Bild. Ausgangspunkt für die Deutsche Nationalbibliothek bei der Neukonzipierung ihrer technischen Abläufe war eine Situation, in der für die Verarbeitung von Online-Dokumenten bereits eine Vielzahl von mehr oder weniger halbautomatischen Verfahren für Netzpublikationen, Online-Dissertationen und weitere Materialien existierte. Diese historisch gewachsenen Strukturen standen nebeneinander, d.h. – nicht untypisch für Gedächtnisorganisationen im öffentlichen Kontext – die einzelne Objektklasse war der definitorische Ausgangspunkt für einen hochspezialisierten Workflow. Ziel war und ist daher die Schaffung eines automatischen, einheitlichen Verfahrens mit der Übergabe der Archivobjekte an das im Rahmen des Projekts kopal entstandene Archivsystem und die dort entstandenen Verfahren.[14] Davon betroffen sind sowohl der Ingest wie aber auch der Zugriff auf die Objekte: Aus der Langzeitarchivlösung kopal werden Objekte an den Arbeitsplatzrechner übergeben oder über das in der Konzeptionsphase befindliche Bereitstellungssystem zur Verfügung gestellt. Dabei sind zahlreiche Arbeitsbereiche in der DNB involviert: neben dem bibliographischen System sind dies die Fachbereiche, externe Ablieferer,

14 kopal (2008)

aber auch die für die digitalen Dienste der DNB Verantwortlichen. Insofern ist hier vieles noch offen und ein Werkstattbericht mag dies illustrieren:[15]

Für den Transfer und das Angebot von Objekten auf elektronischen Materialien auf physischen Datenträgern (d.h. CD- bzw. DVD-Veröffentlichungen) existiert ein älterer, segmentierter Workflow, der nun aufgrund der Anforderungen seitens Archivsystem und künftiger Bereitstellung anzupassen ist. Nach Erstellung der Images der Daten und einer Analyse des vorhandenen Materials wurde daher ein Änderungs- und Ergänzungsvorschlag für den integrierten Workflow dieser Materialgruppe erarbeitet.

Ebenso wird der Workflow für genuin online vorliegende Netzpublikationen unter Einbeziehung der Anforderungen der Langzeitarchivierung neu gestaltet und auf die Schnittstellen des Archivsystems angepasst. Dabei ergeben sich eine ganze Reihe von Problemen: So entsprechen fortlaufende Publikationen (vor allem elektronische Zeitschriften-Artikel) und die künftigen zu archivierenden Objekte häufig nicht der aktuellen Abbildung im Online-Katalog. Bibliografische Metadaten von Archivobjekten müssen aber künftig im bibliografischen System abgebildet werden, um einen einheitlichen Zugang zu gewährleisten. Dazu müssen eine Festlegung von Erschließungsvarianten und ein Mapping von Archivobjekten auf Katalogobjekte erfolgen, letztlich also eine klare Definition der Granularität von Objekten und ihrer Abbildung gefunden werden.

Das URN-Management in der DNB wurde bereits erweitert und vor allem technisch so weiterentwickelt, dass eine Einbindung in andere Arbeitszusammenhänge/Module erfolgen kann. Da jedes Objekt zum Einspielen in das Archiv einen Persistent Identifier benötigt, erfolgt für bereits gesammelte Objekte ohne URN eine retrospektive Vergabe der URN. Alle neuen Objekte müssen entweder mit URN geliefert werden bzw. bei Eingang/Bearbeitung einen URN erhalten, was dem künftigen Verfahren entspricht.

Wesentliche Voraussetzungen für die Einbindung des Archivs in die Geschäftsumgebung der Institution liegen mittlerweile vor oder werden gerade geschaffen. Insbesondere die Kernelemente des Produktionssystems laufen, das produktive Einspielen von Material wurde und wird erprobt, nötige Weiterentwicklungen (z.B. noch fehlende Module zur Auswertung von Dateiformaten) wurden und werden ermittelt und Änderungen / Anpassungen in diversen Workflows der traditionellen Bearbeitung wurden bereits angestoßen. Weitere Aufgaben betreffen in hohem Maße die Übergabe des kopal-Systems, die Eta-

15 Wollschläger (2007), S. 18ff.

blierung einer ständigen Arbeitseinheit sowie die retrospektive Aufarbeitung des früher bereits in die Bibliothek gelangten Materials.

Hinter diesen Bemühungen steht der Anspruch, die neuen, mit der Gesetzesnovelle übernommenen Aufgaben, die weit über das Arbeitsfeld der Langzeitarchivierung hinausgehen, in einem ganzheitlichen technischen Workflow abzubilden. In dessen Mittelpunkt stehen aktuell die Übernahme von elektronischen Objekten mit möglichst breiter Nachnutzung vorhandener Metainformationen und die Integration der Abläufe in die Arbeitsumgebung der DNB.

Die praktischen Erfahrungen an der DNB insbesondere für diesen Bereich belegen den besonderen Bedarf für eine bewusste Modellierung der Geschäftsprozesse, die in der Vergangenheit häufig nur unvollkommen gelungen ist. Im Ergebnis standen isolierte, von nur wenigen Personen bediente und bedienbare Abläufe mit einem hohen manuellen Eingriffs- und Fehlerbehandlungsbedarf. Ohne dass heute bereits ein komplettes Profil der zukünftigen technischen Workflow-Umgebung zitierfähig vorliegt, kann doch gesagt werden, dass ein methodisch bewusstes, in enger Kooperation von Bedarfsträger und Informationstechnik ablaufendes Vorgehen zu deutlich klareren Vorstellungen darüber führt, wie die wesentlichen Arbeitsschritte exakt aussehen und wie sie adäquat so abgebildet werden, dass die entstehenden Services auch langfristig und damit über ihren aktuellen Entstehungshintergrund hinaus genutzt werden.

Dass dabei für eine technische Arbeitsumgebung besondere Anforderungen an die Flexibilität und die Orientierung an offenen Standards gelten, liegt auf der Hand und hat wesentlich die Entwicklungsleitlinien für kopal mitbestimmt.[16]

Quellenangaben

Clifton, Gerard: *Safe Havens In A Choppy Sea: Digital Object Management Workflows At The National Library of Australia* (2005), Beitrag zur iPRES - International Conference on Preservation of Digital Objects, Göttingen (September 15, 2005). In: http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Safe%20Havens%20In%20A%20Choppy%20Sea%20Digital%20Object%20Management%20Workflows%20At%20The%20National%20Library%20of%20Australia%20-%20Gerard%20-0Clifton.pdf

16 kopal (2008a)

Keller, Gerhard / Nüttgens, Markus / Scheer, August-Wilhelm (1992): *Semantische Prozessmodellierung auf der Grundlage „Ereignisgesteuerter Prozessketten (EPK)*. In: A.-W. Scheer (Hrsg.): Veröffentlichungen des Instituts für Wirtschaftsinformatik, Heft 89, Saarbrücken. Online in: http://www.iwi.uni-sb.de/Download/iwihefte/heft89.pdf

Königliche Bibliothek der Niederlande (KB): *How the e-Depot works* (2008) In: http://www.kb.nl/dnp/e-depot/dm/werking-en.html

Königliche Bibliothek der Niederlande (KB): *The e-Depot system (DIAS)* (2008) In: http://www.kb.nl/dnp/e-depot/dias-en.html

Kopal (2008): Projekthompage. In: http://kopal.langzeitarchivierung.de/

Kopal (2008a): *kopal: Ein Service für die Langzeitarchivierung digitaler Informationen*. In: http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf

Martin, Norbert (1999): *Und wie kommt die Dissertation auf den Server? Gedanken zum Workflow. Vortrag auf der IuK-Tagung „ Dynamic Documents „ vom 22.-24.3.1999 in Jena*. In: http://edoc.hu-berlin.de/epdiss/jena3/workflow.pdf

OMG Infrastructure (2007). *UML Infrastructure Specification, v2.1.2*. OMG document formal/07-11-04. In: http://www.omg.org/docs/formal/07-11-04.pdf

OMG Superstructure (2007). *UML Superstructure Specification, v2.1.2*. OMG document formal/07-11-02. In: http://www.omg.org/docs/formal/07-11-02.pdf

Stapel, Johan: *The KB e-Depot. Workflow Management in an Operational Archiving Environment* (2005). Beitrag zur iPRES - International Conference on Preservation of Digital Objects, Göttingen (September 15, 2005). In: http://rdd.sub.uni-goettingen.de/conferences/ipres05/download/Workflow%20Management%20In%20An%20Operational%20Archiving%20Environment%20-%20Johan%20Stapel.pdf

Verwaltungslexikon (2008) *Management und Reform der öffentlichen Verwaltung* (2008) In: http://www.olev.de/w.htm

Wollschläger, Thomas (2007): *„kopal goes live"*. In: Dialog mit Bibliotheken 19 (2007), H.2, S. 17 – 22

Workflow Management Coalition (2008) – Website. In: http://www.wfmc.org/

12.4 Systematische Planung von Digitaler Langzeitarchivierung

Carmen Heister, Hannes Kulovits, Christoph Becker und Andreas Rauber

Durch ständige technologische Veränderungen weisen digitale Objekte eine geringe Lebensdauer auf. Digitale Langzeitarchivierung ist somit zu einer dringlichen Aufgabe geworden. Zur langfristigen Bewahrung digitaler Objekte müssen diese mit Tools zur Langzeitarchivierung bearbeitet werden. Die Wahl eines spezifischen Tools für die Format-Migrationen oder Emulationen und die Einstellung spezifischer Parameter ist jedoch eine sehr komplexe Entscheidung. Die Evaluierung, ob und zu welchem Grad potentielle Alternativen spezifische Anforderungen erfüllen und die Erstellung eines soliden Plans zur Erhaltung einer bestimmten Gruppe von Objekten lässt sich als „Planung von Langzeitarchivierung" zusammenfassen. Derzeit wird die Langzeitarchivierungsplanung manuell, meist ad-hoc, mit wenig oder keiner Softwareunterstützung durchgeführt. Dieses Kapitel stellt einen Workflow vor, der hilft, diesen Planungsprozess zu systematisieren.

Einführung

Es gibt eine Reihe von Strategien und Tools, welche die digitale Langzeitarchivierung unterstützen, jedoch fehlt oftmals eine Entscheidungshilfe für die Auswahl der optimalen Lösung. Für die Wahl einer geeigneten Archivierungsstrategie und eines konkreten Tools müssen komplexe Anforderungen bedacht werden. Sorgsame Dokumentation und gut definierte Vorgehensweisen sind nötig um sicherzustellen, dass das Endergebnis zur Planung von Erhaltungsmaßnahmen den Anforderungen der jeweiligen Einrichtung, insbesondere den Nutzern der Objekte („Designated Community") entspricht. Dies ist auch eine der Kernaufgabe von TRAC[17] und nestor[18].

Eine sorgfältige Planung der digitalen Langzeitarchivierung unterstützt den Entscheidungsprozess zur Auswahl der optimalen Lösung, indem im Planungsprozess verfügbare Lösungsmöglichkeiten gegen klar definierte und messbare Kriterien evaluiert werden. Sie stellt eine Kerneinheit des Open Archival Information System (OAIS) Referenzmodells dar[19], insbesondere im Funktionsmodell Preservation Planning – siehe Kapitel 4. Die Planung besteht aus einem konsistenten Workflow, der idealerweise zu einem konkreten Langzeitarchivie-

17 OCLC (2007)
18 nestor (2006)
19 CCDS (2007)

rungsplan *(„preservation plan")* führt. Für die Planung der digitalen Langzeitarchivierung muss der Planungsbeauftragte über mögliche Lösungswege, die auf die betreffenden Objekte anwendbar sind, informiert sein. Es wird ein vorzugsweise automatisierter Vergleich von Dokumenten und Objekten vor und nach der Verwendung einer Archivierungsstrategie (z.B. einer Migration oder Emulation) benötigt, um die Qualität der verwendeten Erhaltungsmaßnahme *(„preservation action")* zu evaluieren. Der Prozess der zur Auswahl der Erhaltungsmaßnahme geführt hat, sollte darüber hinaus wiederholbar und auch gut dokumentiert sein, um die Nachvollziehbarkeit sowohl der zu Grunde liegenden Entscheidungen als auch der Gründe für die Wahl der Erhaltungsmaßnahme zu gewährleisten. Der im Folgenden vorgestellte Workflow stellt eine Konkretisierung der funk-

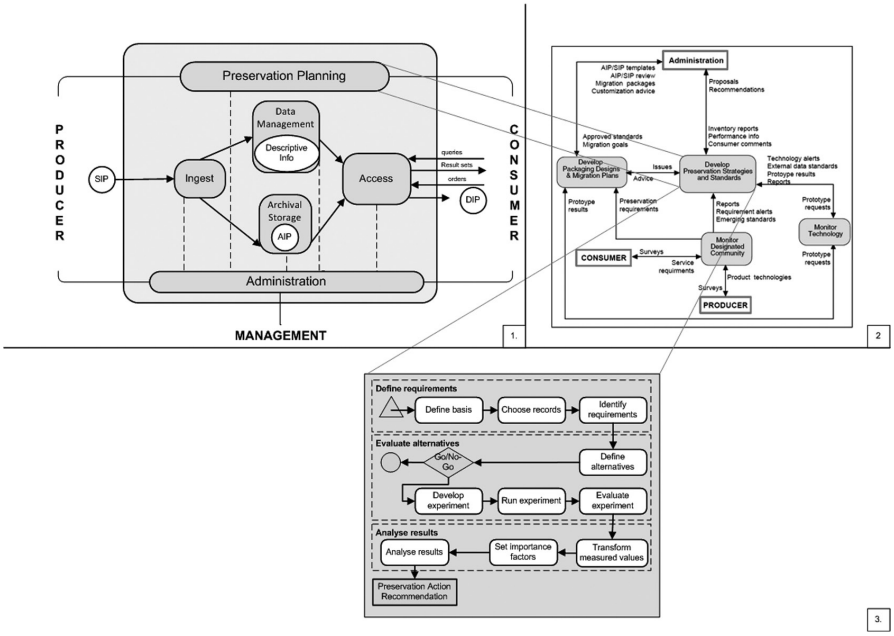


Abbildung 1: OAIS-Modell < Funktionale Entität „Preservation Planning" < Planungs-Workflow

tionalen Komponente *„Develop Preservation Strategies and Standards"* aus dem als ISO 14721 verabschiedeten OAIS Modell „Preservation Planning" dar (Abbildung 1). Der Workflow wurde ursprünglich im Rahmen des Preservation

Clusters des EU NoE DELOS[20] (Network of Excellence on Digital Libraries)[21] konzipiert und nachfolgend im Rahmen des EU Projektes Planets[22] (Preservation and Long-Term Access via Networked Services) verfeinert.[23] Der Workflow basiert auf der Nutzwert-Analyse, einem Verfahren ähnlich der Kosten-Nutzen-Rechnung, kombiniert mit experimenteller Evaluierung.[24]

Der PLANETS Workflow zur Langzeitarchivierung

Anforderungserhebung („Define requirements")

Die Phase 1 des Planungsverfahrens ist die Anforderungserhebung. Dazu gehören das Sammeln von Anforderungen von einer möglichst breiten Nutzergruppe (Abbildung 2), sowie der Faktoren der institutionellen Umgebung, welche die Langzeitarchivierung beeinflussen.

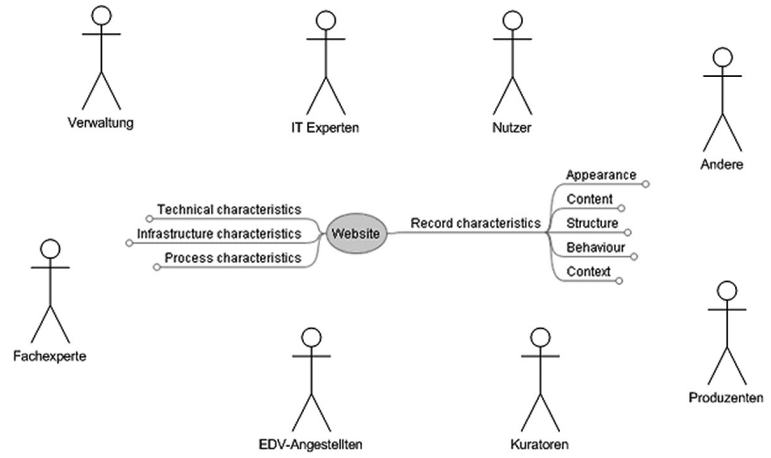


Abbildung 2: Experten, die die Anforderungen auswählen

20 http://www.delos.info/
21 Strodl (2006)
22 http://www.planets-project.eu
23 Farquhar (2007)
24 Rauch (2004)

Evaluierung der Alternativen („Evaluate alternatives")

Die zweite Phase besteht in der Auswahl der in Frage kommenden Strategien, ihrer experimentellen Anwendung auf ausgewählte Beispielobjekte und der Evaluierung der Alternativen bezüglich der definierten Anforderungen.

Analyse der Ergebnisse („Analyse results")

In der dritten Phase werden die Alternativen in ihren Stärken und Schwächen verglichen und analysiert. Auf dieser Basis sind dann fundierte und gut dokumentierte Entscheidungen zur Auswahl der optimalen Strategie möglich.

Erstellen eines Plans zur Langzeitarchivierung
(„Build preservation plan")

Der Plan zur Langzeitarchivierung wird in der vierten Phase in der funktionalen Entität „*Develop Packaging Designs & Migration Plans*" im OAIS-Model nach Genehmigung der empfohlenen Strategie in „*Administration*" erstellt. Er legt fest, welche Archivierungsmaßnahmen wie und von wem durchgeführt werden sollen. Änderungen an den Objekten, eine veränderte Umgebung oder neue Technologien machen es unter Umständen notwendig den Plan anzupassen. Eine Überwachung dieser Parameter und daraus resultierende Veränderungen am Plan bewirken einen ständigen Kreislauf im Planungsprozess.

Detaillierte Beschreibung des Workflows

Im folgenden Abschnitt wird auf die drei Kernphasen des Workflows genauer eingegangen, da sich dieses Kapitel auf die Planungsphasen konzentriert.

Festlegen der Grundlagen („Define basis")

Im ersten Schritt der Phase 1 wird der Kontext des Planungsvorhabens dokumentiert. Dies beinhaltet den Namen des Planes sowie den Namen der Planungsverantwortlichen. Es wird der organisatorische Rahmen dokumentiert, welche Planungsziele die jeweilige Institution hat, was der Planungsgrund ist, welche Zielgruppe angesprochen wird, welche institutionellen Richtlinien zur Langzeitarchivierung existieren (vgl. Kap. 4.2) und welche rechtlichen Bedingungen, personellen sowie finanziellen Ressourcen und organisatorischen Einschränkungen für die Planung wichtig sind.

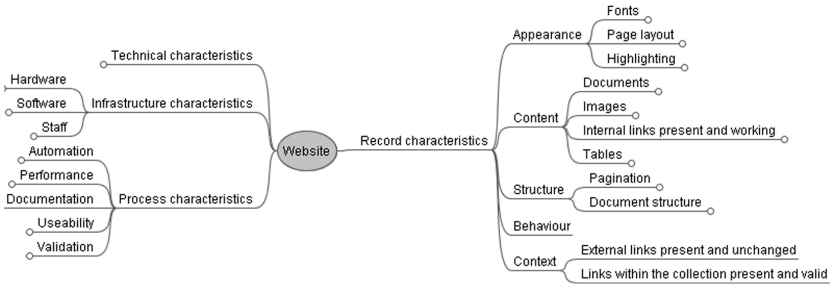


Abbildung 3: Anforderungsform als Mindmap

Auswahl der Datensätze („Choose records")

Im zweiten Schritt werden repräsentative Beispielobjekte ausgewählt, welche die essenziellen Charakteristiken der gesamten Kollektion abdecken. In einem Planungsszenario für die Langzeiterhaltung von digitalen Dissertationen wären das beispielsweise: Eine Dissertation, die sehr viele Abbildungen enthält, eine sehr große bzw. sehr kleine Datei, eine Dissertation, die mathematische Gleichungen/Abbildungen enthält, und eine Datei, die interaktive Elemente beinhaltet. Diese Beispielobjekte werden im späteren Verlauf zur Evaluierung potenzieller Lösungen herangezogen. Üblicherweise werden drei bis fünf Objekte ausgewählt.

Identifizieren der Anforderungen („Identify requirements")

Das Ziel dieses entscheidenden Schrittes ist die Dokumentation der Anforderungen für eine Archivierungsstrategie in klarer und eindeutiger Weise. Diese explizite Definition z.B. der bewahrten Eigenschaften ist eine Kernfordernis sowohl des nestor Kriterienkataloges (Punkt 9.3) [25] als auch des „TRAC"-Kataloges (Punkt B 2.1.)[26]. Allgemeine Ziele und detaillierte Anforderungen werden in einer Baumstruktur, dem sogenannten Kriterienbaum („*Requirements tree*", „*Objective tree*"), gesammelt und strukturiert (Abbildung 3). Die Inhalte des Kriterienbaumes bilden die Basis der quantitativen Evaluierung und ermöglichen dadurch eine objektive Entscheidungsfindung. Während sich die Kriterien im Allgemeinen je nach Kontext und Szenario relativ stark unterscheiden, können einige allgemein gültige Prinzipien festgehalten werden - so hat es sich als

25 nestor (2006)
26 OCLC (2007)

zweckmäßig erwiesen, die Bäume auf der obersten Ebene in vier Hauptkategorien zu unterteilen:

- *Objekteigenschaften („Object characteristics")* beschreiben Anforderungen, inwieweit visuelle und inhaltliche Erlebnisse des Benutzers bei der Betrachtung eines digitalen Objektes erhalten bleiben. Zur Beschreibung der wesentlichen Eigenschaften werden primär fünf Aspekte eines digitalen Objektes herangezogen: Inhalt, Aussehen, Struktur, Verhalten und inhaltlicher Kontext (Metadaten). Bei der nachfolgenden experimentellen Analyse wird gemessen, wie gut diese Eigenschaften bei Anwendung der Erhaltungsmaßnahme erhalten bleiben.
- *Datensatzeigenschaften („Record characteristics")* beschreiben den technischen Kontext der Objekte beziehungsweise die verteilten Strukturen. So können z.B. in Powerpoint-Präsentationen Foliensätze, Videos oder Bilder über eine Verlinkung eingebettet sein oder Webseiten aus zahlreichen Komponenten wie z.B. „Styles-sheets", „Bildern", etc. aufgebaut sein. Diese Kriterien werden unter Umständen vom Anwender nicht unmittelbar wahrgenommen, wenn er mit dem digitalen Objekt zu tun hat. Trotzdem sind sie notwendig, um das Objekt in den Originalzustand zu überführen und damit seine ursprüngliche Wirkung und integrierte Funktionalität wieder herzustellen.
- *Prozesseigenschaften („Process characteristics")* beziehen sich auf den Prozess beziehungsweise das Tool selbst. Sie beinhalten unter anderem die Skalierbarkeit eines Verfahrens auf große Datenmengen oder die Komplexität eines Verfahrens, aber auch Aspekte der Automatisierbarkeit, inwieweit manuelle Eingriffe notwendig sind, etc.
- *Kosten („Costs")* spielen normalerweise eine wichtige Rolle beim Treffen der Entscheidungen. Sie können im Prinzip bei den jeweiligen Bereichen im Baum aufgeführt werden. Aus Gründen der besseren Gliederung werden sie aber meist in einem eigenen Zweig gebündelt. Sie können in technische Kosten und Personalkosten unterteilt werden sowie in Initialkosten und laufende Ausgaben.

Ein Kriterienbaum unterstützt die Identifikation von Kriterien und wird jeweils an einzelnen Stellen erweitert, an anderen reduziert, falls das eher den Vorstellungen einer Organisation entspricht. Zur vollständigen Identifikation der Kriterien ist meist eine ausführliche Recherche relevanter Literatur für das jeweilige Anwendungsgebiet sowie eine detaillierte Analyse der zu bewahrenden Objekte und Anforderungen erforderlich.

Die Kriterienbäume werden üblicherweise in Workshops erstellt, bei denen Anwender, Techniker und Archivierungsexperten zusammenarbeiten, um die relevanten Anforderungen zu ermitteln und zu strukturieren. Ein zentrales Element der Anforderungsanalyse in diesem Zusammenhang ist stets die quantitative Natur der Nutzwertanalyse. Jede Anforderung sollte soweit als möglich objektiv messbar gemacht werden. Daher wird jedem Kriterium in der untersten Ebene eine Skala zugewiesen, auf der die Erfüllung dieses Kriteriums gemessen wird. Soweit wie möglich sollten diese Kriterien objektiv und automatisch messbar sein, z.B. in Euro pro Jahr oder als prozentuelle Abweichung von der ursprünglichen Auflösung eines Bildes in Bildpunkten. In manchen Fällen müssen jedoch (semi-) subjektive Ordinalskalen zum Zuge kommen. Ein Beispiel dafür ist die Akzeptanz oder der Grad der Offenheit und Standardisierung eines Dateiformates.

Der erstellte Baum ist unabhängig von den betrachteten Alternativen; er dokumentiert die individuellen Anforderungen einer Institution oder Person in Bezug auf die langfristige Archivierung einer bestimmten Kollektion digitaler Objekte. Typischerweise enthalten entsprechende Bäume zwischen 20 und 150 Kriterien auf 3 bis 5 Ebenen. Die Anzahl der Kriterien hängt vor allem von der Art der Objekte ab – je komplexer die Inhalte sind, die in den Objekten abgebildet sind, desto aufwändiger ist die Erstellung des Baumes. Diese Komplexität spiegelt sich dann auch in der Anzahl der Kriterien wider.

Alternativen definieren („Define alternatives")

Dieser Schritt betrachtet in Frage kommende Alternativen, wie beispielsweise Migration (vgl. Kapitel 12.2) oder Emulation (vgl. Kapitel 12.3). In diesem Schritt werden die verfügbaren Tools für die in Frage kommenden Strategien ausgewählt. Die Alternativen werden in diesem Schritt ausführlich beschrieben: Name der Alternative, Beschreibung der Alternative, Gründe, warum sie gewählt wurde, Konfigurierungsumgebung und Ressourcen, die für die Ausführung und Evaluierung nötig sind. Wichtig sind insbesondere die Versionsnummer eines Tools, die Parameter-Einstellungen, das installierte Betriebssystem, die Schriftarten, Programmbibliotheken etc.

Fortfahren / Abbruch („Go/No-Go")

Unter Berücksichtigung der definierten Anforderungen, der Alternativen und einer Einschätzung der benötigten Ressourcen wird in diesem Schritt entschieden, ob der Prozess der Evaluierung fortgesetzt, abgebrochen oder verschoben werden soll. Außerdem wird entschieden, welche der aufgelisteten Alternativen

evaluiert werden sollen. Pro Alternative wird dokumentiert, weshalb sie in die engere Wahl gekommen ist oder verworfen wird. Beispielsweise kann es sein, dass für eine Alternative Hardware benötigt wird, die in der Anschaffung für die jeweilige Institution von vornherein viel zu teuer ist: Aus Kostengründen kann diese Alternative nicht evaluiert werden. Dieser Grund für die Entscheidung wird dann dokumentiert. Eine weitere Möglichkeit kann sein, dass eine neue Version eines Tools in naher Zukunft verfügbar sein wird. Diese Alternative kann dann in die Liste aufgenommen, die Evaluierung jedoch auf einen späteren Zeitpunkt verschoben werden („*Deferred-go*").

Experiment entwickeln („Develop experiment")

Um reproduzierbare Ergebnisse zu gewährleisten, wird in diesem Schritt ein Entwicklungsplan für jede Alternative spezifiziert, die das Experiment-Umfeld und die Art und Weise der Evaluierung mit einschließt. Dies umfasst die Rechnerumgebung, auf der die Experimente durchgeführt werden, die Konfiguration und das Aufsetzen der Messinstrumente (Zeitmessung etc.). Im Idealfall ist eine standardisierte Test-Umgebung vorhanden.

Experiment durchführen („Run experiment")

Die betrachteten Alternativen werden nun in einem kontrollierten Experiment auf die gewählten Beispielobjekte angewandt. Das heißt, die Objekte werden mit den ausgewählten Tools migriert oder in den jeweiligen Emulatoren geöffnet. Dabei anfallende Fehlermeldungen bzw. Zeitmessungen sowie Ausgaben in Protokolldateien werden erfasst. Auch dieser Schritt kann durch die Verwendung von in zentralen Registries erfassten Tools, die über Webservices standardisiert aufgerufen werden können, drastisch vereinfacht werden.

Experimente evaluieren („Evaluate experiments")

Um festzustellen, zu welchem Grad die Anforderungen im Kriterienbaum von den einzelnen Alternativen erfüllt werden, werden die Ergebnisse der Experimente evaluiert. Hierfür wird jedes einzelne Blatt im Kriterienbaum für jedes Objekt evaluiert. Die Evaluierung kann zum Teil automatisiert durch Analysetools unterstützt werden, welche die signifikanten Eigenschaften der Objekte vor und nach der Anwendung der Tools vergleichen und die Ergebnisse dokumentieren.

Umwandeln/Gleichsetzung der gemessenen Werte („Transform measured values")

Nach der Evaluierung der Kriterien im Kriterienbaum sind diese in unterschiedlichen Skalen (z.B. EURO, Sekunden, Farbe: ja/nein) definiert. Damit die Kriterien vergleichbar und aggregierbar werden, wird pro Skala eine Transformationstabelle spezifiziert, welche die Werte der Messskala auf eine einheitliche Zielskala, den sogenannten Nutzwert abbildet. Die Zielskala ist üblicherweise eine Zahl zwischen 0 und 5, wobei 5 der beste Wert ist, während 0 ein nicht akzeptables Ergebnis darstellt.[27]

Das Kriterium „Proprietäres Dateiformat" mit einer Boolean Skala „Yes/No" kann je nach Szenario unterschiedlich transformiert werden. Bei einer Transformation von „No" auf den Wert „eins" und „Yes" auf den Wert „fünf", wäre ein proprietäres Dateiformat zwar akzeptabel aber niedrig bewertet. Jedoch bei einer Transformation von „No" auf den Wert „null" (und „Yes" auf den Wert „fünf") wäre ein proprietäres Dateiformat ein Ausschlusskriterium für die gesamte Alternative.

Alternative	Total Score Weighted Sum	Total Score Weighted Multiplication
PDF/A (Adobe Acrobat 7 prof.)	4.52	**4.31**
PDF (unchanged)	**4.53**	0.00
TIFF (Document Converter 4.1)	4.26	3.93
EPS (Adobe Acrobat 7 prof.)	4.22	3.99
JPEG 2000 (Adobe Acrobat 7 prof.)	4.17	3.77
RTF (Adobe Acrobat 7 prof.)	3.43	0.00
RTF (ConvertDoc 4.1)	3.38	0.00
TXT (Adobe Acrobat 7 prof.)	3.28	0.00

Abbildung 4: Evaluierungsergebnisse elektronischer Dokumente

Wertigkeiten festlegen („Set importance factors")

Die Kriterien, die im Kriterienbaum festgelegt worden sind, haben nicht alle die gleiche Wertigkeit für den Planenden. In diesem Schritt wird daher eine relative Gewichtung der Kriterien auf allen Ebenen durchgeführt, um der unterschiedlichen Bedeutung der einzelnen Ziele Rechnung zu tragen. Sind beispielsweise für eine Institution die Kosten sehr wichtig, werden sie in der Gewichtung höher gestuft als beispielsweise bestimmte Objekteigenschaften. Eine Institution, die beispielsweise eine sehr große Anzahl an Objekten migrieren muss, wird auf der höchsten Ebene des Kriterienbaums die Prozesseigenschaften etwas höher

27 Becker (2007)

gewichten als die übrigen. Folgende Gewichtung wäre denkbar: Objekteigenschaften (20%), Datensatzeigenschaften (20%), Prozesseigenschaften (40%) und Kosten (20%). Damit haben gute bzw. schlechte Prozesseigenschaften einen größeren Einfluss auf das Endergebnis.

Evaluierungsergebnisse analysieren („Analyse evaluation results")

Im abschließenden Schritt werden die Ergebnisse aller Alternativen berechnet und aggregiert, um eine Kennzahl zu schaffen, die zum Vergleich der Alternativen herangezogen werden kann. Dabei können verschiedene Aggregationsmechanismen verwendet werden. Die wichtigsten Aggregationsmechanismen sind die Aufsummierung und die Multiplikation. Bei der Aufsummierung werden die transformierten Ergebniswerte jeder Alternative mit dem relativen Gewicht des entsprechenden Kriteriums multipliziert und über die Hierarchie des Baumes hinweg aufsummiert. Dadurch ergibt sich auf jeder Ebene eine Kennzahl zwischen null und fünf, die dem Erfüllungsgrad der entsprechenden Anforderung durch die betrachtete Alternative entspricht. Bei der Multiplikation dagegen werden die transformierten Werte mit dem relativen Gewicht potenziert und über die Hierarchie des Baumes hinweg multipliziert. Wiederum ergibt sich auf jeder Ebene eine Kennzahl zwischen null und fünf. Der wesentliche Unterschied zur Aufsummierung besteht darin, dass ein einzelnes nichtakzeptiertes Kriterium zu einem Totalausfall der Alternative führt, da durch die Multiplikation der Wert „null" bis in den Wurzelknoten durchschlägt. Das Ergebnis sind aggregierte Ergebniswerte für jeden Teilbaum des Kriterienbaumes und für jede Alternative. Eine erste Reihung der Alternativen kann auf den aufsummierten und multiplizierten Kennzahlen geschehen. Abbildung 4 zeigt die Bewertung von verschiedenen Alternativen mit Hilfe der zwei Aggregationsmethoden „Gewichtete Summe" und „Gewichtete Multiplikation". Der Hauptunterschied dieser zwei Aggregationsmethoden liegt in der Einflussnahme von nicht erfüllten Kriterien auf das Bewertungsergebnis der Alternative. Bei der Multiplikation scheiden Alternativen aus, d.h. sie werden mit 0 bewertet, falls ein oder mehrere Mindestkriterien nicht erfüllt werden. Die Alternativen RTF und TXT scheiden beispielsweise aus, weil sie große Nachteile in der Erhaltung der Struktur des Dokuments aufweisen. Die Alternative PDF („unchanged") scheidet bei der Aggregationsmethode Multiplikation aus, da das essentielle Kriterium der Verhinderung von eingebetteten Skripten nicht erfüllt wird. Bei Aufsummierung wird die Alternative PDF („unchanged") mit 4.53 knapp am höchsten bewertet, da nicht erfüllte Mindestkriterien kein Ausscheiden der Alternative verursachen, sondern normal in die Berechnung ein-

fließen. Unter Berücksichtigung der Ergebnisse der beiden Aggregationsmethoden kann eine genaue Analyse der Stärken und Schwächen jeder Alternative durchgeführt werden.

Das Ergebnis dieses Planungsprozesses ist eine konzise, objektive und dokumentierte Reihung in Frage kommender Alternativen für ein betrachtetes Archivierungsproblem unter Berücksichtigung der spezifischen situationsbedingten Anforderungen. Welche Lösung tatsächlich umgesetzt wird, hängt von den begleitenden Umständen ab. Aus der Nutzwertanalyse lässt sich jedoch eine klare Empfehlung ableiten, die mit direkt sichtbaren Argumenten hinterlegt und sorgfältig abgewogen ist und sich daher sehr gut als Entscheidungsgrundlage eignet. Durch die Darstellung sowohl allgemeiner als auch detaillierter Ergebniszahlen aus standardisierten und reproduzierbaren Testbedingungen wird eine solide Basis geschaffen, auf der wohlüberlegte und dokumentierte Entscheidungen getroffen werden können.

Das Planungstool Plato

Das EU-Projekt PLANETS entwickelt eine verteilte, serviceorientierte Architektur mit anwendbaren Services und Tools für die digitale Langzeitarchivierung[28]. Plato (PLANETS Preservation Planning Tool) (vgl. Kapitel 13.2) ist ein in PLANETS entwickeltes Planungstool, das den oben beschriebenen, in drei Phasen unterteilten Workflow implementiert und zusätzlich externe Services integriert, um den Prozess zu automatisieren.[29]

Eines dieser Services ist DROID (Digital Record Object Identification) von den National Archives UK. Damit kann automatisch die Bezeichnung des Dateiformats, die Version, der MIME-Type (Multipurpose Internet Mail Extensions) und der PUID (PRONOM Persistent Unique Identifier) ermittelt werden. Ein weiteres integriertes Service ist die Beschreibung des digitalen Objektes im XCDL-Format. Dieses Service wurde von der Universität Köln entwickelt und wandelt die ausgewählten Objekte in ein XCDL-Format um, welches für die spätere Evaluierung notwendig ist [5]. Zudem integriert Plato mehrere Registries, aus denen zu den Beispielobjekten passende Erhaltungsmaßnahmen ausgewählt und automatisch auf die Beispielobjekte angewendet werden können. Bestimmte Objekteigenschaften können automatisch gemessen und evaluiert werden.

Durch die Zuhilfenahme von frei verfügbaren Frameworks wie z.B. Java Server Faces und AJAX wurde Plato als eine J2EE-Web-Applikation entwickelt,

28 Becker (2008b)

29 Becker (2008a) Strodl, (2007)

die frei verfügbar für Planungsvorhaben zur digitalen Langzeitarchivierung genutzt werden kann. [30]

Zusammenfassung

In diesem Kapitel wurde der Planets Workflow zur Planung digitaler Langzeitarchivierungsvorhaben vorgestellt. Dieser Workflow ist die konkrete Ausarbeitung der Kerneinheit „Preservation Planning" des mit dem ISO Standard 14721 verabschiedeten OAIS-Modells. Der Workflow erfüllt nach derzeitigem Wissenstand in den entsprechenden Bereichen die Anforderungen von Initiativen zur Zertifizierung und Validierung von vertrauenswürdigen Archiven, insbesondere nach TRAC[31] und dem nestor - Kriterienkatalog für vertrauenswürdige digitale Langzeitarchive[32].

Literaturverzeichnis

Becker, Christoph, Rauber, Andreas (2007): *Langfristige Archivierung digitaler Fotografien*. Wien.

Becker, Christoph et al. (2008a): *Distributed Preservation Services: Integrating Planning and Actions*. In: Christensen-Dalsgaard, Birte et al.: Research and Advanced Technology for Digital Libraries. Proceedings of the 12th European Conference on Digital Libraries (ECDL 2008). Berlin, Heidelberg: Springer-Verlag. S. 25-36.

Becker, Christoph et al. (2008b): *Plato: a service-oriented decision support system for preservation planning*. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. 2008. S. 367-370.

Becker, Christoph et al. (2008c): *A Generic XML Language for Characterising Objects to Support Digital Preservation*. In: Proceedings of the 2008 ACM symposium on Applied computing. 2008. S. 402-406.

CCDS Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference model for an open archival information system (OAIS)* / Consultative Committee for Space Data Systems. public.ccsds.org/publications/archive/650x0b1.pdf

Farquhar, Adam., Hockx-Yu, Helen (2007) *Planets: Integrated services for digital preservation*. In: International Journal of Digital Curation, 2. (2007). S. 88-99.

30 http://www.ifs.tuwien.ac.at/dp/plato

31 OCLC (2007)

32 nestor (2006)

National Library of Australia, Unesco. Information Society Division (Hrsg.) (2005): *Guidelines for the preservation of digital heritage. Prepared by the National Library of Australia*. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2006): *Kriterienkatalog vertrauenswürdige digitale Langzeiarchive*. Version 2. (nestor-Materialien 8). Frankfurt am Main: nestor. www.langzeitarchivierung.de/downloads/mat/**nestor**_mat_08.pdf

OCLC Online Computer Library Center, CRL The Center for Research Libraries (Hrsg.) (2007): *Trustworthy Repositories Audit & Certification* (TRAC): Criteria and Checklist. Chicago, Dublin: Center for Research Libraries, OCLC Online Computer Library Center. http://www.crl.edu/PDF/trac.pdf

Rauch, Carl, Rauber, Andreas (2004): *Preserving digital media: Towards a preservation solution evaluation metric*. In: Chen, Zhaoneng et al.: Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004). Berlin: Springer. S. 203-212.

Strodl, Stephan et al. (2006): *The DELOS Testbed for Choosing a Digital Preservation Strategy*. In: Lecture Notes in Computer Science: Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL 2006). Berlin, Heidelberg: Springer. S. 323-332.

Strodl, Stephan et al. (2007): *How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure*. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries. 2007. S. 29 - 38.

13 Tools

13.1 Einführung

Stefan Strathmann

Die Langzeitarchivierung digitaler Objekte ist eine überwältigend große Herausforderung.

Viele Gedächtnisinstitutionen verfügen über umfangreiche digitale Bestände, die sie auch künftig für Ihre Nutzer bereitstellen möchten. Es liegt auf der Hand, dass die vielen Arbeitsschritte, die durchgeführt werden müssen um eine sichere und langfristige Bereitstellung zu gewährleisten, möglichst nicht manuell erledigt werden sollten. Die digitale Langzeitarchivierung ist dringend auf automatisierte oder zumindest technik-gestützte Abläufe angewiesen.

Schon bei der Planung der digitalen LZA können computerbasierte Werkzeuge die Aufgaben erheblich erleichtern. Die dann später auf diese Planungen aufbauende Praxis der LZA ist ohne automatisierte Abläufe und entsprechende Werkzeuge kaum vorstellbar. Beispielsweise ist die dringend notwendige Erhe-

bung technischer Metadaten ein Prozess, der sich hervorragend zur Automatisierung eignet.

Mit dem Etablieren einer Praxis der digitalen LZA entstehen auch zunehmend mehr Werkzeuge, die genutzt werden können, um die anfallenden Aufgaben automatisiert zu bewältigen. Diese Werkzeuge sind häufig noch in den frühen Stufen der Entwicklung und speziell an die Bedürfnisse der entwickelnden Institution angepaßt. Sie werden aber zumeist zur Nutzung an die LZA-Community weitergegeben und entwickeln sich mit beeindruckender Geschwindigkeit weiter.

Das Kapitel 13 Tools stellt einige der vorhandenen Werkzeuge vor bzw. erläutert deren Benutzung. Insbesondere werden Werkzeuge zur Metadatenextraktion, zum Erstellen von Archivpaketen und zur Planung von LZA-Aktivitäten vorgestellt.

Die Herausgeber wünschen sich, dass dieses Kapitel in den folgenden Neuausgaben des nestor Handbuches deutlich erweitert werden kann.

13.2 Plato

Carmen Heister, Hannes Kulovits, Andreas Rauber

Die Planung digitaler Langzeitarchivierungsmaßnahmen und deren Dokumentation, wie im OAIS Referenzmodell vorgesehen, sowie von der Zertifizierungsinitiativen TRAC und nestor vorgeschrieben, stellen einen relativ komplexen und aufwändigen Prozess dar. Um diesen Ablauf schrittweise zu automatisieren, sowie um Unterstützung beim Durchlaufen der einzelnen Planungsschritte zu bieten, wurde Plato, das Planning Tool entwickelt, welches als Web-Applikation frei verfügbar ist. Plato führt den Anwender durch die einzelnen Schritte des Workflows zur Erstellung eines Langzeitarchivierungsplanes („Preservation Planning"), dokumentiert die Planungskriterien und Entscheidungen, und ermittelt teilautomatisiert die optimale Lösung für die jeweiligen spezifischen Anforderungen einer Institution. In diesem Kapitel wird ein detaillierter Überblick über Plato sowie seine Bedienung gegeben, und vor allem auch auf die bereits integrierten Services verwiesen, welche helfen, den Planungsablauf zu automatisieren.

Einführung

Plato[1] (Planning Tool) ist ein Planungstool, welches im Zuge des EU-Projekt PLANETS[2] entwickelt wurde. Das PLANETS Projekt arbeitet an einer verteilten, serviceorientierten Architektur mit anwendbaren Services und Tools für die digitale Langzeitarchivierung.[3] Das Planungstool implementiert den Planets Workflow zur Planung von Langzeitarchivierung.[4] Es können damit solide Entscheidungen für die Auswahl einer Planungsstrategie getroffen werden, die zu einer optimalen Planung von Langzeitarchivierung der betreffenden digitalen Objekte führt. Wie in Kapitel 20.1 ausführlich beschrieben besteht der PLANETS Preservation Planning Workflow im Kern aus drei Phasen: Die Definition des Planungskontextes (Archivierungsumgebung, Archivierungsgut) sowie der Anforderungen, die Auswahl und Evaluierung potentieller Maßnahmen („actions") anhand gewählter Beispielobjekte, sowie die Analyse der daraus resultierenden Ergebnisse. All diese Schritte werden mit Hilfe der Web-Applikation Plato unterstützt, um einzelne Prozess-Schritte zu automatisieren, sowie

1 http://olymp.ifs.tuwien.ac.at:8080/plato/website/intro.html
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

2 http://www.planets-project.eu/

3 Farquhar, 2007

4 Strodl 2007

um eine automatische Dokumentation jeden Schrittes sicherzustellen.[5] In Plato ist es außerdem möglich einen Aktionsplan („Preservation Action Plan“) zu erstellen, der auf die in der dritten Phase erhaltenen empirischen Ergebnisse aufbaut und einen aufführbaren Workflow zur Durchführung der Langzeitarchivierungsschritte beinhaltet.

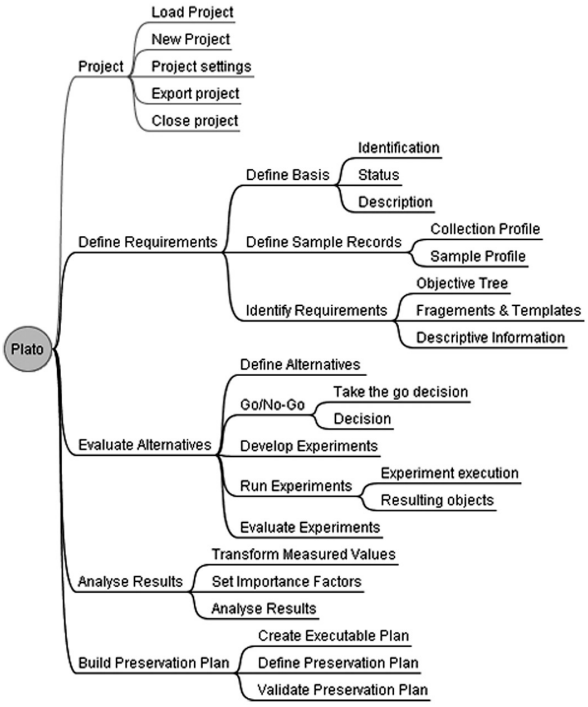


Abbildung 1: Aufbau von Plato

Das Ergebnis des Planungsdurchlaufs mit Plato ist ein Preservation Plan, der für eine konkrete Gruppe von digitalen Objekten die optimale Langzeitarchivierungsmaßnahme (samt Begründung für deren Auswahl) dokumentiert und entsprechende Anleitungen zur Durchführung der Maßnahme sowie deren erneute Evaluierung definiert. Dieser Plan wird in einer Registry abgespeichert, kann aber auch lokal als XML und PDF Dokument abgelegt und somit ebenfalls in ein Langzeitarchiv übernommen werden.

Plato ist über den link http://www.ifs.tuwien.ac.at/dp/plato als Web-Applikation frei verfügbar. Auf der Startseite wird eingangs informiert, was Plato ist

5 Becker, 2008a

und welche Neuerungen in der Entwicklung von Plato hinzugekommen sind. Ein Register weißt zudem auf weiterführende Literatur *(„Documentation")*, Fallstudien *(„Case Studies")* und Veranstaltungen *(„Events")* hin, auf denen Plato vorgestellt und präsentiert wird und wurde. Auf der „Documentation"-Webseite wird eine Liste einführender Literatur zu Plato und dem Planungsworkflow angeboten. Außerdem werden alle wissenschaftlichen Publikationen, die zu Plato veröffentlicht wurden, sowie die Projektberichte zur Verfügung gestellt. Auf der „Case Studies"-Webseite kann Einblick in fertig gestellte Beispielpläne genommen werden. Unter anderem sind hier Case Studies zur Erhaltung von Video Spielen, Interaktive Multimediale Kunst und elektronische Diplomarbeiten und Dissertationen zu finden. Diese können als hilfreiche Vorlage für einen eigenen Preservation Plan dienen. Bei der Entwicklung von Plato wurde besonders auf eine benutzerfreundliche Bedienung im Web-Interface geachtet, die auf allen gängigen Browsern immer wieder ausführlich getestet wird.

Die Schritte in Plato

Um einen eigenen Preservation Plan in Plato zu erstellen, muss sich der Anwender als erstes ein Konto *(„Account")* anlegen. Nach erfolgreicher Anmeldung öffnet sich eine Seite, die vorab die Möglichkeit bietet, einen existierenden Plan aus einer angebotenen Liste zu öffnen, einen neuen Plan zu kreieren, einen „Demo-Plan" zu erstellen oder aber einen schon existierenden Plan in Plato zu importieren. Der Demo-Plan dient zum Testen von Plato. Es kann hierbei durch einen fertig gestellten Plan beliebig durchgeklickt und auch verändert werden. Abbildung 1 bietet einen Überblick über die gesamte Menüstruktur von Plato und die einzelnen Phasen des Planungsprozesses, die in den folgenden Abschnitten detailliert erläutert werden.

Um einen neuen Plan zu erstellen, muss als erstes der Bestand, für den er erstellt werden soll, definiert werden. Üblicherweise handelt es sich dabei um eine mehr oder weniger konsistente Sammlung von digitalen Objekten, die mit Hilfe einer bestimmten Langzeitarchivierungsmaßnahme (z.B. einem bestimmten Migrationstool) behandelt werden sollen, da sie konsistente technische (z.B. Dateiformat, Struktur, Metadaten) und oft auch konzeptionelle Eigenschaften (Verwendungszweck, Zielgruppe) aufweisen. Zudem sollten die Risiken für die Langzeitarchivierung im Vorhinein bekannt sein, welchen mit Hilfe des Preservation Plans begegnet werden soll. Die aufklappbare Navigationsleiste im oberen Bereich des Bildschirms gibt im ersten Menüpunkt die Möglichkeit das Planungsvorhaben zu verwalten. Die weiteren Menüpunkte stehen für die einzelnen Phasen der Planungsworkflows. Der Übersichtlichkeit halber wurden die

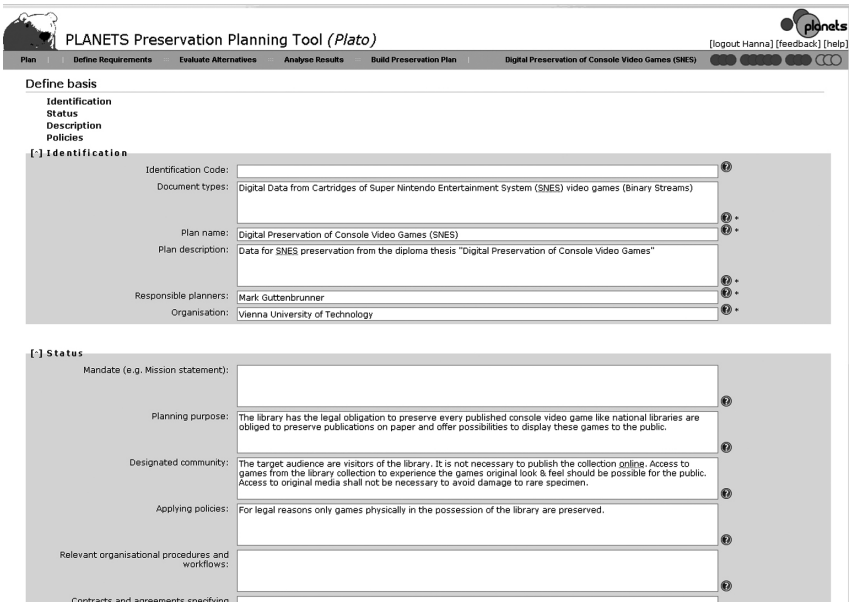
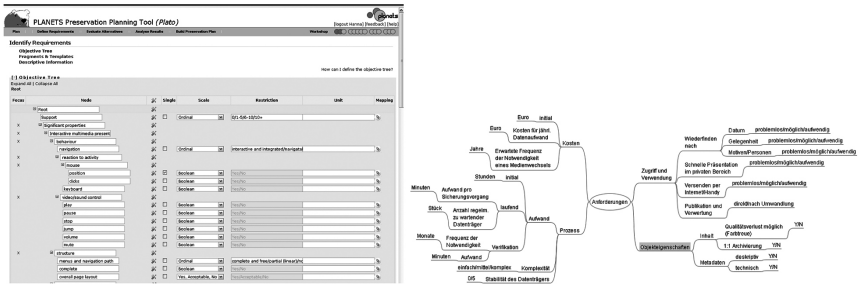


Abbildung 2: Phase 1/ Schritt 1 in Plato

Begrifflichkeiten des Planungsworkflow in der Navigationsleiste übernommen. Auf der rechten Seite der Navigationsleiste zeigt ein Verlaufsanzeiger in Form von gefüllten Kreisen den Status des Planes. Wurde mit den Planungsphasen angefangen, kann leicht durch die einzelnen fertiggestellten Schritte navigiert werden. Es sollte jedoch bei Änderungen in vorhergehenden Phasen darauf geachtet werden, dass diese gespeichert werden. Wird eine Änderung in einer vorhergehenden Phase oder in einem vorhergehenden Schritt einer Phase durchgeführt und diese dann auch gespeichert, wird der Status des Planungsprozesses bis zu dieser Änderung zurückgesetzt, da sich die Voraussetzungen bis zu diesem Status verändert haben und so die nachfolgenden Schritte dementsprechend angepasst werden müssen. Dies bedeutet aber nicht, dass alle nachfolgenden Informationen automatisch gelöscht werden. Damit Änderungen am Preservation Plan jederzeit nachvollzogen werden können, protokolliert Plato intern die letzte Änderung mit. Diese wird mit Datum und dem Benutzernamen der Person, die die Änderung vorgenommen hat, gespeichert und kann im Analyseschritt eingesehen werden.

Phase 1: Definition der Anforderungen

Im zweiten Menüpunkt der Navigationsleiste wird die erste Phase des Planungsworkflows „Festlegen der Anforderungen" („Define requirements") in einzelne Untermenüpunkte, die den einzelnen Schritten innerhalb der Phasen entsprechen, aufgeschlüsselt. In „Define requirements" sollen im ersten Schritt („Define basis") Informationen und Daten zum Planungsvorhaben sowie zum Planungskontext dokumentiert werden (Abbildung 2). Dies beinhaltet zum einen die Dokumentation über den Plan selber („Identifikation"), z.B. wer der Planungsbeauftragte ist und um welche Dokumentenarten es sich handelt. Zum anderen sollen der Status, angewendete Rahmenbedingungen („Policies"), die Zielgruppe und das Mandat (z.B. gesetzliche Verpflichtungen) erfasst werden. Außerdem werden hier die Auslöser („Trigger"), deretwegen dieser Plan erstellt wird, vermerkt. Hierzu ist eine Reihe von Auslösern vordefiniert, wie z.B. die Behandlung eines neuen Bestandes, ein geändertes Langzeitarchivierungsrisiko für ein bestehendes Dateiformat, neue Anforderungen von Seiten der Anwender, etc.



(a) (b)

Abbildung 3: Anforderungsbaum (a) in Plato und (b) als Mindmap

Im zweiten Schritt werden repräsentative Beispielobjekte vom Anwender ausgewählt und in Plato hochgeladen und gespeichert. Hier werden konkret einzelne Objekte aus dem Bestand (oder aus einer Sammlung von Referenzobjekten) ausgewählt, anhand derer die jeweiligen Tools zur Langzeitarchivierung getestet werden sollen. Bei der Auswahl sollte darauf geachtet werden, dass man das Spektrum der technischen und intellektuellen Eigenschaften der Objekte innerhalb des Bestandes erfasst, also z.B. sowohl ein sehr kleines als auch ein sehr großes Objekt auswählt, eines mit Makros, Bildern, mit bestimmten Formatierungen, etc. Anschließend muss beschrieben werden, um welche Objekte es sich dabei handelt – sowohl intellektuell als auch technisch. Für die Beschreibung

der technischen Eigenschaften bietet Plato automatische Unterstützung. Die Formate der Beispielobjekte werden durch den in Plato integrierten Identifizierungsservice DROID[6] automatisch identifiziert und mit Informationen zum PUID (Pronom Persistent Unique Identifier)[7], zum Namen des Formats, der Version sowie des MIME-Type (Multipurpose Internet Mail Extensions-Type)[8] im Plan gespeichert. Dazu werden die Dateien via Webservice an DROID geschickt, welches entsprechende technische Metadaten erhebt und zurückliefert, die daraufhin in den Preservation Plan in Plato übernommen werden. (An der Integration weiterer Analysewerkzeuge für detailliertere Beschreibungen wird derzeit gearbeitet.) Darüber hinaus soll die ursprüngliche technische Umgebung möglichst genau beschrieben werden (verwendete Software, Betriebssystem sowie Art der Verwendung).

Im nächsten Schritt („Identify Requirements“) lässt das Tool den Anwender die Anforderungen zur Planung der Langzeitarchivierung definieren. Dies ist einer der aufwändigsten Schritte in der Erstellung des Plans. Es sollte gewährleistet sein, dass möglichst die Sicht aller Stakeholder (Anwender, Techniker und Archivexperten) in diesem Schritt berücksichtigt wird. Deshalb bietet sich an, die Liste der Anforderungen in einem Workshop zu erstellen. Meist wird in diesen Workshops mit Post-it Notes oder mit Mind-Mapping Software gearbeitet und die Liste in Form eines Baumes strukturiert, um die einzelnen Anforderungen nach inhaltlichen Gesichtspunkten zu strukturieren (Abbildung 3 (b)). Mindmaps, die in der frei verfügbaren Software FreeMind[9] oder in Mindmeister[10] erstellt wurden, können in Plato importiert angezeigt werden. Ferner können die Kriterienbäume natürlich auch innerhalb von Plato mit Hilfe des Web Interface editiert werden (Abbildung13.2.3 (a)). Plato bietet außerdem eine Bibliothek mit Vorlagen in „*Show the template library*“, die unterteilt ist in „Öffentliche Vorlagen“ *(„Public Templates*“), „Eigene Vorlagen“ *(„My Templates*“), „Öffentliche Fragmente“ *(„Public Fragements“)* und „Eigene Fragmente“ *(„My Fragements“).* Diese enthalten vordefinierte Zweige von Kriterien, die in verschiedenen Standardszenarien immer wieder auftauchen und daher nicht jedes Mal von neuem manuell definiert werden müssen, sondern einfach aus der Bibliothek übernommen werden können. Plato beinhaltet öffentlich verfügbare Templates beispielsweise für die Langzeitarchivierung von Diplomarbeiten und Dis-

6 http://droid.sourceforge.net

7 http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm

8 http://www.iana.org/assignments/media-types/

9 http://freemind.sourceforge.net/wiki/index.php/Main_Page

10 http://www.mindmeister.com/ Webversion eines Mindmapping Tools, welches Mindmaps als FreeMind File importiert und exportiert.

sertationen und die Langzeitarchivierung von Internetseiten. Erarbeitet wurden diese Templates aus verschiedenen umfangreichen Case Studies und beinhalten detaillierte Anforderungen an die Langzeitarchivierung der jeweiligen Sammlung. Die Vorlage für die Langzeitarchivierung von Internetseiten enthält genaue Anforderungen an das Aussehen, den Inhalt, die Struktur und das Verhalten von Webseiten. Wird ein Template als Anforderungsbaum übernommen kann dieser problemlos angepasst werden – Teilbäume, die nicht zutreffen, gelöscht und weitere Anforderungen eingefügt werden. Es existiert ebenfalls eine allgemeine Vorlage, die der vorgeschlagenen Grundstruktur eines Anforderungsbaumes entspricht: „Objekteigenschaften", „Technische Eigenschaften", „Infrastruktureigenschaften", „Prozesseigenschaften" und „Kontext". Diese kann herangezogen werden, wenn der Anforderungsbaum von Grund auf neu erstellt werden soll. Es gibt außerdem die Möglichkeit, die Bibliothek mit eigenen Fragmenten oder Vorlagen zu erweitern um sie so an wiederkehrende Anforderungen in der eigenen Institution anzupassen.

In letzter Konsequenz soll mit Hilfe von Plato objektiv ermittelt werden, wie gut einzelne Tools diese Kriterien erfüllen. Zu diesem Zweck muss jedem einzelnen Kriterium nach Möglichkeit ein objektiver Messwert zugewiesen werden. So kann zum Beispiel der Durchsatz bei Migrationstools in MB pro Sekunde gemessen werden; die Bewahrung der eingebetteten Metadaten in einem Objekt mit „Ja / Nein"; die Verfügbarkeit einer Dateiformatdefinition als „freier Standard", „Industriestandard", „proprietäres Format". Das Tool bietet eine Vielzahl von Messskalen („Boolean", „Ordinal", „Yes", „Acceptable", „No", „Integer", „Number" etc.) an, die unabhängig ausgewählt und genutzt werden können.

Am Ende der ersten Phase sind somit die Anforderungen an die optimale Lösung für den gesuchten Preservation Plan definiert, sowie Beispielobjekte ausgewählt, anhand derer einzelne Tools getestet werden sollen.

Phase 2: Evaluierung der Alternativen

In der zweiten Phase „Evaluierung der Anforderungen" *(„Evaluate alternatives")* kann der Anwender Langzeiterhaltungsmaßnahmen definieren, welche er überprüfen beziehungsweise testen will. Alternativen sind hierbei Tools *(„preservation action services")*, die den gewünschten Endzustand des Beispielobjektes erzeugen sollen. Dazu können Tools aus den verschiedensten digitalen Erhaltungsstrategien („Migration", „Emulation", „Beibehaltung des Status quo") (Kapitel 8) verglichen werden. Bei Textdateien kann beispielsweise Formatmigration oder die Beibehaltung des Status quo evaluiert werden. In anderen Fällen, beispiels-

weise bei Videospielen wird eher in Richtung Emulation der Systemumgebung evaluiert. Es können auch sämtliche Erhaltungsstrategien in einem Plan evaluiert werden.

Die Auffindung passender Alternativen kann sich je nach Bestand unterschiedlich aufwändig gestalten. Häufig müssen intensive Recherchephasen eingeplant werden, um herauszufinden, welche Tools überhaupt für die gewünschte Erhaltungsstrategie und den betreffenden Objekttyp derzeit verfügbar sind. Bei der Recherche nach geeigneten Tools können Service Registries helfen, die Tools für die Langzeitarchivierung zu listen. In Plato wurden deshalb *„Service Registries"* wie „CRIB"[11], „Planets Service Registry" oder „Planets Preservation Action Tool Registry" implementiert, welche die Suche nach geeigneten Tools automatisch durchführen und diese dem Anwender vorschlagen. Dazu wird in den Registries nach Tools gesucht, die auf den vorliegenden Beispieldateitypen operieren können. Je nach Art der Registry werden dabei auch komplexere Lösungen wie z.B.: Migrationspfade in mehreren Schritten (von TeX (Typesetting System) über DVI (Device Independent File Format) zu PDF (Portable File Document)) ermittelt. Der Anwender kann dann entscheiden, welche Vorschläge er übernehmen will. Will der Anwender Tools testen, die nicht von den Service Registries vorgeschlagen wurden, können diese manuell angegeben werden. Der Nachteil ist hierbei, dass diese Tools lokal installiert und gemessen werden müssen.

Im nächsten Schritt *„Go/No-Go"* gibt Plato erneut die Möglichkeit zu überlegen, welche der aufgelisteten Alternativen im Planungsprozess evaluiert werden sollen. In diesem Schritt sind Alternativen abwählbar, die beispielsweise interessant wären, aber in der Anschaffung zu kostspielig sind. Andererseits kann auch die Evaluierung eines bestimmten Tools vorerst aufgeschoben werden *(„Deferred go"),* samt Definition, wann bzw. unter welchen Bedingungen die Evaluierung nachgeholt werden soll, sofern z.B. ein bestimmtes Tool erst in naher Zukunft verfügbar sein wird. Die Gründe, die für oder gegen eine Alternative sprechen, können in Plato dokumentiert werden.

Bevor die Experimente zu den einzelnen Alternativen durchgeführt werden, muss der Anwender für jede einzelne Alternative die Konfiguration der Tools definieren und dokumentieren. Sind die Alternativen aus den Service Registries entnommen, erfolgt die Beschreibung automatisch. Sind die Szenarien für die einzelnen Experimente der einzelnen Alternativen und deren Rahmen (Personal, Tools etc.) vollständig und dokumentiert, können die Experimente im Schritt „Run Experiment" durchgeführt werden. Die Ausführung erfolgt

11 http://crib.dsi.uminho.pt/

wieder automatisch, sofern der Anwender die Services von den in Plato angebotenen Service Registries genutzt hat. Hierbei werden die einzelnen Beispielobjekte an die Webservices geschickt, die diese je nach Erhaltungsstrategie migrieren oder in einem Emulator wie z.B. GRATE[12] (andere Emulatoren können manuell aufgerufen werden) emulieren. Teilweise werden Messungen (Zeit, etc.) automatisch erhoben und Logfiles wie auch Fehlermeldungen übernommen. Die entstandenen Ergebnis-Dateien im Falle eines Migrationsprozesses können heruntergeladen werden. Die Ergebnisdateien werden in Plato gespeichert und bilden – gemeinsam mit den ursprünglichen Beispielobjekten – Teil des Plans und der Dokumentation der Experimente, die es erlauben, die Evaluierung jederzeit zu einem späteren Zeitpunkt zu wiederholen bzw. alte Ergebnisse mit jenen neuer Tools zu vergleichen. Bei manuellen Experimenten muss der Anwender die Tools mit den Beispielobjekten selbst aufrufen und die Experimente selbst durchführen sowie die Ergebnisse hochladen, so dass auch diese in Plato gespeichert sind.

Im fünften Schritt („Evaluate Experiments“) der zweiten Phase werden die Experimente auf Basis der Kriterien der Anforderungsliste bzw. des Anforderungsbaum evaluiert. Es werden hierbei für jedes einzelne Ergebnis (also z.B.: für jedes Migrationsergebnis eines jeden Beispielobjekts mit jedem einzelnen Tool) alle Kriterien des Anforderungsbaumes auf Blattebene evaluiert, um die Ergebnisse der einzelnen Experimente empirisch für jede Alternative zu erheben. Auch hier kann mit Hilfe von automatischen Tools („Preservation characterization tools“) ein Teil der Arbeit automatisiert werden. Diese „Characterization Tools“ analysieren den Inhalt der Dateien und erstellen eine abstrakte Beschreibung, die es erlaubt, in vielen Bereichen die Unterschiede vor und nach der Migration zu erheben. Beispiele für solche Beschreibungssprachen sind JHOVE[13] oder XCDL[14]. (An Tools, die einen automatischen Vergleich von Emulationsergebnissen erlauben, wird derzeit gearbeitet). Werte, die nicht automatisch erhoben werden können (wie z.B. eine subjektive Beurteilung des Qualitätsverlustes bei Kompressionsverfahren in der Videomigration), müssen manuell ins System eingegeben werden.

Am Ende der zweiten Phase ist somit für jedes einzelne Beispielobjekt bekannt, wie gut jedes einzelne der Preservation Action Tools die im Kriterienbaum definierten Anforderungen erfüllt.

---

12 http://planets.ruf.uni-freiburg.de/

13 http://hul.harvard.edu/jhove/

14 Becker, 2008

Phase 3: Analyse der Ergebnisse

Die dritte Phase „Consider results“ zielt nun darauf ab, die Ergebnisse aus den Experimenten zu aggregieren, um die optimale Lösung auszuwählen. Um dies zu tun, muss der Erfüllungsgrad der einzelnen Anforderungen durch die verschiedenen Tools erfasst und verglichen werden. Nachdem die Maßzahlen allerdings in den unterschiedlichsten Einheiten erhoben wurden, müssen diese zuerst in eine einheitliche Skala, den sogenannten Nutzwert („Utility value“), transformiert werden. Dazu werden Transformationsskalen festgelegt, welche die aufgetretenen Messwerte jeweils auf einen einheitlichen Wertebereich (z.B. angelehnt an das Schulnotensystem zwischen null und fünf) festlegen. Der Wert „null“ steht für ein unakzeptables Ergebnis, welches, kommt er in einer Anforderung zu einer Alternative vor, dazu führt, dass diese Alternative ausgeschlossen wird. Andererseits bedeutet „fünf“ die bestmöglichste Erfüllung der Anforderung. Beispielsweise kann eine Bearbeitungszeit von 0-3 Millisekunden pro Objekt mit dem Wert „fünf“ belegt werden, 3-10ms mit „vier“, 10-50ms mit „drei“, 50-100ms mit „zwei“, 100-250ms mit „eins“, und jeder Wert über 250ms als unakzeptabler Wert mit „null“ definiert werden. „Ja/nein“ Messwerte können entweder auf „fünf/eins“ oder „fünf/null“ abgebildet werden, je nachdem, ob die Nichterfüllung einen Ausschließungsgrund darstellen würde oder nicht.

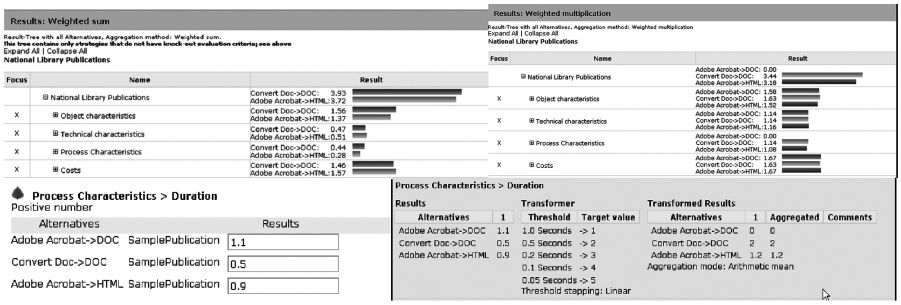


Abbildung 4: Evaluierungsergebnisse elektronischer Dokumente

Nachdem nun alle Messwerte in einheitliche Nutzwerte transformiert worden sind, kommt der optionale Schritt der Gewichtung. In der Grundeinstellung werden alle Kriterien innerhalb einer Ebene des Kriterienbaumes als gleich wichtig betrachtet. Sollte es nun der Fall sein, dass für das Planungskonsortium manche Kriterien von essentieller Bedeutung sind, während andere nur eine untergeordnete Rolle spielen, kann in diesem Schritt jedem Kriterium ein ei-

genes Gewicht relativ zu den anderen Kriterien gegeben werden. So kann z.B. der Erhalt des visuellen Erscheinungsbildes eine viel höhere Bedeutung haben als z.B. der Durchsatz, wenn die Anzahl der zu bearbeitenden Dateien nicht immens groß ist. In diesem Fall können die Prozessmesswerte geringer gewichtet werden, während die entsprechenden Kriterien die das Aussehen und den Erhalt der Funktionalität messen, höher gewichtet werden. In Plato kann dies mit Hilfe von Schiebereglern erfolgen, wo für jede Ebene des Baums die relative Gewichtung jedes einzelnen Knotens nach Wunsch angepasst und fixiert werden kann. Nicht veränderte Gewichte werden danach automatisch entsprechend angepasst.

Sind diese Schritte durchgeführt, kann Plato einen kumulierten Nutzwert für jede Alternative ausrechnen, d.h. wie gut jede einzelne Alternative die Gesamtheit aller Kriterien erfüllt. In der Folge können die Alternativen nach Gewichtung gereiht werden. Dazu stehen eine Reihe von Aggregierungsfunktionen zur Verfügung, von denen üblicherweise zwei im Kern relevant sind, nämlich die additive und die multiplikative Zusammenführung. Letztere zeichnet sich dadurch aus, dass ein Tool, das in einem einzigen Kriterium eine nicht akzeptable Leistung aufweist (also einmal den Nutzwert „null" zugewiesen bekam), auch im Gesamtranking mit „null" gewichtet wird und somit aus der Evaluierung ausscheidet. Hier kann so noch einmal gesondert die Beurteilung der einzelnen Messungen hinterfragt und angepasst werden.

Plato bietet dem Anwender dazu auch eine graphische Darstellung der Ergebnisse, damit die spezifischen Stärken und Schwächen jeder einzelnen potentiellen Maßnahme *„Preservation action"* vom Anwender auf Anhieb gesehen werden können. Abbildung 4 zeigt die Darstellung des Endergebnisses in Plato aus einem Planungsprozess zur Langzeitarchivierung wissenschaftlicher Arbeiten, die ursprünglich in PDF vorliegen, ähnlich dem Beispiel in (Becker 2007b) (Siehe dazu auch Abb. 20.2.4 im Kapitel 20.1, wo Ergebnisse einer ähnlichen Studie in tabellarischer Form zusammengefasst wurden.) Als Alternativen wird eine Reihe von Migrationstools evaluiert. Ferner wird zusätzlich die Null-Hypothese evaluiert, d.h. das Resultat unter der Annahme, dass man keine Langzeitarchivierungsmaßname setzt. In der Abbildung 4 sind die Nutzwerte unter Verwendung der beiden Aggregationsmethoden „Gewichtete Summe" und „Gewichtete Multiplikation", wie in Kapitel 20.1 beschrieben, dargestellt. Innerhalb von Plato kann man zwischen den beiden Aggregationsmethoden wechseln und der Ursache für die unterschiedlichen Rankings auf den Grund gehen. Indem der Baum expandiert wird, kann der Anwender erkennen, in welchen Kriterien die Leistung der einzelnen Tools mit „nicht akzeptabel" bewertet wurde.

Es wird außerdem erkennbar, dass die Alternativen „Migration in RTF (Rich Text Format) mit Adobe Acrobat“ und „Migration in TXT (Text File) mit Adobe Acrobat“ bei den Kriterien *„Apearance“* – „*Structure“* – „*Structure Tables“* und *„Conten“*„ – „*Figure Content*“ jeweils mit „Null“ bewertet wurden. Die „ConvertDoc Migration in RTF“ scheidet wiederum z.B. im Kriterium „*Technical Characteristics*“ – „*Tool“* – „*Makrosupport*“ aus. Die Null-Hypothese PDF *(„unchanged“)* scheidet bei der Aggregationsmethode Multiplikation aus, da das essentielle Kriterium der Verhinderung von eingebetteten Skripts „*Behaviour“* – *„Script blocking*“ nicht erfüllt wird. Durch Klicken auf das jeweilige Kriterium kann unmittelbar zu den entsprechenden Messwerten gesprungen werden. Hier können dann die Gründe für die unterschiedlichen Bewertungen nachvollzogen werden, sowie zu jedem späteren Zeitpunkt die entsprechend migrierten Dokumente geöffnet und deren Bewertung verglichen werden.

Am Ende der dritten Phase liegt nun eine Reihung der einzelnen alternativen „Preservation Action Tools“ vor, die es erlaubt, das am besten geeignete Tool auszuwählen sowie zu begründen, warum dieses Tool besser ist als die anderen. Darüber hinaus kann evaluiert werden, in welchen Bereichen es Schwächen aufweist, und so eine eventuelle Kombinationsstrategie empfohlen werden, d.h. es können unter Umständen zwei Tools kombiniert werden, von denen eines eher das Aussehen, das andere die interne Struktur und den Inhalt bewahrt, oder beispielsweise Elemente (z.B. Metadaten), die bei einer anderweitig hervorragenden Migration verloren gehen, durch separate Extraktion und Speicherung gerettet werden.

Phase 4: Aufbau des Durchführungsplans

Nachdem sich der Anwender am Ende der dritten Phase auf Basis der Analyseergebnisse für eine Alternative entschieden hat, erfolgt die Erstellung des Preservation Plans in der vierten Phase („Build Preservation Plan“). Diese umfasst nicht mehr den eigentlichen Planungsprozess, sondern die Vorbereitung der operativen Umsetzung eines Plans nach dessen Genehmigung. Sie wird hier daher nur verkürzt beschrieben. In dieser vierten Phase werden die notwendigen organisatorischen Maßnahmen definiert, die zur Integration der Erhaltungsmaßnahmen in die Organisation notwendig sind, dazu gehören ein detaillierter Arbeitsplan mit definierten Verantwortungen und Ressourcenzuteilungen zur Installation von notwendiger Hardware und Software. Zusätzlich werden Kosten und Überwachungskriterien für die Erhaltungsmaßnahmen definiert bzw. berechnet.

Im ersten Schritt der vierten Phase erstellt der Anwender einen Arbeitsplan inklusive der technischen Einstellungen wie den Speicherort der Daten, auf die die Maßnahme angewendet werden soll sowie die dafür notwendigen Parametereinstellungen für das Tool. Für die Qualitätssicherung werden Mechanismen geplant, welche die Qualität des Ergebnisses der Maßnahme überprüfen. Der zweite Schritt der Planerstellung beschäftigt sich mit den Kosten der getroffenen Erhaltungsmaßnahmen und der Überwachung des Planes. Die Kosten können entweder nach dem LIFE Kostenmodell[15] oder dem TCO Modell[16] (**T**otal **C**ost of **O**wnership Modell) aufgeschlüsselt werden. Um die laufende Aktualität des Planes sicherzustellen, werden Überwachungskriterien *(„Trigger conditions")* definiert, die festlegen, wann der Plan neu überprüft werden muss. Beispielsweise kann eine geänderte Objektsammlung eine Überprüfung erfordern um eventuell neu zutreffende Alternativen berücksichtigen zu können. Der letzte Schritt zeigt dann den vollständigen Preservation Plan mit empfohlenen Maßnahmen zur Erhaltung einer Sammlung von digitalen Objekten. Nachdem der Plan einer letzten Prüfung unterzogen wurde, wird er von einer berechtigten Person in Plato bewilligt und damit von diesem Zeitpunkt an als gültig festgelegt.

Der Preservation Plan

Im letzten Schritt der vierten Phase gibt Plato den gesamten Preservation Plan aus, welcher dann zur Archivierung als PDF exportiert werden kann. Der Preservation Plan enthält alle Informationen, die der Anwender eingegeben hat, sowie die Ergebnisse der Evaluierung der einzelnen Alternativen als Balkendiagramme. Die Evaluierungsergebnisse der Alternativen werden ebenfalls in einer Baumstruktur dargestellt, wodurch diese zu allen Anforderungen auf allen Ebenen angezeigt werden können. Der Preservation Plan ist wie folgt aufgebaut:

- Identifikation des Planes
- Beschreibung der organisatorischen Einrichtung
- Auflistung aller Anforderungen
- Beschreibung der Alternativen
- Aufbau der Experimente
- Evaluierung der Experimente
- Transformationstabellen
- Resultate (Summe und Multiplikation)

15 Shenton, 2007

16 http://amt.gartner.com/TCO/index.htm

```
<plan>
  <basis>
  </basis>
  <sampleRecords>
    <record shortName="Publikation"
        fullname="publ.pdf"
        contentType="application/pdf">
      <data>JVBERi0xLjQNJeLjz9MNCjc2IDAgb2
            IDw8L0xpbmVhcml6ZWQgMS9MIDQ4Nzg
            yNC9PIDc5L0Ug...
      </data>
      <formatInfo puid="fmt/18"
         name="Portable Document Format"
         version="1.4"
         mimeType="application/pdf"
         defaultExtension="pdf">
    </record>
    <record>
      ...
    </record>
  </sampleRecords>
  <alternatives>
    <alternative discarded="false"
        name="Adobe Acrobat to DOC">
      <description>...</description>
      <experiment>
        <description>...</description>
        <runDescription></runDescription>
        <uploads>
          <upload fullname="publ.doc"
            contentType="application/msword">
            <data>e1xydGYxXGFuc2lcYW5zaWNwZzEy
                  NTJcdWMxXGRlZmYIHtcZm9udHRib
                  HtcZjBcZnN3aXNzXGZj...
            </data>
          </upload>
        </uploads>
      </experiment>
    </alternative>
    <alternative>...</alternative>
  </alternatives>
  <decision>...</decision>
  <tree>
     ...
     <node>
       <leaf name="Encoding" weight="0.4">
         <ordinalScale unit=""/>
         <ordinalTransformer>
           <mappings>
             <mapping ordinal="Original"
                target="5.0"/>
             <mapping ordinal="Changed"
                target="3.0"/>
             <mapping ordinal="None"
                target="1.0"/>
         </ordinalTransformer>
         <evaluation>
           <alternative>...</alternative>
           <alternative>...</alternative>
         </evaluation>
     ...
     </node>
```

⇒ Auflistung der Beispielobjekte auf Basis derer die Experimente durchgeführt wurden. Dieser Block enthält die gesamte Datei inklusive Metainformation und Inhalt (uuencoded).

⇒

Auflistung der gewählten Alternativen und den dazugehörigen Experimenten.

Die Experimente enthalten neben der genauen Beschreibung auch die Ergebnisdateien (Resultate) als upload-Block. Ebenfalls genau beschrieben mit Metadaten und Inhalt (uuencoded).

⇒

Der <tree> Block stellt den umfangreichsten im XML Dokument dar. Er enthält:

- Die einzelnen Knoten und Blätter (Anforderungen)
- Die gewählte Skala der Anforderung
- Die Gewichtungen der einzelnen Knoten und Blätter
- Die Transformationstabellen
- Die Evaluierung pro Alternative.

- Entscheidung für eine Strategie
- Kosten
- Überwachung
- Bewilligung

Plato unterstützt auch den Export des Preservation Plans als XML Datei, welche einem definierten, öffentlich verfügbaren Schema entspricht. Diese Datei enthält alle Daten um den Preservation Plan auf einem anderen System reproduzieren zu können. Neben der Basisinformation, die der Benutzer während der Planerstellung eingegeben hat, sind auch alle Beispielobjekte, auf Basis derer die Evaluierung erfolgte, in die XML Datei eingebettet. Ebenfalls enthalten sind Metadaten über diese Beispielobjekte (z.B. Pronom Unique Identifier), die detaillierten Transformationstabellen, Evaluierungsergebnisse der einzelnen Experimente und die Ergebnisse. Die XML Datei ist wie folgt aufgebaut:

Zusammenfassung

Um einen Preservation Plan in Plato zu erstellen bedarf es viel Erfahrung. In diesem Kapitel wurde das Planungstool Plato vorgestellt, das Institutionen bei der Erstellung von Langzeitarchivierungsplänen unterstützt, die optimal auf ihre Bedürfnisse zugeschnitten sind. Plato implementiert den Planungsprozess, wie er in Kapitel 20.1 vorgestellt wird. Neben der automatischen Dokumentation aller Planungsschritte sowie der durchgeführten Experimente unterstützt es den Prozess vor allem durch die Integration von Services, welche Schritte wie die Beschreibung der ausgewählten Objekte, das Auffinden geeigneter Tools oder die Durchführung und Analyse der Ergebnisse von Langzeitarchivierungsmaßnahmen automatisieren. Durch den Zwang zu exakten Definitionen zu den zu bewahrenden Eigenschaften *(„Significant properties")* (und damit auch automatisch jener Aspekte, die vernachlässigt werden können bzw. verloren gehen dürfen) sowie der Anforderungen an den Langzeitarchivierungsprozess selbst bietet die Erstellung des Kriterienbaumes *(„Objective tree")* einen enormen Verständnisgewinn. Hierbei wird häufig erstmals bewusst und offensichtlich, was digitale Langzeitarchivierung insgesamt bedeutet. Der Anwender muss (und wird dadurch) ein Verständnis für die spezifischen Eigenschaften des zu archivierenden Bestandes entwickeln, um richtige Anforderungen und Entscheidungen treffen zu können.

Plato ist ein Planungstool, welches laufend weiterentwickelt wird. Erweiterungen betreffen vor allem die Einbindung zusätzlicher Services, die einzelne Schritte innerhalb des Planungsworkflows weiter automatisieren. Darüber

hinaus werden die Library Templates und Fragements laufend durch die Zusammenarbeit mit Bibliotheken, Archiven, Museen und anderen Dokumentationseinrichtungen erweitert. Bisher werden diese nur eingeschränkt zur Verfügung gestellt, da diese sonst zu „selbsterfüllenden Prophezeiungen" führen könnten, weil diese ohne Überarbeitung und kritische Prüfung übernommen werden würden. Zum jetzigen Zeitpunkt wird in laufenden Case Studies überprüft, ob Institutionen gleicher Größe, mit ähnlichen Anforderungen ähnliche Bäume erstellen, die im positiven Falle als Templates verfügbar gemacht werden können.

Um mit Plato selbstständig arbeiten zu können, wurden neben den wissenschaftlichen Veröffentlichungen eine Reihe frei verfügbarer Tutorials, Case Studies[17] und ein Handbuch erstellt, welche unter http://www.ifs.tuwien.ac.at/dp/plato/intro_documentation.html abgerufen werden können. Außerdem bestehen derzeit Überlegungen, bei Bedarf das derzeit nur in Englisch verfügbare Webinterface auch in andere Sprachen zu übersetzen.

17 z.B. Becker, 2007a; Becker 2007b

Literaturverzeichnis

Becker, Christoph et al. (2007a) *Preserving Interactive Multimedia Art: A Case Study in Preservation Planning*. In: Goh, Dion Hoe-Lian et al.: Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. Proceedings of the Tenth Conference on Asian Digital Libraries (ICADL'07). Berlin / Heidelberg: Springer. S. 257-266.

Becker, Christoph et al. (2007b) *Long-Term Preservation of Electronic Theses and Dissertations:* A Case Study in Preservation Planning. In: Proceedings of the Ninth Russian National Research Conference on Digital Libraries: Advanced Methods and Technologies, Digital Collections. 2007.

Becker, Christoph et al. (2008a) *Distributed Preservation Services: Integrating Planning and Actions*. In: Christensen-Dalsgaard, Birte et al.: Research and Advanced Technology for Digital Libraries. Proceedings of the 12th European Conference on Digital Libraries (ECDL'08). Berlin, Heidelberg: Springer-Verlag. S. 25-36.

Becker, Christoph et al. (2008b) *A Generic XML Language for Characterising Objects to Support Digital Preservation*. In: Proceedings of the 2008 ACM symposium on Applied computing. 2008. S. 402-406.

Farquhar, Adam., Hockx-Yu, Helen (2007) *Planets: Integrated services for digital preservation*. In: International Journal of Digital Curation, 2. (2007). S. 88-99.

nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2006): *Kriterienkatalog vertrauenswürdige digitale Langzeiarchive. Version 2*. (nestor-Materialien 8). Frankfurt am Main: nestor. www.langzeitarchivierung.de/downloads/mat/nestor_mat_08.pdf

Shenton, H., et. al.(2007): *How much does it cost? The LIFE Project - Costing Models for Digital Curation and Preservation*. In: LIBER Quarterly. The Journal of European Research Libraries. 17. 2007. http://liber.library.uu.nl/publish/issues/2007-3_4/index.html?000210

Strodl, Stephan et al. (2007) *How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure*. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries. 2007. S. 29 - 38.

13.3 Das JSTOR/Harvard Object Validation Environment[18] (JHOVE)

Stefan E. Funk

Einführung

Wie in den vorangehenden Kapiteln bereits besprochen wurde, ist es für eine langfristige Erhaltung von digitalen Objekten dringend erforderlich, zu wissen und zu dokumentieren, in welchem Dateiformat ein solches digitales Objekt vorliegt. Zu diesem Zweck sind auch Informationen von Nutzen, die über das Wissen über den Typ eines Objekts hinausgehen, vor allem detaillierte technische Informationen. Zu wissen, dass es sich bei einem digitalen Bild um ein TIFF-Dokument in Version 6.0 handelt, reicht evtl. nicht aus für eine sinnvolle Langzeiterhaltung. Hilfreich können später Daten sein wie: Welche Auslösung und Farbtiefe hat das Bild? Ist es komprimiert? Und wenn ja, mit welchem Algorithmus? Solche Informationen – technische Metadaten – können aus den Daten des Objekts selbst (bis zu einem gewissen Grad, welcher vom Format der Datei abhängt) automatisiert extrahiert werden.

Anwendung

Mit JHOVE wird im Folgenden ein Werkzeug beschrieben, das außer einer Charakterisierung einer Datei (Welches Format liegt vor?) und einer Validierung (Handelt es sich um eine valide Datei im Sinne der Format-Spezifikation?) zu guter Letzt auch noch technische Metadaten extrahiert. JHOVE kann entweder mit einem grafischen Frontend genutzt werden – wobei eine Validierung oder Extraktion technischer Metadaten von vielen Dateien nicht möglich ist, oder als Kommandozeilen-Tool. Ebenso kann JHOVE auch direkt als Java-Anwendung in eigene Programme eingebunden werden, was für eine automatisierte Nutzung sinnvoll ist. Letzteres ist jedoch dem erfahrenen Java-Programmierer vorbehalten. Als Einführung wird hier das grafische Frontend kurz erklärt sowie eine Nutzung auf der Kommandozeile beschrieben.

---

18 JHOVE – JSTOR/Harvard Object Validation Environment: http://hul.harvard.edu/jhove/

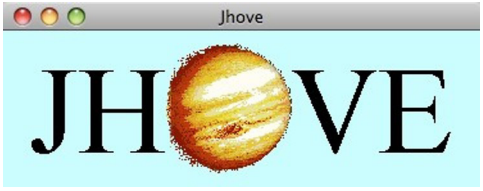
Anforderungen

Für die Nutzung von JHOVE wird eine Java Virtual Machine benötigt, auf der JHOVE Projektseite bei Sourgeforge.net[19] wird Java in Version 1.6.0_12 empfohlen.

Das grafische Frontend JhoveView

Download

Nach dem Herunterladen des .zip oder .tar.gz Paketes von der Sourceforge Projektseite – beschrieben wird hier die Version 1.2 vom 10. Februar 2009 – wird das Paket in ein beliebiges Verzeichnis entpackt. Zum Starten des grafischen Frontends starten Sie bitte das Programm JhoveView.jar im Verzeichnis ./bin/ – entweder durch Doppelklick oder von der Kommandozeile per java -jar bin/JhoveView.jar (nach dem Wechsel in das Verzeichnis, indem sich JHOVE befindet).



Menü-Optionen

Die beiden vorhandenen Menü-Optionen "File" und "Edit" sind schnell erklärt:

- Unter "File" kann eine Datei aus dem Internet oder vom Dateisystem geöffnet werden, das sogleich von JHOVE untersucht wird.
- Unter "Edit" kann gezielt ein JHOVE-Modul gewählt werden, mit dem eine Datein untersucht werden soll. Nicht die Einstellung "(Any)" zu benutzen – für eine automatische Erkennung des Formats – kann zum Beispiel dann Sinn machen, wenn eine TIFF-Datei nicht automatisch als solche erkannt wird, weil sie vielleicht nicht valide ist. Dann kann JHOVE dazu bewegt werden, dieses Bild mit dem TIFF-Modul zu untersuchen, um so eine entsprechende – und weiter helfende – Fehlermeldung zu bekommen. Weiterhin kann hier die Konfigurationsdatei editiert werden (um neue Module einzubinden).

19 http://sourceforge.net/projects/jhove/

Dateien untersuchen

Wählt man nun eine Datei aus, für erste Tests sollten die vorhandenen Module berücksichtigt werden, wird diese Datei von JHOVE untersucht. Im Folgenden wird ein Fenster angezeigt, in dem alle von JHOVE extrahierten Informationen angezeigt werden. Hier kann nach Belieben durch den Baum geklickt werden. An erster Stelle wird das Modul und dessen Versionsnummer angezeigt, mit dem die Datei untersucht wurde. Wird hier als Modulname "BYTESTREAM" angezeigt, heißt das, dass JHOVE kein passendes Modul gefunden hat, das Bytestream-Modul wird dann als Fallback genutzt. Hier hilft es unter Umständen – wie oben erwähnt – das Modul per Hand einzustellen.

JHOVE Ausgaben anzeigen und speichern

Die Speicheroption, die nun zur Verfügung steht, kann genutzt werden, um die Ergebnisse wahlweise als Text oder als XML zu speichern und in einem anderen Programm zu nutzen/anzusehen. So können die Informationen beispielsweise in einem XML- oder Texteditor bearbeitet oder anderweitig genutzt werden. Im Folgenden ein Beispiel einer Untersuchung einer Textdatei im Zeichensatz UTF-8:

```
JhoveView (Rel. 1.1, 2008-02-21)
       Date: 2009-03-03 10:33:31 CET
       RepresentationInformation:
       /Users/fugu/Desktop/nestor-hand
       buch-kapitel-13_2009-03-03/test.txt
       ReportingModule: UTF8-hul, Rel. 1.3 (2007-08-30)
       LastModified: 2009-03-03 10:33:12 CET
       Size: 64
       Format: UTF-8
       Status: Well-Formed and valid
MIMEtype: text/plain; charset=UTF-8
UTF8Metadata:
       Characters: 60
       UnicodeBlocks: Basic Latin, CJK Unified Ideographs
```

Als XML-Repräsentation sieht das Ergebnis aus wie folgt und kann somit maschinell sehr viel genauer interpretiert werden.

```
<?xml version="1.0" encoding="utf-8"?>
<jhove xmlns:xsi="http://www.w3.org/2001/XMLS-
chema-instance" xmlns="http://hul.harvard.edu/
```

```
       ois/xml/ns/jhove" xsi:schemaLocation="http://
       hul.harvard.edu/ois/xml/ns/jhove http://hul.
       harvard.edu/ois/xml/xsd/jhove/1.5/jhove.xsd"
       name="JhoveView" release="1.1" date="2008-02-21">
       <date>2009-03-03T10:40:00+01:00</date>
       <repInfo
       uri="/Users/fugu/Desktop/nestor-hand-
buch-kapitel-13_2009-03-03/test.txt">
       <reportingModule release="1.3"
date="2007-08-30">UTF8-hul</reportingModule>
       <lastModified>2009-03-03T10:33:12+01:00</lastModified>
       <size>64</size>
       <format>UTF-8</format>
       <status>Well-Formed and valid</status>
       <mimeType>text/plain; charset=UTF-8</mimeType>
       <properties>
              <property>
                     <name>UTF8Metadata</name>
                     <values arity="List" type="Property">
                            <property>
                                   <name>Characters</name>
                                   <val-
ues arity="Scalar" type="Long">
                                          <value>60</value>
                                   </values>
                            </property>
                            <property>
                                   <name>UnicodeBlocks</name>
                                   <val-
ues arity="List" type="String">

<value>Basic Latin</value>

<value>CJK Unified Ideographs</value>
                                   </values>
                            </property>
                     </values>
              </property>
       </properties>
       <note>Additional representation in-
formation includes the line endings:
              CR, LF, or CRLF</note>
       </repInfo>
       </jhove>
```

Eine genauere Dokumentation des grafischen Frontends, des Kommandozeilentools, sowie zu JOHVE allgemein findet sich auf der JHOVE-Homepage (auf Englisch) unter "Tutorial", aktuelle Informationen zur Distribution und die neueste Version derselben auf der JHOVE SourceForge-Projektseite.

JHOVE auf der Kommandozeile

Die Möglichkeit, ganze Verzeichnisse zu untersuchen und kurz mal zu schauen, wieviele valide Dateien darin enthalten sind, ist – neben allen Möglichkeiten des grafischen Frontends – ein großer Vorteil des Kommandozeilentools, das JHOVE zur Verfügung stellt.

Konfiguration

Um das Kommandozeilentool nutzen zu können, ändern Sie bitte zunächst den Namen der Datei jhove.tmpl in jhove (Linux/Unix) oder jhove_bat.tmpl in jhove.bat (Windows). Ändern Sie bitte noch – den Anweisungen in diesen Dateien zufolge – den Pfad zu Ihrem JHOVE-Verzeichnis in diesen Skripten. Haben Sie beispielsweise das JHOVE-Paket in /home/ kopiert, lautet der Pfad /home/jhove (Linux/Unix), arbeiten Sie auf einem Windows-System, tragen Sie für das Verzeichnis C:\Programme\ bitte C:\Programme\jhove ein. Sollte der Pfad zu Ihrer Java-Installation nicht stimmen, passen Sie bitte auch diesen noch an. Wenn Sie alles richtig konfiguriert haben, bekommen Sie durch Tippen von ./jhove bzw. jhove.bat detaillierte Informationen zu Ihrer JHOVE-Installation.

Verzeichnisse rekursiv untersuchen

Wenn Sie nun beispielsweise alle XML-Dateien untersuchen möchten, die sich im Beispiel-Verzeichnis der JHOVE-Installation befinden, rufen Sie JHOVE folgendermaßen auf:

```
./jhove -h audit examples/xml/
```

Die Ausgabe enthält folgendes und beschreibt in Kürze, welche Dateien untersucht wurden, ob und wie viele davon valide sind:

```
<?xml version="1.0" encoding="UTF-8"?>
<jhove xmlns:xsi="http://www.w3.org/2001/XMLSche-
ma-instance" xmlns="http://hul.harvard.edu/
ois/xml/ns/jhove" xsi:schemaLocation="http://
hul.harvard.edu/ois/xml/ns/jhove http://hul.
harvard.edu/ois/xml/xsd/jhove/1.5/jhove.xsd"
name="Jhove" release="1.1" date="2008-02-21">
```

```
<date>2009-03-03T11:27:27+01:00</date>
<audit home="/Users/Fugu/Desktop/jhove">
<file mime="text/xml" status="well-formed">
examples/xml/build.xml</file>
<file mime="text/plain; charset=US-ASCII" status="valid">
examples/xml/external-parsed-entity.ent</file>
<file mime="text/plain; charset=US-ASCII" status="valid">
examples/xml/external-unparsed-entity.ent</file>
<file mime="text/xml" status="well-formed">
examples/xml/jhoveconf.xml</file>
<file mime="text/plain; charset=US-ASCII" status="valid">
examples/xml/valid-external.dtd</file>
</audit>
</jhove>
<!-- Summary by MIME type:
text/plain; charset=US-ASCII: 3 (3,0)
text/xml: 2 (0,2)
Total: 5 (3,2)
-->
<!-- Summary by directory:
/Users/Fugu/Desktop/jhove/examples/xml: 5 (3,2) + 0,0
Total: 5 (3,2) + 0,0
-->
<!-- Elapsed time: 0:00:02 >
```

Weitere Parameter

Als weitere Parameter können unter anderem Handler und Module genauer spezifiziert werden sowie Ausgabe-Dateien und Encoding konfiguriert werden. Hier darf nach Belieben probiert, getestet und gespielt werden, um zu probieren, technische Metadaten zu extrahieren und Dateien zu validieren.
Im Folgenden noch eine kurze Beschreibung des Nutzung des Kommandozeilentools...

```
jhove  [-c config] [-m module [-p param]]
       [-h handler [-P param]]
       [-e encoding][-H handler] [-o output] [-x saxclass]
       [-t tempdir] [-b bufsize] [[-krs] dir-file-or-uri […]]
```

...und die Bedeutung der wichtigsten:

`-c config`	–	Pfad zur JHOVE-Konfigurationsdatei.
`-m module`	–	Name des Moduls, möglich sind hier: AIFF-hul, ASCII-hul, BYTESTREAM, GIF-hul, HTML-hul, JPEG-hul, JPEG2000-hul, PDF-hul, TIFF-hul, UTF8-hul, WAVE-hul und XML-hul.
`-p param`	–	Modul-spezifische Parameter.
`-h handler`	–	Name des Output-Handlers (Grundeinstellung: TEXT).
`-P param`	–	Handler-spezifische Parameter.
`-o output`	–	Name der Ausgabe-Datei (Grundeinstellung: stdout).
`-x saxclass`	–	SAX-Parser-Klasse (Grundeinstellung: J2SE 1.4 default).
`-t tempdir`	–	Temporäres Verzeichnis, in dem temporäre Dateien erzeugt werden.
`-b bufsize`	–	Puffergröße für gepufferte I/O Operationen (Grundeinstellung: J2SE 1.4 default).
`-k`	–	Berechnet CRC32, MD5, und SHA-1 Checksummen.
`-r`	–	Zeigt rohe Data Flags an, nicht die textlichen Äquivalente.
`-s`	–	Format-Identifikation basiert nur auf internen Signaturen.
`dir-file-or-uri`	–	Verzeichnis, Pfadname oder URI der zu untersuchenden Dateien.

13.4 Die kopal Library for Retrieval and Ingest (koLibRI)

Stefan E. Funk

Einführung

Die *kopal Library for Retrival and Ingest* ist ein Framework zur Integration eines Langzeitarchivs wie dem IBM Digital Information Archiving System[20] (DIAS) in die Infrastruktur einer Institution. Insbesondere organisiert koLibRI das Erstellen und Einspielen von Archivpaketen in DIAS und stellt Funktionen zur Verfügung, um diese abzurufen und zu verwalten. koLibRI stellt eine Bibliothek von Java-Tools dar, die im Projekt kopal entwickelt wurden. Sie wurde bewusst so angelegt, dass sie als Ganzes oder in Teilen auch in anderen Zusammenhängen nachnutzbar ist.

An dieser Stelle soll der Teil koLibRIs beschrieben und beispielhaft dargestellt werden, der für das Erstellen von Archivpaketen verantwortlich ist; eine ausführliche Beschreibung der gesamten Funktionalität der kopal Library for Retrieval and Ingest sowie weitere technische Details ist in deren Dokumentation[21] zu finden und auch auf der Internetseite des Projekts kopal[22].

Funktionsweise

Im einfachen Fall generiert koLibRI aus den mit dem zu archivierenden Objekt gelieferten Metadaten sowie den von JHOVE[23] maschinell extrahierten Metadaten eine XML-Datei nach METS Schema, verpackt diese zusammen mit dem Objekt in einer Archivdatei (.zip oder .tar) und liefert diese Datei als Submission Information Package (SIP) an das DIAS.

So gesehen ist koLibRI für eine vollständige Langzeitarchivierungslösung mit DIAS entwickelt worden. Jedoch kann koLibRI auch als eigenständige Software zur Generierung der METS Dateien oder kompletten SIPs nach dem Universellen Objektformat[24] vollkommen ohne DIAS eingesetzt werden. Die auf diese Weise generierten XML-Metadatendateien oder die kompletten SIPs kön-

20 http://www-5.ibm.com/nl/dias/
21 http://kopal.langzeitarchivierung.de/kolibri/koLibRI_v1_0_dokumentation.pdf
22 http://kopal.langzeitarchivierung.de/
23 JSTOR/Harvard Object Validation Environment (JHOVE): http://hul.harvard.edu/jhove/
24 http://kopal.langzeitarchivierung.de/downloads/kopal_Universelles_Objektformat.pdf

nen für den Datenaustausch zwischen verschiedenen Institutionen verwendet werden; ein Aspekt, der bei der Entwicklung des UOF besonders im Vordergrund stand. Alternativ kann durch den modularen Aufbau der koLibRI auch mit vertretbarem Aufwand an ein anderes Archivsystem oder ein anderes Metadatenformat angepasst werden, da die Schnittstellen ausreichend spezifiziert sind.
Mit koLibRI kann ein für das Erstellen der Archivpakete benötigter Workflow abgebildet werden. Dieser Workflow kann den eigenen Bedürfnissen angepasst und erweitert werden. Die koLibRI-Infrastruktur nutzt prinzipiell vier Konstrukte, um Workflows abzubilden und zu verarbeiten:

- Zunächst sammelt der sogenannte *ProcessStarter* die einzuspielenden Dateien/Daten ein, die als kleinste Einheit definiert wurden – in unserem Beispiel wird dies ein Verzeichnis mit beliebigen Dateien sein.
- Jede Einheit wird vom ProcessStarter an die *ProcessQueue* angehängt, die dann der Reihe nach (oder auch nebenläufig) abgearbeitet werden.
- In den *ActionModules* werden einzelne Aufgaben implementiert, die für ein jedes Objekt in der ProcessQueue durchgeführt werden sollen. Für den Beispiel-Workflow werden hier folgende Module genutzt: FileCopyBase, MetadataExtractorDmd, MetadataGenerator, MetsBuilder und Zip. Weitere Module werden mit koLibRI geliefert und können integriert werden.
- Die Reihenfolge, in der die ActionModules für jedes dieser Objekte in der ProcessQueue verarbeitet werden, wird als *Policy* konfiguriert.

Installation und Konfiguration

Download

Zunächst wird das koLibRI-Paket von der kopal-Homepage benötigt, bitte laden Sie diese von der Internetseite des Projekts kopal. Benötigt wird das gepackte Programmpaket „kopal Library for Retrieval and Ingest“ (http://kopal.langzeitarchivierung.de/kolibri/koLibRI_v1_0.zip), das Sie bitte in ein beliebiges Verzeichnis entpacken.

Anforderungen

Da die kopal Library for Retrieval and Ingest komplett in Java implementiert wurde, sollte die Software prinzipiell auf jeder Plattform laufen, die eine Java Virtual Machine in der Version 1.5 zur Verfügung stellt. Alle weiteren

erforderlichen Java Software-Bibliotheken sind in dem Paket enthalten und vorkonfiguriert.

Konfiguration des Workflowtool Skriptes

Zunächst müssen die folgenden Werte in den beiden Startskripten `workflowtool` (Linux/Unix) oder `workflowtool.bat` (Windows) an die lokalen Verhältnisse angepasst werden:

- `KOLIBRI_HOME`

Hier tragen Sie bitte den Pfad zu Ihrer koLibRI-Installation ein, zum Beispiel `/home/funk/kolibri_v1_0` (Linux/Unix) bzw. `C:\Programme\kolibri_v1_0` (Windows).

- `JAVA_HOME`

Sollte hier der Pfad zu Ihrer Java-Installation nicht stimmen, passen Sie diesen bitte ebenfalls noch an.

Konfiguration der Policies-Datei

Die Datei policies.xml im Verzeichnis config/ wird um die folgenden Zeilen ergänzt; vor dem letzten schließenden Tag – </policies> – fügen Sie bitte die folgenden Zeilen ein:

```
<policy name="example_lza_handbuch">
  <step class="FileCopyBase">
    <step class="XorFileChecksums">
      <step class="MetadataExtractorDmd">
        <step class="MetadataGenerator">
          <step class="MetsBuilder">
            <step class="Zip">
              <step class="CleanPathToContentFiles"/>
            </step>
          </step>
        </step>
      </step>
    </step>
  </step>
</policy>
```

Konfiguration der Konfigurations-Datei

In der Datei config.xml im Verzeichnis config/ werden folgende Werte gesetzt:

- Der Wert der Eigenschaft des Feldes `<field>defaultPolicyName</field>` wird in den Wert `<value>example_lza_handbuch</value>` geändert, so wird unsere vorher hinzugefügte Policy genutzt.
- Die Werte der Felder `logfileDir`, `destinationDir`, `workDir` und `tempDir` werden jeweils mit dem Pfad zu den jeweiligen Verzeichnissen ersetzt. Bitte legen Sie diese vorher an, am besten direkt in Ihren kolibri-Verzeichnis (Beispielsweise als „log“, „dest“, „work“ und „temp“): `<value>./log/</value>`, `<value>./dest/</value>`, `<value>./work/</value>` und `<value>./temp/</value>`.
- Schließlich wird noch ein Verzeichnis als Hotfolder benötigt, aus dem die zu behandelnden Dateien hineinkopiert werden. Bitte legen Sie ein weiteres Verzeichnis „./hotfolder“ an, dessen Wert sollte bereits in der Konfigurations-Datei eingetragen sein.

Starten von koLibRI

Zum Starten des Workflowtools wechseln Sie bitte in das Verzeichnis der koLibRI-Installation – oder bleiben gleich dort, sollten Sie schon da sein – und tippen

```
./workflowtool -c config/config.xml (Linux/Unix)
```

bzw.

```
workflowtool /c config\config.xml (Windows)
```

Sie bekommen nun – wenn nun alles richtig konfiguriert ist – eine Ausgabe auf der Konsole, die mit den folgenden Zeilen endet:

```
[INFO]       Checking hotfolder /Users/fugu/Desk-
top/koLibRI_v1_0/./hotfolder for new content

[INFO]

All current files scheduled, waiting for more
```

Nun können Sie testweise ein beliebiges Verzeichnis in dieses Hotfolder kopieren (bitte zu Anfang mit nicht allzuviel Inhalt!), und koLibRI fängt an zu arbeiten.

Ergebnis

Als Ergebnis erhalten Sie im Verzeichnis `dest` eine .zip-Datei, in der sich zum einen Ihre im Hotfolder befindlichen Dateien befinden und außerdem eine METS-Datei im Universellen Objektformat mit dem Namen `mets.xml`. Diese enthält neben den in der Template-XML-Datei `config/uof_template.xml` enthaltenen Daten – die für die METS-Datei als Vorlage genommen wird – technische Metadaten für jede einzelne Datei (extrahiert von JHOVE) im LMERfile-Format, sowie Metadaten zu dem gesamten Objekt im LMERobject-Format[25].

Weitere Konfigurationsmögichkeiten der kopal Library for Retrieval and Ingest – von denen es noch viele gibt – sowie eine ausführliche Beschreibung der Nutzung auch mit dem DIAS, und weitere Nutzungsszenarien und Erweiterungsmöglichkeiten der koLibRI sind in der ausführlichen Dokumentation nachzulesen. Weiterhin gibt es die Möglichkeit, über die koLibRI-Internetseite den Entwicklern Rückmeldungen zu Erfahrungen mit koLibRI mitzuteilen.

Literatur

Funk, Stefan; Kadir Karaca Koçer, Sabine Liess, Jens Ludwig, Matthias Neubauer: *kopal Library for Retrieval and Ingest – Dokumentation –. 2007*. http://kopal.langzeitarchivierung.de/kolibri/koLibRI_v1_0_dokumentation.pdf

25 http://www.d-nb.de/standards/lmer/lmer.htm

14 Geschäftsmodelle

14.1 Einführung

Achim Oßwald

Neben der vor dem Hintergrund neuerer Verfahren und Erfahrungen weiterhin relevanten Frage, auf welche Weise Langzeitarchivierung optimal realisiert werden könnte und sollte, drängt sich eine weitere Frage in den Vordergrund: Wie können die ausgewählten Verfahren finanziert und in Geschäftsmodelle eingebunden werden?

Nur in den LZA-Anfängen ist optimistisch über die Frage der Kosten spekuliert worden. Zu dieser Zeit bestand die Hoffnung, die Sicherung digitaler Objekte könne günstiger ausfallen als beispielsweise jene von Druckwerken. Schon bald jedoch wurde deutlich, dass die gewählten Maßnahmen zur Erhaltung bzw. Erneuerung von Daten, Datenträgern und Wiedergabeumgebungen relativ aufwändig und teuer sind. Dies aber bedeutet, dass die mit einem Verfahren der Langzeitarchivierung und Langzeitverfügbarkeit verbundenen Kosten – je nach gewähltem Archivierungskonzept – gänzlich oder zumindest in Teilen unregelmäßig wiederkehrend anfallen.

Zu ermitteln und zu analysieren, welche Kostenfaktoren überhaupt bei der LZA zum Tragen kommen und welche konkreten Kosten damit aus heutiger Sicht verbunden sein werden, ist Gegenstand von einigen wenigen Projekten. Bis auf weiteres stellt diese Frage für alle Einrichtungen, die mit der Aufgabe der Langzeitarchivierung aufgrund gesetzlicher Bestimmungen und sonstiger Vereinbarungen betraut sind, einen nur begrenzt abgesicherten Aspekt dar.

Kapitel 14.2 „Kosten" gibt einen Einblick in die aktuelle Diskussion zum Thema und einen Überblick zu den Kostenfaktoren, die voraussichtlich mit den Aktivitäten für die Langzeitarchivierung und Langzeitverfügbarkeit verbunden sein werden.

Kapitel 14.3. „Service- und Lizenzmodelle" zeigt auf, welche Optionen sich auf dieser Grundlage für dienstleistende Organisationen und Einrichtungen abzeichnen und welche Servicemodelle derzeit von den als Dienstleister aktiven Organisationen angeboten werden könnten.

Niemand weiß heute, ob die derzeit praktizierten Verfahren zur Langzeitarchivierung und Langzeitverfügbarkeit mit ihren z. T. sehr unterschiedlichen Geschäftsmodellen ihrerseits wiederum langzeitfähig sind. Dies wird sich zukünftig erweisen. Umso größere Sorgfalt und Professionalität ist notwendig, wenn Verfahren und Strategien für die Langzeitarchivierung und Langzeitverfügbarkeit ausgewählt und in öffentlichem oder privatwirtschaftlichem Auftrag realisiert werden. Dauerhafte Finanzierungskonzepte sind dabei eine unabdingbare Voraussetzung, um die methodischen und technischen Überlegungen dauerhaft zum Tragen kommen lassen zu können.

14.2 Kosten

Thomas Wollschläger und Frank Dickmann

In diesem Kapitel werden Kostenfaktoren benannt, die für den Betrieb eines digitalen Langzeitarchivs von Bedeutung sind. Des Weiteren werden Ansätze vorgestellt, wie die individuellen Kosten der LZA in einer Institution ermittelt werden können.

Kostenfaktoren bei Einrichtung und Unterhaltung eines Langzeitarchivs

Abhängig vom konkreten Langzeitarchivierungskonzept der jeweiligen Einrichtung werden folgende Kostenfaktoren zu berücksichtigen sein:

Initiale Kosten

- Informationsbeschaffung über LZA-Systeme
- Erhebung von Bestand, Zugang und gewünschten Zugriffsoptionen für digitale Materialien im eigenen Haus
- Erhebung von vorhandenen Personal- und Technikressourcen im eigenen Haus
- Projektplanung, ggf. Consulting, Ausschreibung(en)

Beschaffungskosten

- Hardware: Speichersysteme und *sämtliche* infrastrukturellen Einrichtungskosten (Serveranbindungen, Datenleitungen, Mitarbeiterrechner usw.)
- Ggf. Lizenz(en) für Software-Systeme oder Beitrittskosten zu Konsortien
- Weitere Aufwendungen: z.B. Anpassungsentwicklungen von Open Source Software-Produkten, Entwicklung/Anpassung von Schnittstellen, Erstellung von physischen und digitalen Schutzmaßnahmen (auch solche aus rechtlichen Gründen)
- Ggf. Einstellung neuer Mitarbeiter und/oder Schulung vorhandener Mitarbeiter

Betriebskosten

- Dateningest des bisher vorhandenen Materials
- Dateningest des neu eingehenden Materials
- Laufende Storage-Kosten

- Sonstige Dauerbetriebskosten: z.B. Strom, Datenleitungskosten, *sämtliche* Sicherheitsmaßnahmen, Backups, regelmäßige Wartung(en) und Tests, Software-Upgrades
- Zukauf von weiteren Speichereinheiten
- Hard- und Software-Komplettersatz in Intervallen
- Ggf. laufende Lizenzkosten und/oder laufende Beitragszahlungen bei Konsortien

Die konkreten Kosten sind dabei jeweils abhängig von

- der Zahl und Komplexität der Workflows bei einer Institution
- der Menge, Heterogenität und Komplexität der zu archivierenden Objekte und ihrer Metadaten
- den gewünschten Zugriffsmöglichkeiten und Schnittstellen sowie ggf.
- den Anforderungen Dritter an die archivierende Institution bzw. Verpflichtungen der Institution gegenüber Dritten

Die Ermittlung von Kosten für die Langzeitarchivierung

Die tatsächliche Ermittlung der Kosten, die auf eine Einrichtung für die Langzeitarchivierung ihrer digitalen Dokumente zukommen, gestaltet sich in der Praxis noch relativ schwierig. Viele LZA-Unternehmungen befinden sich derzeit noch im Projektstatus oder haben noch nicht lange mit dem produktiven Betrieb begonnen. Daher liegen noch wenige Erfahrungswerte vor, wie sich insbesondere der laufende Betrieb eines solchen Archivs kostenmäßig erfassen lässt. Außerdem befindet sich nach wie vor die zunehmende Menge und Varianz insbesondere der Internet-Publikationen in einem Wettlauf mit den technischen Möglichkeiten, die von Gedächtnisorganisationen zur Einsammlung und Archivierung eingesetzt werden können.

Einen begrenzten Anhaltspunkt können die angesprochenen Unternehmungen zumindest in der Hinsicht liefern, was die Ersteinrichtung eines digitalen Langzeitarchivs betrifft. BMBF und DFG haben eine ganze Reihe von solchen Projekten gefördert und verschiedene Institutionen haben Projekte aus eigenen Mitteln finanziert.[1] Das bisher am umfangreichsten geförderte LZA-Vorhaben in Deutschland war das Projekt kopal mit einem Fördervolumen von 4,2 Mio. Euro.[2] Diese Kosten schließen die vollständige Entwicklung eines Ar-

1 Siehe dazu die Projektübersicht in der nestor-Informationsdatenbank http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?zeig=10
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

2 Vgl. Wollschläger (2007), S. 247.

chivsystems einschließlich Objektmodell, Aufbau von Hard- und Softwareumgebungen in mehreren Einrichtungen und mehrjährige Forschungsarbeiten ein. Zum Projektende hat kopal allerdings in einem Servicemodell konkrete Kosten für den Erwerb eines vollständigen Archivs zum Eigenbetrieb vorgelegt. Wenn das kopal-Archivsystem unter Zukauf von Beratung und ggf. Entwicklung eigenständig betrieben wird, soll ein Erstaufwand für Hard- und Software eines Systems mittlerer Größe von ca. 750.000 € anfallen. Hiervon entfielen 40% auf Softwarelizenzen und 60% auf Systembereitstellung und -betrieb.[3] Wiewohl solche Angaben nur exemplarisch sein können, kann dennoch davon ausgegangen werden, dass die Kosten für die Ersteinrichtung eines LZA-Systems in einer Einrichtung einen gewissen Schwellenwert nicht unterschreiten werden.

Die Zahl der Ansätze, die bisher versucht haben, Modelle für die Betriebskostenermittlung digitaler LZA zu entwickeln, ist begrenzt. Nennenswert ist hierbei der Ansatz des LIFE-Projekts aus Großbritannien. „The LIFE Project" war ein einjähriges Projekt (2005/2006) der British Library (BL) in Zusammenarbeit mit dem University College London (UCL) mit dem Hauptziel, ein Kostenmanagement für die Langzeiterhaltung elektronischer Ressourcen zu entwickeln. Es wurde eine Formel zur Ermittlung der Archivierungskosten entwickelt. Manche Fragen mussten noch offen blieben, so war es z.B. bislang nicht adäquat möglich, im Rahmen des Projektes die Kosten der Langzeiterhaltung von gedruckten und elektronischen Veröffentlichungen zu vergleichen. Die Formel lautet: $L_T=Aq+I_T+M_T+Ac_T+S_T+P_T$. Dabei stehen die Werte für folgende Parameter.[4]

- L = complete lifecycle cost over time 0 to T.
- Aq = Acquisition
- I = Ingest
- M = Metadata
- Ac = Access
- S = Storage
- P = Preservation

Jeder der Parameter kann weiter in praktische Kategorien und Elemente aufgeteilt werden. Alle Schritte können entweder, wenn der Prozess direkt kalkulierbar ist, als Kostenfaktor berechnet werden oder, wenn nötig, jeweils auch noch

3 Siehe kopal (2007), S. 2.

4 Vgl. McLeod/Wheatley/Ayris (2006), S. 6. Das tiefergestellte (T) in der Formel bedeutet als Attribut der Parameter „over time".

in beliebig viele Unterpunkte untergliedert werden. So kann die Berechnung für die jeweilige Institution individuell angepasst werden. Innerhalb des LIFE-Projekts wurden zum einen beispielhafte Berechnungen der LZA-Kosten des Projektmaterials vorgenommen und dabei Kosten für „the first year of a digital asset's existence" und „the tenth year of the same digital assets' existence" vergleichbar ermittelt[5] und exemplarisch auch die Kosten pro Speichermenge. Zum anderen hat das Projekt die entwickelten Formelwerke zur Verfügung gestellt, so dass interessierte Institutionen selbst Berechnungen anhand der Individualparameter vornehmen können.

In Anbetracht der aktuellen Erkenntnisse fokussieren Kostenschätzungen für die LZA hauptsächlich auf die Kosten pro Speichermenge, wie z.B. in LIFE. Bei eingehender Betrachtung der Prozesse, die im Rahmen eines implementierten LZA-Systems anfallen, sind jedoch die Speicherkosten nicht der Hauptkostenfaktor. Vielmehr lässt sich aus den Prozessen Acquisition, Metadata und Preservation eine hohe Personalintensität ableiten, insbesondere aus dem Grund, da diese Prozesse nur sehr eingeschränkt automatisierbar sind. Daher sind es vielmehr die Personalkosten, die langfristig den höchsten Anteil an den LZA-Kosten haben werden.[6]

Ebenso von Bedeutung ist die Anzahl unterstützter Formate, da diese Anzahl und der Personaleinsatz eng miteinander verknüpft sind. Jedes zusätzliche Format erfordert zusätzlichen Aufwand durch qualifiziertes Personal. Demgegenüber steht aber der Nutzen, den das Angebot eines Formats liefert. Hierzu wurde festgestellt, dass weniger häufig verwendete Formate ein unterdurchschnittliches Kosten-Nutzen-Verhältnis aufweisen. Beispielsweise haben in LIFE die Formate PDF, TXT und HTML ca. 85% aller Dokumente abgedeckt, allerdings nur 7% der Kosten verantwortet, während die 12 am wenigsten verwendeten Formate 0,1% der Dateien abgedeckt, aber ca. 41% der Kosten hervorgerufen haben. Grundlage dabei sind die Gesamtkosten über einen Zeitraum von 20 Jahren.[7] Aus diesem Grund ist eine Einschränkung der Formatvielfalt ein empfehlenswertes Kostensteuerungsinstrument für die LZA.

Eine bedeutende Frage für die Festlegung der Archivierungsstrategie – nämlich für das eigentliche „Preservation Planning", die Erhaltungsmaßnahmen über die Lebenszeit eines digitalen Objekts – einer Institution ist, ob auf Dauer Migrationen oder Emulationen kostengünstiger sind. Hierzu sind noch keine abschließenden Aussagen möglich. Generell verbreitet ist die Auffassung, dass Migration der kostengünstigere Weg sei. Innerhalb von LIFE wurden dazu An-

5 Vgl. ebenda, S. 3.
6 Vgl. Ashley (1999), S. 123.
7 Vgl. Björk, B.-C. (2007), S. 23.

sätze formuliert, die jedoch hauptsächlich sehr exemplarische Migrationen behandeln und noch nicht repräsentativ sind.[8] Andere Studien kommen dagegen zu dem Schluss, dass Emulationen auf längere Sicht kostengünstiger seien:

> *While migration applies to all objects in the collection repetitively, emulation applies to the entire collection as a whole. This makes emulation most cost-effective in cases of large collections, despite the relatively high initial costs for developing an emulation device. When considering the fact that only small fragments of digital archives need to be rendered in the long run, it may turn out that from a financial perspective emulation techniques will be more appropriate for maintaining larger archives.*[9]

Da die bestehenden Langzeitarchive gerade erst dabei sind, die ersten „echten" Maßnahmen von Preservation Planning umzusetzen, wird hier auf Erfahrungswerte zu warten sein, die entsprechende Ergebnisse unterstützen können.

Konsequenzen für die Gedächtnisorganisationen

Angesichts der zu erwartenden nicht unerheblichen Kosten für die *Ersteinrichtung* eines LZA-Systems dürften kleinere Einrichtungen nicht umhin kommen, zwecks Einrichtung eines solchen Systems mit anderen Institutionen zu kooperieren bzw. sich einem bestehenden System anzuschließen und/oder sich den Zugang dazu über Lizenzen zu sichern. Selbst größere Institutionen werden für die Einrichtung eines LZA-Systems oft kooperative Formen wählen, um hohe Ersteinrichtungskosten aufzuteilen, die sich sonst nicht auf mehrere Schultern verteilen lassen. Ebenso könnte angesichts der noch bestehenden Unsicherheit, wie sich künftig die Kosten für den Dauerbetrieb des Langzeitarchivs und das Preservation Planning entwickeln werden, die Entscheidung zugunsten der Variante ausfallen, sich in bestehende Systeme einzukaufen oder über kostenpflichtige Lizenzen Teilnehmer an einem kommerziell ausgerichteten System zu werden. Letzteres macht in der Regel Zugeständnisse an die gewünschte Preservation Policy notwendig, so dass eine Gedächtnisorganisation abwägen muss, welche Kosten – Lizenzen für ein kommerzielles System oder eigene Entwicklungskosten, z.B. für die Anpassung von Open Source Software – die jeweils lohnendere Investition ist.

Die Teilnahme an kooperativen Formen der Langzeitarchivierung ist unter Kostenaspekten in jedem Fall empfehlenswert. Hierbei können Institutionen über z.B. gemeinsame Speichernutzung bzw. gegenseitiges Backup, gegenseitige Nutzung von Entwicklungsergebnissen, gemeinsame Adressierung

8 Vgl. Ebenda, S. 10.

9 Zitiert nach Oltmans/Kol (2005), #5 – Conclusion.

übergreifender Herausforderungen oder kooperative Verwaltung von Open Source Software Synergien schaffen und erhebliche Ressourceneinsparungen ermöglichen.

Quellen und Literatur

Ashley, K. (1999): *Digital Archive Costs: Facts and Fallacies*, in: *Proceedings of the DLM-Forum on Electronic Records (DLM '99)*, DLM-Forum, Brussels, 1999, S. 123, http://ec.europa.eu/archives/ISPO/dlm/fulltext/full_ashl_en.htm

Björk, B.-C. (2007): *Economic evaluation of LIFE methodology*, LIFE Project, London, UK, 2007, URL: http://eprints.ucl.ac.uk/7684/1/7684.pdf

Kopal (2007): *kopal: Ein Service für die Langzeitarchivierung digitaler Informationen.* Projekt kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen), 2007, http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf

McLeod, Rory; Wheatley, Paul; Ayris, Paul (2006): *Lifecycle Information for E-literature* : A summary from the LIFE project ; Report Produced for the LIFE conference 20 April 2006. LIFE Project, London (via http://www.ucl.ac.uk/ls/lifeproject/ or directly under http://eprints.ucl.ac.uk/archive/00001855/01/LifeProjSummary.pdf)

Oltmans, Erik; Kol, Nanda (2005): *A Comparison Between Migration and Emulation in Terms of Costs*. In: RLG DigiNews, Volume 9, Number 2, 15.04.2005 (http://worldcat.org/arcviewer/1/OCC/2007/07/10/0000068902/viewer/file1.html#article0)

Wollschläger, Thomas (2007): *kopal – ein digitales Archiv zur dauerhaften Erhaltung unserer kul-turellen Überlieferung.* In: Geschichte im Netz : Praxis, Chancen, Visionen ; Beiträge der Tagung .hist2006, Berlin: Clio-online und Humboldt-Universität zu Berlin, 2007, S. 244 – 257 (Historisches Forum 10 (2007), Teilband I).

Siehe außerdem die Einträge in der nestor-Informationsdatenbank zum Thema „Kosten" unter http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?show=8

14.3 Service- und Lizenzmodelle

Thomas Wollschläger und Frank Dickmann

In den wenigsten Fällen werden Langzeitarchivierungssysteme von einer einzigen Institution produziert und genutzt. Schon bei einer zusätzlichen Nutzer- oder Kundeninstitution für das hergestellte und/oder betriebene Archivsystem müssen Lizenz- oder Geschäftsmodelle aufgestellt sowie Servicemodelle für zu leistende Langzeitarchivierungs-Dienstleistungen definiert werden.

Lizenzmodelle

Lizenzkosten fallen in der Regel für die Nutzung kommerzieller Softwareprodukte an. Dabei gibt es unterschiedliche Möglichkeiten. Zum einen können solche Produkte lizenziert und eigenständig in der eigenen Institution eingesetzt werden. Dabei ist die Hersteller- oder Vetriebsfirma neben den (einmalig oder regelmäßig) zu zahlenden Lizenzgebühren zumeist durch Support- und Updateverträge mit der Nutzerinstitution verbunden. Beispiele hierfür sind etwa das System *Digitool* der Firma Exlibris[10] oder das *DIAS*-System von IBM.[11]

Zum anderen besteht bei einigen Produkten die Möglichkeit, dass eine Betreiberinstitution (die nicht identisch mit dem Hersteller oder Systemvertreiber sein muss) das Archivsystem hostet und eine Nutzung für Dritte anbietet. Hierbei werden Lizenzkosten meist vom Betreiber auf die Kunden umgelegt oder fließen in die Nutzungskosten für die Archivierung ein. Ein Beispiel hierfür ist das insbesondere auf die Archivierung von e-Journals ausgerichtete System *Portico*. Hierbei erfolgt eine zentrale, an geografisch auseinander liegenden Orten replizierte Archivierung. Die Kosten von Portico richten sich für eine Bibliothek nach dem verfügbaren Erwerbungsetat. Der jährliche Beitrag für die Nutzung des Systems kann daher je nach dessen Höhe zwischen 1% des Erwerbungsetats und maximal 24.000 US-$ liegen.[12]

Neben den kommerziellen Produkten gibt es eine Reihe von Open Source-Lösungen im Bereich der Archivierungssysteme. Durch die Nutzung von Open Source-Lizenzen[13] fallen oft keine Lizenzgebühren bzw. -kosten für die Nutzerinstitutionen an, sondern zumeist nur Aufwands- und Materialkosten. Zudem sind Archivinstitutionen, die eine Open Source-Software oder ein Open Sour-

10 Siehe http://www.exlibrisgroup.com/digitool.htm
11 Siehe http://www-05.ibm.com/nl/dias/
12 Vgl. http://www.portico.org/libraries/aas_payment.html
13 Siehe hierzu v.a. http://www.opensource.org/licensez

ce-Netzwerk nutzen, dahingehend gefordert, durch eigene Entwicklungsbeiträge das Produkt selbst mit weiterzuentwickeln.[14] Beispiele für verbreitete Open Source-Lösungen sind das System *DSpace*[15] und die *LOCKSS*- bzw. *CLOCKSS*-Initiative.[16] Die LOCKSS-Technologie will die langfristige Sicherung des archivierten Materials dadurch sicherstellen, dass jedes Archivobjekt mit Hilfe des Peer-to-Peer-Prinzips bei allen Mitgliedern gleichzeitig gespeichert wird. Jedes Mitglied stellt einen einfachen Rechner exklusiv zur Verfügung, der im Netzwerk mit den anderen Mitgliedern verbunden ist und auf dem die LOCKSS-Software läuft.

Neben der Nutzung reiner kommerzieller Lösungen und reiner Open Source-Lösungen gibt es auch Mischformen. Dabei kann es von Vorteil sein, nur für Teile des eigenen LZA-Systems auf kommerzielle Produkte zurückzugreifen, wenn sich dadurch beispielsweise die Höhe der anfallenden Lizenzkosten begrenzen lässt. Andererseits erwirbt man mit vielen Lizenzen zumeist auch Supportansprüche, die etwa bei geringeren eigenen Entwicklungskapazitäten willkommen sein können. Ein Beispiel für eine solche LZA-Lösung ist das *kopal*-System. Hierbei wird das lizenz- und kostenpflichtige (modifizierte) Kernsystem DIAS verwendet, während für den Ingest und das Retrieval die kostenfreie Open Source-Software *koLibRI* zur Verfügung gestellt wird.[17]

Eine Institution muss somit abwägen, welches Lizenzmodell für sie am vorteilhaftesten ist. Kommerzielle Lizenzen setzen den Verwendungs- und Verbreitungsmöglichkeiten der Archivsysteme oft enge Grenzen. Open Source-Lizenzen bieten hier in der Regel breitere Möglichkeiten, verbieten aber ggf. die Exklusivität bestimmter Funktionalitäten für einzelne Institutionen. Hat sie ausreichende Entwicklungskapazitäten und Hard- bzw. Softwareausstattung, kann die Nutzung von Open Source-Lösungen ein guter und gangbarer Weg sein. Dies gilt beispielsweise auch, wenn sich die Institution als Vorreiter für leicht nachnutzbare Entwicklungen sieht oder im Verbund mit anderen Einrichtungen leicht konfigurierbare Lösungen erarbeiten will. Hat sie jedoch nur geringe Enwicklungsressourcen und decken die kommerziellen Lizenzen alle benötigten Services ab, so kann trotz ggf. höherer Lizenzkosten die Wahl kommerzieller Produkte bzw. von standardisierten Services seitens LZA-Dienstleistern angeraten sein.

---

14 Vgl. hierzu insbesondere das Kapitel „Kostenrelevante Eigenschaften einer ungewöhnlichen Organisationsform“, in: Lutterbeck/Bärwolff/Gehring (2007), S. 185 – 194.

15 Siehe http://www.dspace.org/

16 Siehe http://www.lockss.org/

17 Siehe http://kopal.langzeitarchivierung.de/index_koLibRI.php.de.

Servicemodelle

Wie bereits dargestellt, bestehen die wesentlichen Faktoren für die Entscheidung einer Institution für bestimmte Lizenz- und Geschäftsmodelle in den von ihr benötigten Services zur Langzeitarchivierung.[18] Entscheidungskriterien für die Wahl der Einrichtung und/oder Nutzung bestimmter LZA-Services können sein:

Auftrag und Selbstverständnis

- Liegt ein (z.B. gesetzlicher) Auftrag vor, dass die Institution digitale Dokumente eines bestimmten Portfolios sammeln und (selbst) langzeitarchivieren muss?
- Gilt dieser Auftrag auch für Materialien Dritter (z.B. durch Pflichtexemplarregelung)?
- Hat die Institution den Anspruch oder das Selbstverständnis, LZA-Services selbst anbieten oder garantieren zu wollen?
- Liegt eine rechtliche Einschränkung vor, Materialien zwecks LZA Dritten zu übergeben?

Ausstattung und Ressourcen

- Hat die Institution die benötigte Hardware- und/oder Softwareausstattung bzw. kann sie sie bereitstellen, um LZA betreiben zu können?
- Tritt die Institution bereits als Datendienstleister auf oder ist sie selbst von Datendienstleistern (z.B. einem Rechenzentrum) abhängig?
- Stehen genügend personelle Ressourcen für den Betrieb, den Support (für externe Nutzer) und für nötige Entwicklungsarbeiten zur Verfügung?
- Lassen die Lizenzen des genutzten Archivsystems / der Archivsoftware eine Anbindung Dritter an die eigene Institution zwecks LZA zu?

18 Selbstverständlich spielen auch die technischen Möglichkeiten des eingesetzten Archivsystems selbst eine wesentliche Rolle. Einen Kriterienkatalog zur technischen Evaluierung von Archivsystemen bietet z.B. das Kapitel *Software Systems for Archiving* bei Borghoff (2003), S. 221 – 238.

Je nachdem, wie diese Fragen beantwortet werden, stehen für die Wahl des Servicemodells potentiell viele Varianten zur Verfügung. Diese drehen sich im Wesentlichen um die folgenden Konstellationen:

- Die Institution stellt einen LZA-Service (nur) für digitale Dokumente aus eigenem Besitz bereit.
- Die Institution stellt diesen LZA-Service auch für Dritte zur Verfügung.
- Die Institution stellt selbst keinen LZA-Service bereit, sondern nutzt die Services eines Dritten für die Archivierung der eigenen Daten.

Dabei ist jeweils zusätzlich und unabhängig von der Frage, welche Institution den *Service* an sich anbietet, relevant, ob die Daten bzw. respektive die Hardware-/Storage-Umgebung von der Service-Institution selbst oder von Dritten gehostet wird. Beispielsweise kann eine Institution verpflichtet sein, selbst einen LZA-Service anzubieten. Dennoch mag der Umfang des jährlich anfallenden Materials den aufwändigen Aufbau einer solchen Hardware-/Storage-Umgebung sowie entsprechender Betriebskompetenzen nicht rechtfertigen. Hier könnte die Institution entscheiden, zwar einen LZA-Service aufzubauen – und ggf. sogar Dritten gegenüber ein entsprechendes Geschäftsmodell anzubieten –, das Datenhosting jedoch an einen geeigneten Dienstleister abzugeben. Ein Beispiel für ein solches Servicekonzept ist das *kopal*-Projekt. Die Hauptmandanten betreiben zwar gemeinschaftlich das Archivsystem kopal und stellen ihre Dienstleistungen (zumeist kleineren) Nutzerinstitutionen zur Verfügung, die eigentliche Datenhaltung wird jedoch bei einem Rechenzentrum betrieben, wo die gemeinschaftlich genutzte Hardware zentral gehostet und per Fernzugriff genutzt wird.[19]

Zu den einzelnen Dienstleistungen, die im Rahmen eines LZA-Service-Modells von einer Institution angeboten werden können, gehören beispielsweise folgende:

- Der Betrieb des LZA-Systems und Annahme von Archivmaterial
- Durchführung von Erhaltungsmaßnahmen (von Bitstream-Preservation bis zur Migration von Material)
- Zurverfügungstellung von Datenkopien bei Datenverlusten seitens der Ablieferinstitution
- Bereitstellen eines Pseudonymisierungsdienstes[20], wobei die personenspezifischen Daten und die inhaltlichen Daten (z.B. medizinische Daten)

19 Siehe Kopal (2007), S. 1-2.

20 Reng et. al. (2006), S. 49 f.

an jeweils anderen Standorten durch andere LZA-Services gespeichert werden (diese Service-Variante spielt im Hinblick auf die LZA von Forschungsdaten aus dem biomedizinischen Bereich eine entscheidende Rolle für die Akzeptanz der LZA)
- Installation des Systems bzw. von Zugangskomponenten für Remote Access vor Ort
- Beratungsleistungen, z.B. zum Geschäftsmodell, zum Einsatz der Archivsoftware, zur Speicherverwaltung etc.
- Support und Schulungen
- Weiterentwicklung des Archivsystems bzw. von gewünschten Komponenten

Handelt es sich bei dem Dienstleister, der von einer Archivinstitution in Anspruch genommen wird, um einen reinen Datenhost, könnten folgende Dienstleistungen relevant werden:

- Hardwarehosting und -betreuung
- Hosting und Betreuung von Standardsoftware
- Sichere Datenhaltung (z.B. durch Mehrfachbackups)
- Zurverfügungstellung von Datenkopien bei Datenverlusten seitens der Ablieferinstitution
- Notfall- und Katastrophenmanagement
- Beratungsleistungen, z.B. zur Speicherverwaltung

Gerade im Hinblick auf ein Commitment sind ebenso Service-Levels bezüglich der Aufbewahrungsdauer sinnvoll. Ebenso können mit derartigen Service-Levels die Wünsche von Nutzern feingranularer adressiert werden:

- Aufbewahrung bis zu 5 Jahre als Backup-Lösung
- Aufbewahrung bis zu 10 Jahre zur Realisierung guter wissenschaftlicher Praxis
- Aufbewahrung bis zu 30 Jahre zur Erfüllung gesetzlicher Anforderungen und langfristiger Speicherung
- Aufbewahrung für mehr als 30 Jahre als „richtige“ Langzeitarchivierung

Entsprechend einer Kostenkalkulation muss dann jeder LZA-Dienstleister Preise für die einzelnen Service-Levels definieren, die zum einen die eigenen Vollkosten decken und zum anderen die nachhaltige Entwicklung der LZA – im Sinne der Preservation – ermöglichen. Letzteres bezieht sich insbesondere auf

die Entwicklungskosten für technische Maßnahmen durch einen LZA-Dienstleister und auf Kostensteigerungen, wie z.B. durch höhere Personalkosten in Folge neuer Tarifverträge.

Jede Institution muss die eigenen Möglichkeiten bezüglich des Angebots von LZA-Services sorgfältig evaluieren. Hat sie einmal damit begonnen, insbesondere für Dritte solches Services anzubieten, werden dadurch Verpflichtungen eingegangen, die durch künftige technische Entwicklungen ggf. nur erschwert eingehalten werden können. Daher kann es ratsam sein, LZA-Services koordiniert oder kooperativ mit anderen Einrichtungen anzubieten bzw. zu nutzen. Lassen sich die Dienstleistungen von externen Anbietern nutzen und ist dies auch unter Kostengesichtspunkten der wirtschaftlichere Weg, kann es auch für Teile des digitalen Bestands einer Einrichtung sinnvoll sein, diese durch den Service eines solchen Anbieters archivieren zu lassen. Eine andere Möglichkeit bietet sich in dem beschriebenen Hardware-Hosting bzw. Storage-Betrieb durch einen ausgewiesenen Dienstleister.

Quellen und Literatur

Borghoff, Uwe M. et al. (Hrsg.) (2003): *Long-Term Preservation of Digital Documents* : Principles and Practices. Heidelberg [u.a.] : Springer

Kopal (2007): *kopal: Ein Service für die Langzeitarchivierung digitaler Informationen*. Projekt kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen), 2007 (s. http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf)

Lutterbeck, Bernd / Bärwolff, Matthias / Gehring, Robert A. (Hrsg.) (2007): *Open Source Jahrbuch 2007* : Zwischen freier Software und Gesellschaftsmodell. Berlin : Lehmanns Media, 2007 (s. http://www.opensourcejahrbuch.de/download/jb2007)

Reng, C. et al.. (2006): Generische Lösungen zum Datenschutz für die Forschungsnetze in der Medizin. Berlin : Medizinisch-Wissenschaftliche Verlagsgesellschaft.

15 Organisation

15.1 Einführung

Sven Vlaeminck

Die Organisation der digitalen Langzeitarchivierung (LZA) ist eine vielschichtige Aufgabe, die zahlreiche Handlungsfelder aufweist: So ist „die Informationsübernahme in ein digitales Langzeitarchiv […] nicht nur ein technischer Transfer zwischen zwei Systemen, sondern sie ist insbesondere ein Prozess mit vielen organisatorischen Anforderungen, an dessen Ende die Übernahme der Verantwortung durch das digitale Langzeitarchiv steht."[1]

Aufgrund dieser zahlreichen organisatorischen Anforderungen haben sich in den letzten Jahren verschiedene Arbeitsgruppen mit der Problematik der

1 nestor-Arbeitsgruppe Standards für Metadaten, Transfer von Objekten in digitale Langzeitarchive und Objektzugriff (Hg.): Wege ins Archiv. Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv - Version I - zur öffentlichen Kommentierung, nestor-materialien 10, Göttingen/Koblenz, November 2008, S.2. Verfügbar unter: http://nbn-resolving.de/urn:nbn:de:0008-2008103009
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

Organisation der digitalen Langzeitarchivierung beschäftigt: Allein aus dem nestor-Netzwerk entstanden zwei Publikationen, die diese Fragestellung als wesentliches Thema behandeln. Hierbei handelt es sich um den Kriterienkatalog vertrauenswürdige digitale Langzeitarchive[2] und, ergänzend dazu, um den Ratgeber Wege ins Archiv. Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv[3]. Bereits im Jahr zuvor wurden mit der Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist[4] eine ähnliche Publikation aus den USA veröffentlicht. Zudem entwickelte die Deutsche Initiative für Netzwerkinformation e. V. (DINI) bereits für das Jahr 2007 das so genannte DINI-Zertifikat[5]. Dieses zielt darauf ab, detailliert die technischen, organisatorischen und prozessualen Anforderungen an einen Dokumenten- und Publikationsservice zu beschreiben, wünschenswerte Entwicklungsmöglichkeiten im technischen und organisatorischen Bereich aufzuzeigen sowie die Einhaltung von Standards und Empfehlungen zu gewährleisten.[6]

Organisatorische Anforderungen an ein digitales Langzeitarchiv

Teil der organisatorischen Anforderungen an digitale Langzeitarchive ist es, dass sinnvolle Arbeitsabläufe oder Workflows entwickelt werden. Diese müssen sowohl den Weg eines digitalen Objekts vom Produzenten in das digitale Langzeitarchiv hinein abdecken, als auch – da der archivierte Inhalt auch von den Nutzern des digitalen Langzeitarchivs abgerufen und nachgenutzt werden soll – im umgekehrter Richtung den Weg aus dem digitalen Langzeitarchiv zu den jeweiligen Nutzern der Daten.

---

2 nestor-Arbeitsgruppe Vertrauenswürdige Archive - Zertifizierung (Hg.): Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 2, nestor-materialien 8, Frankfurt am Main, November 2008. Verfügbar unter: http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_08.pdf

3 nestor-Arbeitsgruppe Standards für Metadaten, Transfer von Objekten in digitale Langzeitarchive und Objektzugriff, 2008.

4 The RLG - National Archives and Records Administration Digital Repository Certification Task Force (Ed.): Trustworthy Repositories Audit & Certification: Criteria and Checklist, Version 1.0, February 2007. Verfügbar unter: http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91

5 Deutsche Initiative für Netzwerkinformation e.V. (DINI), Arbeitsgruppe „Elektronisches Publizieren“ DINI – Zertifikat. Dokumenten- und Publikationsserver 2007, Version 2.0, 2006. Verfügbar unter: http://edoc.hu-berlin.de/series/dini-schriften/2007-3/PDF/3.pdf

6 http://www.dini.de/service/dini-zertifikat/

Eine besondere Herausforderung besteht darin, dass „Daten […]aus üblicherweise heterogenen technischen und organisatorischen Kontexten so übernommen werden [müssen], dass sie trotzdem in ganz anderen, zukünftigen Kontexten verstehbar und nutzbar sein werden.“[7]

Für eine erfolgreiche Organisation der digitalen Langzeitarchivierung müssen unterschiedliche Prozesse definiert und umgesetzt sein:[8] So ist es zunächst notwendig, dass das digitale Langzeitarchiv seine Ziele klar definiert hat. Dazu zählt sowohl die Erklärung der Übernahme der Verantwortung für den dauerhaften Erhalt der in digitalen Objekten repräsentierten Information als auch die Bestimmung der Zielgruppe(n) sowie die Entwicklung von Kriterien zur Auswahl digitaler Objekte (etwa durch von Sammelrichtlinien, Auswahl- und Bewertungskriterien oder Kriterien der Überlieferungsbildung).

Zudem muss das digitale Langzeitarchiv seiner Zielgruppe bzw. seinen Zielgruppen eine angemessene Nutzung der durch die digitalen Objekte repräsentierten Informationen ermöglichen. Eine solche Nutzung von Informationen ist jedoch nur möglich, wenn bereits bei der Planung und Entwicklung des digitalen Langzeitarchivs Maßnahmen zum Erhalt, zur Verfügbarkeit sowie zur Interpretierbarkeit der digitalen Objekte getroffen wurden.

Zur angemessenen Nutzung eines Langzeitarchivs zählt ferner, dass Nutzer und Nutzerinnen angemessene Recherchemöglichkeiten vorfinden und die Bedürfnisse der Nutzer-Community auch bei den Dienstleistungsportfolios Berücksichtigung finden. Die transparente Darstellung der Nutzungsbedingungen und ggf. anfallender Kosten ist ebenfalls Teil der Ermöglichung einer angemessenen Nutzung.

Auch die Gewährleistung des Agierens auf der Basis rechtlicher Regelungen zählt zu den organisatorischen Fragestellungen. Diese betreffen sowohl den Bereich der Übernahme der digitalen Objekte als auch deren Archivierung und Nutzung.

Zur Schaffung von Planungs- und Rechtssicherheit sollte das digitale Langzeitarchiv nach Möglichkeit formale Übereinkünfte mit den Produzenten bzw.

---

7 nestor-Arbeitsgruppe Standards für Metadaten, Transfer von Objekten in digitale Langzeitarchive und Objektzugriff, 2008, S.2.

8 Ausführlich werden die zu lösenden organisatorischen, technischen und finanziellen Fragestellungen im „Kriterienkatalog digitale Langzeitarchive“ beschrieben. Darüber hinaus werden ina dieser Publikation der Umgang mit Objekten sowie Fragen der Infrastruktur und der Sicherheit behandelt.

Lieferanten digitaler Objekte schließen.[9] Zudem ist darauf zu achten, dass sowohl bei der Archivierung (Archivablage, Einsatz von Maßnahmen zur Langzeiterhaltung, etc.) wie auch bei der Nutzung der fraglichen Daten auf rechtliche und geschlossene vertragliche Regelungen (z.B. Urheberrecht, Datenschutz, Schutzfristen…) geachtet wird.

Eine bedeutende Herausforderung für die Organisation des digitalen Langzeitarchivs liegt in der Angemessenheit der Organisationsform und der Organisationsstrukturen des digitalen Langzeitarchivs. Dazu zählt, dass die Ziele des digitalen Langzeitarchivs kurz-, mittel- und langfristig erfüllt werden können. Die Finanzierung des digitalen Langzeitarchivs muss dazu ebenso sichergestellt sein, wie die ausreichende Präsenz qualifizierten Personals für die anfallenden Aufgaben.

Das digitale Langzeitarchiv ist dazu angehalten, langfristig zu planen und sicherzustellen, dass die übernommenen Aufgaben notfalls auch über das Bestehen des digitalen Langzeitarchivs sichergestellt werden können. Zudem ist es notwendig, dass organisatorische Maßnahmen getroffen werden, um auf wichtige Veränderungen in technischen, organisatorischen oder rechtlichen Bereichen schnell und angemessen reagieren zu können.

Abschließend ist es für die Organisation des digitalen Langzeitarchivs notwendig, ein angemessenes Qualitätsmanagement durchzuführen. Dieses ist vor allem durch die Definition und Dokumentation aller Prozesse und Verantwortlichkeiten zu gewährleisten. Als Ausgangsbasis zur Definition von Kernprozessen können beispielsweise die funktionalen Entitäten des OAIS[10], wie Aufnahme (Ingest), Archivablage (Archival Storage) und Nutzung (Access) herangezogen werden. Anhand dieser Kernprozesse können dann Unterstützungs- und Managementprozesse definiert werden, etwa in den Bereichen Datenmanagement und Qualitätsmanagement. Eine Dokumentation anhand definierter Verfahren empfiehlt sich darüber hinaus auch für Ziele, Konzepte, Spezifikationen, Implementationen, Prozesse, Software, Objekte und Metadaten etc. [11]

---

9 Ein Mustervertrag findet sich beispielsweise unter: http://www.babs-muenchen.de/content/netzpublikationen/einzelbewilligung.pdf

10 Consultative Committee for Space Data Systems (Ed.): Recommendation for Space Data System Standards. Reference Model for an Open Archival Information System (OAIS). Blue Book, January 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf

11 Vgl. nestor-Arbeitsgruppe Vertrauenswürdige Archive - Zertifizierung, 2008, S.19.

Literaturverzeichnis

nestor-Arbeitsgruppe Vertrauenswürdige Archive - Zertifizierung (Hg.): *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive.* Version 2, nestor-materialien 8, Frankfurt am Main, November 2008, http://nbn-resolving.de/urn:nbn:de:0008-2008021802

The RLG – National Archives and Records Administration Digital Repository Certification Task Force (Ed.): *Trustworthy Repositories Audit & Certification:* Criteria and Checklist, Version 1.0, February 2007, http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91

nestor-Arbeitsgruppe Standards für Metadaten, Transfer von Objekten in digitale Langzeitarchive und Objektzugriff (Hg.): *Wege ins Archiv. Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv* – Version I – zur öffentlichen Kommentierung, nestor-materialien 10, Göttingen/Koblenz, November 2008, S.2, http://nbn-resolving.de/urn:nbn:de:0008-2008103009

Deutsche Initiative für Netzwerkinformation e.V. (DINI), Arbeitsgruppe „Elektronisches Publizieren" (Hg.): *DINI – Zertifikat. Dokumenten- und Publikationsserver 2007*, Version 2.0, 2006, http://edoc.hu-berlin.de/series/dini-schriften/2006-3/PDF/3.pdf

Lyon, Liz (Ed.): *Dealing with Data: Roles, Rights, Responsibilities and Relationships Consultancy Report*, Bath, 19. Juni 2007, S. 55f, http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Consultative Committee for Space Data Systems (Ed.): *Recommendation for Space Data System Standards.* Reference Model for an Open Archival Information System (OAIS). Blue Book, January 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf

15.2 Perspektiven der Beschreibung

Christian Keitel

Die Organisation der Archivierung wird aus unterschiedlicher Perspektive angedacht. Die verschiedenen Ansätze werden vorgestellt. Mögliche übergreifende Dienstleistungen werden ebenso beschrieben wie die Aufgaben, die sich den Phasen im Lebenszyklus der Unterlagen – Produktion, Archiv, Benutzung – zuordnen lassen. Abschließend werden einige konkrete Beispiele skizziert.

Wie kann die Archivierung digitaler Objekte organisatorisch umgesetzt werden? Zwei Ansätze versuchen diese Frage auf unterschiedlichem Weg zu beantworten. Dabei liegt der Hauptunterschied in der Ausgangsposition, von der aus der Gedankengang entwickelt wird. Der eher traditionell zu nennende Ansatz geht von den bei der Archivierung insgesamt anfallenden Aufgaben aus und beschreibt dann, von wem sie umgesetzt werden können. Die entsprechenden Studien setzen zumeist das OAIS-Funktionsmodell an den Anfang ihrer Überlegungen. Der Ansatz findet sich eher bei den Einrichtungen, die bereits ein klares Mandat für die Archivierung der Objekte besitzen. Als Beispiel können hier die staatlichen Archive genannt werden, die seit jeher für die in den Behörden entstandenen Unterlagen zuständig sind. Daneben gibt es seit wenigen Jahren vor allem im Bereich der Digitalisierung und der Archivierung naturwissenschaftlicher Daten Versuche, zunächst nach den Personen oder Einrichtungen zu fragen, die mit der Archivierung im weitesten Sinne befasst sind oder befasst sein müssten. Fragebögen stehen denn auch häufig am Anfang der Studien, die diesen Ansatz verfolgen. 2007 konnte Liz Lyon so überzeugend sechs unterschiedliche Rollen zusammen mit ihren Rechten, Verantwortlichkeiten und wechselseitigen Beziehungen beschreiben.[12] Konkret unterscheidet sie zwischen Wissenschaftler, Institution, Datenzentrum, Benutzer, Träger (Geldgeber) und Verleger. Die Vertreter dieses Ansatzes fragen dann in einem zweiten Schritt, welche Aufgaben von den Beteiligten in Zukunft übernommen werden sollten. Ziel ist es, auf diese Weise ein tragfähiges Geschäftsmodell zu erstellen.[13] Dieser zweite Ansatz findet sich eher bei den Objektarten, für die

12 Lyon (2007).

13 Im Rahmen der DFG Aktionslinie:„Entwicklung von Organisations- und Geschäftsmodellen zur Langzeitarchivierung“ sind hier sowohl das in Göttingen angesiedelte KolaWiss-Projekt (Kooperative Langzeitarchivierung für Wissenschaftsstandorte, http://kolawiss.uni-goettingen.de/) als auch das von der Bayerischen Staatsbibliothek und der Universität der Bundeswehr in München getragene DFG-Projekt „Entwicklung von Organisations- und Geschäftsmodellen für die

bislang keine Zuständigkeiten geklärt sind. So konstatiert etwa der Zwischenbericht der Blue Ribbon Task Force als eines der zentralen Probleme digitaler Langzeitarchivierung „Confusion and/or lack of alignment between stakeholders, roles, and responsibilities with respect to digital access and preservation.“[14] Die beiden Modelle ergänzen sich komplementär, sie können als zwei Seiten einer Medaille angesehen werden. Die Stärke des Rollenmodells liegt in der gegenseitigen Abgrenzung einzelner Einrichtungen. Dagegen macht es das OAIS möglich, die Aufgabenverteilung innerhalb eines Archivs auf abstrakte Weise zu beschreiben.

Übergreifende Dienstleistungen

Drei Bereiche sind es, die in der einschlägigen Fachliteratur als Kandidaten für übergreifende Dienstleistungen genannt werden: Die Rede ist von fachlichen, technischen und administrativen Diensten. Sie bieten eine Möglichkeit, konkrete Schritte in Richtung einer arbeitsteiligen kooperativen Umsetzung der Aufgabe vorzunehmen. Auf der anderen Seite werden diese Bereiche vom OAIS-Standard als Bestandteile eines Archivsystems beschrieben. Auch werden ihre Funktionen teilweise durch bereits bestehende digitale Archive ausgeübt.

Fachliche Dienstleistungen für den ganzen Lebenszyklus werden seit einigen Jahren gerne unter dem Stichwort *digital curation* zusammengefaßt. Bereits 2003 haben Philip Lord und Alison MacDonald hierzu organisatorische Überlegungen angestellt.[15] Das digitale Archiv selbst wird von ihnen in einen größeren Rahmen mit drei weiteren Komponenten eingebettet. Sowohl die Produktion als die Benutzung sollen durch disziplinspezifische und operative kuratorische Dienstleistungen unterstützt werden, welche ihrerseits durch allgemeine disziplinbezogene zentrale Dienste unterstützt werden, die als nationale Kompetenzzentren gedacht werden können. Den Abschluss bilden disziplinübergreifende Dienste wie Schulungen oder die Überwachung von Technologie und Dateiformaten. Die Autoren sehen hier ein weiteres nationales Kompetenzzentrum und setzen dieses explizit mit dem Digital Curation Centre in Edinburgh gleich.

Ein übergreifender technischer Support ist in verschiedener Hinsicht denkbar. Zunächst müssen die Daten während aller Phasen physisch erhalten werden. *Bitstream-Preservation* stellt durch den ganzen Lebenszyklus hindurch erst

Langzeitarchivierung der Digitalisate aus DFG-geförderten Projekten“, s. Beinert et al. (2008).

14 Sustaining the digital investment (2008), S. 2.

15 Lord (2003).

die Grundlage, auf der alle anderen Maßnahmen aufbauen können. Auch die Bereitstellung der notwendigen Hardware sowie der Netzinfrastruktur kann übergreifend geregelt werden. Bei der Software lassen sich vergleichbar allgemeine Aussagen nur schwer treffen.

Übergreifende administrative Dienstleistungen lassen sich schließlich in größeren verteilten Forschungseinrichtungen wie z.B. Universitäten beschreiben. Das KolaWiss-Projekt[16] benennt hier zunächst das Präsidium und die Hochschulleitung, dann den Datenschutzbeauftragten. Diese Stellen kennen eine Zuständigkeit sowohl für die kurzfristige Sphäre der Produktion als auch die des langfristig angelegten Bereiches der Archivierung.

15.2.1 Aufgaben im Lebenszyklus

Produktion

Wer erstellt die interessierenden Daten? Bereits hier geben die einschlägigen Studien unterschiedliche Antworten. Das KolaWiss-Projekt unterscheidet, ob die Informationen durch einzelne Forscher oder ein Institut, instituts- bzw. organisationsübergreifend oder auch nur zeitlich befristet zusammengestellt werden. Dagegen ist bei Lyon die Produktion nur auf den Wissenschaftler selbst bezogen. Sie trennt stattdessen im Bereich der Produktion eine weitere Rolle ab. Die Institution gilt ihr nicht als Produzent, sie steht für den kurzfristigen Erhalt der Daten, bevor diese Aufgabe dann zur langfristigen Aufbewahrung an ein Datenzentrum übergeht. Obwohl sie bei ihrer Aufgabenbeschreibung davon ausgeht, dass sowohl Produzent als auch Datenzentrum (Archiv) kuratorische Aufgaben vorzunehmen haben, obwohl sie also das Konzept der *digital curation* reflektiert, trennt sie beide Bereiche doch klar und in der Tradition des OAIS voneinander. Tatsächlich sind die meisten digitalen Archive organisatorisch vom Produzenten getrennt. Dennoch wird gerade diese Trennung manchmal relativiert oder auch aufgehoben:

Archivierung durch die Produzenten (1): 1996 wurde den australischen Behörden nach der Theorie des *records continuum* auferlegt, alle alten, im Dienst nicht mehr benötigten Dokumente dauerhaft selbst zu verwahren. Den Archiven kam dabei die Rolle zu, das Funktionieren des Konzepts sicherzustellen, also eine Art „Archivierungspolizei" zu spielen. Bereits 2000 kehrte das Australische Nationalarchiv wieder zu seiner traditionellen Politik zurück, d.h. zur Übernahme dieser Dokumente. Begründet sein dürfte diese Rückkehr in dem Umstand,

16 http://kolawiss.uni-goettingen.de/

dass nur Archive und Bibliotheken ein genuines Interesse an der Erhaltung von Informationen haben, die in den Augen ihrer Ersteller „veraltet" sind. Erst dieses Interesse gewährleistet, dass vermeintlich uninteressante Daten weiterhin gepflegt werden.

Archivierung durch die Produzenten (2): Die Systeme der Umweltbeobachtung verwahren aktuell produzierte Daten zusammen mit den Daten vergangener Jahrzehnte. Die einzelnen Informationen sollen dauerhaft im selben System und unter denselben Namen aufgefunden und angesprochen werden, die systemische Einheit dieser Daten ist über einen langen Zeitraum hinweg erwünscht. Die Information veraltet also im Gegensatz zum beschriebenen australischen Beispiel theoretisch nie. Vergleichbare Systeme werden derzeit in vielen Naturwissenschaften aufgebaut. Diese Form der Digital Curation geht über die von Lyon vertretene Rollenverteilung weit hinaus.

Archivisches Engagement bei den Produzenten: Seit über 15 Jahren engagieren sich die klassischen Archive in den Behörden bei der Einführung elektronischer Akten und anderer digitaler Systeme. Ihr Motiv: Bei der Einführung eines Systems werden die Grundlagen dessen gelegt, was dann später im Archiv ankommt. Danach ist es weniger aufwändig, in der Behörde Dinge grundsätzlich zu regeln, als später jedes Objekt einzeln nachbearbeiten zu müssen. Im DOMEA-Konzept (Dokumentenmananagement und Elektronische Archivierung) werden die beiden Bereiche auch begrifflich zusammengezogen. In eine ähnliche Richtung gehen auch Überlegungen, in größeren Organisationen, die gleichermaßen für die Produktion und Erhaltung digitaler Objekte zuständig sind (z.B. Universitäten) einen *Preservation Officer* anzustellen, der durch den gesamten Lebenslauf der Objekte angesprochen werden kann.

Archive werden zu Produzenten: Durch die Digitalisierungsstrategien der Archive und Bibliotheken mutieren diese klassischen Gedächtnisinstitutionen auf einmal selbst zu Datenproduzenten. Zunächst bedarf es zusätzlicher Qualitätssicherungsmaßnahmen für die Digitalisate. Mittel- und langfristig muss auch das Verhältnis von Produktion und Archiv neu bestimmt werden.

Archiv

Interne Organisationsmodelle zur Arbeitsteilung in einem digitalen Archivs wurden bislang kaum veröffentlicht. Teilweise dürfte dies darin begründet sein, dass noch immer viele Aktivitäten nur einen zeitlich befristeten Projektstatus besitzen. Es bietet sich daher an, die veröffentlichten Details einem abstrakteren Rahmen einzufügen, wie ihn das OAIS-Funktionsmodell anbietet. Gesondert beschrieben wird im Anschluss das Modell der Koninklijke Bibliotheek der

Niederlande, das eine alternative Darstellung der in einem Archiv anfallenden Aufgaben anbietet. Schließlich werden noch weitere Faktoren genannt, die bei der Organisation eines digitalen Archivs zu berücksichtigen sind.

Darstellung nach Aufgaben (OAIS)

Im *Ingest* werden die Übernahmepakete (SIPs) definiert, entgegen genommen, überprüft und in Archivierungspakete (AIPs) umgewandelt. Auch bei einer festen Trennung zwischen Produzenten und Archiv können die einzelnen Aufgaben sehr unterschiedlich aufgeteilt werden. Hierzu gehören die Auswahl der Objekte, ihre Ausstattung mit Metadaten und die ggf. erforderliche Migration der Objekte in ein archivierungsfähiges Format. Entsprechend kann sich die dem Archiv verbleibende Ingest-Aufgabe vor allem administrativ gestalten (es gibt dem Produzenten die entsprechenden Vorgaben) oder zunehmend auch technische Komponenten enthalten (es setzt diese Punkte selbst um). Die Entscheidung für eine der beiden Optionen ist wesentlich von der Gleichartigkeit der Objekte abhängig: Erst wenn sich die Objekte sehr stark gleichen, kann die Zahl der Vorgaben so weit reduziert werden, dass eine entsprechende Automatisierung auch erfolgreich umgesetzt werden kann. Bei stark differierenden Objekten lassen sich diese Regeln nicht in einer vergleichbar umfassenden Weise aufstellen, weshalb die Aufgaben vom Archiv selbst übernommen werden müssen, was dessen Aufwand entsprechend erhöht. Im letztgenannten Fall können dann weitere Teilaufgaben gebildet werden. Beispielsweise kann die Metadatenerfassung in zwei aufeinanderfolgende Schritte aufgespaltet werden: a) Anlegen erster identifizierender Metadaten und b) nähere Beschreibung im Zuge der weiteren Bearbeitung.

Im Bereich *Archival Storage* werden die AIPs über einen langen Zeitraum gespeichert. Der Zustand der Speichermedien wird kontinuierlich überwacht, im Bedarfsfall werden einzelne Medien ersetzt, regelmäßig werden auch ganze Medien-Generationen in neuere Speichertechnologien migriert. Neben Hardware und Software sind hier also vor allem IT-Kenntnisse erforderlich. Es ist daher auch der Bereich, der am ehesten von den klassischen Gedächtnisinstitutionen an externe Rechenzentren ausgelagert wird. Andererseits unterscheiden sich die Anforderungen der digitalen Archivierung z.T. erheblich von denen, die gewöhnlich an Rechenzentren gestellt werden. Die National Archives and Records Administration (NARA) der Vereinigten Staaten hat daher Anfang der 1990er Jahre den Bereich wieder ins Haus geholt.

In den meisten Gedächtniseinrichtungen kann der Bereich des *Data Management* auf eine lange Tradition zurückblicken. Hier werden die identifizierenden, beschreibenden und administrativen Metadaten gepflegt. Bibliotheken spre-

chen von Katalogen, Archive von Findmitteln und –büchern. Sofern nicht ein eigenes Recherchesystem für die digitalen Objekte aufgebaut wird, liegt es nahe, die Verantwortung für diesen Bereich an die Organisationseinheiten zu delegieren, die bereits für die Beschreibung der analogen Objekte zuständig sind.

Eine zentrale Rolle kommt schließlich dem Bereich der Digitalen Bestandserhaltung, also des *Preservation Planning* zu. Digitale Archivierung erfordert eine kontinuierliche aktive Begleitung der archivierten Objekte. Wesentlich ist die Terminierung und Koordination der einzelnen Erhaltungsprozesse. Schnittstellen bestehen zu den Bereichen Ingest, Archival Storage und Data Management. Zu diesem Bereich wurden bislang nur wenige organisatorische Überlegungen veröffentlicht.

Darstellung nach Kompetenzbereichen

Die Koninklijke Bibliotheek der Niederlande hat einen zu OAIS alternativen Entwurf einer Archivbeschreibung vorgelegt.[17] Die Autoren der Studie kennen nicht nur fünf Kompetenzbereiche, sie unterscheiden in diesen jeweils noch in die Ebenen der Anweisung, Kontrolle und Ausführung (*direct, control, execute*). *Service management* enthält so auf der obersten Ebene eine release strategy und einen distribution plan. Auf der mittleren Ebene fließen diese strategischen Vorstellungen in die Rechteverwaltung ein. Im operativen Bereich finden sich dann Lesesaal, Internet und andere praktische Dienstleistungen. Weitere Bereiche sind das *Collection Management*, *Preservation Management*, *Business Management* und *IT Management*. Im operativen Bereich besitzen Collection und Preservation Management drei gemeinsame Aufgaben: *Characterisation*, *Validation* und *Cataloguing*. Die anderen Aufgaben sind klar voneinander getrennt.

Weitere Faktoren

Über die Aufgaben und Kompetenzbereiche hinaus können noch weitere Faktoren genannt werden, die bei der Organisation der digitalen Archivierung zu berücksichtigen sind. Genannt werden können die Größe der Einrichtung, ihre sonstigen Aufgaben und die Qualifikation ihres Personals. Sehr große Archive können zu jeder Einheit des OAIS-Funktionsmodells mindestens eine administrative Einheit bilden. Zusätzlich kann noch ein Forschungsbereich ausgegliedert werden. Kleinere Archive sind dagegen gezwungen, mit weniger administrativen Einheiten auszukommen. Bei klassischen Gedächtniseinrichtungen stellt sich die Frage, welche Aufgaben unabhängig vom Medientyp bearbeitet werden können. Sollen z.B. digitale und analoge Objekte, sollen Datenbanken

17 Van Diessen (2008).

und Publikationen im PDF-Format zusammen beschrieben werden? In zahlreichen Bereichen sind zudem sowohl die Kenntnisse traditionell ausgebildeter Archivare oder Bibliothekare als auch ausgeprägte IT-Kenntnisse erforderlich. Die Organisation ist daher auch von dem bereits bestehenden Personalbestand der Einrichtung und den Möglichkeiten zur Neueinstellung abhängig.

Benutzung

Ebenso wie die Produktion lässt sich auch dieser Bereich nicht allgemeingültig von dem des Archivs abgrenzen. Wo wird recherchiert? Ist dies noch, wie OAIS vermutet, innerhalb des Data Managements des Archivs oder greift der Benutzer auf ein institutionenübergreifendes Internetportal zu? Recherchiert er also innerhalb oder außerhalb des Archivs? Ähnliche Fragen lassen sich auch bei der Benutzung selbst anstellen: Muss der Benutzer in den Lesesaal oder auf die Internetseiten des Archivs kommen oder bekommt er ein Datenpaket ausgehändigt, das er dann an einem beliebigen Ort einsehen kann? Zwar greifen Recherche und Benutzung letztlich auf Daten zu, die im Archiv vorgehalten werden. Dennoch ist es im Sinne der verteilten Rollen denkbar, dass diese Rolle auch von einer archivübergreifenden zentralen Recherche- und Benutzungsstelle ausgeübt werden kann. In diesem Fall wäre der Begriff des Archivs neu zu überdenken.

15.2.2 Beispiele/Umsetzung in die Praxis

Centre national d'études spatiales (CNES)

Die französische Raumfahrtagentur CNES archiviert fast ausschließlich digitale Daten. Es wurden drei administrative Einheiten gebildet: a) Ingest, b) Archival Storage und c) Data Management und Access. Im Ingest arbeiten Archivare und Computerspezialisten zusammen. Der Archivar definiert die zu übernehmenden Objekte, überprüft sie auf ihre Vollständigkeit und strukturiert sie. Der Computerspezialist definiert Daten und Metadaten, nimmt die physische Übernahme und die Validierung vor und entwickelt entsprechende Tools. Das neue Berufsbild des *Digital Data Manager* kann auf beiden Gebieten des Ingest tätig werden. Beim Archival Storage werden ausschließlich Computerspezialisten eingesetzt. Seit 1994 wird dieser Bereich vom STAF (Service de Transfert et d'Archivage de Fichiers) ausgeführt. Die OAIS-Bereiche Data Management und Access werden beim CNES zusammengezogen. Im Vordergrund stehen Datenbank-, Retrieval- und Internettechnologien, daneben werden vertiefte

Kenntnisse über das Archiv benötigt. Das Funktionieren des Archivs wird durch eine Koordinationsstelle, bewusst klein konzipierte Überlappungsbereiche und die weitgehende Unabhängigkeit der einzelnen Einheiten gewährleistet.

The National Archives (UK)

Die National Archives haben mehrere objektspezifische Ansätze zur digitalen Archivierung entwickelt, die zusätzlich von zentralen Systemen (z.B. die Formatdatenbank PRONOM) unterstützt werden. Seit 2001 ist zudem für die Erhaltung von *born digital material* nicht mehr das Records Management Department sondern das neu eingerichtete Digital Preservation Department zuständig. Für strukturierte Daten wurde 1997 eine Kooperationsvereinbarung mit dem Rechenzentrum der Londoner Universität (University of London Computer Centre) geschlossen, in deren Folge das National Digital Archive of Datasets (NDAD) 1998 in Betrieb genommen werden konnte. Die National Archives sind für die Auswahl der Daten und die Definition der Service-Levels zuständig, NDAD für alle weiteren Aufgaben (explizit unterschieden werden Ingest, Preservation und Access). Im NDAD arbeiten zwölf Personen in vier Disziplinen: Die Project Archivists treffen zentrale Entscheidungen über die Organisation des Archivs, Katalogisierung und Indexierung und leiten die Computer-Spezialisten an. Die Archive Assistants sind für die Benutzerbetreuung zuständig. Sie unterstützen die Project Archivists z.B. durch Einscannen der Papierdokumentation. Die Data Specialists sind für die technische Umsetzung der getroffenen Entscheidungen zuständig. Der Systems Support Staff stellt schließlich das Funktionieren von Hard- und Software sicher. Für die Archivierung elektronischer Records (Akten) wurde in den National Archives Mitte der 1990er Jahre das EROS-Projekt aufgesetzt, das dann im Seamless-Flow-Programm fortgesetzt wurde. Gleichzeitig werden im 2003 in den National Archives gegründeten *Digital Archive* bereits Records übernommen und Erfahrungen aufgebaut. Für die Archivierung von Internetseiten haben sich die National Archives 2003 mit der British Library, den Nationalbibliotheken von Wales und Schottland, JISC und dem Wellcome Trust zum UK Web Archiving Consortium zusammengeschlossen, um eine gemeinsame Infrastruktur zur Web-Archivierung aufzubauen.

Deutsche Nationalbibliothek (DNB) und Staats- und Universitätsbibliothek Göttingen (SUB)

Die Deutsche Nationalbibliothek und die Staats- und Universitätsbibliothek Göttingen haben ihre Lösung zur Archivierung digitaler Objekte im Projekt KOPAL[18] gemeinsam mit der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und der IBM Deutschland entwickelt. Die Partner gehen von einem arbeitsteiligen Vorgehen aus: Die Übernahme und Aufbereitung der AIPs liegt in den Händen der beteiligten Bibliotheken und erfolgt durch eine OpenSource-Software. Die fertigen AIPs werden dann per Fernzugriff zentral im Rechenzentrum der GWDG gespeichert. Dabei kommt das durch die IBM entwickelte DIAS-System zu Einsatz. Die Benutzung erfolgt dann wiederum durch Fernzugriff bei den beiden Bibliotheken. Weitere Aufschlüsse soll für Göttingen das Projekt KolaWiss erarbeiten.

Literatur

Beinert, Tobias et al. (2008): *Development of Organisational and Business Models for the Long-Term Preservation of Digital Objects,* http://www.bl.uk/ipres2008/presentations_day1/04_Lang.pdf

Brown, Adrian (2006 a): *Archiving Websites. A Practical Guide for Information Management Professionals,* London.

Brown, Adrian (2006 b): *Developping practical approaches to active preservation,* in: Proceedings of the 2nd International Conference on Digital Curation, Glasgow.

van Diessen, Raymond J. / Sierman, Barbara / Lee, Christopher A. (2008): *Component Business Model for Digital Repositories: A Framework for Analysis,* http://www.bl.uk/ipres2008/presentations_day1/van_Diessen_a03.pdf

DOMEA-Konzept: Das Organisationskonzept, die Erweiterungsmodule und weitere Informationen finden sich auf den Seiten http://www.verwaltung-innovativ.de unter dem Stichwort "Organisation"

Huc, Claude (2004): *An organisational model for digital archive centres*, http://www.erpanet.org/events/2004/amsterdam/presentations/erpaTraining-Amsterdam_Huc.pdf

Jones, Richard (2006): Theo Andrew, John MacColl, *The Institutional Repository,* Oxford 2006

KolaWiss-Projekt (Kooperative Langzeitarchivierung für Wissenschaftsstandorte), http://kolawiss.uni-goettingen.de/

18 http://kopal.langzeitarchivierung.de/

Lord, Philip/Macdonald/Alison (2003): *e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*, http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lyon, Liz (2007): *Dealing with Data: Roles, Rights, Responsibilities and Relationships,* http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Reference Model for an Open Archival Information System (OAIS), Blue Book 2002, http://www.ccsds.org/publications/archive/650x0b1.pdf

Sleeman, Patricia (2004): It's Public Knowledge: *The National Digital Archive of Datasets*. In: Archivaria 58 (2004), S. 173 – 200.

Sustaining the digital investment: *Issues and Challanges of Economically Sustainable Digital Preservation* (2008), http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

16 Recht

16.1 Einführung

Mathias Jehn

Die Langzeitarchivierung digitaler Dokumente stellt Gedächtnisinstitutionen aber nicht nur in technischer Hinsicht vor ganz neue Herausforderungen. Auch in juristischer Hinsicht ist die Archivierung von gedrucktem Material ganz anders zu beurteilen als die Archivierung von digitalen Daten. Diesen juristischen Aspekten der Langzeitarchivierung ist der Beitrag von Arne Upmeier gewidmet.

Während es Gedächtnisorganisationen bisher mit Objekten zu tun hatten, deren Eigentümer sie waren und deren Benutzung und Erhaltung sie als Eigentümer allein verantworteten, ist die Situation bei unkörperlichen, digitalen Objekten rechtlich eine völlig andere. Im digitalen Raum ist bereits jede technische Aktivierung von Inhalten als Vervielfältigungsakt urheberrechtlich relevant. Die Entscheidung etwa, ob ein Buch aufgeschlagen werden darf, kann ein Eigentümer des Buches alleine treffen (ohne also Autor oder Verlag um Zustimmung

bitten zu müssen). Liegt der gleiche Text aber in elektronischer Form vor, ist das dem Aufschlagen entsprechende Aufrufen auf dem Computer eine urheberrechtlich relevante Vervielfältigung, die die Rechte von Autor oder Verlag tangieren kann. In ähnlicher Weise kann der Eigentümer eines historischen Dokuments alleine entscheiden, ob das Papier einer chemischen Entsäuerung zugeführt werden soll, um es der Nachwelt zu erhalten. Digitale Quellen können aber nur für die Nachwelt bewahrt werden, wenn sie regelmäßig vervielfältigt und gegebenenfalls auch (z.B. durch Formatänderungen) in ihrer Datenstruktur verändert werden. Im Gegensatz zu einer Papierentsäuerung berühren auch diese Tätigkeiten das Urheberrecht.

Last not least ist die Archivierung von Daten, egal welcher Art, kein Selbstzweck. Die Daten sollen irgendwann, irgendwem in irgendeiner Form wieder präsentiert werden. Die rechtlichen Voraussetzungen (und Möglichkeiten), wann und wie digitale Dokumente wieder zugänglich gemacht werden dürfen, sind teilweise ganz andere als bei den vertrauten analogen Objekten.

16.2 Rechtliche Aspekte

Arne Upmeier

Die Langzeitarchivierung digitaler Dokumente stellt Gedächtnisinstitutionen nicht nur in technischer Hinsicht vor ganz neue Herausforderungen.[1] Auch in juristischer Hinsicht ist die Archivierung von gedrucktem Material ganz anders zu beurteilen als die Archivierung von digitalen Daten. Während es Gedächtnisorganisationen bisher mit Objekten zu tun hatten, deren Eigentümer sie waren und deren Benutzung und Erhaltung sie als Eigentümer allein verantworteten, ist die Situation bei unkörperlichen, digitalen Objekten rechtlich eine völlig andere. Im digitalen Raum ist bereits jede technische Aktivierung von Inhalten als Vervielfältigungsakt urheberrechtlich relevant. Die Entscheidung etwa, ob ein Buch aufgeschlagen werden darf, kann ein Eigentümer des Buches alleine treffen (ohne also Autor oder Verlag um Zustimmung bitten zu müssen). Liegt der gleiche Text aber in elektronischer Form vor, ist das dem Aufschlagen entsprechende Aufrufen auf dem Computer eine urheberrechtlich relevante Vervielfältigung, die die Rechte von Autor oder Verlag tangieren kann. In ähnlicher Weise kann der Eigentümer eines historischen Dokuments alleine entscheiden, ob das Papier einer chemischen Entsäuerung zugeführt werden soll, um es der Nachwelt zu erhalten. Digitale Quellen können aber nur für die Nachwelt bewahrt werden, wenn sie regelmäßig vervielfältigt und gegebenenfalls auch in ihrer Datenstruktur verändert werden (Migrationen). Im Gegensatz zu einer Papierentsäuerung berühren auch diese Tätigkeiten das Urheberrecht.

Die sich aus der gewachsenen Bedeutung des Urheberrechtes ergebenden Spannungen zwischen Archivierungsinteressen und betroffenen Urheberrechten sind kein ausschließlich deutsches Phänomen, sondern bereiten Langzeitarchivierungsprojekten weltweit zunehmende Schwierigkeiten.[2] Die prak-

1 Zum ganzen Thema ausführlicher: Euler, Ellen: Zur Langzeitarchivierung digital aufgezeichneter Werke und ihrer urheberrechtlichen Einordnung und Beurteilung. In: AfP 2008/5, S. 474-482. Im Projekt nestor gibt es innerhalb der „AG Kooperative Langzeitarchivierung" eine „TaskForce Recht", die sich speziell mit Rechtsfragen der Langzeitarchivierung beschäftigt.

2 Stellvertretend auch für viele kleinere Projekte und Initiativen weltweit sei hier eine große gemeinsame Studie der Library of Congress, des JISC (Vereinigtes Königreich), des OAK Law Projects (Australien) und der SURFfoundation (Niederlande) aus dem Juli 2008 erwähnt: „International Study on the Impact of Copyright Law on Digital Preservation" (http://www.digitalpreservation.gov/library/resources/pubs/docs/digital_preservation_final_report2008.pdf)
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

tischen Schwierigkeiten werden noch verschärft durch die – für den juristischen Laien kaum noch zu durchschauende – **Kompliziertheit** des Urheberrechts. Sehr vieles hängt von den konkreten Umständen im Einzelfall ab und lässt sich nicht generalisieren. Auch die folgenden Ausführungen bleiben daher notwendig allgemein und vieles – im Einzelfall Entscheidendes – muss außen vor bleiben.

Was darf archiviert werden?

Ein digitales Objekt muss über eine bestimmte Schöpfungshöhe verfügen, um überhaupt im Sinne des Urheberrechts schutzwürdig zu sein, d.h. es muss über einen bestimmten geistigen Inhalt, der in einer bestimmten Form Ausdruck gefunden hat und eine gewisse Individualität verfügen. Nicht jeder Text oder jedes Musikstück unterliegt daher automatisch dem Urheberrecht. Auch eine ungeordnete Sammlung von wissenschaftlichen Rohdaten ist im Regelfall nicht urheberrechtlich geschützt. Digitale Objekte, die danach gar nicht dem Urheberrecht unterliegen, können im Allgemeinen unproblematisch archiviert werden.

Rechtlich unproblematisch sind auch Dokumente, die aus dem einen oder anderen Grunde **gemeinfrei** sind. Hierzu zählen beispielsweise amtliche Werke § 5 Urheberrechtsgesetz (UrhG), wie etwa Gesetze oder Verordnungen und auch alle Werke, deren Urheberrechtsschutz bereits abgelaufen ist. Dies ist in der Regel siebzig Jahre nach dem Tode des Urhebers der Fall (§ 64 UrhG).[3]

Gesetzlich bisher nur sehr unzureichend geregelt ist der Umgang mit sogenannten „verwaisten Werken" (*orphan works*) bei denen der Urheber nicht mehr zu ermitteln ist oder bei denen es aus anderen Gründen schwierig oder gar unmöglich ist, die genaue Dauer des Urheberrechtsschutzes zu bestimmen.[4]

Juristisch betrachtet, ist die Archivierung von digitalen Objekten vor allen Dingen deswegen problematisch, weil die Objekte im Normalfall für die Archivierung **kopiert** werden müssen. Für das Kopieren von Werken stellt das deutsche Urheberrecht aber bestimmte Hürden auf.

Unter bestimmten Umständen dürfen auch urheberrechtlich geschützte Werke kopiert und archiviert werden. Der einfachste Fall ist das Vorliegen einer ausdrücklichen oder konkludenten Zustimmung des Urheberrechtsinhabers.

---

3 In Einzelfällen kann es auch bei gemeinfreien Werken und digitalen Objekten, die nicht dem Urheberrecht unterliegen rechtliche Hindernisse geben, die eine freie Verwertung untersagen (z.B. aus dem Wettbewerbsrecht.). Die sollen an dieser Stelle aber nicht weiter diskutiert werden. Näher dazu: Rehbinder: Urheberrecht, Rn. 126, 534.

4 Spindler, Gerald / Heckmann, Jörn: Retrodigitalisierung verwaister Printpublikationen – Die Nutzungsmöglichkeiten von „orphan works" de lege lata und ferenda. In: GRUR Int 2008/4, S. 271-284.

Bei Internetpublikationen ist das häufig der Fall, etwa wenn auf bestimmte Lizenzmodelle Bezug genommen wird (GNU *GPL*, Creative Commons etc.). Aus dem bloßen Einstellen von Inhalten im Internet alleine kann aber nicht auf eine konkludente Zustimmung geschlossen werden. Alleine aus der Tatsache, dass jemand etwas öffentlich zugänglich macht, kann nämlich nicht geschlossen werden, dass er auch damit einverstanden ist, wenn sein Angebot kopiert und dauerhaft gespeichert wird (und die Kopie womöglich seinem weiteren Zugriff entzogen ist). Zudem sind Anbieter und Urheber eines Internetangebots oft nicht identisch. Dann kann der Anbieter einem Dritten schon deswegen kein Recht zur Vervielfältigung einräumen, weil er selbst im Zweifel dieses Recht nicht hat. Anders ausgedrückt: Es ist ohne zusätzliche Zustimmung nicht erlaubt, eine interessant erscheinende Website zu Archivierungszwecken zu kopieren. Ausnahmen können sich aber ergeben, wenn zugunsten der archivierenden Institution eine spezialgesetzliche Ermächtigung besteht. Dies kann beispielsweise im Bundesarchivgesetz oder im Gesetz über die Deutsche Nationalbibliothek der Fall sein.[5]

Wie darf gesammelt werden?

Digitale Langzeitarchive lassen sich im Prinzip auf zweierlei Weisen füllen. Zum einen können analoge oder digitale Objekte, die sich bereits im Besitz einer archivierenden Institution befinden, ins Archiv übernommen werden. Im Regelfall setzt dies die vorherige Anfertigung einer Archivkopie oder, im Falle von analogen Objekten, deren Digitalisierung voraus. Zum anderen können auch Objekte, die sich nicht im Besitz der Institution befinden (sondern beispielsweise frei zugänglich im Internet) in das Archiv übernommen werden. Beide Wege sind nur innerhalb bestimmter rechtlicher Grenzen erlaubt. Das Problem ist auch hier jeweils, dass das Anfertigen von Vervielfältigungen nicht gemeinfreier Werke regelmäßig einer Zustimmung des Urheberrechtsinhabers bedarf. Es gibt jedoch wichtige Ausnahmen.

Anfertigung von Archivkopien

Auf den ersten Blick erscheint es naheliegend, von ohnehin vorhandenen digitalen Objekten Kopien anzufertigen, um diese dauerhaft zu archivieren.

---

5 Vgl. Heckmann, Jörn / Weber, Philipp: Elektronische Netzpublikationen im Lichte des Gesetzes über die Deutsche Nationalbibliothek. In: AfP 2008/3, S. 269-276; Steinhauer, Eric: Pflichtablieferung von Netzpublikationen. Urheberrechtliche Probleme im Zusammenhang mit der Pflichtablieferung von Netzpublikationen an die Deutsche Nationalbibliothek. In: K&R 2009/3, S. 161-166.

Ebenso naheliegend scheint es, analoge Objekte, die sich sowieso im Besitz der archivierenden Institution befinden, zu digitalisieren und die Digitalisate zu archivieren.

Die wichtigste Norm im Urheberrecht, die eine Anfertigung von solchen Archivkopien auch ohne Zustimmung eines Urhebers erlaubt, steht in § 53 Abs. 2 Satz 1 Nr. 2 UrhG. Demnach sind Vervielfältigungen (und darum handelt es sich bei einer Digitalisierung) gestattet, wenn die Vervielfältigung ausschließlich zur Aufnahme in ein eigenes Archiv erfolgt. Dies gilt aber nur mit wichtigen Einschränkungen:

- Die Vervielfältigung darf ausschließlich der Sicherung und internen Nutzung des vorhandenen Bestandes dienen (Archivierungszweck). Unzulässig ist hingegen die Verfolgung sonstiger Zwecke, wie etwa einer Erweiterung des eigenen Bestandes.
- Als Kopiervorlage muss ein „eigenes Werkstück" dienen. Für jede einzelne Archivierung ist dabei jeweils ein Original im Eigentum der archivierenden Institution erforderlich, selbst dann, wenn die ansonsten identischen Kopien nur unter anderen Schlagworten abgelegt werden sollen.[6]
- Es muss sich um ein Archiv handeln, das im öffentlichen Interesse tätig ist und keinerlei wirtschaftlichen Zweck verfolgt. Gewerbliche Unternehmen, anders als beispielsweise gemeinnützige Stiftungen, sind also nicht privilegiert und dürfen ohne ausdrückliche Zustimmung der Urheberrechtsinhaber keine elektronischen Archive anlegen. Ihnen bleibt nur die analoge Archivierung, beispielsweise durch Mikroverfilmung.
- Von „Datenbankwerken" dürfen keine Archivkopien angefertigt werden (§ 53 Abs. 5 UrhG). „Datenbankwerke" sind Sammlungen von „Werken, Daten oder anderen unabhängigen Elementen, die systematisch oder methodisch angeordnet und einzeln mit Hilfe elektronischer Mittel oder auf andere Weise zugänglich sind" (§ 87a Abs. 1 UrhG)[7]. Hierzu zählen auch komplexere Webseiten.[8]
- Technische Kopierschutzverfahren dürfen nicht entfernt oder umgangen werden. Befindet sich beispielsweise eine kopiergeschützte CD-ROM im Besitz einer Gedächtnisorganisation und will diese die darauf befindlichen Daten archivieren, dann darf der Kopierschutz nicht ohne weiteres umgangen werden (§ 95a UrhG). Die Gedächtnisorganisation hat allerdings einen Anspruch darauf, dass der Rechteinhaber (z.B. der Her-

6 BGHZ 134, 250 – CB-Infobank I.

7 Die Unterscheidung des Gesetzgebers zwischen „Datenbankwerken" (§ 4 UrhG) einerseits und „Datenbanken" (§ 87a ff. UrhG) andererseits ist in diesem Fall unbeachtlich.

8 Vgl. z.B. LG Köln NJW-COR 1999, 248 L; LG Köln CR 2000, 400 – kidnet.de.

steller der CD-ROM), die zur Umgehung des Schutzes erforderlichen Mittel zur Verfügung stellt, wenn die geplante Archivkopie ansonsten erlaubt ist (§ 95b UrhG). Größere Institutionen können auch mit der herstellenden Industrie pauschale Vereinbarungen treffen.[9]

Harvesting[10]

Vor besondere rechtliche Probleme stellt das Harvesting von Internetangeboten, und zwar unabhängig davon, ob nach bestimmten Selektionskriterien (etwa bestimmten Suchworten) oder unspezifisch (etwa eine ganze Top-Level-Domain) gesammelt wird. Obwohl Harvesting ein gängiges Verfahren im Internet ist (vgl. etwa die Angebote von Google Cache oder archive.org), ist es nach derzeitiger Rechtslage in Deutschland nicht unproblematisch. Das Harvesting ist jedenfalls dann zulässig, wenn die Zustimmung des Urhebers vorliegt (wenn beispielsweise die Betreiber einer museal interessanten Homepage einem Museum gestatten, in regelmäßigen Abständen ein automatisiertes Abbild der Homepage zu machen und dieses zu archivieren). Ohne Zustimmung des Urhebers darf keine Archivkopie angefertigt werden.

In einigen Rechtsgebieten, insbesondere den USA, kann von einer Zustimmung ausgegangen werden, wenn einer Speicherung nicht ausdrücklich widersprochen wurde und auch im Nachhinein kein Widerspruch erfolgt.[11] Nach deutscher Rechtslage reicht dies nicht aus. Die Zustimmung muss eindeutig sein. Ausnahmen, die ein Harvesting durch bestimmte Gedächtnisorganisationen gestatten, sind nur über spezielle Bundesgesetze möglich. Beispielsweise soll nach dessen amtlicher Begründung das Gesetz über die Deutsche Nationalbibliothek dieser den Einsatz von Harvesting-Verfahren ermöglichen.[12]

---

9 Vgl. die Vereinbarung zwischen dem Bundesverband der phonographischen Wirtschaft, dem Deutschen Börsenverein und der Deutschen Nationalbibliothek: http://www.ddb.de/wir/recht/vereinbarung.htm.

10 Dazu näher: Euler, Ellen: Web-Harvesting vs. Urheberrecht : was Bibliotheken und Archive dürfen und was nicht. In: Computer und Recht 2008/1, S. 64-68.

11 „Google Cache", „Archive.org" und vergleichbare Harvester respektieren robots.txt Dateien über die eine Speicherung untersagt wird. Zudem werden auf Antrag des Rechteinhabers Seiten aus dem Archiv gelöscht. Zur Rechtslage in den USA vgl. das Urteil „Blake A. Field v. Google Inc. (No. 2:04-CV-0413, D.Nev)" (Online unter: http://www.linksandlaw.com/decisions-148-google-cache.htm)

12 Vgl. die amtliche Begründung zu § 2 Nummer 1 des DNBG: http://www.ddb.de/wir/pdf/dnbg_begruendung_d.pdf [6.3.2009]. Ungeachtet dieser amtlichen Begründung erlaubt auch das Gesetz über die Deutsche Nationalbibliothek kein flächendeckendes Harvesting (Euler, oben Fn. 10, S. 66 und Steinhauer, oben Fn. 5, S. 164).

Wann und wie dürfen Archivobjekte verändert werden?

Migration und Emulation

Im Sinne einer langfristigen Verfügbarkeit der archivierten Objekte müssen diese gelegentlich migriert oder emuliert werden. Bei jeder Migration und, in eingeschränkterem Maße, auch bei jeder Emulation[13] kommt es zu gewissen qualitativen und/oder quantitativen Änderungen am jeweiligen Objekt. Das Wesen von Migrationen und Emulationen besteht gerade darin, die Interpretation digitaler Daten, die aufgrund ihres veralteten Formats wertlos sind, zu sichern, um sie weiterhin nutzen zu können. Diesem Ziel wird aber nur entsprochen, wenn die neuen Dateien trotz etwaiger Veränderungen denselben Kern von Informationen aufweisen wie die veralteten. Dieser wesentliche Informationskern stellt sicher, dass die neue Datei durch dieselben schöpferischen Elemente geprägt sein wird wie die alte.

Entgegen gewichtigen Stimmen in der juristischen Literatur[14], handelt es sich bei den notwendigen Änderungen im Erscheinungsbild des Objekts in aller Regel noch nicht um eine – zustimmungspflichtige – Bearbeitung / Umgestaltung im Sinne des § 23 UrhG, sondern um eine Vervielfältigung (§ 16 UrhG). Zum einen sind die Änderung eines Dateiformates oder das Öffnen einer Datei in einer emulierten EDV-Umgebung rein mechanische Vorgänge, die nicht von einem individuellen Schaffen desjenigen geprägt sind, der diese Vorgänge technisch umsetzt. Zum anderen kommt es bei (rechtlich unproblematischeren) Vervielfältigungen ebenfalls häufig zu kleineren Abweichungen. Solange die Vervielfältigungsstücke jedoch ohne eigene schöpferische Ausdruckskraft geblieben sind, sie noch im Schutzbereich des Originals liegen und ein übereinstimmender Gesamteindruck besteht,[15] reichen auch gewisse Detailabweichungen vom Original nicht, um von einer Bearbeitung/Umgestaltung auszugehen.

Mit anderen Worten: Soweit eine Institution das Recht hat, Kopien anzufertigen (z.B. aus dem erwähnten § 53 Abs. 2 UrhG), darf sie auch migrieren oder emulieren. Nur in den Ausnahmefällen, in denen die Migration zu einer deut-

---

13 Es kommt dabei nicht darauf an, ob der Bitstream des ursprünglichen Objekts selbst verändert wurde, um die Abbildung auf einem neueren System zu ermöglichen. Entscheidend ist vielmehr das Erscheinungsbild für den Nutzer. In einer ganz anderen Hard- und Softwareumgebung kann im Einzelfall auch ein Objekt, dessen Daten selbst vollkommen unverändert geblieben sind, so anders erscheinen, dass von einer Umgestaltung des ursprünglichen Objekts gesprochen werden kann.

14 Hoeren: Rechtsfragen zur Langzeitarchivierung, S. 7-9; Euler, oben Fn. 10, S. 475f.; Steinhauer, oben Fn. 5, S. 164.

15 BGH GRUR 1988, 533, 535; Schulze-Dreier/Schulze: UrhG, § 16 Rn. 10.

lichen Abweichung vom Original führt, bedarf es einer zusätzlichen Zustimmung des Urhebers.

In bestimmten Fällen wird von der archivierenden Institution aber mehr verlangt als bloße Konformität mit dem Urheberrechtsgesetz. Gerade im juristischen oder auch medizinischen Zusammenhang (z.B. bei der Archivierung von beweiskräftigen Dokumenten oder Patientenakten) können erhöhte Ansprüche an Authentizität und Integrität der Archivobjekte gestellt werden. Auch hier ist zu vieles rechtlich ungeklärt, als dass an dieser Stelle näher darauf eingegangen werden könnte.

Wer darf von wo auf die archivierten Objekte zugreifen?

Der Archivbegriff der Informationswissenschaften unterscheidet sich wesentlich von dem des Urheberrechts. Während in den Informationswissenschaften auch und gerade die Erschließung und Zugänglichmachung der archivierten Materialien im Vordergrund stehen, ist der Archivbegriff in § 53 Abs. 2 UrhG deutlich enger. Hier werden ausschließlich die Sammlung, Aufbewahrung und Bestandssicherung als Archivzwecke angenommen. Ein Archiv, dessen Zweck in der Benutzung durch außenstehende Dritte liegt, ist daher kein Archiv im Sinne des § 53 UrhG. Damit sind die meisten klassischen Gedächtnisorganisationen, die ihre Aufgabe in der Informationsversorgung ihrer Nutzer und weniger im Sammeln und Sichern der Bestände sehen, auf den ersten Blick von der Privilegierung des § 53 ausgenommen. Sie dürften ohne ausdrückliche Zustimmung der jeweiligen Rechteinhaber keine Vervielfältigungen anfertigen. Eine Langzeitarchivierung digitaler Daten ohne – unter praktischen Vorzeichen oft nur schwer zu erlangende – Zustimmung wäre damit *de facto* unmöglich.

Die Berechtigung, Archivkopien anzufertigen, hängt wesentlich davon ab, ob und inwiefern außenstehende Nutzer Zugang zu den Archivmaterialien erlangen sollen. Hier sind grundsätzlich drei Varianten denkbar: rein interne Nutzung, eingeschränkte Nutzung und eine offene Nutzung.

Interne Nutzung

Noch verhältnismäßig unproblematisch ist eine rein interne Nutzung. Wenn Daten aus einem digitalen Archiv ausschließlich von den Mitarbeitern des Archivs im Rahmen des Archivzweckes eingesehen werden, ist dies gestattet. Schwierig wird es jedoch bereits, wenn Mitarbeiter, zum Beispiel per Download oder Computerausdruck, weitere Vervielfältigungen herstellen. Hier muss jeweils erneut geprüft werden, ob diese Vervielfältigungen auch ohne Zustimmung des Urhebers erlaubt sind (z.B. aus Gründen der wissenschaftlichen Forschung – § 53 Abs. 2 S. 1 Nr. 1 UrhG).

Nutzung durch einen begrenzten Nutzerkreis

§ 52b UrhG gestattet es öffentlichen Bibliotheken, Museen und Archiven, ihren Bestand an eigens dafür eingerichteten elektronischen Leseplätzen zugänglich zu machen. Analoge Bestände dürfen zu diesem Zweck digitalisiert werden und bereits vorhandene Archivdigitalisate in den gesteckten Grenzen öffentlich zugänglich gemacht werden.
§ 52b UrhG enthält aber auch wichtige Beschränkungen, die es zu beachten gilt.

- Privilegiert werden nur nichtkommerzielle öffentliche Bibliotheken, Museen und Archive. Nicht-öffentliche Bibliotheken, wie Schul-, Forschungseinrichtungs- oder Institutsbibliotheken oder gewerbliche Archive dürfen sich nicht auf § 52b UrhG berufen.
- Die Anzahl der erlaubten Zugriffe an den eingerichteten Leseplätzen richtet sich grundsätzlich nach der Zahl des in der Gedächtnisorganisation vorhandenen Bestandes.
- Vertragliche Vereinbarungen (etwa Datenbanklizenzen) gehen vor. Wenn die Nutzung durch Dritte vertraglich ausgeschlossen worden ist, kann dies nicht unter Berufung auf § 52b UrhG umgangen werden.

Ähnlich wie bei einer internen Nutzung ist zu entscheiden, ob und wann Nutzer downloaden oder ausdrucken dürfen (s.o.).

Wenn aus einem der genannten Gründe § 52b UrhG nicht greift (etwa, weil es sich bei der archivierenden Institution um eine nicht-öffentliche Forschungsbibliothek handelt), bleibt die Frage, inwieweit die Institution ihren Nutzern Zugang zu den archivierten Materialien gewähren darf. Dies ist in bestimmten Fällen möglich. Beispielsweise ist die Zugänglichmachung von kleinen Teilen von Werken, kleineren Werken und einzelnen Zeitungs- oder Zeitschriftenbeiträgen durch (eng) abgrenzte Personengruppen, wie etwa einzelnen Forscherteams oder den Teilnehmern eines Universitätsseminars, erlaubt, soweit die Nutzung dabei zum Zwecke der wissenschaftlichen Forschung oder zu Unterrichtszwecken (§ 52a UrhG) erfolgt.[16]

Offene externe Nutzung

Es gehört zum Charme der neuen Medien und insbesondere des Internets, dass sie im Prinzip einen weltweiten Zugriff ermöglichen. Der Gesetzgeber hat aber die Entscheidung darüber, ob ein digitales Objekt einer breiten Öffentlichkeit zugänglich gemacht werden soll, alleine dem Urheber übertragen. Ohne Zu-

16 Das gilt auch für den Zugang zu Vervielfältigungsstücken, die zu Archivzwecken angefertigt worden sind (§ 53 Abs. 2 S. 1 Nr. 2 UrhG).

stimmung des Urhebers darf also keine Gedächtnisorganisation urheberrechtlich geschütztes Material ortsungebunden öffentlich zugänglich machen.

Wer haftet für die Inhalte?

Wenn eine Gedächtnisorganisation in großem Umfang digitale Objekte der mehr oder weniger breiten Öffentlichkeit anbietet, besteht die Gefahr, dass einige der Objekte durch ihren Inhalt gegen Rechtsnormen verstoßen. Volksverhetzende oder pornografische Inhalte lassen sich durch entsprechende Filtersoftware und im Idealfall eine intellektuelle Sichtung des Materials noch relativ leicht erkennen. Oft ist es aber nahezu unmöglich, ehrverletzende Behauptungen oder Marken- und Patentverletzungen zu identifizieren. Es ist also eine wichtige Frage, welche Sorgfaltspflichten eine Gedächtnisorganisation zu beachten hat, die ihr digitales Archiv öffentlich zugänglich machen will.

Leider ist auch hier so vieles vom konkreten Einzelfall abhängig, dass es sich nicht mehr wirklich sinnvoll in einer kurzen Zusammenfassung darstellen lässt. Eine ausführlichere Darstellung würde den hier vorgegebenen Rahmen aber sprengen. Nur ganz allgemein lässt sich Folgendes sagen:

Die in diesem Bereich wichtigsten Normen stehen in den §§ 7 - 10 Telemediengesetz (TMG). Danach ist zu unterscheiden, ob es sich bei den veröffentlichten Inhalten um eigene oder fremde handelt. Eine straf- und zivilrechtliche Verantwortung für die Richtigkeit und Rechtmäßigkeit der Inhalte trifft die anbietende Organisation nur im ersten Fall. Ob die Inhalte im Einzelfall der Organisation als eigene zugerechnet werden, richtet sich dabei nicht nach Herkunft oder Eigentum der Objekte, sondern nach der Sicht der Nutzer.[17] Nur wenn ein Nutzer aus den Gesamtumständen eindeutig erkennen konnte, dass es sich bei dem Angebot nicht um ein eigenes Informationsangebot der betreffenden Organisation handelt, ist die Haftung eingeschränkt. Eine Gedächtnisorganisation, die fremde Daten allgemein zugänglich macht, sollte daher darauf achten, dass die „fremden" Angebote im Layout hinreichend deutlich von den eigenen abgegrenzt sind. Außerdem sollte deutlich darauf hingewiesen werden,

17 Das ist im Falle von Gedächtnisorganisationen schwierig, handelt es sich doch um Material aus eigenen Archiven. In einem bestimmten Sinne ist also auch das angebotene Archivmaterial „eigen" und wird insbesondere nicht „für einen Nutzer" (§ 10 TMG) gespeichert. Trotzdem ist es klar ersichtlich und ergibt sich meist auch aus dem (oft gesetzlichen) Auftrag der Gedächtnisorganisation, dass sie sich die angebotenen Inhalte nicht zu Eigen machen will und kann. Eine Haftung als Content-Provider wäre daher unbillig. Vielmehr ist § 10 TMG zugunsten der jeweiligen Gedächtnisorganisation analog anzuwenden, wenn die Abgrenzung der Inhalte, die im engeren Sinne „eigen" sind und denjenigen, die als „fremde" zur Verfügung gestellt werden, hinreichend deutlich ist.

dass sich die Gedächtnisorganisation nicht mit den Inhalten der angebotenen Publikationen oder verlinkten Seiten identifiziert und eine Haftung für diese Inhalte ausgeschlossen ist. Hiermit stellt sie klar, dass sie lediglich dann zur Haftung herangezogen werden kann, wenn sie falsche oder rechtswidrige Inhalte trotz Kenntnis oder Evidenz nicht beseitigt.

Auch wenn deutlich gemacht wurde, dass die zugänglich gemachten Inhalte keine eigenen sind, müssen bestimmte Sorgfaltspflichten beachtet werden. Vor allen Dingen muss bei Bekanntwerden einer Rechtsverletzung der Zugang unverzüglich gesperrt werden (§ 7 Abs. 2 TMG). Eine weitere Speicherung des Objektes bleibt aber – von wenigen Ausnahmen abgesehen – möglich, denn nur die Zugänglichmachung muss unterbunden werden.

Literatur

Dreier, Thomas / Schulze, Gernot: *Urheberrechtsgesetz: Urheberrechtswahrnehmungsgesetz, Kunsturhebergesetz; Kommentar*. 3. Auflage. München: Beck, 2008

Euler, Ellen: *Zur Langzeitarchivierung digital aufgezeichneter Werke und ihrer urheberrechtlichen Einordnung und Beurteilung*. In: AfP 2008/5, S. 474-482

Euler, Ellen: *Web-Harvesting vs. Urheberrecht : was Bibliotheken und Archive dürfen und was nicht*. In: Computer und Recht 2008/1, S. 64-68

Goebel, Jürgen W. / Scheller, Jürgen: *Digitale Langzeitarchivierung und Recht*; nestor-Materialien 01: urn:nbn:de:0008-20040916022

Heckmann, Jörn / Weber, Philipp: *Elektronische Netzpublikationen im Lichte des Gesetzes über die Deutsche Nationalbibliothek*. In: AfP 2008/3, S. 269-276

Hoeren, Thomas: *Rechtsfragen zur Langzeitarchivierung (LZA) und zum Anbieten von digitalen Dokumenten durch Archivbibliotheken unter besonderer Berücksichtigung von Online-Hochschulschriften*: urn:nbn:de:0008-20050305016

Library of Congress. National Digital Information Infrastructure and Preservation Program / Joint Information Systems Committee (UK) / Queensland University of Technology. Open Access to Knowledge (OAK) Law Project / Surf Foundation (Netherlands): International Study on the Impact of Copyright Law on Digital Preservation: http://www.digitalpreservation.gov/library/resources/pubs/docs/digital_preservation_final_report2008.pdf

Rehbinder, Manfred: *Urheberrecht: Ein Studienbuch*. 15. Auflage, München: Beck, 2008

Schack, Haimo: *Dürfen öffentliche Einrichtungen elektronische Archive anlegen?* In: AfP – Zeitschrift für Medien- und Kommunikationsrecht 1/2003, S. 1-8

Spindler, Gerald / Heckmann, Jörn: *Retrodigitalisierung verwaister Printpublikationen – Die Nutzungsmöglichkeiten von „orphan works" de lege lata und ferenda.* In: GRUR Int 2008/4, S. 271-284

Steinhauer, Eric: *Pflichtablieferung von Netzpublikationen. Urheberrechtliche Probleme im Zusammenhang mit der Pflichtablieferung von Netzpublikationen an die Deutsche Nationalbibliothek.* In: Kommunikation & Recht 2009/3, S. 161-166

16.3 Langzeitarchivierung wissenschaftlicher Primärdaten

Gerald Spindler und Tobias Hillegeist

Neben der Langzeitarchivierung von Büchern und Zeitschriften gewinnt die Langzeitarchivierung von wissenschaftlichen Primärdaten (sog. Rohdaten) in jüngster Zeit eine immer bedeutendere Rolle, da immer mehr Hochschulen und Forschungseinrichtungen dazu übergehen, die von ihnen gewonnenen Daten zu archivieren. Dabei sollen die Daten in den meisten Fällen nicht nur archiviert, sondern auch Dritten, wie beispielsweise anderen Forschungseinrichtungen oder einzelnen Fremdforschern zur Verfügung gestellt werden. Aus rechtlicher Sicht ist dabei vor allem entscheidend, ob die Archivierung dieser Daten eine urheberrechtliche Relevanz aufweist, die Daten also urheberrechtlich geschützt sind und, sofern dies zutrifft, wer Inhaber der erforderlichen Nutzungsrechte ist. Des Weiteren stellt sich für Forschungseinrichtungen die Frage, ob es in ihrem Ermessen liegt, die von ihnen gewonnenen Daten zu archivieren oder ob diesbezüglich unter Umständen sogar eine gesetzliche Verpflichtung besteht. Hinsichtlich der Archivierung personenbezogener Daten können sich darüber hinaus datenschutzrechtliche Probleme stellen, was vor allem für Universitätskliniken relevant ist.

Urheberrechtlicher Schutz an einzelnen Daten

Sofern an wissenschaftlichen Primärdaten ein urheber- oder leistungsrechtlicher Schutz besteht, dürften diese nur archiviert werden, sofern die archivierende Einrichtung Inhaber der erforderlichen Nutzungsrechte wäre bzw. der jeweilige Rechteinhaber der Einrichtung die Archivierung gestatten würde. Ein urheberrechtlicher Schutz würde gem. § 2 Abs. 2 UrhG voraussetzen, dass die einzelnen Daten eine persönliche geistige Schöpfung darstellen. Da es, wie oben bereits festgestellt, bei wissenschaftlichen Primärdaten jedoch an der für einen urheberrechtlichen Schutz notwendigen geistigen Schöpfungshöhe fehlt, unterliegen zumindest die einzelnen Daten grundsätzlich nicht dem Schutz des Urheberrechtsgesetzes[18].

18 Siehe dazu bereits oben S. 16:3.

Urheberrechtlicher Schutz gem. § 4 UrhG bzw. §§ 87a ff. UrhG

Etwas anderes könnte jedoch dann gelten, wenn Daten in Tabellen oder auf andere Art zusammengefasst werden. In diesen Fällen könnte nämlich ein Datenbankwerk nach § 4 Abs. 2 UrhG und/oder eine Datenbank gem. § 87a UrhG vorliegen.

Die Entstehung eines urheberrechtlich geschützten Datenbankwerkes im Sinne des § 4 UrhG wird regelmäßig an der dafür erforderlichen geistigen Schöpfungshöhe scheitern, die nach § 4 Abs. 2 UrhG in der individuellen Auswahl oder Anordnung der enthaltenen Daten bestehen muss. Eine solche Individualität wird in den bei Sammlungen von wissenschaftlichen Primärdaten nämlich grundsätzlich nicht vorliegen, da die Anordnung nach logischen Gesichtspunkten erfolgen wird und es damit im Regelfall an einer besonderen Struktur hinsichtlich der Auswahl oder Anordnung der Daten fehlen wird[19].

Im Gegensatz zum urheberrechtlichen Schutz nach § 4 UrhG setzt der leistungsrechtliche Schutz der §§ 87a ff. zwar keine geistige Schöpfungshöhe voraus. Allerdings wird der sui-generis-Schutz der Datensammlung nach § 87a UrhG für Datenbanken, in denen wissenschaftliche Primärdaten enthalten sind, in den meisten Fällen daran scheitern, dass für die Erstellung dieser Datenbanken keine wesentliche Investition im Sinne des § 87a UrhG erforderlich ist. Investitionen werden vielmehr bei der Datenerhebung, also beispielsweise der Durchführung der Forschungsreihe oder eines Experimentes getätigt werden. Die Investitionen für die Datengewinnung sind jedoch im Rahmen des § 87a UrhG nach überwiegender Ansicht in Rechtsprechung und Literatur gerade nicht zu berücksichtigen[20].

Datenbankhersteller im Sinne des § 87a UrhG

Sofern für eine Datenbank mit wissenschaftlichen Primärdaten im Einzelfall doch eine wesentliche Investition erforderlich wäre, wäre gem. § 87a UrhG diejenige Person bzw. die Einrichtung Datenbankhersteller und damit Inhaber der an der Datenbank bestehenden Nutzungsrechte, die diese Investition getätigt

---

19 BGH GRUR 2005, 857, 858 – HIT BILANZ; OLG Nürnberg GRUR 2000, 607; *Dreier* in Dreier/Schulze, § 4 Rn. 12; *Czychowski* in Fromm/Nordemann, § 4 Rn. 12; *Loewenheim* in Loewenheim, § 9 Rn. 229; *ders.* in Schricker, § 4 Rn. 8.

20 EuGH GRUR 2005, 254, 256 Tz. 40 ff. – Fixtures-Fußballspielpläne II; EuGH C-46/02 Tz. 44 ff.; EuGH GRUR 2005, 252, 253 – Fixtures-Fußballspielpläne I; siehe auch Erwägungsgrund 9, 10 und 12 der RL96/9/EG; *Vogel* in Schricker, § 87a Rn. 30; ; a.A. *Czychowski* in Fromm/Nordemann, § 87a Rn. 19.

hat[21]. Dies wird im Regelfall die Hochschule oder das Forschungsinstitut sein, in dessen Einrichtungen die Daten zusammengestellte wurden. Damit würden sich hinsichtlich einer Langzeitarchivierung der Daten keine urheberrechtlichen Probleme ergeben. Zu beachten ist jedoch, dass in Fällen, in denen Forschungsprojekte durch sogenannte Drittmittel finanziert werden, die finanzierende Einrichtung wohl Trägerin der wesentlichen Investition und damit Datenbankherstellerin im Sinne des § 87a UrhG wäre, so dass ihr die zur Langzeitarchivierung erforderlichen Nutzungsrechte zustünden. In diesen Fällen könnten Forschungseinrichtungen in ihren Verträgen mit den Drittmittelgebern im Vorfeld vereinbaren, dass eventuell entstehende Nutzungsrechte an Datenbanken, die im Rahmen des finanzierten Forschungsprojektes erstellt werden, der Forschungseinrichtung zumindest als einfache Nutzungsrechte eingeräumt werden. Auf diese Weise wäre sichergestellt, dass die Forschungseinrichtung die anfallenden Daten auch archivieren und Dritten zugänglich machen zu dürfen.

Inhaber der Nutzungsrechte an einem Datenbankwerk

Sollte eine Datensammlung im Einzelfall doch einem urheberrechtlichen Schutz gem. § 4 UrhG unterliegen, wäre der Urheber Inhaber der Nutzungsrechte. Im Gegensatz zum Datenbankhersteller, der auch eine juristische Person sein kann[22], kann Urheber jedoch nur eine natürliche Person sein[23]. Die Forschungseinrichtung wäre damit also nicht automatisch Inhaberin der Nutzungsrechte an einem Datenbankwerk. Eine gesetzliche Schranke würde zugunsten der Hochschule bzw. Forschungseinrichtung jedoch, wie oben bereits festgestellt, nicht eingreifen[24]. Die Forschungseinrichtung müsste sich demnach die zur Archivierung erforderlichen Nutzungsrechte vom Nutzungsrechtsinhaber vertraglich einholen.

---

21 *Dreier* in Dreier/Schulze, § 87a Rn. 19; *Czychowski* in Fromm/Nordemann, § 87a Rn. 25; *Vogel* in Schricker, § 87a Rn. 45.

22 *Kotthoff* in Dreyer/Kotthoff/Meckel, § 87a Rn. 40; *Czychowski* in Fromm/Nordemann, § 87a Rn. 25, 27; *Dreier* in Dreier/Schulze, § 87a Rn. 20.

23 *Katzenberger/Loewenheim* in Schricker, § 7 Rn. 2; *W. Nordemann* in Fromm/Nordemann, § 7 Rn. 9; *Schulze* in Dreier/Schulze, § 7 Rn.2

24 Siehe bereits oben S. 16:5 f.

Erlangung der Rechte aufgrund eines bestehenden Arbeitsverhältnisses mit dem Rechteinhaber

Unter Umständen erlangt die Hochschule die Rechte jedoch bereits aufgrund eines bestehenden Arbeitsverhältnisses mit dem Rechteinhaber. Dies wäre grundsätzlich der Fall, wenn der Urheber des betreffenden Datenbankwerkes in einem Angestelltenverhältnis zur Universität stünde. Aus diesem folgt nämlich die Pflicht des Arbeitnehmers, dem Arbeitgeber die Nutzungsrechte zu übertragen, die er in Erfüllung seiner aufgrund des Arbeits- oder Dienstverhältnisses geschuldeten Tätigkeit erlangt hat[25]. Dabei erfolgt die Einräumung der Nutzungsrechte regelmäßig im Voraus bei Abschluss des Arbeits- oder Dienstvertrages[26], spätestens jedoch mit Ablieferung des Werkes[27]. Sofern der Urheber eines Werkes bzw. der Datenbankhersteller oder Lichtbildner in einem Angestellten- oder Dienstverhältnis zur Universität stand, wäre er also gegenüber der Universität grundsätzlich zur Übertragung der Nutzungsrechte verpflichtet. Zu beachten ist jedoch, dass aufgrund der durch Art. 5 Abs. 3 GG verfassungsrechtlich garantierten Wissenschaftsfreiheit diese Grundsätze nicht auf Hochschul-, Honorar- oder Gastprofessoren übertragen werden können, da die Veröffentlichung von Forschungsergebnissen nicht mehr zu deren Aufgabenbereich gehört[28]. Handelt es sich bei dem Urheber des Datenbankwerkes oder dem Datenbankherstellers also um einen Professor, so wird die Universität nicht aufgrund des bestehenden Arbeitsverhältnisses Inhaberin der entsprechenden Nutzungsrechte. Aus diesem Grund sollte in den von Hochschulen oder Forschungseinrichtungen geschlossenen Arbeitsverträgen grundsätzlich eine Klausel enthalten sein, wonach die Vertragspartner ihrem künftigen Arbeitgeber die Rechte, die sie im Rahmen ihrer Forschungstätigkeit erlangen, zumindest als einfache Nutzungsrechte einräumen. Hinsichtlich des Inhalts einer solchen Klausel ist zu beachten, dass diese aufgrund der sogenannten

25 BGH GRUR 1991, 523, 525; 1952, 257, 258 – Krankenhauskartei; *Dreier* in Dreier/Schulze, § 43 Rn. 18; *Dreyer* in Dreyer/Kotthoff/Meckel, § 43 Rn. 7, 13; *A. Nordemann* in Fromm/Nordemann, § 43 Rn. 1; *Rojahn* in Schricker, § 43 Rn. 37; *Wandtke*, GRUR 1999, 390, 392.

26 *Dreier* in Dreier/Schulze, § 43 Rn. 19; *Rojahn* in Schricker, § 43 Rn. 46.

27 BGH GRUR 1974, 480, 483 – Hummelrechte; *A. Nordemann* in Fromm/Nordemann, § 43 Rn. 30.

28 BGH GRUR 1991, 523, 525 – Grabungsmaterialien; 1985, 529, 530 – Happening; OLG Karlsruhe GRUR 1988, 536, 537 – Hochschulprofessor; *Dreier* in Dreier/Schulze, § 43 Rn. 12; *Rojahn* in Schricker, § 43 Rn. 31, 65; *A. Nordemann* in Fromm/Nordemann, § 43 Rn. 43.

Zweckübertragungslehre nicht pauschal abgefasst sein darf, sondern vielmehr die genauen Nutzungsrechte und -arten bezeichnen muss.

Pflicht zur Archivierung

Des Weiteren ist im Rahmen der Langzeitarchivierung von wissenschaftlichen Primärdaten relevant, ob für Forschungseinrichtungen eine gesetzliche Pflicht besteht, die erhobenen Daten zu archivieren. Grundsätzlich besteht dabei keine Verpflichtung zur Archivierung der erhobenen Daten. Ausnahmen ergeben sich jedoch im Bereich der Buchführung, bei Personalsachen, bei Bankunterlagen, Akten der Verwaltung, Gerichtsakten und für medizinische Dokumentationen. Im Rahmen der Langzeitarchivierung von wissenschaftlichen Primärdaten sind dabei vor allem Aufbewahrungs- und Archivierungspflichten von medizinischen Dokumentationen relevant. Diese ergeben sich hauptsächlich aus den §§ 28 Röntgenverordnung (RöntgV), 42 Strahlenschutzverordnung (StrlSchV), sowie 1 Gentechnikaufzeichnungsverordnung (GenTAufzV). Darüber hinaus können sich standesrechtliche Dokumentationspflichten aus den Landesberufsordnungen der Ärzte ergeben. Die genannten Vorschriften schreiben dabei zwar nicht ausdrücklich eine elektronische Archivierung vor, sondern lediglich, dass die Daten generell dokumentiert werden müssen. Dabei wird eine Dokumentation aufgrund des technischen Fortschrittes aber wohl in der überwiegenden Zahl der Fälle elektronisch erfolgen.

Verantwortliche Personen für die ordnungsgemäße Archivierung

Dies wirft die Frage auf, wer für die Durchführung der Archivierung verantwortlich ist, sofern eine Archivierungspflicht besteht.
Verantwortlich für die Dokumentation ist dabei grundsätzlich der gesetzliche Vertreter der Forschungseinrichtung, die in den Anwendungsbereich der oben genannten Normen fällt. Sofern der Anwendungsbereich der RöntgV eröffnet ist, sind daneben gem. § 15 Abs. 2 RöntgV ebenfalls die Strahlenschutzbeauftragten der Einrichtung verantwortlich. Für die Archivierung von Behandlungs- und Untersuchungsdaten ist neben dem gesetzlichen Vertreter der Klinik, an der diese erhoben worden sind, ebenfalls der jeweilige behandelnde Arzt aufgrund des Behandlungs- bzw. Krankenhausvertrages für die ordnungsgemäße Archivierung der Behandlungs- und Untersuchungsverantwortlich.

Verhinderung der Weitergabe archivierter Daten durch Dritte

Sofern die archivierende Einrichtung ihre Daten Dritten, wie zum Beispiel Fremdforschern oder anderen Forschungseinrichtungen zur Verfügung stellt, hat sie unter Umständen ein Interesse daran, dass der Empfänger der Daten diese nicht unbefugt an Dritte weitergibt. Dies gilt vor allem für medizinische Forschungsdaten, da diese in der Regel personenbezogen sind und ihre Verwendung damit den Vorschriften des BDSG bzw. der Landesdatenschutzgesetze unterliegt. Da die Daten aber grundsätzliche keinem urheberrechtlichen oder leistungsrechtlichen Schutz unterliegen werden und ein Unterlassungsanspruch nach § 97 Abs. 1 UrhG damit ausscheidet, muss die unbefugte Weitergabe auf andere Weise verhindert werden. Aus diesem Grund sollte mit Dritten, denen ein Zugriff auf die archivierten Daten gewährt wird, ein Lizenzvertrag geschlossen werden, der die zulässige Nutzung der Daten durch den Fremdforscher regelt und eine Verschwiegenheitsklausel beinhaltet, die die Fremdforscher verpflichtet, die Daten nicht unbefugt weiterzugeben. Für den Fall, dass gegen diese Vereinbarung verstoßen wird, sollte in der Vereinbarung außerdem eine Vertragsstrafe vorgesehen werden.

Sicherstellung der Authentizität und Integrität der archivierten Daten

Eine rechtsgültige Authentizität und Integrität der archivierten Forschungsdaten kann durch Verwendung einer qualifizierten elektronischen Signatur erreicht werden. Dabei besteht grundsätzlich keine Pflicht, die Authentizität und Integrität der Daten sicherzustellen. Eine Ausnahme gilt jedoch für medizinische Forschungsdaten. Bei diesen ist aufgrund der Anforderungen des Bundesdatenschutzgesetzes sowie der einzelnen Landesdatenschutzgesetze, die im Falle der Archivierung medizinischer Forschungsdaten einschlägig sein können, eine Verpflichtung zur Gewährleistung der Integrität und Authentizität anzunehmen[29]. Aber auch in den übrigen Fällen ist die Verwendung einer qualifizierten elektronischen Signatur anzuraten. Da die Daten nicht nur archiviert, sondern unter Umständen auch fremden Forschungsstellen zur Verfügung gestellt werden sollen, liefe die archivierende Forschungseinrichtung andernfalls Gefahr, das Vertrauen anderer Forschungsstellen in die Authentizität seiner Daten zu verlieren. Darüber hinaus könnten Schadensersatzansprüche anderer Forschungsstellen entstehen, sofern diesen aufgrund von manipulierten Daten

29 Vgl. Anlage zu § 9 BDSG.

ein Schaden entstünde und die archivierende Forschungseinrichtung keine entsprechenden Vorkehrungen gegen eine derartige Manipulation getroffen hat.

Zulässigkeit der Archivierung personenbezogener Daten

Neben den Vorschriften des Urheberrechtes könnte sich ein Verbot der Langzeitarchivierung von Daten ebenfalls aus dem BDSG bzw. den Landesdatenschutzgesetzen ergeben, sofern es sich um personenbezogene Daten handelt. In diesen Fällen müsste sich die archivierende Einrichtung das Einverständnis der Personen einholen, auf die sich die Daten beziehen[30]. Eine rechtswirksame Einwilligung eines Probanden bedarf dabei sowohl nach den Landesdatenschutzgesetzen als auch dem Bundesdatenschutzgesetz der Schriftform[31]. Aus diesem Grund empfiehlt es sich, vom Probanden gleich mehrere Einwilligungserklärungen unterschreiben zu lassen, damit für den Fall der Beschädigung oder Zerstörung eines Exemplars noch mindestens eine weitere formgerechte Erklärung als Ersatz vorhanden ist. Die Anfertigung von Kopien genügt hingegen nicht, da eine Kopie oder auch ein elektronischer Scan nicht der Schriftform des BGB, sondern lediglich der Textform entsprechen[32]. Auch wenn das Gesetz Ausnahmefälle vorsieht, in denen die Einwilligung auch formlos möglich ist, sollte aus Gründen der Rechtssicherheit stets eine schriftliche Einwilligung eingeholt werden, da die Beurteilung bzw. der Beweis vor Gericht, dass die Schriftform im Einzelfall entbehrlich war, mitunter schwierig sein kann. Darüber hinaus ist der Proband gezielt darauf aufmerksam zu machen, dass er in die Verwertung seiner Daten einwilligt. Dies kann dadurch erreicht werden, dass die Einwilligung visuell hervorgehoben oder im Dokument explizit auf diese hingewiesen wird. Der Proband muss ferner vor Abgabe ausdrücklich darüber informiert werden, welche seiner Daten auf welche Art verarbeitet oder genutzt werden sollen. Insbesondere ist er darauf hinzuweisen, dass seine Daten eventuell anderen Fremdforschern zugänglich gemacht werden. Hinsichtlich des Inhalts der Erklärung muss diese ebenfalls genau spezifizieren, hinsichtlich welcher Daten der Proband seine Einwilligung erteilt und auf welche Arten die Daten genutzt werden dürfen. Schließlich muss die Einwilligung des Probanden auf dessen freier Entscheidung beruhen. Sofern eine wirksame datenschutzrechtliche Einwilligung des Probanden vorliegt, ist darin außerdem gleichzeitig

30 Vgl. z.B. § 4 Abs. 1 Nr. 2 NDSG sowie § 4 BDSG.

31 Siehe nur § 4a BDSG und § 4 Niedersächsisches Datenschutzgesetz (NDSG).

32 *Ellenberger* in Palandt, Bürgerliches Gesetzbuch, 68. Aufl. 2009, § 126 Rn. 8; *Wendtland* in Bamberger/Roth (Hrsg.), Kommentar zum Bürgerlichen Gesetzbuch Bd. 1, 2. Aufl. 2007, § 126 Rn. 6, 8, 11; *Simitis* in Simitis, § 4a Rn. 38.

eine (zumindest konkludent erteilte) Entbindung des Arztes von seiner ärztlichen Schweigepflicht zu sehen. Die Entbindung von der Schweigepflicht entspricht dabei konsequenterweise in ihrer Reichweite dem Umfang, in welchem der Proband auch der datenschutzrechtlich relevanten Nutzung seiner Daten eingewilligt hat.

Rechtliche Anforderungen an die Archivierung personenbezogener Daten

Sofern die Forschungseinrichtung das Einverständnis des Betroffenen zur Archivierung seiner Daten eingeholt hat, treffen Sie im Rahmen der Nutzung und Verarbeitung dieser Daten gewisse Pflichten hinsichtlich der zu treffenden organisatorischen und technischen Maßnahmen[33]. So hat sie unter anderem dafür Sorge zu tragen, dass Unbefugte keinen Zutritt zu den Datenverarbeitungsanlagen erhalten. Die archivierende Forschungseinrichtung hat also festzulegen, welche Personen in welchem Umfang Zugang zu ihren Verarbeitungsanlagen und deren IT-Systemen haben dürfen und muss die Bedingungen und die Form der Identifikation und Authentisierung der Zugriffsberechtigten festzulegen[34]. Hinsichtlich des Zugriffs und der Bearbeitung der Daten ist darüber hinaus genau festzulegen, wie die Authentisierung und Identifikation von Mitarbeitern und Zugriffsgeräten zu erfolgen hat und welche Aktionen bei einer nicht erfolgreichen Authentisierung zu erfolgen haben. Dies kann unter anderem erreicht werden, indem für den Zugriff auf den Datenkatalog und die Eingabe neuer bzw. die Veränderung bereits gespeicherter Daten spezielle Zugriffsrechte entsprechend den Aufgabenfeldern der einzelnen Mitarbeiter zugeteilt werden[35]. Damit könnte nur ein begrenzter und möglichst kleiner Kreis von Mitarbeitern Eingaben vornehmen und die gespeicherten personenbezogenen Daten ändern. Dabei sind die Zugriffsrechte nur insoweit zu erteilen, als die Inhaber der Zugriffsrechte diese auch tatsächlich ihrem Tätigkeitsfeld entsprechend benötigen. So könnten separate Nutzungsrechte für den Zugang zu den Daten, der Eingabe von neuen Daten, der Übertragung der Daten an einen anderen Speicherort, der Veränderung sowie der Löschung der Daten erteilt werden. Die Vergabe, Änderung oder Entziehung dieser Nutzungsrechte darf dabei nur durch autorisierte Personen erfolgen und ist genau zu dokumentieren, damit

33 Siehe z.B. § 9 BDSG und § 7 NDSG

34 *Bundesamt für Sicherheit in der Informationstechnik BSI* (Hrsg.), Handbuch für die sichere Anwendung der Informationstechnik, 1992, 11.2.4; abrufbar unter: http://www.bsi.bund.de/literat/kriterie.htm; siehe auch Kommentar zum NDSG, § 7 Zu Abs. 2 Nr. 5.

35 *BSI*, 1992, 11.2.4; 11.2.5; 11.2.6.

stets Klarheit darüber herrscht, wie groß der Personenkreis ist, der Zugriff auf die Daten hat und welche Personen er umfasst. Die Datenbestände sind infolge dessen so aufzubereiten, dass bei einer Eingabe in den Datenbestand zunächst geprüft wird, ob die jeweilige Person auf die Daten zugreifen darf, bzw. ob und inwieweit sie Änderungen an den Datensätzen vornehmen darf. Dies kann beispielsweise durch die Einrichtung von Zugriffssicherungen in Form von Passwörtern und durch chipkartenbasierte oder biometrische Identifikationsverfahren geschehen. Des Weiteren empfiehlt sich in diesem Zusammenhang die Installation eines physikalischen Schreibschutzes, damit die Daten nicht nachträglich manipuliert werden können[36]. Daneben sollten Regeln für die Aufbewahrung von Datenträgern, wie etwa CD-ROM oder Festplatten, aufgestellt werden, auf denen sich personenbezogene Daten befinden. Neben diesen Maßnahmen, die eine unbefugte Veränderung bzw. einen unbefugten Zugriff verhindern sollen, sollten die archivierten Daten unter Umständen mit einer qualifizierten elektronischen Signatur versehen werden, um so etwaige Manipulationen von Datenbeständen möglichst schnell aufzufinden, die trotz aller getroffenen Sicherheitsvorkehrungen unter Umständen nicht verhindert werden können. Ferner sollten Ereignisse im Zusammenhang mit den personenbezogenen Daten protokolliert werden, um so feststellen zu können, zu welchem Zeitpunkt welche Daten von welchem Zugriffsgeräte aufgerufen bzw. verändert worden sind. Eine Protokollierung sollte ferner hinsichtlich der erteilten Zugriffsrechte erfolgen. Zu beachten ist dabei, dass die vorgenommenen Protokollierungen vollständig und klar aufgebaut sind, um im Ernstfall tatsächlich nachvollziehen zu können, zu welchem Zeitpunkt welche Veränderung von welchem Arbeitsplatz vorgenommen wurde[37].

36 *BSI,* 11.2.5.

37 So auch Kommentar zum NDSG, § 7 Zu Abs. 2 (Nr.6) Nr. 7; *BSI*, 11.2.4.

Literatur

Bamberger, H. G./Roth, H. (Hrsg.), *Kommentar zum Bürgerlichen Gesetzbuch,* Bd. 1, 2. Aufl. 2007.

Bundesamt für Sicherheit in der Informationstechnik BSI (Hrsg.), *Handbuch für die sichere Anwendung der Informationstechnik,* 1992, abrufbar unter: http://www.bsi.bund.de/literat/kriterie.htm.

Der Landesbeauftragte für den Datenschutz Niedersachsen, *Kommentar zum NDSG*; abrufbar unter: http://www.lfd.niedersachsen.de/master/C52351301_N38588372_L20_D0_I560.html.

Dreier, T./Schulze, G., *Urheberrechtsgesetz: Kommentar*, 3. Aufl. 2009.

Dreyer, G./Kotthoff, J./Meckel, A. (Hrsg.), *Heidelberger Kommentar zum Urheberrecht*, 2. Aufl. 2009.

Fromm, F.K./Nordemann, W. (Hrsg.), *Urheberrecht: Kommentar zum Urheberrechtsgesetz,* 10. Aufl. 2009.

Loewenheim, U. (Hrsg.), *Handbuch des Urheberrechts,* 2003.

Schricker, G. (Hrsg.), *Urheberrecht: Kommentar*, 3. Aufl. 2006.

Simitis , S. (Hrsg.), *Bundesdatenschutzgesetz*, 6. Aufl. 2006.

Wandtke, A., *„Reform des Arbeitnehmerurheberrechts?“*, GRUR 1999, 390.

17 Vorgehensweise für ausgewählte Objekttypen

17.1 Einführung

Regine Scheffel

Die vorangegangenen Kapitel über Strategien, Modelle, Standards u.a. vermitteln den (derzeitigen) Kenntnisstand, der notwendig ist, um kompetent Probleme der Langzeitarchivierung und Langzeitverfügbarkeit anzupacken. Vielfach treten jedoch Anforderungen zutage, die Praktikerinnen und Praktiker in (Kulturerbe-)Institutionen nicht kurzfristig selbst klären, ändern oder erfüllen können (z.B. policies, Organisationsmodelle oder Hardwareumgebung). Dennoch stehen sie unter Handlungsdruck, um die digitalen Objekte in ihrem Verantwortungsbereich nutzbar zu erhalten. Hier setzt das folgende Kapitel an, das konkrete Anwendungsfelder der genannten Aspekte (z.B. Formate) in der Praxis vorstellt.

Diese Anwendungsfelder beziehen sich nicht auf Handlungsfelder in Bibliotheken, Museen, Archiven oder Forschungseinrichtungen (z.B. Publikationen), sondern auf den Umgang mit den unterschiedlichen Medienarten wie Text, Bild und Multimedia in seinen diversen Ausprägungen. Darüber hinaus wer-

den Langzeitarchiverung und Langzeitverfügbarkeit komplexer digitaler Materialsammlungen thematisiert, die über den Medienmix hinaus weitere spezifische Anforderungen stellen, z.B. Websites, wissenschaftliche Rohdaten oder Computerspiele.

17.2 Textdokumente

Karsten Huth

Der langfristige Erhalt von Textdokumenten ist nur scheinbar einfacher als der Erhalt von anderen digitalen Objekten. In digitalen Textdokumenten vereint sich Fachwissen aus der Kunst des Buchdrucks, der Linguistik und der Semiotik. Nur der Aspekt, dass wir vom frühen Kindesalter ans Lesen, Schreiben und Textverstehen herangeführt werden, sodass Texte wie selbstverständlich unseren Alltag prägen, lassen uns die komplexen Kenntnisse und kognitiven Fähigkeiten, die der Mensch dafür benötigt, kaum mehr wahrnehmen. Das optische Erscheinungsbild eines digitalen Textes besteht auf der Datenebene aus zwei wesentlichen Komponenten. Die Wichtigste ist der Zeichensatz, über den numerische Werte Textzeichen zugewiesen werden. Die zweite Komponente ist der Font, d.h. ein kompletter Satz von Bildern der benötigten Schriftzeichen. Dieses Kapitel klärt in einführender Weise über die Abhängigkeiten dieser Komponenten auf.

Definition

Die Definition des Begriffs Textdokument im Bereich der Langzeitarchivierung bzw. die Antwort auf die Frage: "Was ist ein Textdokument?", ist nicht einfach zu beantworten. Kommen doch zwei Ebenen eines digitalen Objekts für eine Definitionsgrundlage in Frage[1]. Auf der konzeptuellen Ebene liegt ein Textdokument genau dann vor, wenn das menschliche Auge Text erkennen, lesen und interpretieren kann. Diese Anforderung kann auch eine Fotografie, bzw. das Bild eines Textes erfüllen. Auf der logischen Ebene eines digitalen Objektes, der Ebene der binären Codierung und Decodierung liegt ein Textdokument genau dann vor, wenn innerhalb des Codes auch Textzeichen codiert sind und dadurch Gegenstand von Operationen werden (z.B. Kopieren und Verschieben, Suchen nach bestimmten Worten und Wortfolgen, Ersetzen von bestimmten Zeichenfolgen usw.).

Da ein Archiv seine Archivobjekte generell auf der konzeptuellen Ebene betrachten muss, insbesondere da sich die technikabhängige logische Ebene im Laufe der Zeit durch Migration grundsätzlich ändert,[2] soll für dieses Kapitel die erste Definition zur Anwendung kommen: *Ein Textdokument liegt genau dann vor, wenn das menschliche Auge Text erkennen, lesen und interpretieren kann.*

---

1 Vgl. Funk, Stefan, Kap 7.2 Digitale Objekte und Formate

2 Vgl. Funk, Stefan, Kap 8.3 Migration

Diese Definition ermöglicht die Verwendung von Dateiformaten zur Speicherung von Bildinformationen ebenso wie die speziell auf Textverarbeitung ausgerichteten Formate. Welchen Formattyp ein Archiv zur Speicherung wählt, hängt von den wesentlichen Eigenschaften des Archivobjekts ab. Die wesentlichen Eigenschaften eines digitalen Archivobjekts müssen vom Archiv bei oder bereits vor der Übernahme des Objekts in das Archiv festgelegt werden und ergeben sich gemäß den Vorgaben des Open Archive Information System (OAIS) größtenteils aus den Ansprüchen und Möglichkeiten der Archivnutzer.[3]

Archivierung von Textdokumenten mit Bildformaten:

Die Archivierung von Textdokumenten in Bildformaten empfiehlt sich genau dann, wenn der optische Eindruck eines Textdokuments eine der wesentlichen Eigenschaften des Archivobjekts ist, welches auf das Genaueste erhalten werden muss. Solche Fälle ergeben sich z.B. bei der Digitalisierung von amtlichem Schriftgut, bei der anschließend das Original aus Papier vernichtet wird, während man die digitale Fassung archiviert. Da bei diesen Objekten das originale Schriftbild sowie von Hand aufgetragene Zeichen (z.B. Anmerkungen, Unterschriften und Paraphen) für die dauerhafte korrekte Interpretation des Archivobjektes unbedingt erhalten werden müssen, ist die Speicherung als Bild der beste Weg. In einem Bildformat sind in der Regel nur Informationen über Bildpunkte und ihre jeweiligen Farb- und Helligkeitswerte in einem Raster verzeichnet (Bitmap-Grafik). Diese Formate beinhalten von sich aus keinerlei Informationen über den abgebildeten Text. Deshalb kann man in einer solchen Datei nicht nach bestimmten Textstellen suchen, Textinhalte herauskopieren oder verschieben. Die Unveränderlichkeit der inhaltlichen und optischen Darstellung ist für ein Archiv von Vorteil.

Eine Abhandlung zu möglichen Bildformaten im Archiv befindet sich im Kapitel 15.2 „Bilddokumente“.[4] Bildformate werden in diesem Kapitel nicht weiter thematisiert.

---

3 Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information System: Blue Book*. Washington, DC. Page 3-4

4 Für eine kurze Übersicht über Bildformate s. Rohde-Enslin, Stefan (2004): *nestor - Ratgeber - Nicht von Dauer: Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen*. Berlin: nestor, IfM . S. 12ff : urn:nbn:de:0008-20041103017

Archivierung von Textdokumenten mit Textformaten:

Die Archivierung von Textdokumenten in Textformaten empfiehlt sich genau dann, wenn die Erhaltung der Textinformation des Objektes im Vordergrund steht. Bei der Archivierung von Textformaten sind grundsätzliche technische Abhängigkeiten zu beachten.

- Abhängigkeit 1: der Zeichensatz (Character Set)
  Einen Zeichensatz kann man sich als Tabelle vorstellen, in der ein numerischer einem Zeicheninhalt zugeordnet wird. Die Maschine nimmt den Wert der Zahl und sieht in der Zeichensatztabelle an der entsprechenden Stelle nach, in welches Zeichen die Zahl decodiert werden muss. Dieser Vorgang hat noch nichts mit der Darstellung eines Zeichens auf dem Bildschirm oder beim Druckvorgang zu tun.[5]
  Beispiel: Beim American Standard Code for Information Interchange (ASCII) Zeichencode entspricht der Wert 65 (binär 01000001) dem Zeichen „A".
- Abhängigkeit 2: Schriften (Font)
  Fonts geben den Zeichen eine Gestalt auf dem Bildschirm oder beim Druck. Dem Zeichen eines Zeichensatzes ist innerhalb eines Fonts ein Bild (oder mehrere Bilder) zugeordnet. Bekannte Schrifttypen sind z.B. Arial, Times New Roman usw.

Die korrekte Darstellung eines Textes ergibt sich demnach aus einer Kette von Abhängigkeiten. Um ein Textdokument mitsamt dem Schriftbild (d.h. Formatierungen, Absätze und Font) erhalten zu können, benötigt ein Archiv den korrekten Zeichensatz und den korrekten Font. Dies ist ein Problem für den dauerhaften Erhalt, denn die meisten Dateiformate, die im Bereich der Textverarbeitung verwendet werden, sind von Zeichensätzen und Fonts abhängig, die außerhalb der Textdatei gespeichert werden. Insbesondere die Zeichensätze sind oft ein Teil des Betriebssystems. Das Textverarbeitungsprogramm leistet die Verknüpfung von Code – Zeichen – Schriftzeichen und sorgt für die korrekte Darstellung des Textdokuments.

---

5 Für eine gelungene Einführung in das Gebiet der Zeichensätze s. Constable, Peter (2001): *Character set encoding basics. Understanding character set encodings and legacy encodings.* In: Implementing Writing Systems: An introduction. 13.06.2001. http://scripts.sil.org/IWS-Chapter03

Konsequenzen für das Archiv

Für die langfristige Darstellbarkeit eines Textes muss das Archiv zumindest den oder die verwendeten Zeichensätze kennen. Die Informationen über die Zeichensätze sollten mit Bezug auf die jeweiligen Dateien in den Metadaten des Archivs fest verzeichnet sein.

Bei Neuzugängen sollte das Archiv Dateiformate wählen, die weit verbreitete und standardisierte Character Sets unterstützen. Der älteste (seit 1963) Zeichensatz ASCII kann beinahe auf allen Plattformen decodiert und dargestellt werden. Leider wurde dieser Zeichensatz allein für den amerikanischen Raum entwickelt, so dass er keinerlei Umlaute und kein „ß" enthält. Damit ist ASCII für deutsche Textdokumente nicht ausreichend. Für Archive besonders zu empfehlen sind Dateiformate, die Unicode,[6] speziell Unicode Transformation Format (UTF)-8 (Unicode encoding Form neben UTF-16 und UTF-32) unterstützen. UTF-8 ist auch der empfohlene Zeichensatz für Dokumente im HTML-, XML- oder SGML-Format. Weit verbreitet und für Archive geeignet ist der Zeichensatz „Latin-1, Westeuropäisch" ISO 8859-1, der auch ASCII-Texte darstellen kann.

Die gewissenhafte Dokumentation der verwendeten Zeichensätze sollte ein Archiv zumindest vor dem Verlust der reinen Textinformation bewahren. Um auch die ursprüngliche optische Form zu erhalten, sollten die technischen Informationen über die verwendeten Schriftsätze (Fonts) ebenso vom Archiv in den Metadaten nachgewiesen werden.

Bei bereits bestehenden Beständen an Textdokumenten, sollte mit geeigneten technischen Werkzeugen der zugrundeliegende Zeichensatz ermittelt werden. Sollte der ermittelte Zeichensatz nicht den oben erwähnten weit verbreiteten Standards entsprechen, empfiehlt sich auf lange Sicht wahrscheinlich eine Migration, vorausgesetzt die geeigneten technischen Werkzeuge sind im Archiv vorhanden.

Besonders geeignete Dateiformate für Archive

Da das Archiv alle Informationen über die verwendeten Zeichensätze und Fonts sammeln und erschließen muss, sollten nur Dateiformate verwendet werden, aus denen diese Informationen auch gewonnen werden können. Dies ist

6 Whistler, Ken/ Davis, Mark/ Freytag, Asmus (2004): *Unicode Technical Report #17. Character Encoding Model. Revision 5*. In: Unicode Technical Reports. 09.09.2004. http://www.unicode.org/reports/tr17/
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

bei Dateiformaten der Fall, wenn ihr technischer Aufbau öffentlich (entweder durch Normung oder Open Source) beschrieben ist. Ein Archiv sollte Textformate meiden, deren technischer Aufbau nicht veröffentlicht wurde (proprietäre Formate), da dann der Zugriff auf die für die Langzeitarchivierung wichtigen technischen Informationen kompliziert ist.

Ein Beispiel für ein offenes Dokumentformat ist das „Open Document Format" (ODF). Der gesamte Aufbau einer ODF-Datei ist öffentlich dokumentiert. Eine Datei besteht im wesentlichen aus mehreren komprimierten XML-Dateien, die alle mit dem Zeichensatz UTF-8 gespeichert wurden. Die von ODF-Dateien verwendeten Schriftsätze sind kompatibel zu UTF-8 und in den XML-Dateien angegeben. Sollte eine ODF-Textdatei im Archiv mit den vorhandenen technischen Mitteln nicht mehr darstellbar sein, dann kann mindestens der Textinhalt und die Struktur des Dokuments über die zugrundeliegenden XML-Dateien zurückgewonnen werden.

Ein Textformat, das speziell für die Archivierung entwickelt wurde, ist das PDF/A-Format. Das Dateiformat wurde so konzipiert, dass Zeichensatz und die verwendeten Fonds in der jeweiligen Datei gespeichert werden. Ein Textdokument im PDF/A Format ist somit unabhängiger von der jeweiligen Plattform, auf der es dargestellt werden soll.

Literatur

Consultative Committee for Space Data Systems (Hrsg.) (2002): *Reference Model for an Open Archive Information System*. Blue Book. Washington, DC. Page 3-4

Constable, Peter (2001): *Character set encoding basics. Understanding character set encodings and legacy encodings.* In: Implementing Writing Systems: An introduction. 13.06.2001. http://scripts.sil.org/IWS-Chapter03

Rohde-Enslin, Stefan (2004): *nestor - Ratgeber - Nicht von Dauer: Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen.* Berlin: nestor, IfM . S. 12ff : urn:nbn:de:0008-20041103017

Whistler, Ken/ Davis, Mark/ Freytag, Asmus (2004): *Unicode Technical Report #17. Character Encoding Model. Revision 5*. In: Unicode Technical Reports. 09.09.2004. http://www.unicode.org/reports/tr17/

17.3 Bilddokumente

Markus Enders

Digitale Bilddokumente (auch Images genannt) sind seid einigen Jahrzehnten in Gebrauch. Digitale Fotos, gescannte Dokumente oder anderweitig digital erzeugte Bilddokumente sind und werden millionenfach erstellt. Gedächtnisorganisationen müssen mit diesen Daten umgehen und sie langfristig archivieren können. Diese Aufgabe bietet verschiedene Herausforderungen. Unterschiedliche Datenformate, deren Benutzung und Unterstützung durch Softwareapplikationen bestimmten Moden unterliegen, sind nur ein Problem. Es stellt sich ferner die Frage, welche Metadata für Bilddokumente generiert werden können, wo diese gespeichert werden sollten und mit welchen Hilfsmitteln diese erzeugt bzw. extrahiert werden.

Seitdem Anfang der 1990er Jahre Flachbettscanner nach und nach in die Büros und seit Ende der 1990er Jahre auch zunehmend in die Privathaushalte einzogen, hat sich die Anzahl digitaler Bilder vervielfacht. Diese Entwicklung setzte sich mit dem Aufkommen digitaler Fotoapparate fort und führte spätestens seit der Integration kleiner Kameramodule in Mobiltelefone und Organizer sowie entsprechender Consumer-Digitalkameras zu einem Massenmarkt.
Heute ist es für Privatleute in fast allen Situationen möglich, digitale Images zu erzeugen und diese zu verschiedenen Zwecken weiterzubearbeiten. Der Markt bietet unterschiedliche Geräte an: von kleinen Kompaktkameras bis zu hochwertigen Scanbacks werden unterschiedliche Qualitätsbedürfnisse befriedigt.

Entsprechend haben sich auch Softwarehersteller auf diesen Markt eingestellt. Um Bilddokumente nicht im Dateisystem eines Rechners verwalten zu müssen, existieren heute unterschiedliche Bildverwaltungsprogramme für Einsteiger bis hin zum Profifotografen.

Diese Entwicklung kommt auch den Gedächtnisorganisationen zugute. Vergleichsweise günstige Preise ermöglichen es ihnen, ihre alten, analogen Materialien mittels spezieller Gerätschaften wie bspw. Scanbacks, Buch- oder Microfilmscannern zu digitalisieren und als digitales Image zu speichern. Auch wenn Texterfassungsverfahren über die Jahre besser geworden sind, so gilt die Authentizität eines Images immer noch als höher, da Erkennungs- und Erfassungsfehler weitestgehend ausgeschlossen werden können. Das Image gilt somit als „Digitales Master", von dem aus Derivate für Online-Präsentation oder Druck erstellt werden können oder deren Inhalt bspw. durch Texterkennung / Abschreiben für Suchmaschinen aufbereitet werden kann.

Datenformate

Digitale Daten müssen immer in einer für den Computer lesbaren und interpretierbaren Struktur abgelegt werden. Diese Struktur wird Datenformat genannt. Eine Struktur muss jedoch jeden Bildpunkt nicht einzeln speichern. Vielmehr können Bildpunkte so zusammengefasst werden, dass eine mehr oder weniger große Gruppe von Punkten als ein einzige Einheit gespeichert werden. Anstatt also jeden Bildpunkt einzeln zu speichern, belegen mehrere Bildpunkte denselben Speicherplatz. Die Art und Weise, wie diese Punkte zusammengefasst werden, wird als Komprimierungsalgorithmus bezeichnet. Dieser kann fest mit einem bestimmten Datenformat verbunden sein.

Sowohl das Datenformat als auch der Komprimierungsalgorithmus können bestimmte technische Beschränkungen haben. So kann durch das Format bspw. die Farbtiefe oder maximale Größe eines Bildes eingeschränkt sein. Der Komprimierungsalgorithmus kann bspw. nur auf reine schwarz-weiss Bilder angewendet werden.

In den letzten zwei Jahrzehnten wurde eine Vielzahl von Datenformaten für Bilddaten entwickelt. Zu Beginn der Entwicklung wirkten technische Faktoren stark limitierend. Formate wurden im Hinblick auf schnelle Implementierbarkeit, wenig Resourcenverbrauch und hohe Performanz während des Betriebs entwickelt. Dies führte zu vergleichsweise einfachen Lösungen, die auf einen bestimmten Anwendungszweck zugeschnitten waren. Teilweise wurden sie so proprietär, dass entsprechende Dateien nur von der Herstellersoftware, die zu einem Scanner mitgeliefert wurde, gelesen und geschrieben werden konnten. Der Austausch von Daten stand zu Beginn der Digitalisierung nicht im Vordergrund, so dass nur ein Teil der Daten zu Austauschzwecken in allgemein anerkannte und unterstützte Formate konvertiert wurden.

Heute ermöglicht das Internet einen Informationsaustausch, der ohne standardisierte Formate gar nicht denkbar wäre. Der Begriff „Standard“ ist aus Sicht der Gedächtnisorganisationen jedoch kritisch zu beurteilen, da „Standards“ häufig lediglich so genannte “De-facto“-Standards sind, die nicht von offiziellen Standartisierungsgremien erarbeitet und anerkannt wurden. Ferner können derartige Standards bzw. deren Unterstützung durch Hard- und Softwarehersteller lediglich eine kurze Lebenserwartung haben. Neue Forschungsergebnisse können schnell in neue Produkte und damit auch in neue Datenformate umgesetzt werden.

Für den Bereich der Bilddokumente sei hier die Ablösung des GIF-Formats durch PNG (Portable Network Graphics) beispielhaft genannt. Bis weit in die 1990er Jahre hinein war GIF der wesentliche Standard, um Grafiken im Inter-

net zu übertragen und auf Servern zu speichern. Dieses wurde aufgrund leistungsfähigerer Hardware, sowie rechtlicher Probleme durch das JPEG- und PNG-Format abgelöst. Heute wird das GIF-Format noch weitestgehend von jeder Software unterstützt, allerdings werden immer weniger Daten in diesem Format generiert. Eine Einstellung der GIF-Format-Unterstützung durch die Softwarehersteller scheint damit nur noch eine Frage der Zeit zu sein.

Ferner können neue Forschungsansätze und Algorithmen zu neuen Datenformaten führen. Forschungsergebnisse in dem Bereich der Wavelet-Komprimierung[7] sowie die Verfügbarkeit schnellerer Hardware führten bspw. zu der Erarbeitung und Implementierung des JPEG2000 Standards, der wesentlich bessere Komprimierungsraten bei besserer Qualität liefert als sein Vorgänger und zeigt, dass heute auch hohe Komprimierungsraten bei verlustfreier Komprimierung erreicht werden können.

Verlustfrei ist ein Komprimierungsverfahren immer dann, wenn sich aus dem komprimierten Datenstrom die Quelldatei bitgenau rekonstruieren lässt. Verlustbehaftete Komprimierungsverfahren dagegen können die Bildinformationen lediglich annäherungsweise identisch wiedergeben, wobei für das menschliche Auge Unterschiede kaum oder, je nach Anwendung, überhaupt nicht sichtbar sind.

Trotz eines starken Anstiegs der Übertragungsgeschwindigkeiten und Rechengeschwindigkeiten sind auch heute noch bestimmte Datenformate für spezifische Einsatzzwecke im Einsatz. Ein allgemeines Universalformat existiert nicht. Dies hat mitunter auch mit der Unterstützung dieser Formate durch gängige Internetprogramme wie Web-Browser, Email-Programme etc. zu tun. Nachfolgend sollen die gängigsten derzeit genutzten Datenformate kurz vorgestellt werden:

PNG (Portable Network Graphics): Dieses Datenformat wurde speziell für den Einsatz in Netzwerken entwickelt, um Bilder schnell zu übertragen und anzuzeigen. Entsprechend wurde ein Komprimierungsalgorithmus mit einem guten Kompromiss zwischen Dateigröße und Performanz gewählt. Dieser komprimiert das Bild verlustfrei. Überwiegend kommt dieses Format für die Anzeige von kleineren Images im Web-Browser zum Einsatz.

JPEG: Das JPEG Format komprimiert im Gegensatz zu PNG verlustbehaftet. D.h. das ursprüngliche Ergebnis-Bild lässt sich nach der Dekomprimierung

7 Weitere, einführende Informationen zu Wavelets finden sich unter: Graps, Amara (o.J.): An Introduction to Wavelets, http://www.amara.com/ftpstuff/IEEEwavelet.pdf

nicht mehr genau reproduzieren. Dadurch lässt sich ein wesentlich höherer Komprimierungsfaktor erreichen, der zu kleineren Dateien führt. Speziell für den Transfer von größeren Farbbildern in Netzwerken findet dieses Format Anwendung.

TIFF (Tagged Image File Format): TIFF wurde als universelles Austauschformat in 1980ern von Aldus (jetzt Adobe) entwickelt. Obwohl letzte Spezifikation zwar schon aus dem Jahr 1992 datiert,[8] ist es heute immer noch in Gebrauch. Dies liegt überwiegend an dem modularen Aufbau des Formats. Das Format definiert sogenannte Tags, die über eine Nummer typisiert sind. Entsprechend dieser Nummer enthalten diese Tags unterschiedliche Informationen. Somit ließen sich mit der Zeit neue Tags definieren, um neuartige Daten abzuspeichern. Auch die Art und Weise, wie die Bilddaten komprimiert werden ist nicht eindeutig definiert. Vielmehr definiert TIFF eine Liste unterschiedlicher Komprimierungsalgorithmen, die zum Einsatz kommen können. Darunter ist neben einigen verlustfreien Algorithmen auch dasselbe verlustbehaftete Verfahren zu finden, welches auch im JPEG Format angewandt wird. Als eines der wenigen Datenformate erlaubt TIFF auch die unkomprimierte Speicherung der Bilddaten. Aus diesem Grund wurde TIFF lange als einziges Format für die Speicherung der Archivversion eines digitalen Bildes (Master-Image) angesehen, auch wenn es nicht sehr effizient mit dem Speicherplatz umgeht. Dieser relativ große Speicherbedarf trug allerdings auch dazu bei, dass TIFF nicht als geeignetes Format für die Übertragung von Bilddaten im Internet angesehen wurde und mit der Entwicklung alternativer Formate wie GIF oder PNG begonnen wurde. Auch wenn bei heutigen Ressourcen und Bandbreiten dies nicht mehr ein so grosses Problem wäre, können TIFF-Dateien von keinem Web-Browser angezeigt werden.

JPEG2000: Ursprünglich wurde JPEG2000 als „Nachfolgeformat" für JPEG entwickelt. Hierbei wurde versucht Nachteile des JPEG Formats gegenüber TIFF unter Beibehaltung hoher Komprimierungsraten auszugleichen. Dies gelang durch die Anwendung neuartiger sogenannter Wavelet basierter Komprimierungsalgorithmen. Neben einer verlustbehafteten Komprimierung unterstützt JPEG2000 auch eine verlustfreie Komprimierung. Aufgrund des neuartigen Komprimierungsalgorithmus sind die erzeugten Dateien wesentlich kleiner als bei TIFF. Dies ist nicht zuletzt auch der Grund, warum JPEG2000

8 O.V.:TIFF 6.0 Specification. http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf

neben TIFF als Datenformat für das „Digital Master“ eingesetzt wird, wenn es um das Speichern großer Farbbilder geht. Ähnlich des TIFF Formats können JPEG2000 Bilder derzeit nicht von einem Web-Browser angezeigt werden. Als Auslieferungsformat im Internet ist es daher derzeit nicht brauchbar.

Aus Perspektive der Langzeitarchivierung kommen also generell die Datenformate TIFF und JPEG2000 als Datenformat für das „Digital Master“ in Frage. Allerdings sind beide Formate so flexibel, dass diese Aussage spezifiziert werden muss.

Beide Formate können unterschiedliche Arten der Komprimierung nutzen. Diese ist entscheidend, um die Eignung als „Digital Master“-Format beurteilen zu können. So ist bspw. die LZW-Komprimierung für TIFF Images nach Bekanntwerden des entsprechenden Patents auf den Komprimierungsalgorithmus aus vielen Softwareprodukten verschwunden. Als Folge daraus lassen sich LZW-komprimierte TIFF Images nicht mit jeder Software einlesen, die TIFF unterstützt. Die Verlustbehaftete Komprimierung von JPEG2000 ist ebenfalls nicht als Format für das „Digital Master“ geeignet. Da hierbei Bytes nicht originalgetreu wieder hergestellt werden können, kommt für die Archivierung lediglich die verlustfreie Komprimierung des JPEG2000-Formats zum Einsatz.

Ferner spielt auch die Robustheit gegenüber Datenfehlern eine Rolle. So genannter „bitrot“ tritt mit der Zeit in fast allen Speichersystemen auf. Das bedeutet das einzelne Bits im Datenstrom kippen – aus der digitalen „1“ wird also eine „0“ oder umgekehrt. Solche Fehler führen dazu, dass Bilddateien gar nicht oder nur teilweise angezeigt werden können. Verschiedene Komprimierungsalgorithmen können entsprechend anfällig für einen solchen „bitrot“ sein. Datenformate können auch zusätzliche Informationen enthalten (sogenannte Checksums), um solche Fehler aufzuspüren oder gar zu korrigieren.

JPEG2000 bietet aufgrund seiner internen Struktur und des verwendeten Algorithmus einen weitreichenden Schutz gegen „bitrot“. Eine Fehlerrate von 0.01% im Bilddatenstrom (bezogen auf die Imagegesamtgröße) führt zu kaum sichtbaren Einschränkungen, wohingegen unkomprimierte TIFF-Dateien zu einzelnen fehlerhaften Zeilen führen können. Komprimierte TIFF-Dateien sind ohnehin wesentlich stärker von Bitfehlern beeinträchtigt, da der Bilddatenstrom nicht mehr vollständig dekomprimiert werden kann.[9]

---

9 Buonora, Paolo / Liberati, Franco: A Format for Digital Preservation – a study on JPEG 2000 File Robustness in: D-Lib Magazine, Volume 14, Number 7/8, http://www.dlib.org/dlib/july08/buonora/07buonora.html

Die Farbtiefe eines Bildes ist ebenfalls ein wichtiges Kriterium für die Auswahl des Datenformats für das „Digital Master". Rein bitonale Bilddaten (nur 1 bit pro Pixel, also reines Schwarz oder reines Weiß) können nicht im JPEG2000-Format gespeichert werden. Diese Bilddaten können jedoch im TIF-Format[10] durch die Verwendung des optionalen FaxG4-Komprimierungsalgorithmus sehr effizient gespeichert werden, welches verlustfrei komprimiert.

Den oben genannten Datenformaten ist gemein, dass sie von der Aufnahmequelle generiert werden müssen. Digitalkameras jedoch arbeiten intern mit einer eigenen an den CCD-Sensor angelehnten Datenstruktur. Dieser CCD-Sensor erkennt die einzelnen Farben in unterschiedlichen Sub-Pixeln, die nebeneinander liegen, wobei jedes dieser Sub-Pixel für eine andere Farbe zuständig ist. Um ein Image in ein gängiges Rasterimageformat generieren zu können, müssen diese Informationen aus den Sub-Pixeln zusammengeführt werden – d.h. entsprechende Farb-/Helligkeitswerte werden interpoliert. Je nach Aufbau und Form des CCD-Sensors finden unterschiedliche Algorithmen zur Berechnung des Rasterimages Anwendung. An dieser Stelle können aufgrund der unterschiedlichen Strukturen bereits bei einer Konvertierung in das Zielformat Qualitätsverluste entstehen. Daher geben hochwertige Digitalkameras in aller Regel das sogenannte „RAW-Format" aus, welches von vielen Fotografen als das Master-Imageformat betrachtet und somit archiviert wird. Dieses so genannte „Format" ist jedoch keinesfalls standardisiert.[11] Vielmehr hat jeder Kamerahersteller ein eigenes RAW-Format definiert. Für Gedächtnisinstitutionen ist diese Art der Imagedaten gerade über längere Zeiträume derzeit nur schwer zu archivieren. Daher wird zumeist auch immer eine TIFF- oder JPEG2000-Datei zusätzlich zu den RAW-Daten gespeichert.

Die Wahl eines passenden Dateiformats für die Images ist, gerade im Rahmen der Langzeitarchivierung, also relativ schwierig. Es muss damit gerechnet werden, dass Formate permanent auf ihre Aktualität, d.h. auf ihre Unterstützung durch Softwareprodukte, sowie auf ihre tatsächliche Nutzung hin überprüft werden müssen. Es kann davon ausgegangen werden, dass Imagedaten von Zeit zu Zeit in neue Formate überführt werden müssen, wobei unter Umständen auch ein Qualitätsverlust in Kauf genommen werden muss.

---

10 TIFF-Image oder TIFF-Datei aber TIF-Format, da in TIFF bereits „Format" enthalten ist (Tagged Image File Format).

11 Zu den Standarisierungsbestrebungen siehe http://www.openraw.org/info sowie http://www.adobe.com/products/dng/

Metadaten für die Archivierung

Ziel der Langzeitarchivierung ist das dauerhafte Speichern der Informationen, die in den Bilddokumenten abgelegt sind. Das bedeutet nicht zwangsläufig, dass die Datei als solche über einen langen Zeitraum aufbewahrt werden muss. Es kann bspw. erforderlich werden Inhalte in neue Formate zu überführen. Eine sogenannte Migration ist immer dann erforderlich, wenn das Risiko zu hoch wird ein bestimmtes Datenformat nicht mehr interpretieren zu können, weil kaum geeignete Soft- oder Hardware zur Verfügung steht.

Neben dem dauerhaften Speichern der Bilddaten ist es ebenfalls wichtig den Kontext der Bilddaten zu sichern. Unter Kontext sind in diesem Fall alle Informationen zu verstehen, die den Inhalt des Bilddokuments erst zu- und einordnen lassen. Dies ist in aller Regel der Archivierungsgegenstand. So ist bspw. eine einzelne als Bild digitalisierte Buchseite ohne den Kontext des Buches (= Archivierungsgegenstand) nicht einzuordnen. Im dem Fall eines Katastrophenszenarios, in dem auf zusätzliche Informationen, wie sie in etwa ein Repository oder ein Katalog enthält, nicht zugegriffen werden kann, weil entweder das System nicht mehr existiert oder aber die Verknüpfung zwischen System und Bilddokument verloren gegangen ist, können zusätzliche Metadaten, die in dem Bilddokument direkt gespeichert werden, den Kontext grob wieder herstellen.

Deskriptive Metadaten in Bilddokumenten

Diese sogenannten deskriptiven Metadaten, die den Archivierungsgegenstand und nicht das einzelne Bilddokument beschreiben, können direkt in jedem Bilddokument gespeichert werden. Jedes Datenformat bietet dazu eigene proprietäre Möglichkeiten.

Frühe Digitalisierungsaktivitäten haben dazu bspw. die TIFF-Tags PAGENAME, DOCUMENTNAME und IMAGEDESCRIPTION genutzt, um entsprechende deskriptive Metadaten wie Titelinformation und Seitenzahl abzubilden.[12] Diese sind mitunter auch heute noch in Digitalisierungsprojekten gebräuchlich. Eine weniger proprietäre Lösung ist die von Adobe entwickelte Extensible Metadata Plattform (XMP).[13] Zum Speichern von deskriptiven

12 O.V.: Bericht der Arbeitsgruppe Technik zur Vorbereitung des Programms „Retrospektive Digitalisierung von Bibliotheksbeständen“ im Förderbereich „Verteilte Digitale Forschungsbibliothek“, Anlage 1, http://www.sub.uni-goettingen.de/ebene_2/vdf/anlage1.htm

13 O.V.: XMP Specification, September 2005, http://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf

Metadaten verwendet XMP das Dublin Core Schema. XMP-Daten können sowohl zu TIFF und JPEG2000 hinzugefügt werden als auch zu PDF und dem von Adobe entwickeltem Bilddatenformat für RAW-Bilddaten DNG.

Im Falle eines Katastrophenszenarios im Rahmen der Langzeitarchivierung lässt sich mittels dieser XMP-Daten ein entsprechender Kontext zu jedem Bilddokument wieder aufbauen.

Technische Metadaten für Bilddokumente

Jede Datei hat aufgrund ihrer Existenz inhärente technische Metadaten. Diese sind unabhängig vom verwendeten Datenformat und dienen bspw. dazu die Authentizität eines Images zu beurteilen. Checksummen sowie Größeninformationen können Hinweise darauf geben, ob ein Image im Langzeitarchiv modifiziert wurde.

Darüber hinaus gibt es formatspezifische Metadaten. Diese hängen direkt vom eingesetzten Datenformat ab und enthalten bspw. allgemeine Informationen über ein Bilddokument:

- Bildgröße in Pixel sowie Farbtiefe und Farbmodell
- Information über das Subformat – also bspw. Informationen zum angewandten Komprimierungsalgorithmus, damit der Datenstrom auch wieder entpackt und angezeigt werden kann.

Mittels Programmen wie bspw. JHOVE[14] lassen sich eine Vielzahl von technischen Daten aus einer Datei gewinnen. Gespeichert wird das Ergebnis als XML-Datei. Als solche können die Daten in Containerformate wie bspw. METS eingefügt und im Repository gespeichert werden. Aufgrund der Menge der auszugebenden Informationen sind diese allerdings kritisch zu bewerten. Entsprechende Datensätze bspw. für ein digitalisiertes Buch sind entsprechend groß. Daher wird nur in seltenen Fällen der komplette Datensatz gespeichert, sondern bestimmte technische Metadaten ausgewählt. Für Bilddokumente beschreibt NISO Z39.87 ein Metadatenschema für das Speichern von technischen Metadaten.[15] Eine entsprechende Implementierung in XML steht mit MIX ebenfalls bereit.[16]

---

14 JHOVE – JSTOR/Harvard Object Validation Enviroment, http://hul.harvard.edu/jhove/

15 O.V.: Data Dictionary – Technical Metadata for Digital Still Images, http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c

16 http://www.loc.gov/standards/mix/

Es ist anzunehmen, dass zukünftig Migrationsprozesse vor allem bestimmte Sub-Formate betreffen werden, also bspw. nur TIFF-Dateien mit LZW-Komprimierung anstatt alle TIFF-Dateien. Für die Selektion von entsprechenden Daten kommt dem Format also eine große Bedeutung zu. Mit PRONOM steht eine Datenbank bereit, die Dateiformate definiert und beschreibt. Dabei geht die Granularität der Datenbank weit über gängige Formatdefinitionen, wie sie bspw. durch den MIME[17]-Type definiert werden, hinaus. TIFF-Dateien mit unterschiedlicher Komprimierung werden von PRONOM[18] als unterschiedliche Formate verstanden. Um diese Formatinformationen aus den Bilddokumenten zu extrahieren steht mit DROID[19] ein entsprechendes Tool zur Verfügung.

Herkunftsmetadaten für Bilddokumente

Für die Langzeitarchivierung sind neben technischen Metadaten auch Informationen über die Herkunft der Bilddateien wichtig. Informationen zur eingesetzten Hard- und Softwareumgebung können hilfreich sein, um später bestimmte Gruppen zur Bearbeitung bzw. Migration (Formatkonvertierungen) auswählen oder aber um Bilddokumente überhaupt darstellen zu können.

Im klassischen Sinn werden Formatmigrationen zwar anhand des Dateiformats ausgewählt. Da jedoch Software selten fehlerfrei arbeitet, muss bereits bei der Vorbereitung der Imagedaten Vorsorge getroffen werden, entsprechende Dateigruppen einfach selektieren zu können, um später bspw. automatische Korrekturalgorithmen oder spezielle Konvertierungen durchführen zu können.

Ein nachvollziehbares und in der Vergangenheit real aufgetretenes Szenario ist bspw. die Produktion fehlerhafter PDF-Dateien auf Basis von Images durch den Einsatz einer Programmbibliothek, die sich im nachhinein als defekt erwies. In der Praxis werden diese nur zugekauft, sodass deren Interna dem Softwareanbieter des Endproduktes unbekannt sind. Tritt in einer solchen Programmbibliothek ein Fehler auf, so ist dieser eventuell für den Programmierer nicht auffindbar, wenn er seine selbst erzeugten Dateien nicht wieder einliest (bspw. weil Daten nur exportiert werden). Ein solcher Fehler kann auch nur in einer bestimmten Softwareumgebung (bspw. abhängig vom Betriebssystem) auftreten. Kritisch für die Langzeitarchivierung wird der Fall dann, wenn einige Softwareprodukte solche Daten unbeanstan-

17 Freed, N; Borenstein, N (1996): Multipurpose Internet Mail Extensions (MIME) part one, RFC2045, http://tools.ietf.org/html/rfc2045

18 http://www.nationalarchives.gov.uk/pronom/

19 Digital Record Object Identification (DROID): http://droid.sourceforge.net

det laden und anzeigen, wie in diesem Fall der Adobe PDF-Reader. „Schwierigkeiten" hatten dagegen OpenSource Programme wie Ghostscript sowie die eingebauten Postscript-Interpreter einiger getesteter Laserdrucker.

Trotz gewissenhafter Datengenerierung und Überprüfung der Ergebnisse kann es also dazu kommen, dass nicht konforme Bilddokumente über Monate oder Jahre hinweg produziert werden. Entsprechende Informationen zur technischen Laufzeitumgebung erleichtern jedoch die spätere Identifikation dieser „defekten" Daten im Langzeitarchivierungssystem.

Eine weitere Aufgabe der Herkunftsmetadaten ist es den Lebenszyklus eines Dokuments aufzuzeichnen. Durch Verweise auf Vorgängerdateien können Migrationsprozesse zurückverfolgt werden. Dies gewährleistet, dass auch auf frühere Generationen als Basis für eine Migration zurückgegriffen werden kann. Im Fall von „defekten" Daten ist das eine wesentliche Voraussetzung, um überhaupt wieder valide Inhalte generieren zu können.

Sowohl technische als auch Herkunftsmetadaten werden als eigenständige Metadatenrecords unter Verwendung spezifischer Metadatenschemata gespeichert. Für Bilddokumente bietet sich MIX für die technischen Metadaten an. Da Herkunftsmetadaten nicht spezifisch auf Bilddokumente zugeschnitten sind, stellen allgemeine Langzeitarchivierungsmetadatenschemata wie bspw. PREMIS[20] entsprechende Felder bereit.

Um die unterschiedlichen Metadaten zusammen zu halten, kommt darüber hinaus ein Containerformat wie METS[21] oder MPEG-21 DIDL[22] zum Einsatz.

Ausblick

Sollen Bilddokumente entsprechend der oben skizzierten Anforderungen für die Langzeitarchivierung vorbereitet werden, ist es aus praktischer Sicht unerlässlich aktuelle Werkzeuge und Geschäftsprozesse zu evaluieren. Viele Werkzeuge sind bspw. nicht in der Lage entsprechende Metadaten wie bspw. XMP in einem Bilddokument zu belassen. Ein Speichern des Bilddokuments sichert zwar den entsprechenden Bilddatenstrom, lässt die deskriptiven Metadaten außen vor.

---

20 http://www.loc.gov/premis

21 Siehe Kap. 6.2 Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten

22 Bekart, Jeroen; Hochstenbach, Patrick; Van de Sompel Herbert (2003): Using MPEG-21 DIDL to represent complex objects in the Los Alamos National Laboratory Digital Library In: D-Lib Magazine, Band 9, November 2003, http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

Das Vorbereiten der Bilddokumente für die Langzeitarchivierung ist in aller Regel ein mehrstufiger Prozess. Dieser Prozess muss wohl dokumentiert und gesteuert werden, um eine gleichbleibende Qualität sicherzustellen. Ein „spontan" durchgeführtes Laden und Abspeichern eines Images könnte dazu führen, dass sich bspw. technische Metadaten wie die Checksumme ändern, da eigene, zusätzliche Metadaten durch die Software eingefügt wurden. In der Praxis hat sich für die Aufbereitung von Bilddokumente folgender, hier stark vereinfachter Workflow als sinnvoll erwiesen:

- Einfügen der deskriptiven Metadaten in das Bilddokument
- Validieren des Datenformates des Bilddokuments
- Extrahieren der Formatinformation (JHOVE) inkl. der Formatbestimmung (DROID)
- Extrahieren der allgemeinen technischen Metadaten (Checksummen)
- Generierung der technischen und Herkunftsmetadaten (MIX und PREMIS) aus den Formatinformationen
- Einfügen der technischen und Herkunftsmetadaten in ein Containerformat des Repositories.

Aufgrund der Menge an Bilddokumenten ist dieser Prozeß nur automatisiert durchführbar. Um Fehler zu vermeiden und auch auf nachträglich notwendige Korrekturen reagieren zu können, ist der Einsatz spezieller Software zur Steuerung von Geschäftsprozessen sinnvoll. Dadurch wird eine gleichbleibende Qualität gewährleistet. Ferner ist zu hoffen, dass damit Zeitaufwand und Kosten für die Langzeitarchivierung von Bilddokumenten sinken.

17.4 Multimedia/Komplexe Applikationen

Winfried Bergmeyer

Die Anforderung für den Erhalt und die Nutzung von multimedialen und komplexen Applikationen werden bestimmt durch die Vielfältigkeit der integrierten Medien und den oft nichtlinearen, da benutzergesteuerten Ablauf. Kern der Langzeitarchivierung ist daher der Erhalt der Progammlogik und -ausführung, der nur durch eine angemessene Dokumentation und Bereitstellung der notwendigen Laufzeitumgebung gewährleistet werden kann. Mit der Bewahrung dieser Programme wird der Umgang mit digitalen Daten in unserer Gesellschaft für spätere Generationen dokumentiert.

Bis zum Beginn des 20. Jahrhunderts bestanden die kulturellen Erzeugnisse, die ihren Weg in Bibliotheken, Archive und Museen fanden, in der Regel aus Büchern, Handschriften, Plänen, Gemälden und anderen Medien, deren Nutzung ohne technische Hilfsmittel erfolgen konnte. Mit Erfindung der Fotografie, des Films und der Tonaufzeichnung hat sich das Spektrum der kulturellen Produktion um Medien erweitert, die das Kulturschaffen bzw. dessen Aufzeichnung revolutionierten, dabei aber technische Hilfsmittel, beispielsweise in Form von Tonbandgeräten oder Schallplattenspielern, für deren Nutzung erforderlich machten. Zum Ende des ausgehenden 20. Jahrhunderts erlebten wir mit der Revolution der Informationstechnologie eine weitere, tiefgreifende Veränderung. Nicht nur, dass mit dem Internet und dem Aufkommen moderner audiovisueller Anwendungen neuartige Kommunikations- und Ausdrucksformen entstanden, auch wurden und werden analoge Objekte zum Zweck der Langzeitbewahrung und der Langzeitverfügbarkeit in das digitale Format überführt. Diese digitalen Objekte sind ohne entsprechende Interpretation der Datenströme durch den Computer nicht nutzbar und damit verloren. Der Auftrag zur Bewahrung des kulturellen Erbes[23] erfordert angesichts dieser Abhängigkeiten neue Konzepte für die Sicherung und Nutzbarkeit digitaler Zeugnisse unserer Kultur in Bibliotheken, Archiven und Museen.

Der Begriff „*Multimedia*“ bedarf in diesem Zusammenhang einer genaueren Definition.[24] Entsprechend des eigentlichen Wortsinnes beinhalten multimediale Objekte zumindest zwei unterschiedliche Medien, z.B. Ton und Bildfolgen.

---

23 http://portal.unesco.org/ci/en/files/13367/109966596613Charter_ge.pdf/Charter_ge.pdf

24 Das Wort „Multimedia“ wurde 1995 durch die Gesellschaft für deutsche Sprache zum „Wort des Jahres“ erklärt. 1995 stand der Begriff vor allem für die interaktiven Innovationen im Bereich der Computertechnologie.

Mittlerweile ist dieser Begriff allerdings für die Bezeichnung von Objekten mit nichttextuellen Inhalten gebräuchlich. Wir werden den Begriff hier in diesem letztgenannten Sinne verwenden.

Vor allem im Audio- und Videobereich steht die technische Entwicklung in Abhängigkeit von der permanenten Erschließung neuer kommerzieller Märkte. Damit ergibt sich, angeschoben durch den Innovationsdruck des Marktes, das Problem der Obsoleszens von Hardware, Software und Dateiformaten. Ein Blick auf den Bereich der Tonaufzeichnung zeigt z.B. im Hardwarebereich seit den frühen Wachszylindern ein vielfältiges Entwicklungsspektrum über Schallplatte, Tonband, Kassette, Diskette, CD-Rom und DVD, deren Innovationszyklen sich sogar beschleunigen. Ein Ende der technischen Fort- und Neuentwicklung ist nicht in Sicht. Die Bewahrung der so gespeicherten kulturellen Erzeugnisse erfordert für die kulturbewahrenden Institutionen erhebliche finanzielle, technische und personelle Anstrengungen. In der Bestandserhaltung rücken die inhaltserhaltenden Maßnahmen, beschleunigt durch den Trend zur digitalen Herstellung von Publikationen, sowie zur Digitalisierung von analogem Material, immer stärker in den Mittelpunkt.[25]

Seit den 1990er Jahren wurden beispielsweise CD-Roms mit multimedialen Inhalten auf den Markt gebracht, die sich das neue Medium und seine interaktiven Möglichkeiten zunutze machten. Bereits heute sind die ersten Exemplare auf aktuellen Computern nicht mehr nutzbar. Zwar ist das Medium (CD-Rom) nicht veraltet, aber die digitalen Informationen können nicht interpretiert werden, da die notwendigen Programme auf aktuellen Betriebssystemen nicht mehr lauffähig sind. Ganze Sammlungen dieser multimedialen Publikationen drohen unbrauchbar zu werden und somit als Teil unseres kulturellen Erbes verloren zu gehen. In diesem Rahmen sei auch auf die zahlreichen Disketten, CD-Roms und DVDs verwiesen, die als Beilagen zu Publikationen in die Bibliotheken Eingang finden. Hier stellt sich zusätzlich die Aufgabe, die darauf enthaltenen Informationen (Programme, Texte, Bilder, Videos etc.) zu bewahren und darüber hinaus den Verweis auf die gedruckte Publikation, ohne die die digitalen Inhalte oft unverständlich sind, zu erhalten.

Mit den sich verändernden Distributionsformen (Video-on-demand, Filesharing u.a.) entstehen neue Notwendigkeiten für die Sicherung der Urheber- und Verwertungsrechte in Form des „Digital Rights Mangement" mit Nutzungslimitierungen, die weitreichende Folgen für die Langzeitarchivierung, vor allem im Bereich der audiovisuellen Medien, mit sich bringen.

---

25 Royan, Bruce/Cremer, Monika: *Richtlinien für Audiovisuelle und Multimedia-Materialien in Bibliotheken und anderen Institutionen*, IFLA Professional Reports No. 85, http://www.ifla.org/VII/s35/index.htm#Projects

Unter einer komplexen Applikation wird eine Datei oder eine Gruppe von Dateien bezeichnet, die als Computerprogramm ausgeführt werden können. Dies kann ein Anwendungsprogramm, ein Computerspiel ebenso wie eine eLearning-Anwendung sein. Multimediale Elemente sind dabei oftmals Bestandteil dieser Applikationen. Anders als bei den oben besprochenen, nichttextuellen Objekten ist bei den Applikationen oftmals eine direkte Abhängigkeit der Nutzbarkeit vom Betriebssystem und/oder Hardware gegeben.[26] Erst die diesen Applikationen inhärenten Programmabläufe inklusive der Einbettung multimedialer Elemente erfüllen die intendierten Aufgaben und Ziele. Interaktive Applikationen verlangen daher Langzeitarchivierungsstrategien in Form der Emulation[27] oder aber der „Technology preservation", der Archivierung der Hardware und Betriebssysteme. Eine Migration der Daten für die Nutzung auf anderen Betriebssystemen wird hier nur in wenigen Fällen über die Compilierung des Quellcodes (falls vorhanden) möglich sein.

Ein wesentliches Element von komplexen Applikationen ist der Verzicht auf lineare Abläufe, d.h. die Nutzer können selbstbestimmt mit dem Programm interagieren. Im Gegensatz zur Erhaltung der Digitalisate von Einzelobjekten oder Objektgruppen ist ein wesentlicher Bestandteil interaktiver Applikationen, dass hier nicht das Einzelobjekt und seine Metadaten im Vordergrund stehen, sondern die Verarbeitung von Daten, die entweder Teil der Applikation sind (z.B. in einer Datenbank) oder aber über Schnittstellen importiert oder manuell eingegeben werden.

Eine in diesem Zusammenhang immer wieder gestellte Frage ist die nach der Zulässigkeit dieser Emulations- und Migrationskonzepte im Bezug auf Kunstwerke und deren Authentizität.[28] Die zunehmenden Interaktions- und Modifikationsmöglichkeiten durch den Rezipienten, die Einfluß auf das künstlerische „Objekt" (Anwendung) haben und haben sollen, werfen zusätzliche Fragen auf,

---

26 Ein Beispiel aus der Praxis der Langzeiterhaltung von multimedialen CD-Roms bietet Martin, Jeff: *Voyager Company CD-ROMs: Production History and Preservation Challenges of Commercial Interactive Media.* In: http://www.eai.org/resourceguide/preservation/computer/pdf-docs/voyager_casestudy.pdf

27 Rothenberg, Jeff: *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* In: http://www.clir.org/PUBS/reports/rothenberg/contents.html. Er fordert die Einbindung digitaler Informationen in die Emulatoren, so dass es möglich wird, originäre Abspielumgebungen zu rekonstruieren.

28 Als Beispiel siehe die Diskussion um das Projekt "The Earl King". Rothenberg , Jeff/ Grahame Weinbren/Roberta Friedman: *The Erl King, 1982–85*, in: Depocas ‚Alain/Ippolito, Jon/Jones, Caitlin (Hrsg.) (2003): *The Variable Media Approach - Permanence through Change.* New York, S. 101 – 107. Ders.: *Renewing The Erl King*, January 2006, in: http://bampfa.berkeley.edu/about/ErlKingReport.pdf

die im Rahmen der Langzeitarchivierung und der Langzeitverfügbarkeit beantwortet werden müssen.[29] Hier besteht eine Verbindung zu den Erhaltungsproblematiken von Computerspielen, da auch hier über die reine Nutzbarkeit der Programme hinaus das „Look and Feel"-Erlebnis, das u.a. auch vom Einsatz originaler Hardwareumgebungen abhängig ist, elementarer Bestandteil der Anwendung ist.[30]

Insbesondere für komplexe Applikationen gilt, dass für die Erhaltung und Nutzungsfähigkeit beiliegendes Material in Form von Verpackungen, Handbüchern, Dokumentation etc. ebenfalls archiviert werden muss. Für die weitere Nutzung der Programme ist die Sicherung der Installationsanweisungen, Programmierungsdokumentationen und Bedienungsanleitungen notwendig.[31] Diese Aufgabe stellt somit erhöhte Anforderungen an die Erstellung und Anwendung von umfassenden Archivierungskonzepten, z.B. auf Grundlage des OAIS (Open Archival Information System).[32]

Die Bedeutung der unterschiedlichen Arten von Metadaten im Rahmen der Langzeitarchivierung komplexer Applikationen wird u.a. in vielen Projekten zur Archivierung von Medienkunst deutlich.[33] Es sind nicht nur die zusätzlich anfallenden Informationen wie Handbücher, Installationsanweisungen etc sondern

---

29 Rinehart, Richard: *The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Century*. In: http://switch.sjsu.edu/web/v6n1/article_a.htm

30 Die Kombination unterschiedlicher Verfahren der Langzeitarchivierung wurde von der University of Queensland durchgeführt. Hunter, Jane/Choudhury, Sharmin: *Implementing Preservation Strategies for complex Multimedia Objects*. In: http://metadata.net/panic/Papers/ECDL2003_paper.pdf

31 Duranti, Luciana: *Preserving Authentic Electronic Art Over The Long-Term: The InterPARES 2 Project*, Presented at the Electronic Media Group, Annual Meeting of the American Institute for Conservation of Historic and Artistic Works, Portland, Oregon, June 14, 2004. Die Projekte InterPares und InterPares2 und aktuell InterPares3 setzen sich u.a. mit den Anforderungen zur Langzeitarchivierung aktueller Werke der bildenden und darstellenden Künste auseinander. Siehe dazu http://www.interpares.org/ip2/ip2_index.cfm

32 Siehe als Beispiel der Implementierung des OAIS das Projekt „Distarnet" der Fachhochschule Basel. Melli, Markus: *Distartnet. A Distributed Archival Network*. In: http://www.distarnet.ch/distarnet.pdf und Margulies, Simon: *Distranet und das Referenzmodell OAIS*. In: http://www.distarnet.ch/distoais.pdf. Das europäische Projekt *CASPAR* (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) befasst sich u .a. mit der Implementierung des OAIS in der Archivierungsprozess. http://www.casparpreserves.eu/

33 Mikroyannidis, Alexander/Ong, Bee/Ng, Kia/Giaretta, David: *Ontology-Driven Digital Preservation of Interactive Multimedia Performances*. In: http://www.leeds.ac.uk/icsrim/caspar/caspar-data/AXMEDIS2007-caspar20071022-v1-4-a.pdf. Rinehart, Richard: *A System of Formal Notation for Scoring Works of Digital and Variable Media Art*. In: http://aic.stanford.edu/sg/emg/library/pdf/rinehart/Rinehart-EMG2004.pdf

auch die integrierten Medien (Audio, Video, Grafik etc.) mit ihren unterschiedlichen Nutzungsanforderungen und der Bezug zur Applikationslogik, die eine umfangreiche, strukturierte Metadatensammlung erfordern. Vorhandene Standards für Metadatenschemata, die für multimediale und interaktive Applikationen im Rahmen der Langzeitarchivierung Verwendung finden können, sind mit *PREMIS*[34] und *LMER*[35] bereits vorhanden, darüber hinaus wird in vielen Projekten *METS* (Metadata Encoding and Transmission Standard) für die Kapselung (packaging) der Beschreibung der digitalen Objekte und deren Metadaten verwendet.[36]

Den Umgang und die Nutzung digitaler Informationen in unserer Gesellschaft und Kultur auch für folgende Generationen zu dokumentieren, ist für die Bewahrung komplexer Applikationen das entscheidende Argument. Nicht allein die Produktion digitaler Daten hat unsere Welt verändert, sondern vor allem der Umgang mit ihnen. Die Bewahrung und Sicherung der Ausführbarkeit dieser Computerprogramme trägt dem Prozess der grundlegenden Veränderungen in vielen Lebensbereichen, die sich durch den Einsatz der neuen Medien revolutioniert haben, Rechnung.

Literatur

Borghoff, Uwe M. /Rödig, Peter/ Scheffczyk, Jan (2003): *Langzeitarchivierung. Methoden zur Rettung digitaler Datenbestände.* Dpunkt Verlag.

Hunter, Jane/Choudhury, Sharmin: *Implementing Preservation Strategies for complex Multimedia Objects*. In: http://metadata.net/panic/Papers/ECDL2003_paper.pdf

Melli, Markus (2003): *Distartnet. A Distributed Archival Network*. In: http://www.distarnet.ch/distarnet.pdf

Rinehart, Richard: *The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Century*. In: http://switch.sjsu.edu/web/v6n1/article_a.htm

34 *PREMIS* (PREservation Metadata: Implementation Strategies) wurde durch das OCLC Online Computer Library Center entwickelt. http://www.oclc.org/research/projects/pmwg/

35 LMER (Langzeitarchivierungsmetadaten für elektronische Ressourcen) ist eine Entwicklung der von der Deutschen Bibliothek auf Basis des „Metadata Implementation Schema" der Nationalbibliothek Neuseelands. http://www.d-nb.de/standards/lmer/lmer.htm

36 http://www.loc.gov/standards/mets/

Rothenberg, Jeff: *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* In: http://www.clir.org/PUBS/reports/rothenberg/contents.html

Steinke, Tobias (Red.) (2005): *LMER. Langzeitarchivierungsmetadaten für elektronische Ressourcen*. Leipzig, Frankfurt a. M., Berlin. http://www.d-nb.de/standards/pdf/lmer12.pdf

17.5 Video

Dietrich Sauter

Videoformate haben in den letzten Jahren an Komplexität stark zugenommen. Ursache ist die fortschreitende Entwicklung immer neuer und effizienterer Kompressionsverfahren, aber auch die starke Integration der Speicherbausteine jeder Art. Die vielfältigen Produktionsformate, die heute zum Einsatz kommen, schlagen aber nicht direkt auf die Speicherung in den Langzeitarchiven durch. Denn für Langzeitarchive gelten andere Regeln, die nicht so kurzatmig sind. Nachdem das Produktionsformat der nächsten Jahre feststeht, ist eine gewisse Ruhe zu erwarten. Der Beitrag gibt einen Überblick über (Standardisierungs-) Ansätze und offene Fragen, die für die Langzeitarchivierung von Video von Bedeutung sind. Zum besseren Verständnis werden in einem umfangreichen Anhang die Videoformate in ihrer Entwicklung vorgestellt. Die Übersicht über die gängigen Speichermedien und das Glossar ermöglichen auch Nicht-Fachleuten die Fachdiskussion nachzuvollziehen und sind als Nachschlagewerk angelegt.

Die Anzahl und Komplexität der Formate im Bereich Video ist in den vergangenen Jahren nicht übersichtlicher geworden: Ständig werden neue Techniken präsentiert, immer kürzer werden die Produktzyklen. Dabei herrschen Zustände, die den Anwendern etliches Fachwissen abverlangen. Für die Produktion und die alltägliche Verwendung hat die EBU (European Broadcast Union) vier Systeme für HDTV und ein System für das herkömmliche SDTV definiert. Die folgende Formatübersicht listet die digitalen Systeme auf.

Zielformat ist das System4 1080p/50, d.h. 1080 Zeilen mit 1920 Pixeln und 50 Vollbildern/sec bei progressiver Abtastung. Da aus Gründen der Übertragungsbandbreite die Kosten heute zu hoch sind, wird das System1 720p/50 für den Übergang in den nächsten ca. 10 Jahren vorgeschlagen.

Produktionsspeicher

Die Datenrate ist die erste Orientierung für die Qualität. Hoch entwickelte Kodiertechniken nutzen dabei die verfügbare Datenrate am effektivsten aus.
Ein Signal mit einer Videodatenrate von 200 Megabit pro Sekunde (Mbit/s) und einem Abtastverhältnis von 4:4:4 kann eine höhere Bildqualität aufzeichnen als eines mit 100 Mbit/s und 4:2:2.

HDTV Systeme EBUTech 3299	Horizontal samples	Active lines	Frame rate	Sub-sampling / Quantisation [Bit]		Net image Bit Rate [Gbit/s]
System 1 720p/50	1280	720	50	4:2:2	10	0,9216
System 2 1080i/25	1920	1080	25	4:2:2	10	1,0368
System 3 1080p/25	1920	1080	25	4:2:2	10	1,0368
System 4 1080p/50	1920	1080	50	4:2:2	10	2,0736
SDTV 576i/25	720	576	25	4:2:2	10	0,207

Abbildung 1: Von der EBU definierte Formate (HDTV-Breitbildformat 16:9, SDTV 4:3)

Eine wichtige Rolle spielt auch das Raster, mit dem das jeweilige Verfahren arbeitet. Bei SD-Verfahren liegt dieses Raster fest: 576 Zeilen mit je 720 Pixeln, bei HD gibt es den Unterschied zwischen 720 Zeilen und je 1280 Pixeln oder 1080 Zeilen mit je 1920 Pixeln.

Weiteren Einfluss hat das Kompressionsverfahren. Ein grundlegender Unterschied besteht darin, ob nur innerhalb eines Bildes komprimiert wird (Intraframe), oder ob mehrere aufeinander folgende Bilder gemeinsam verarbeitet werden (Interframe). Interframe-Verfahren, also auf mehreren Bildern basierend, arbeiten prinzipiell effektiver, erlauben also bei einer vorgegebenen Datenrate höhere Bildqualität als Intraframe-Verfahren. Interframe-Verfahren erfordern jedoch höhere Rechenleistung beim Kodieren und Dekodieren, sie sind dadurch für die Postproduktion ungünstiger, da sich die Nachbearbeitungszeiten wie Rendering verlängern. Je intensiver Material bearbeitet werden soll, umso höher sollte die Datenrate sein und umso besser ist es, nur Intraframe-Kompression zu nutzen.

Archive zur alltäglichen Verwendung und in der Sendung

Die Rundfunkanstalten unterhalten Produktions- und Sendearchive. Diese Archive enthalten z. Zt. meist Metadaten, die auf die Inhalte verweisen. In zunehmendem Maße werden aber digitale Inhalte eingestellt. Die Fortschritte

sind in der Audiowelt wegen der geringeren Datenrate größer als bei den Videoarchiven. Geschlossene Contentmanagementsysteme sind heute überall im Aufbau. Dabei wird immer noch in High- und Low-Resolution-Speicherung unterschieden. Für die Zukunft sind aber integrierte Systeme zwingend. Die trimediale Ausspielung erfordert auch noch ein formatunabhängiges Abspeichern der Inhalte. Die angewandten Kompressionsformate erlauben außer bei JPEG2000 keine skalierte Speicherung, sodass die Formate in der Auflösung diskret vorgehalten werden müssen.
Kriterien für Archivmaterial sind:

- *Festigkeit:* Es muss das wieder heraus kommen, was einmal hineingesteckt wurde.
- *Nutzbarkeit:* was herauskommt, muss auch zu gebrauchen sein.

Langzeitarchive

Die Langzeitarchivierung von Medien erfordert Speichermedien, die möglichst lange gelagert werden können, ohne dass sich die Eigenschaften elementar verändern. Die meisten der heutigen Träger erfüllen diese Anforderung nicht.
Die Folge ist, die Medien müssen ständig ausgetauscht und damit die Inhalte kopiert werden.

In der Informationstechnik verwendete Träger werden deshalb in der Regel ca. alle sieben Jahre kopiert bzw. geklont. Bei häufiger Benutzung der Träger sind auch kürzere Zyklen angesagt. Im Bereich der analogen Aufzeichnung von Videosignalen haben sich die Aufzeichnungsformate alle fünf Jahre verändert. Die Lebensdauer der Formate beträgt aber dennoch ca. 15 Jahre. Jahrzehntelang wurde ein Fehler beim Auslesen eines analogen Videobandes durch das „Concealment“ verdeckt. Eine Zeile konnte durch den Inhalt der vorherigen ersetzt werden und die Wiedergabe lief ohne Störung weiter. Sicher gibt es eine Grenze dabei, wie viel Concealment man tolerieren kann, bevor die Fehlerverdeckung sichtbar wird. Aber ein paar Fehler pro Bild würden sicher akzeptiert werden.

In der digitalen Welt sind IT- Systeme so ausgelegt, dass sie ein File perfekt auslesen. Es gibt zahlreiche Firmen, die diese Fähigkeiten betonen, und es ist alles sehr imponierend, wenn es dann auch funktioniert. Aber wenn etwas versagt, wird in der Regel das ganze File abgelehnt. Was genau passiert, hängt vom File- Management System, vom Operativen System und von den Details der individuellen Anwendungen ab. Es ist aber die Realität, dass häufig ‚cannot read

file‘ oder ‚cannot open file‘ als Fehlermeldungen erscheinen und dann kann nicht mehr weiter gearbeitet werden.

Die MAZ-Fehlerverdeckung war immer vorhanden, arbeitete in Echtzeit immer dann und dort, wo sie gebraucht wurde und hielt so den analogen Betrieb am Laufen.

Für Langzeitarchive gilt generell: es dürfen nur standardisierte Codecs verwendet werden, bei denen alle Parameter offen zugänglich und dokumentiert sind. Dabei sind Codecs vorzuziehen, die auch eine Softwarevariante haben, damit nach Wegfall der Hardware die Inhalte trotzdem wieder erschlossen werden können. Die Metadaten sollen der Essenz beigefügt werden, damit sie nicht verloren gehen. Die dafür notwendigen Auslesemechanismen (Wrapper) sollten möglichst weit verbreitet sein.

Video-Aufzeichnungsformate

Um die Komplexität der Problemstellung bei der Langzeitarchivierung von Video nachvollziehen zu können, muss man ein Verständnis für die Vielfalt der in der Praxis verbreiteten Videoformate entwickeln. Deshalb werden zunächst Formate mit Videokompression genannt und anschließend die marktgängigen Formate in ihrer Entwicklung vorgestellt. Da dieser Part sehr umfangreich ist, wird er im Anhang aufgeführt. An dieser Stelle findet eine Konzentration auf langzeitachivierungsrelevante Formate statt, für die die marktüblichen Videoformate nicht geeignet sind.

Empfehlung zur Anwendung des Fileformats MXF (Material eXchange Format)

Für den Austausch wird das Fileformat MXF entsprechend der aktuellen SMPTE-Spezifikationen eingesetzt. Das übergebene Fileformat MXF hat den vereinbarten Übergaberichtlinien zu entsprechen. Darin ist unter anderem folgendes festgelegt:

Essence Container
Es wird der MXF Generic Container verwendet. Es wird im produzierten Kompressionsverfahren abgegeben. An Kompressionsverfahren werden zugelassen:

- MPEG 422P/ML (SMPTE S356, EBU-D94),
- DV-based/DV-based50 (SMPTE S314),

- DV (schließt MiniDV und DVCAM ein).
- Für die Filmverteilung JPEG2000

Operational Pattern
Es ist vorgesehen zeitlich kontinuierliche Einzelobjekte in separaten Files zu übertragen.

Video, Audio und Daten sind in gemultiplexter Form (compound) auszutauschen. Damit ist sichergestellt, dass die ausgetauschten MXF Files streamingfähig sind. Aus diesem Grund wird im Austausch ausschließlich das Operational Pattern 1a vorgesehen (siehe Abbildung 2). In einem operational pattern können Komplexitäten hinsichtlich folgender Parameter festgelegt bzw. angezeigt werden:

- Zusammensetzung der output timeline
- Anzahl der essence containers
- ob ein oder mehrere output timelines im MXF-File vorhanden sind
- ob der MXF-File für einen Stream-Transfer geeignet ist oder nicht
- ob sich Essenz außerhalb des MXF-Files befindet
- ob alle essence containers des MXF-Files nur einen essence track besitzen oder mindestens ein essence container mehr als einen essence track aufweist
- ob die Essenz durch eine index table indexiert wird.

Bislang sind die operational patterns OP-1a, OP-1b, OP-2a, OP-2b und OP-Atom bei der SMPTE standardisiert. Die Standards der operational patterns OP-1c, OP-2c, OP-3a bis 3c werden gerade erarbeitet (Sohst, Lennart (2006).

Allgemeine Anforderungen an eine Langzeitarchivierung für Videosignale

Ziel eines Formates für die Langzeitarchivierung ist eine möglichst werkgetreue Aufzeichnung und nachträgliche Wiedergabe.

Das Kodierverfahren darf nicht von spezieller Hardware abhängig sein, deshalb kommt kein klassisches Videoformat für die Langzeitarchivierung in Betracht. Filebasierte Aufzeichnungen sind deshalb vorzuziehen, wenn auch eine größere Neigung zu Infizierung mit Viren, Trojanern etc. besteht. Reine softwarebasierte Kodierverfahren können langfristiger eingesetzt werden. Zu den Sicherheitsanforderungen an Langzeitarchive bezüglich Viren, Trojanern, Würmern etc., siehe Oermann, A. (2007).

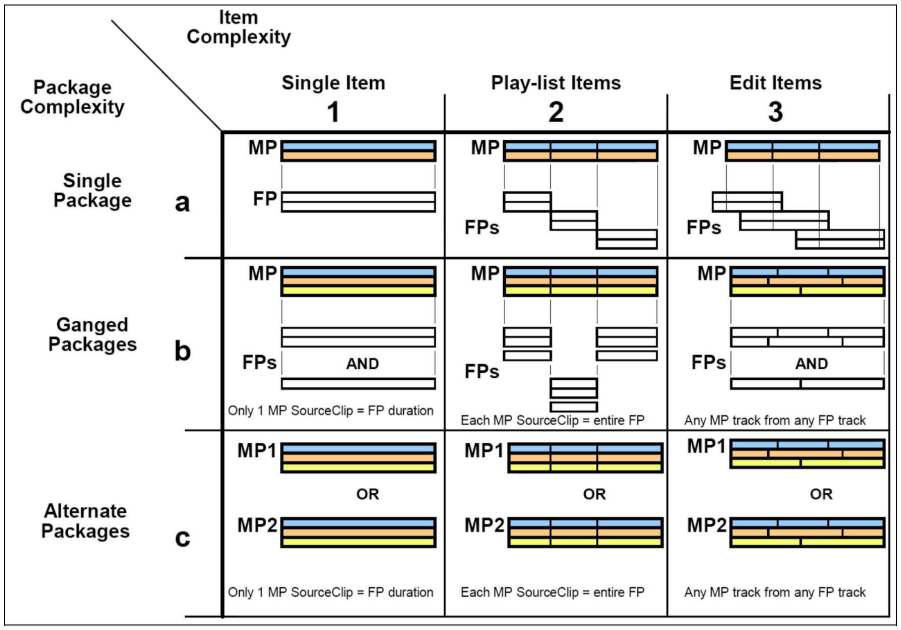


Abbildung 2: Achsen der operational pattems 1-c, 2a-c und 3a-c (FP = file package)(Sohst, Lennart (2006))

Metadaten

Angelpunkt für eine gute Langzeitarchivierung und Nutzung sind die erfassten Metadaten. Die Organisation und Speicherung erfordert genau angepasste Modelle. Das Broadcast Metadata Exchange Format (BMF) ist das Austauschformat mit dem einheitlichen Datenmodell für den Broadcastbereich; es beinhaltet mehrere Bestandteile (siehe Abbildung).

Das grundlegende Dokument zu BMF (Ebner, A. (2005)) beschreibt das zugrundegelegte Klassenmodell. Auf Basis dieses Klassenmodells wurde ein XML-Schema erstellt sowie die KLV-Codierung (Key-Length-Value). Diese wird durch eine in Arbeit befindliche Registrierung in den SMPTE-Dictionaries gewährleistet. Der zweite wesentliche Bestandteil umfasst die Spezifikation der Schnittstelle und deren Semantik.

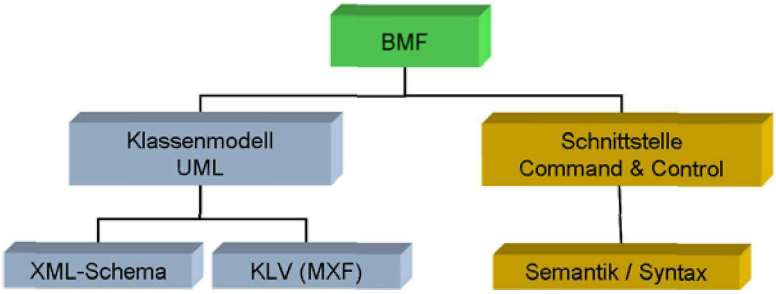


Abbildung. 3: Bestandteile des Broadcast Metadata Exchange Format – BMF

Das Klassenmodell von BMF, welches das geforderte einheitliche Datenmodell repräsentiert, dient einem geregelten und eindeutigen Austausch von Metadaten, es beschreibt nicht deren Speicherung.
Anforderungen, entwickelt aus den Analysen von Anwendungsfällen und Prozessen der öffentlich-rechtlichen Rundfunkanstalten, bilden die Grundlage von BMF. Die Analysen umfassen die gesamte Wertschöpfungskette dieser Rundfunkanstalten. Damit ist es für nahezu den gesamten Produktionsprozess anwendbar. Folgende Produktionsprozesse sind momentan berücksichtigt:

- Idee/Schemaplanung, Programmplanung, Sendeplanung, Sendungsplanung
- Herstellungsplanung, redaktionelle Arbeit (Produktkonzept)
- Akquise, Bearbeitung, Sendevorbereitung, Playout, Archivierung.

Zur Unterstützung dieser Prozesse sind die zur Abwicklung erforderlichen Informationen in BMF berücksichtigt. Eine zeitliche Abfolge der Prozesse wird durch BMF nicht vorgegeben, da diese sich zwischen den Rundfunkanstalten unterscheidet. Bei der Anwendung von BMF ist es jedoch nicht erforderlich für jeden Prozessschritt, bzw. Anwendungsfall das gesamte BMF anzuwenden, sondern nur den jeweils erforderlichen relevanten Anteil.

Bei der Realisierung von BMF sind mehrere Konzepte berücksichtigt worden:

- Konzept für den Austausch (nicht Speicherung)
- Redaktionelles Konzept zur Ausarbeitung und Herstellung des Produkts, welches in Redaktionsmanagement- und Produktionsplanungssystemen bereits angewendet wird
- Konzept von MXF zur Beschreibung von Essenz, das von Herstellern, die Essenz handhaben, bereits implementiert ist
- Konzept eines Schedulers zur Unterstützung der Distribution des Produkts, das in Sendeplanungssystemen und Automationen Anwendung findet
- Konzept der Stratifikation zur Unterstützung der Dokumentation, das in FESADneu/ARCHIMEDES (Archivdatenbanken in der ARD) bereits angewendet wird.

Als weiter Bestandteil von BMF wird die Schnittstelle definiert. Sie beschreibt die Semantik und die Syntax, also die Kommandos, welche über die Schnittstelle zur Verfügung stehen. Dies sind im Wesentlichen Anweisungen, was mit den ausgetauschten Metadaten im Zielsystem geschehen soll. Die Datenstruktur der Schnittstellen ist durch das Klassenmodell und das XML-Schema von BMF definiert.

Austauschformate bei der Wiederverwendung

Die meisten europäischen Rundfunkanstalten haben das Format 720 Zeilen mit je 1280 Pixel, progressive Abtastung mit 50 Bildern/s als Produktions- und Austauschformat vereinbart. Als Kompressionsformate werden heute die Produkte für den Mainstream verwendet. Formate mit möglichst geringer Kompression erleichtern die Wiederverwendung und eine nachträgliche weitere Postproduktion.

Tabelle der möglichen Austauschformate für den Mainstream heute

Sony XDCAM-HD 4:2:2	50 Mbit/s, MPEG-2 Long-GOP,
Panasonic AVC-Intra	100 Mbit/s und 50 Mbit/s, I-Frame

Thomson JPEG2000	100 Mbit/s, 75 Mbit/s und 50 Mbit/s, I-Frame
Avid DNxHD	175 Mbit/s und 115 Mbit/s, I-Frame

Qualitätssicherung beim Kopiervorgang

Da alle Trägermaterialen altern, ist das Kopieren nicht zu vermeiden. Der Kopiervorgang muss damit auf die schonendste Weise mit dem Inhalt und den Metadaten umgehen. Ein Klonen wäre die beste Art. Bei diesem Verfahren wird der Inhalt so dupliziert, dass sich das Original nicht von der Kopie unterscheidet. Ist dieses Verfahren nicht anwendbar, so sind die folgenden Regeln zu beachten: Bei der Langzeitarchivierung ist eine Reduzierung der Bitauflösung nicht zulässig, weil dies zu einer schlechteren Qualität führt. Der Wechsel des Kodierverfahrens ist nur dann zulässig, wenn das Ergebnis nicht verschlechtert wird. Transcodierung sollte also möglichst vermieden werden. Wird eine Transcodierung notwendig, so sollte möglichst, wenn vorhanden die höchste Bitrate verwendet werden. Weiterhin ist darauf zu achten, dass der Farbraum nicht verändert wird, d.h. keine Reduzierung der Auflösung z.B. 8-bit anstelle von 10-bit. *Kommpressionverfahren wie JPEG2000 sind für Video-Langzeitarchive besonders geeignet, da dies eine skalierte Speicherung verschiedener Auflösungen zulässt; so können die Format 4k/2k für Film, HDTV oder SDTV im selben File abgelegt werden.*

Mechanische Stabilität der Träger

Die Träger der Informationen sind bei der Langzeitspeicherung besonders kritisch auszuwählen, da sie möglichst stabil bleiben sollen, um ein ständiges Kopieren vermeiden zu können. Magnetische Träger sind dabei nur bedingt geeignet, da sich die Träger sehr oft verändern. Es treten z.B. Substanzen aus, die ein Verkleben der einzelnen Schichten fördern, oder die Magnetschicht löst sich ab. Mit Filmen hat man bessere Erfahrungen gemacht.

Haltbarkeit von Filmen (analoge Speicherung)

Die Farbstoffe von Farbfilmen bleichen mit der Zeit aus. Sie werden blasser und farbstichig. Schwarzweißfilme sind sehr lange haltbar, wenn sie sachgemäß entwickelt wurden. Ältere Farbmaterialien hatten nach 20 bis 25 Jahren einen deutlichen Farbstich. Heute sind sie bei richtiger Lagerung (Aufbewahrung in Kühlräumen bei ca. 0 bis -2 Grad Celsius nach Angaben der Hersteller 50 bis

100 Jahre haltbar ohne erkennbar an Qualität zu verlieren. Die Erfahrungen können diese optimistische Prognose leider nicht immer bestätigen.

Auf der Filmverpackung ist ein Datum aufgedruckt, bis zu dem der Film *unbelichtet* einwandfrei ist. Er darf nicht wärmer gelagert werden als angegeben, meist um 20° C. Je wärmer er gelagert wird, desto schneller verliert der Film seine garantierten Eigenschaften.

Am längsten halten Filme bei niedrigen Temperaturen. Lagern sie im Kühlschrank, haben sie noch lange nach dem Garantiedatum gute Qualität und zwar um so länger, je niedriger die Temperatur ist. Nach einer Kaltlagerung sollten die Filme so lange in ihrer wasserdichten Verpackung im Zimmer stehen, bis sie Zimmertemperatur haben (eine oder mehrere Stunden), sonst beschlagen sie mit Kondenswasser.

Belichtet man den Film, sollte er möglichst bald entwickelt werden, spätestens nach 4 Wochen, denn das latente Bild hält wesentlich kürzer als ein unbelichteter Film.

Bei zu langer Lagerung verringert sich die Lichtempfindlichkeit. Die Bilder sind dann unterbelichtet, zu dunkel und eventuell farbstichig, weil die einzelnen Farbschichten unterschiedlich auf die lange Lagerung reagieren.

Feuchtigkeit verkürzt die Haltbarkeit von Filmen. Originalverpackte Filme sind gut dagegen geschützt. Ist der Film in der Kamera, sollte man sie vor allem in Gegenden mit hoher Luftfeuchtigkeit (zum Beispiel Tropen) gut schützen.

Schädigende Gase verkürzen die Lebensdauer der Filme. Sind Filme von Pilzen oder Bakterien befallen, lassen sich die entstandenen Schäden nicht mehr beheben. Filme sollten kühl, chemisch neutral, dunkel und bei geringer Luftfeuchtigkeit um 40 % aufbewahrt werden.

Diese Zahlen in der Tabelle sind optimistisch und wurden leider bis heute nur sehr selten erreicht. Das Essigssäuresyndrom (Vinegar Syndrom) bildet sich immer aus und wird durch die Luftfeuchtigkeit und die Lagertemperatur beschleunigt. Diese autokatalytische Reaktion bewirkt, dass die freigesetzte Essigsäure den weiteren Zerfall fördert.

Temperatur °C

20%	1250	600	250	125	60	30	16
30%	900	400	200	90	45	25	12
40%	700	300	150	70	35	18	10
50%	500	250	100	50	25	14	7
60%	350	175	80	40	20	11	6
70%	250	125	60	30	16	9	5
80%	200	100	50	25	13	7	4

Abbildung 4: Haltbarkeit nach Kodak in Jahren (Reilly, J. M. (1993))

Digitale Speichermedien

Digitale Speicher unterscheiden sich von analogen Speichern durch eine geringere Fehleranfälligkeit, die digital gespeicherten Informationen sind in der Regel durch Fehlercodes geschützt. Eine detaillierte Aufstellung der digitalen Speichermedien findet sich in Anhang 3.

Rechtefragen bei unterschiedlichen Produktionsformaten

Die Vielfalt der Produktionsformate ist deshalb so groß, weil verschiedene Ausspielwege bedient werden. Diese verschiedenen Ausspielwege führen zu unterschiedlichen Rechtesituationen. So werden SD-Produktionen, obwohl sie aus HD-Produktionen hervorgingen, meistens rechtlich getrennt verwertet. Zusätzliche Online-Ausspielungen werden in der Regel in einem weiteren Vertrag bewertet und natürlich auch getrennt verrechnet.

Einzelne Bildrechte in den Produktionen werden für unterschiedliche Ausspielwege getrennt bewertet. Hier kann es zu teuren Nachbearbeitungen kommen, weil z.B. einzelne Bilder für eine CD-Ausspielung, für die das Recht nicht erworben wurde, aus der Essenz genommen werden müssen. Bei Internet-Veröffentlichungen gibt es für private Teile starke Einschränkungen. Auch hier muss nachgeschnitten werden. Für die Langzeitarchivierung sind aber nur vollständige Dokumente zu gebrauchen.

Viele Rundfunkanstalten sichern sich die Senderechte und weitere Einstellungen z.B. in Mediatheken durch individuelle Verträge.

Zu befürchten ist, dass sich die Zersplitterung durch immer feingliederige Rechtverwertungsformen fortsetzt. Für die Langzeitarchivierung ist das Recht am Content abschließend zu klären, d.h. bei der Übernahme müssen alle bekannten Verwertungswege im Vertrag erfasst sein.

Literaturhinweise:

Ebner, A. (2005): *Austausch von Metadaten – Broadcast Metadata Exchange Format, BMF,* Technischer Bericht Nr. B 193/2005 IRT-München 2005

Fößel, Siegfried (2009): *Videoakquisitionsformate im Überblick*. In: Fernseh- und Kinotechnik (FKT) 1-2/2009

Gebhard, Ch. / Müller-Voigt, G (2005); Technikreport: *aktuelle digitale Videoformate – im Dickicht der Formate*, www.film-tv-video

Höntsch, Ingo (2004): *MXF – Das Austauschformat für den Fernsehbetrieb*. In: Jahresbericht 2004. Institut für Rundfunktechnik München

Knör, Reinhard (2008): *Aufzeichnungs- und Kompressionsformate für die HD-Produktion*, HD-Symposium Bregenz 2008

Nufer, Christoph (2003): *Analyse von MXF Dateien hinsichtlich ihrer Leistungsmerkmale und Standardkonformität*, Diplomarbeit Fachhochschule Stuttgart, Hochschule der Medien

Oermann, Andrea / Lang, Andreas / Dittmann, Jana (2007): *Allgemeine Bedrohungen von Programmen mit Schadensfunktionen in digitalen Multimedia-Langzeitarchiven.* In: Viren, Trojaner, Würmer in schützenswerten Langzeitarchiven. Symposium im Institut für Rundfunktechnik 2007

Reilly, J. M. (1993): *IPI Storage Guide for Acetate Film*. Image Permance Institut, RIT, Rochester NY

Sohst, Lennart (2006): *Erfassung der MXF-Technologie und Erstellung einer Planungshilfe für MXF-Anwender*. Diplomarbeit FH Oldenburg Ostfriesland Wilhelmshafen Fachbereich Technik Studiengang Medientechnik

Thomas, Peter (2008): *Dateiformate für Archivierung und Programmaustausch.* In: Fernseh- und Kinotechnik (FKT 4/2008)

Anhang

- Anhang 1: Nomenklatur der Videoformate
- Anhang 2: Video-Aufzeichnungsformate
- Anhang 3: Digitale Speichermedien
- Anhang 4: Glossar

1. Nomenklatur der Videoformate

D-1	1986	D-1	SD	Transparent	IEC 61016
D-2	1988	D-2	SD	Composite	IEC 61179
D-3	1989	D-3	SD	Composite	IEC 61327
D-4		Jap. Unglücks-Nr			
D-5	1994	D-5	SD	Transparent	IEC 61835
D-6	1996	Voodoo	HD	Transparent	SMPTE 277M, 278M
D-7	1997/98	DVCPRO	SD	Compression DV-based I-frame only	IEC 62071
D-8		Res. Betacam SX			
D-9	1997	Digital S	SD	Compression MPEG-2 I-frame only	IEC 62156
D-10	2001	IMX	SD	Compression	IEC 62289, EBU D94
D-11	2000	HDCAM	HD	Compression	IEC 62356
D-12	2000	DVCPROHD	HD	Compression	IEC 62447
D-13		US/EU Unglücks-Nr.			
D-14	1994	D-5	SD	Transparent	IEC 61016
D-15	1998	HD-D5	HD	Compression	IEC 62330
D-16	2005	HDCAM-SR	HD	Compression	IEC 62141
DCT	1993	DCT	SD	Compression	---
Digital Betacam	1993	DigiBeta	SD	Compression	IEC 61904
Betacam SX	1997	Betacam SX	SD	Compression MPEG-2 Long GOP	---
DVCAM	2000	DVCAM	SD	Compression DV-based (I-frame only)	---

Tab. 1: Band-basierte Systeme (Knör, Reinhard (2008))

2. Video-Aufzeichnungsformate

XDCAM	Optical Disc	SD	Compression
XDCAM HD	Optical Disc	HD MPEG-2 Long GOP	Compression
P2	Solid state	SD	Compression
P2 HD	Solid state	HD MPEG-4/AVC I-Frame only	Compression
Editcam	HardDisk / Solid state	SD	Compression
Editcam HD	Harddisk/Solid State	HD DNxHD I-Frame only	Compression
Infinity	Harddisk Cartridge / Solid State	SD / HD JPEG2000 I-Frame only	Compression
Consumer HDV	Band bis DVD	SD / HD MPEG-2 Long GOP	Compression
Consumer HD AVC	DVD, Solid State, Harddisk	SD / HD MPEG-4/AVC Long GOP	Compression
XDCAM EX	Solid state	HD MPEG-2 Long GOP	Compression

Tab. 2: Bandlose-Formate (Knör, Reinhard (2008))

Übersicht über die Art der Kompression bei Videoformaten

Die relevanten Video-Magnetaufzeichnungsformate arbeiten fast ausschließlich (Ausnahme D-5 Format) mit implementierter, oft firmenspezifischer Videokompression:

- Digital Betacam — eigene Videokompression,
- D-5 — transparentes Format DSK270 entsprechend ITU-R BT.601
- DVCPRO — DV-based-Kompression mit 25 Mbit/s (SMPTE 314M)
- Betacam SX — eigene Videokompression
- DVCPRO50 — DV-based-Kompression mit 50 Mbit/s (SMPTE 314M)
- IMX — MPEG-2 422P@ML Kompression (EBU D94 und SMPTE 356M)

- DV-home DV Kompression mit Abtastraster 4:2:0 (DIN EN 61834-1,2,4)

Die gängigen Videoformate im Einzelnen

DV

SD-Videobandformat für die Aufzeichnung digitaler Ton- und Bilddaten auf ein ME-Metallband mit einer Breite von 6,35 mm (1/4“). Das DV-Format arbeitet mit einer Auflösung von 8 Bit und komprimiert die Daten mit dem Faktor 5:1. Die Komprimierung findet immer nur innerhalb eines Bildes statt (Intraframe).

Hierfür wird ein mathematisches Verfahren, die diskrete Cosinus-Transformation (DCT), eingesetzt. Mit Hilfe dieser DCT und weiteren Rechenoperationen werden die nicht relevanten Informationen innerhalb eines Bildes erkannt und dann gezielt weggelassen. Zudem werden die Helligkeits- und Farbanteile des Bildsignals nicht im Verhältnis 4:2:2, sondern im Verhältnis 4:2:0 verarbeitet.

DV arbeitet mit einer Videodatenrate von rund 25 Megabit pro Sekunde (Mbit/s) und zeichnet den Ton unkomprimiert im Zweikanal-Modus mit 16 Bit und 48 Kilohertz digital auf. Viele Geräte ermöglichen auch das Arbeiten mit vier Tonkanälen.

DVCPRO “6,35 mm - Schrägspurformat D-7“

Panasonic entwickelte das digitale SD-Videobandformat DVCPRO auf der Basis des DV-Formats. Die technischen Daten des DVCPRO-Standards sind also denen von DV sehr ähnlich.

Allerdings gibt es einige Unterschiede: DVCPRO arbeitet mit einer Spurbreite von 18 µm. Das Band läuft bei DVCPRO mit 33,8 mm/s, also fast doppelt so schnell wie bei DV. Weitere Unterschiede zu DV: als Abtastverhältnis wurde bei DVCPRO in der PAL-Version 4:1:1 festgelegt und nicht 4:2:0. Dadurch wollten die Entwickler vertikale Farbfehler reduzieren, die sich bei DV nach mehrfachem Kopieren als Farbschmieren bemerkbar machen können. DVCPRO arbeitet mit einer Intraframe-Kompression.

DVCPRO50 "6,35 mm - Schrägspurformat D-7"

Das digitale SD-Videobandformat DVCPRO50 ist eine weitere Evolutionsform des DV-Formats. Vom Grundkonzept entspricht dieses Panasonic-Format dem Format DVCPRO. Es unterscheidet sich hauptsächlich durch die 4:2:2-Signalverarbeitung und die Videodatenrate von 50 Mbit/s von DVCPRO und den anderen DV-Formaten, ist also für eine höhere Bildqualität konzipiert. DVCPRO50 arbeitet mit Intraframe-Kompression.

Um die höherwertige Signalverarbeitung und die höhere Datenrate zu erreichen, wird das Band im Vergleich zu DVCPRO mit doppelter Geschwindigkeit bewegt, die Laufzeit pro Kassette reduziert sich im Vergleich zu DVCPRO um die Hälfte.

DVCPRO50 zeichnet in der PAL-Ausführung 24 Spuren pro Bild auf, also doppelt so viele Schrägspuren wie DVCPRO. Beim Ton sieht DVCPRO50 vier anstelle von zwei digitalen Audiokanälen vor. Sie bieten eine Auflösung von 16 Bit/48 kHz.

Alle DVCPRO50-Geräte können auch DVCPRO-Aufnahmen abspielen. Wie DVCPRO wird auch DVCPRO50 in Deutschland bei einigen TV-Sendern eingesetzt (unter anderem ZDF, MDR, SWR). Dieses Format kommt mit seiner höheren Datenrate und der daraus resultierenden höheren Bildqualität bei »höherwertigen« SD-Produktionen zum Einsatz, also bei Dokumentationen und Produktionen, bei denen man glaubt, sie später mehrfach verwerten zu können.

Digital Betacam

SD-Videobandformat für die digitale Komponentenaufzeichnung mit 10 Bit Auflösung. Digital Betacam arbeitet mit einer Datenkompression von 2:1 (DCT-basierend). Es wird immer nur innerhalb eines Bildes komprimiert (Intraframe). Aufgezeichnet wird bei Digital Betacam auf Halbzoll-Reineisenmagnetband (12,7 mm breit). Digital Betacam wurde von Sony entwickelt und ist das älteste der aktuellen Digitalformate mit einer weiten Verbreitung im TV-Produktionsbereich.

IMX "12,65 mm (0,5 in) Schrägspurformat D-10"

MPEG-IMX

Sony wählte MPEG-IMX als Bezeichnung für ein SD-Format, bei dem Videobilder gemäß dem MPEG-Standard mit einer Videodatenrate von 50 Mbit/s komprimiert gespeichert werden. IMX, wie das Format üblicherweise genannt wird, wurde von den Normungsgremien unter der Bezeichnung D10 standardisiert. Bei MPEG-IMX wird immer nur innerhalb eines Bildes komprimiert (Intraframe).

Mit dem IMX-Datenformat arbeiten auch XDCAM-Geräte von Sony. Bandbasierte Geräte im IMX-Format nutzen Kassetten aus der Betacam-Familie. Recorder können so ausgelegt werden, dass sie auch Bänder der Formate Betacam, Betacam SP, Betacam SX und Digital Betacam abspielen können.

Bei disk-basierten Geräten wird IMX ebenfalls als Aufzeichnungsformat angeboten. Hierbei besteht die Möglichkeit, mit verschiedenen Video-Datenraten zu arbeiten, bei den XDCAM-Geräten sind 30, 40 und 50 Mbit/s wählbar.

XDCAM SD

XDCAM nennt Sony seine SD-Produktlinie auf Optical-Disc-Basis für den professionellen Markt. Speichermedium ist dabei die »Professional Disc«, die technisch mit der Blu-ray Disc verwandt, aber damit nicht kompatibel ist.

XDCAM-Geräte können DVCAM- und IMX-Daten aufzeichnen. Dabei werden natürlich die spezifischen Vorteile eines Disc-Speichermediums genutzt, um die Arbeitsabläufe bei der Bearbeitung zu verbessern.

Sony bietet derzeit zwei reine SD-Camcorder im XDCAM-Format an. Diese Geräte können ausschließlich Single-Layer-Discs verarbeiten.

DVCAM

Das digitale SD-Videobandformat DVCAM basiert auf dem DV-Format und wurde von Sony entwickelt. Sony variierte beim DVCAM-Format etliche DV-Parameter: Es blieb bei der 4:2:0 -Abtastung und der Datenreduktion mittels DCT, aber die Spurbreite wurde bei DVCAM auf 15 µm erhöht.

Das Band läuft entsprechend auch mit einer höheren Geschwindigkeit (28,2 mm/s), dadurch sind die Spielzeiten kürzer als bei DV. Sie betragen je nach Kassette 12, 64, 124, 164 oder 184 Minuten. Aufgezeichnet wird wie bei DV auf ME-Bänder. DVCAM-Geräte können auch DV-Kassetten abspielen.

Wie bei DVCPRO sollen auch bei DVCAM die Änderungen gegenüber DV das Format robuster und profitauglicher machen. Es wird immer nur innerhalb eines Bildes komprimiert (Intraframe).

HDV

Canon, Sharp, Sony und JVC haben gemeinsam die Basis-Spezifikationen dieses High-Definition-Formats für den Consumer-Bereich erarbeitet und verabschiedet. HDV basiert auf DV, dem weltweit immer noch am weitesten verbreiteten digitalen Videoproduktionsformat. HDV nutzt als Speichermedium normale DV-Videokassetten und kann mit Auflösungen von 1080i/25 oder 720p/50 Zeilen arbeiten. Damit die Bilder mit höherer Auflösung auf die DV-Kassette

passen, werden Video- und Audio-Signale bei der Aufzeichnung mittels MPEG komprimiert.

Für die Videosignale ist das eine Kodierung gemäß MPEG-2 jeweils über mehrere Bilder hinweg (Interframe Compression, Long-GoP), was die Aufzeichnung und Wiedergabe von HD-Video mit einer Datenrate erlaubt, die der des DV-Formats entspricht, wo aber jeweils nur innerhalb eines Bildes komprimiert wird (Intraframe Compression). Audiosignale werden bei HDV mit einer Abtastfrequenz von 48 kHz und 16-Bit-Quantisierung digitalisiert und gemäß MPEG-1 Audio Layer 2 komprimiert.

Zusammengefasst lässt sich sagen: HDV arbeitet mit der gleichen Videodatenrate wie DV, also mit maximal 25 Mbit/s. Durch die veränderte Form der Kompression erreicht aber HDV eine insgesamt höhere Bildqualität als DV, was die Zeilenzahl betrifft.

Wichtiger Unterschied zwischen HDV und DV: HDV arbeitet immer mit Breitbild (16:9), ist also auch vom Bildsensor her auf Breitbild optimiert, DV arbeitet dagegen originär mit 4:3 und kann nur mit Kompromissen als 16:9-Format betrieben werden.

AVCHD

AVCHD ist ein HD-Format für digitale Camcorder, das die Aufzeichnung von 1080i- und 720p-Signalen auf 8-cm-DVDs und auch auf SD-Speicherkarten normieren soll. Das Format nutzt die MPEG-4-basierte AVC/H.264-Kodierung zur Video-Kompression und Dolby Digital (AC-3) oder Linear-PCM für die Audio-Kodierung. AVCHD-Geräte bieten unterschiedliche Datenraten bis 24 Mbit/s, die bei den Geräten selbst meist nur als Qualitätsstufen mit Kürzeln wie LP, SP und HQ bezeichnet sind. Die maximale AVCHD-Systemdatenrate von 24 Mbit/s entspricht der Videodatenrate der 720p-Variante von HDV, aber das AVCH D-Verfahren gilt als moderner, effektiver und leistungsfähiger als MPEG-2. Allerdings ist der Codec auch so komplex, dass man derzeit in der Postproduktion beim Bearbeiten von AVCHD sehr schnell die Grenze der Echtzeitbearbeitung erreicht.

DVCPROHD

DVCPROHD entwickelte Panasonic zunächst als Bandformat auf der Basis von DVCPRO und somit indirekt aus dem Consumer-Format DV. Es wird immer nur innerhalb eines Bildes komprimiert (Intraframe).

DVCPROHD unterscheidet sich durch die 4:2:2-Signalverarbeitung und die Videodatenrate von 100 Mbit/s, sowie das Raster (1080i oder 720p Zeilen) von

den anderen DV-Formaten. Um die höherwertige Signalverarbeitung und die höhere Datenrate zu erreichen, wird das Band mit vierfacher Geschwindigkeit eines DVCPRO-Bandes bewegt, die Laufzeit pro Kassette reduziert sich im Vergleich zu DVCPRO auf ein Viertel.

DVCPROHD erfordert mit der Videodatenrate von 100 Mbit/s zwar eine höhere Kompression bei der HD-Aufzeichnung als HD-D5 und HDCAM, ermöglicht aber den Bau von preisgünstigeren und aufgrund der kleineren Kassette auch kompakteren Camcordern.

Um im DVCPROHD-Format bei bandbasierter Aufzeichnung längere maximale Spielzeiten pro Kassette realisieren zu können, hat Panasonic die Bandaufzeichnung in diesem Format auf zwei verschiedene Arten variiert.

DVCPROHD EX ist ein Extended-Format, bei dem mit einer Spurbreite von 9 µm aufgezeichnet wird, wodurch längere Spielzeiten von bis zu 124 Minuten auf einem einzigen Band möglich werden. Diese Intraframe-Aufzeichnungsart ist bei jeweils einem Camcorder- und einem Recorder-Modell in der aktuellen Produktpalette von Panasonic im Einsatz.

XDCAM HD

XDCAM HD nutzt das gleiche optische Speichermedium wie XDCAM (Professional Disc), zeichnet aber HD-Bilder in 1080i als MPEG-2 Long-GoP bei einstellbarer Bitrate von 35, 25 oder 18 Mbit/s auf (Interframe-Kompression). Dabei werden verschiedene Varianten der Quantisierung genutzt: bei 18 und 35 Mbit/s wird mit variabler, bei 25 Mbit/s mit konstanter Datenrate gearbeitet. Dadurch sind die Datenströme von XDCAM-HD-Aufnahmen mit 25 Mbit/s kompatibel zu denen von HDV. Von Aufnahmen mit 35 Mbit/s mit variabler Datenrate ist dagegen eine bessere Qualität zu erwarten als von HDV. XDCAM HD arbeitet mit einem Abtastverhältnis von 4:2:0 und einem Raster von 1440 x 1080 Bildpunkten.

XDCAM EX

XDCAM EX nennt Sony die HD-Aufzeichnung von Bild und Ton auf Festspeichermedien mit PC-Express-Abmessungen. Die Speicherkarten tragen die Bezeichnung SxS.

Aber nicht nur das Speichermedium unterscheidet XDCAM EX von XDCAM HD: XDCAM EX nutzt im Unterschied zu XDCAM HD nicht ausschließlich das reduzierte Raster von 1440 x 1080 Bildpunkten, sondern arbeitet in der höchsten einstellbaren Qualitätsstufe mit 1920 x 1080. Es bleibt aber bei XDCAM EX wie schon bei XDCAM HD beim Abtastverhältnis 4:2:0.

XDCAM HD 422

XDCAM HD 422 basiert auf XDCAM HD, arbeitet aber mit höheren Datenraten und nutzt, um auf vernünftige Spielzeiten zu kommen, als Speichermedium eine Dual-Layer-Disc (DL). Die höheren Datenrate von XDCAM HD 422 resultiert aus der 4:2:2-Signalverarbeitung, die das Format bietet. Und ebenfalls nicht ohne Einfluss auf die Datenrate: XDCAM HD 422 arbeitet nicht wie XDCAM HD mit dem reduzierten Raster von 1440 x 1080 Bildpunkten, sondern mit vollen 1920 x 1080. Gemeinsam ermöglichen diese Maßnahmen eine höhere Bildqualität.

HDCAM

Digitales Videoformat für die HD-Aufzeichnung mit 1920 x 1080 Bildpunkten im 16:9-Format. Aufgezeichnet wird auf ein 14 µm dickes Metallpartikelband mit extrem feinen Partikeln (0,125 µm Länge), das in ein Gehäuse aus der Betacam-Kassettenfamilie gespult ist. HDCAM zeichnet pro Bild 12 Spuren mit je 20 µm Breite auf. Das Bandlaufwerk entspricht weitgehend dem einer Digital-Betacam-Maschine. Da bei HD hohe Datenraten von rund 1,5 Gbit/s anfallen, können diese von HDCAM nicht direkt auf das Band geschrieben werden. Deshalb wird im Verhältnis 3:1:1 abgetastet und nach einer Vorfilterung (Reduzierung der horizontalen Auflösung auf 1440 Pixel mit 8-Bit Auflösung) und folgt dann eine Intraframe-DCT-Kompression von rund 4,4:1, so dass die Videodatenrate am Ende bei 185 Mbit/s liegt.

HDCAM SR

Wichtigster Unterschied zu HDCAM: HDCAM SR HQ zeichnet RGB-Daten im 4:4:4-Abtastverhältnis mit 880 Mbit/s auf. Dabei arbeitet HDCAM SR aber nicht unkomprimiert, sondern mit einer MPEG-4-basierten, relativ niedrigen Kompressionsrate von 4,2:1 bei 1080i-Betrieb. Weitere Formate sind 4:2:2 YUV bei einer Datenrate von 440 Mbit/s, was einem Kompressionsverhältnis von 2,3:1 entspricht. Das Gerät kann sowohl 720p50 als auch 1080i25 aufzeichnen und wiedergeben. HDCAM SR soll dort zum Einsatz kommen, wo die mit dem stärker komprimierenden HDCAM erreichbare Qualität nicht ausreicht, etwa bei Kinofilm-Produktionen, bei Special-Effects-Shots, die intensiv nachbearbeitet werden müssen, beim Film-Mastering, aber auch in der Archivierung.

P2HD, AVC-Intra

P2HD nutzt die exakt gleichen Speicherkarten wie P2, es ist im Grunde kein eigenständiges Format, sondern wird von Panasonic für bandlose HD-Geräte

benutzt, seit der Hersteller neben DVCPROHD mit AVC-Intra einen weiteren Codec eingeführt hat, um HD-Signale auf P2-Karten zu speichern. P2-AVC-Intra arbeitet effektiver, als der von Panasonic in HD-Camcordern ebenfalls genutzte DVCPROHD-Codec. Diese höhere Effektivität kann auf zwei Arten eingesetzt werden: Bei der gleichen Videodatenrate wie DVCPROHD (100 Mbit/s) erreicht man demnach mit AVC-Intra eine verbesserte Bildqualität und volle 4:2:2-Abtastung bei 10-Bit-Quantisierung. Alternativ kann mit der gleichen Bildqualität wie bei DVCPROHD gearbeitet werden, dann kommt AVC-Intra mit halber Videodatenrate aus (50 Mbit/s) und es passt doppelt so viel Bildmaterial auf die Speicherkarte wie mit DVCPROHD.

D5-HD

Das eigentlich für die unkomprimierte Aufzeichnung von Standard-Videosignalen konzipierte D5-Format kann auch zur Aufzeichnung von hoch aufgelösten HD-Signalen verwendet werden. Dann werden anstelle der unkomprimierten SD-Videosignale eben komprimierte HD-Videosignale auf das Band geschrieben. D5-HD bietet eine höhere Videodatenrate als das konkurrierende Sony-Format HDCAM und ermöglicht dadurch eine niedrigere Kompressionsrate. 235 Mbit/s bei D5-HD gegenüber 185 Mbit/s bei HDCAM können sich besonders in der Postproduktion und der Archivierung qualitätssteigernd bemerkbar machen. D5-HD-Maschinen können auch im SD-Format D5 aufnehmen.

Infinity, JPEG2000

Thomson Grass Valley hat mit Infinity ein Konzept für ein Videoproduktionssystem entwickelt, bei dem Speichermedium und Codecs weitgehend entkoppelt sind. Bislang gibt es nur ein Camcorder-Modell, das mit diesem Konzept arbeitet. Als Speichermedien können dabei CF-Karten oder Rev-Pro-Wechselfestplatten verwendet werden.

Der Camcorder arbeitet laut Hersteller intern immer im 1080i-Bildmodus mit 4:2:2-Abtastverhältnis bei 10 Bit Farbtiefe. Die maximal erreichbare Datenrate liegt bei 100 Mbit/s.

Als bevorzugten Codec nutzt Infinity JPEG2000, sowohl für HD-, wie für SD-Aufnahmen.

Redcode Raw

Die Kamera Red One des Herstellers Red Digital Cinema wird derzeit besonders von Independent-Filmern stark beachtet: Sie verspricht kinotaugliche Bil-

der mit 4K-Auflösung zu vergleichsweise moderaten Preisen. Die Red One kann auf CF-Speicherkarten oder auf Festplatte aufnehmen. Hierbei verwendet der Hersteller das eigene Kompressionsverfahren Redcode Raw. Dabei werden die Rohdaten des Bildsensors unter Einsatz eines Wavelet-Verfahrens mit variabler Bitrate komprimiert (ähnlich JPEG2000). Die maximale Datenrate kann dabei auf 224 Mbit/s oder auf 288 Mbit/s festgelegt werden, was einer Kompression von 12:1 und 9:1 entspricht, wenn man die Rohdatenrate zugrunde legt, die der Sensor abgibt. Aufgezeichnet werden dabei komprimierte Rohdaten, also keine RGB- oder Videosignale im engeren Sinn. Diese Daten müssen vor der Vorführung in jedem Fall bearbeitet und aufbereitet werden, was eher den Abläufen bei der Filmproduktion entspricht, als der klassischen Videoproduktion.

DNxHD

Der Coder für DNxHD wurde von der Firma Avid für ihre Schnittsysteme entwickelt. Er steht als 8-und auch als 10-Version zur Verfügung. In beiden Versionen ist das Abtastverhältnis 4:2:2. Die notwendige Datenrate liegt bei 175 Mbit/s für 720p/50 und 185 Mbit/s für 1080i/25 bzw. 1080p/25. Es stehen 8 Tonkanäle mit 24 bit 48 kHz zur Bearbeitung bereit. Die Kompression beträgt 5,5:1 bei 10-bit-Signalen.

3. Digitale Speichermedien

P2

P2 steht als Kürzel für Professional Plug-In Card, ein Speichermedium, das Panasonic speziell für den Einsatz in bandlosen Profi-Camcordern entwickelt hat.

Die P2-Speicherkarte ist ein Solid-State-Speichermedium, es gibt also keine bewegten Teile. Jede P2-Karte kombiniert vier SD-Card-Speicherchips in einem PCMCIA-Gehäuse, dadurch wird zumindest in der Theorie die vierfache Transfer- und Schreib-Datenrate erreicht, wie bei einer einzelnen SDHC-Karte: bis zu 640 Mbit/s Transferrate sind theoretisch möglich.

Eine 8-GB-Karte kann 36 Minuten DVCPRO-Material oder 8 Minuten DVCPROHD aufzeichnen.

Aktuell verfügbare P2-Geräte sind mit zwei bis fünf Karten-Slots ausgestattet, die Speicherkapazität soll in den kommenden Jahren weiterhin rasch ansteigen, der Kartenpreis rasch sinken.

SxS

Solid-State-Speicherkarten für XDCAM EX von Sony und SanDisk. Die SxS-Speicherkarten passen in PC-Express-Slots an Laptops und PCs, sowie in XDCAM-EX-Geräte von Sony.Die maximale, theoretische Übertragungsrate gibt Sony mit 800 Mbit/s an.

Professional Disc

Die Professional Disc (PD) hat Sony als Speichermedium für das XDCAM-Format entwickelt. Das optische Speichermedium ist technisch mit der Blu-ray Disc verwandt, aber damit inkompatibel.

In einem zweiten Schritt hatte Sony erstmals zur NAB2007 eine Professional Disc mit höherer Speicherkapazität vorgestellt. Diese 50-GB-Variante der Professional Disc bietet mehr als die doppelte Speicherkapazität gegenüber der zuerst eingeführten Single-Layer-Scheibe.

Die höhere Speicherkapazität wird mit einer zweiten Speicherschicht auf der Scheibe erreicht, einem Verfahren, das es auch bei der DVD und bei Blu-ray gibt. Die 50-GB-Scheibe ist also eine Dual-Layer-Disc (DL). Um diese beschreiben und lesen zu können, sind Laufwerke nötig, deren Schreib/Lese-Einheit die beiden Schichten getrennt beschreiben und auslesen kann. Vor der Einführung der Dual-Layer-Disc ausgelieferte XDCAM-Geräte können das nicht, in alle neuen und zukünftigen Modelle will Sony ausschließlich die neue Technik integrieren.

Die Dual-Layer-Scheibe erhöhte nicht nur die Kapazität von neuen und kommenden Geräten in den Formaten XDCAM und XDCAM HD, sondern ermöglichte auch ein weiteres Format:

CF-Card

CompactFlash-Speicherkarte, die ursprünglich im Fotobereich größere Verbreitung fand, sich später aber auch in anderen Bereichen etablieren konnte. Diesen Speicherkartentyp setzt unter anderem der Hersteller Red Digital Cinema als Speichermedium bei der Kamera Red One ein. Sony bietet für zwei seiner HDV-Camcorder den andockbaren CF-Card-Recorder HVR-MRC1 an, der wahlweise HDV-Files (.m2t) oder DV/DVCAM-Files (.avi/.dv) aufzeichnen kann.

Der CF-Kartentyp Ultra II ist für Datenraten von bis zu 80 Mbit/s ausgelegt und derzeit in einer Größe von maximal 16 GB erhältlich. Extreme III schafft Datenraten von bis zu 160 Mbit/s und ist ebenfalls mit einer Maximal-

kapazität von 16 GB im Handel. Extreme IV soll einen Datenstrom von bis zu 320 Mbit/s verarbeiten.

SD-Card

SD ist das Kürzel für Secure Digital, ein von SanDisk entwickeltes, kompaktes Speicherchip-System. SD-Karten sind kleiner und dünner als CF-Speicherkarten.

Die aktuell leistungsfähigste Version von SD-Speicherkarten sind SDHC-Karten. Panasonic nutzt SDHC-Karten in AVCHD-Camcordern. Die SDHC-Karten sind nach der Transferrate in Klassen unterteilt: Bei Klasse 2 sind das 16 Mbit/s, bei Klasse 4 32 Mbit/s. Um also in der maximalen AVCHD-Qualität von 18 Mbit/s aufnehmen zu können, reichen Klasse-2-SDHCs nicht aus.
Bei vielen Camcordern werden SD-Speicherkarten nicht als Träger der Bild- und Toninformation genutzt, sondern um Camcorder-Einstellungen zu speichern und zwischen Geräten austauschen zu können (Scene Files, Picture Profiles).

Memory Stick

Der von Sony entwickelte Memory-Stick kommt bei aktuellen Camcordern nur vor, um digitale Fotos oder Camcorder-Parameter (Picture Profiles) zu speichern.

GFPak

In Form der GF-Paks bietet Toshiba in Zusammenarbeit mit Ikegami eine weitere Variante von Festspeicher für Videoaufnahmen an. Dieses Speichermedium kommt im jüngsten bandlosen Camcorder von Ikegami zum Einsatz. GFPaks sind deutlich größer als SD- oder CF-Speicherkarten, bieten aber etwa eine integrierte Kapazitätsanzeige und sind mit zwei Schnittstellen ausgestattet, die im IT-Bereich weit verbreitet sind: SATA und USB 2.0 man benötigt also nicht unbedingt einen Reader oder einen Rechner mit speziellen Slots, um das Material von GFPaks kopieren und sichten zu können.

Rev Pro

Rev Pro ist eine Entwicklung von Iomega und Thomson Grass Valley für das Infinity-System. Dabei handelt es sich um spezielle Wechselfestplatten: Die einzelne Cartridge enthält einen Spindelmotor und eine magnetische Disk, alle anderen, teureren Komponenten, die eine normale Festplatte ausmachen, wie etwa Controller, Datenpuffer, Schreib- und Leseköpfe, sind nicht in der Wechseldisk, sondern im zugehörigen Laufwerk enthalten.

Thomson Grass Valley bietet drei Disks an: Die rot markierte Disk bietet eine Kapazität von 35 GB und wird von Thomson zum Netto-Listenpreis von 67,50 US-Dollar angeboten. Die neue goldene Rev Pro XP bietet 40 GB und erreicht eine höhere Schreib- und Leserate: Sie kann laut Hersteller zwei Datenströme mit bis zu 75 Mbit/s gleichzeitig schreiben oder wiedergeben. Der Netto-Listenpreis dieser schnelleren Disk liegt bei rund 70 US-Dollar. Die blaue Rev Pro ER ist auf größere Kapazität optimiert und erreicht laut Thomson 65 GB.

Die goldene XP-Disk kann bis zu 50 Minuten JPEG2000-HD-Material mit einer Datenrate von 75 Mbit/s oder mehr als 40 Minuten mit 100 Mbit/s aufzeichnen. DV-Material mit 25 Mbit/s kann laut Hersteller in sechsfacher Geschwindigkeit übertragen werden, selbst komplexere Postproduction-Aufgaben sollen sich damit direkt auf der Disk realisieren lassen.

Die blaue ER-Disk speichert rund 70 Minuten HD-Material mit einer Datenrate von 100 Mbit/s, oder 90 Minuten mit 75 Mbit/s.

FieldPaks

Dieses spezielle Wechselfestplattensystem nutzt Ikegami bei seinem Editcam-System. Am Camcorder können verschiedene Codecs eingestellt werden, die Daten werden dann im entsprechenden Format auf die FieldPaks geschrieben.

Festplatten, Diskrecorder: Focus Firestore, Sony

Über Schnittstellen wie IEEE-1394 oder USB 2.0 kann heute an viele Camcorder auch ein portabler Diskrecorder angeschlossen werden, der dann parallel oder alternativ zum eingebauten Laufwerk des Camcorders die Bild- und Tondaten speichert. Solche Diskrecorder gibt es von den Camcorder-Herstellern Sony, JVC und Panasonic, sowie von weiteren Anbietern, unter denen Focus Enhancements mit seiner Firestore-Familie zu den populärsten und erfolgreichsten zählt. Die Besonderheit der Firestores besteht darin, dass diese Geräte eine Vielzahl von Dateiformaten unterstützen und es erlauben, die Daten gleich so auf die Platte zu schreiben, dass das jeweils gewünschte Schnittsystem direkt auf die Dateien zugreifen kann.

4. Glossar

Abtastung - 4:2:2 Abtastung: Die beiden Farbsignale (Chrominanz) werden in jeder Zeile nur halb so häufig abgetastet wie das Schwarzweißsignal (Luminanz)
4:2:0 Abtastung: Die beiden Farbsignale (Chrominanz) werden in jeder zweiten Zeile gar nicht abgetastet („0“). In den anderen Zeilen geschieht das Abtasten nur halb so häufig („2“) wie das des Schwarzweißsignals („4“).
4:1:1 Abtastung: Von der Luminanz wird jedes Pixel aufgezeichnet, von der Chrominanz nur jedes vierte (bei DV).
3:1:1 Abtastung: Die beiden Farbsignale (Chrominanz) werden in jeder Zeile nur halb so häufig abgetastet wie das Schwarzweißsignal (Luminanz), insgesamt ist das Signal durch downsampling reduziert, z.B. statt 1920 nur 1440 Pixel je Zeile.
aliasing - (aliasing effect, aliasing error) Aliasing, Alias-Störung, Alias-Effekt, Rückfalt-Effekt: Allgemein könnte man Alias-Störungen als „Erzeugung falscher Signale durch Wahl ungünstiger Frequenzen“ bezeichnen. Beispiele: Generell können Alias-Störungen beim Digitalisieren analoger Daten auftreten: Im Fall einer Unterabtastung des Signals vor der A/D-Wandlung wird ein (falsches) niederfrequentes Signal (Alias) anstatt des korrekten Signals erzeugt. Zu den Alias-Effekten zählen auch Bildschirm-“Unschönheiten“ wie Treppenstufen bei schrägen Linien, die Erzeugung von „falschen“ Mosaikstrukturen bei der Darstellung feiner Muster sowie Crawling, die auf mangelnde Pixel-Auflösung zurückzuführen sind; auch das Rückwärtsdrehen von Rädern im Film zählt dazu.
Analoges Video - Ein von einer unendlichen Anzahl von gleichmäßig kleinen Abstufungen dargestelltes Videosignal zwischen gegebenen Videopegeln.
Aspect Ratio - TV-Bildseitenverhältnis. Beim Standard-TV verhalten sich Breite und Höhe zueinander im Verhältnis 4:3 bzw. 1,33:1, bei Breitbild-TV sind es 16:9 bzw. 1,78:1
Bandbreite - Bandbreite umschreibt in der Analogtechnik den Umfang eines Frequenzbereiches innerhalb eines Signalspektrums, das für die Übertragung eines Signals ohne größere Abweichung von den Nenndaten erforderlich ist. In der Informationstechnologie ist damit die Datenmenge gemeint, die sich pro Zeiteinheit über einen Kanalweg übertragen lässt.
Bildauflösung - Gibt die Zahl der Bildpunkte (Pixel), aus denen sich ein Monitorbild zusammensetzt, als Zahlenpaar an. Zum Beispiel 1920 (waagerechte) mal 1080 (senkrechte Pixelzahl).
BMF – Broadcast Metadata Exchange Format. Vom Institut für Rundfunktechnik auf der Basis von Prozessanalysen in den Rundfunkanstalten entwickelt.

Chroma - Begriff in der Fernsehtechnik für Farbsättigung, „Farbstärke“, wird aber auch für „Farbart“ Farbton plus Farbsättigung gebraucht.
Chroma - Chrominanz, C, Cr, Cb, U, V; Farbanteil des Videosignals. Komponentensignale enthalten ein Signal für die Differenz weiss-rot (Cr oder U) und weiss-blau (Cb oder V).
Chrominanz - Anteil des Videosignals, das die Farbinformationen in sich trägt, (Farbton und Sättigung, aber nicht die Helligkeit). In der Digitaltechnik stellt eine Matrix, ein Block oder ein einzelner Pixel den Farbunterschied dar, der sich auf die Hauptfarben R, G und B bezieht. Die für den Farbanteile verwendeten Bezeichnungen lauten Cr und Cb. Siehe auch YCbCr.
Codec - Kunstwort aus Compression (Kompression, Verdichtung) und Decompression (Dekompression, Wiederaufblasen). Der Begriff kann für einen Software-Algorythmus oder einen Hardware-Chipsatz verwendet werden. Soft- oder Hardware, welche speziell dafür entwickelt wurde, Videos nach bestimmten Kompressionsalgorythmen umzurechnen. Der Compressor verkleinert eine Datei, um sie besser speichern oder übertragen zu können, der Decompressor rechnet die kodierte Datei zur Darstellung in Echtzeit temporär um, ohne diese jedoch zu speichern. Decompressoren werden auch zum Rendern benötigt, da Pixelinformationen nicht in komprimiertes Videomaterial eingerechnet werden können.
Datenrate - Video; Sie entscheidet über die Bildqualität digitaler Fernsehprogramme und wird in Megabit pro Sekunde (Mbit/s) angegeben. Datenraten von 5 bis 6 Mbit/s entsprechen dabei einer Bildqualität, wie sie analoge Fernsehprogramme liefern. Für HDTV werden 12 bis 16 Mbit/s veranschlagt.
DCT - Abkürzung für Discrete Cosine Transform - Diskrete Kosinustransformation. Eine Kompressionsmethode (insbesondere der Bildschirmdaten) aus dem Orts- in den Frequenzbereich, mit der Daten digitalisiert werden. Verbreitete Methode zur Datenkompression von digitalen Videobildern, die durch die Auflösung von Bildblöcken (normalerweise 8x8 Pixel) in Frequenzen, Amplituden und Farben erreicht wird. Dabei erfolgt erst eine Intraframe Kodierung und dann eine Bild-zu-Bild Kodierung. Somit bestehen die erzeugten Daten aus den Informationen des ersten Bildes und danach nur noch aus den Unterschieden von einem Bild zum nächsten. Siehe auch verlustbehaftete Kompression.
EBU - European Broadcast Union
Essenz – Bezeichnung für den Inhalt einer Produktion. Zu unterscheiden vom Content= Essenz und Rechte an dieser
Farbmodelle - Videokameras zeichnen das Bild in drei Farbauszügen Rot, Grün und Blau (RGB) auf. Da das menschliche Auge empfindlicher auf die Helligkeit ist als auf die Farbe, wird das Signal umgerechnet in einen Helligkeitsanteil (Lu-

minanz, Y) und einen Farbanteil (Chrominanz) mit der Rot- und der Blaudifferenz (Cb, Cr). Digitales Video mit 8 bit Auflösung, erlaubt die Werte 16-235 für Luminanz und 0-224 für Chrominanz. Von der Luminanz wird jedes Pixel aufgezeichnet, von der Chrominanz nur jedes zweite (4:2:2 Abtastung bei D1 und Digibeta) oder gar nur jedes vierte (4:1:1 Abtastung bei DV).

Flash Memory - Ein Speicherbaustein, der auch nach Abschalten des Systems die auf ihm gespeicherten Daten dauerhaft behält.

GOP - Abkürzung für Group of Pictures, im System MPEG gebräuchlich zur Definition einer zusammengehörigen Bildergruppe. Nach MPEG-1 und MPEG-2: „Gruppe von Bildern" im hierarchischen Datenstrom, zwischen „Bild" und „Sequenz". Am Anfang dieser Gruppe steht immer ein „Intraframe"-codiertes (Stütz-)Bild (I-Bild, intraframe coding). Darauf folgen P- und B-Bilder (uni-irektional und bi-direktional codierte Bilder).

HDTV - Abkürzung für High Definition TeleVision. Hochauflösendes Fernsehen, Fernsehen in Kinoqualität mit besonders hoher Bild- und Tonqualität: Breitbild 16:9.

HD - Abkürzung für High Definition

Interlace - Synonym für Zwischensprung, Zwischenzeile, Zeilensprung-Verfahren.

International Organization for Standardisation - (ISO) Eine weltweite Vereinigung nationaler Normungsinstitutionen, die internationale Standards erarbeitet. Diese Standards werden von speziellen technischen Ausschüssen, die jeweils für eine bestimmte Norm zuständig sind, zunächst in Form von Entwürfen vorgelegt.

interframe coding - Zwischenbild-Codierung: Kompressionsverfahren, bei dem unter Ausnutzung der zwischen aufeinanderfolgenden Bildern bestehenden Redundanzen lediglich die Unterschiede codiert werden.

intraframe coding - englisch für „Innenbild-Codierung": Kompressions-Codierung unter Ausnutzung der zwischen den Punkten eines Vollbildes bestehenden Redundanzen (jedes Bild wird für sich allein codiert, I-Bild).

ITU - Abkürzung für International Telecommunications Union (UIT). Die International Telecommunications Union ist eine zivile Organisation, die etwa 175 Mitglieds- und Beobachter-Staaten vereinigt und auf eine weltweit standardisierte Telekommunikation hinarbeitet. Der Sitz der ITU ist Genf, zwei bekannte Untergremien der Union sind die CCIR und CCITT.

JPEG - Abkürzung für Joint Photographic Experts Group. Bezeichnung für einen Standard zur nicht verlustfreien Kompression von Bilddaten. Ein ISO/ITU-Standard für das Speichern von Farb- und Schwarzweißbildern in einem komprimierten Format über die diskrete Kosinustransformation (DCT). Es ist

ein kompaktes Bildformat mit variabler Kompressionsrate, das als Alternative zu GIF entwickelt wurde, aber die Anzahl der Farben nicht wie GIF auf 256 reduziert und nutzt die Schwächen des menschlichen Auges aus, indem es Informationen spart, wo das Auge sie nicht bemerkt. Der Komprimierungsgrad lässt sich wählen; je höher er allerdings ist, desto geringer wird die Bildqualität. Ein Komprimierungsverhältnis von 100:1 bedeutet einen erheblichen Verlust und ein Verhältnis von ungefähr 20:1 zeigt nur einen unerheblichen Verlust. Je höher die Kompression, desto kleiner die Datei; allerdings gehen auch mehr Detail-Informationen verloren. Dieser Effekt macht sich als Treppeneffekt an Linien bemerkbar. Dieses Dateiformat wird in vielen Digitalkameras als Speicherformat eingesetzt, um mehr Aufnahmen auf einem Speichermedium unterzubringen. JPEG ist gleichzeitig die Kurzbezeichnung für eine Grafik, die als Datei im JPEG-Format gespeichert wurde. Mit JPEG komprimierte Bilddateien sind an der Endung jpg zu erkennen. Obwohl JPEG Grafiken eines beliebigen Farbraumes codieren kann, werden die besten Kompressionsraten bei Verwendung eines Farbraumes wie Lab erzielt, bei dem jedes Pixel sich aus einer Helligkeits- und zwei Farbkomponenten zusammensetzt. Neben dem „normalen" JPEG-Standard gibt es nun auch JPEG-LS [ISO/IEC 14495-1] zur verlustfreien bzw. nahezu verlustfreien Kompression fotorealistischer Bilder.
Kompression - Reduzierung der File-Größe durch entsprechende Kompressionsalgorithmen. Man kann hier prinzipiell zwischen verlustfreien und nicht verlustfreien Algorithmen unterscheiden.
Mainstream – technische Produktionsform in den Rundfunkanstalten, unter der die alltägliche Arbeit wie Nachrichten, Magazine, Telenovas verstanden wird. Nicht die höchste Stufe der technischen Qualität.
MAZ – Magnetische AufZeichnung von Bild und Ton
Material eXchange Format (MXF) – „Es ist ein Hüllformat, auch Wrapper Format oder Containerformat genannt, welches ein oder mehrere Essenzen (auch *pay load*) in sich kapselt und akkurat beschreibt. Diese Essenzen können Bilder, Ton oder auch Daten sein. Eine MXF-Datei enthält genug Informationen, um zwei Anwendungen den Austausch von Essenzen zu ermöglichen, ohne vorher Informationen ausgetauscht zu haben. Dazu enthält sie so genannte Metadaten, die z.B. Informationen über die Länge der Datei, verwendete Codecs (Kompressionsverfahren) und Timeline-Komplexität bereitstellen.
Im Unterschied zu den bandbasierten Videoformaten (MAZ-Technik) soll die MXF-Definition den dateibasierten Umgang mit professionellen Videoformaten vereinfachen. Durch Standardisierung soll der Weg zum IT-basierten nonlinearen Videoschnitt (NLE) beschleunigt werden, ohne durch gemischte und herstellerspezifische (proprietäre) Datenformate behindert zu werden.

Der Standard wurde vom SMPTE, von der European Broadcasting Union (EBU) und von der Advanced Authoring Format (AAF)-Association vorangetrieben und im Jahre 2003 unter der Normbezeichnung SMPTE 377M verabschiedet. Das Dateiformat ist als ISO-Standard vorgeschlagen." (Quelle. http://www.pro-mpeg.org; nach: http://de.wikipedia.org/wiki/MXF, zitiert in: Höntsch 2004)

MAZ – Magnetische AufZeichnung von Bild und Ton

MPEG - Abkürzung für Moving Pictures Experts Group. Normungsausschuss für Datenkompressionsverfahren bei bewegten Bildern.

MPEG-1 - Bei Audio: Kompressionsstandard für Multimedia-Anwendungen bis zu einer Datenrate von 1,5 MBit/s. System zur datenreduzierten Codierung von bis zu 2 Kanälen.

MPEG-1 - Bei Video: Kompressionsstandard für Multimedia-Anwendungen bis zu einer Datenrate von 1,5 Mbit/s. System zur datenreduzierten Codierung mit niedriger Bildqualität. Verwendet bei CD-I und Video-CD.

MPEG-2 - Bei Audio: MPEG2 Mehrkanal Audio ist, neben Dolby Digital, eines der digitalen Surround-Systeme, die bei DVD eingesetzt werden. System zur datenreduzierten Codierung von bis zu 7+1 Kanälen.

MPEG-2 - Bei Video: MPEG2 ist eine erweiterte Version des MPEG1-Standards bis zu einer Datenrate von 100 MBit/s (Gesamtdatenrate), der bereits für Video CD-Aufzeichnungen eingesetzt wird. MEPG2 wurde 1994 als universelles Video-Kompressionssystem für Übertragung, Kommunikation und Speicherung auf optischen Datenträgern eingeführt. System zur datenreduzierten Codierung mit hoher Bildqualität. Verwendet bei DVD.

MPEG-4 - bietet höhere Bildqualität bei niedrigeren Datenraten und die Möglichkeit der Bildskalierung und der Manipulation. H.264 ist eine Variante für eine bessere Kodierung bei gleichzeitiger Reduzierung der Datenrate.

MXF - Material eXchange Format. Von der SMPTE definierter Austauschcontainer (s.o.)

Operational Patterns- MXF-Container können Daten fast beliebiger Komplexität beinhalten. *Operational Patterns* (OPs) definieren die Komplexität eines MXF-Files. Dies erfolgt über Einschränkungen der Anzahl und Struktur der verwendeten Packages bzw. Clips. Vier dieser neun allgemeinen OPs *(Generalized Operational Patterns)* werden bereits in entsprechenden SMPTE Standards definiert (OP 1a, 1b, 2a, 2b). Darüber hinaus existieren spezielle OPs *(Specialized Operational Patterns),* die eigene Grenzen für den Grad der Komplexität definieren. *„OP Atom"* ist beispielsweise für den Transport und die Speicherung eines einzigen Essence Elementes konzipiert. Zudem gibt es OP für die Codierung von Audio-only Files mit der Absicht, sie für

nicht-MXF-Decoder lesbar zu machen.
Nufer, Christoph (2003), Sohst, Lennart (2006)
Normierung der einzelnen Operational Pattern:

- OP1a: SMPTE 378M
- OP1b: SMPTE 391M
- OP1c: SMPTE 408M
- OP2a: SMPTE 392M
- OP2b: SMPTE 393M
- OP2c: SMPTE 408M
- OP3a: SMPTE 407M
- OP3b: SMPTE 407M
- OP3c: SMPTE 408M
- OP-Atom: SMPTE 390M

PAL - Abkürzung für Phase Alternate (oder Alternation) Line. „Phasenumschaltung von Zeile zu Zeile“: von Prof. Bruch (Telefunken) entwickeltes analoges Farbcodier-Verfahren (für die Farbfernseh-Übertragung), das vor allem in Westeuropa (außer Frankreich), Australien und in einigen anderen Regionen der Erde verbreitet ist: Es handelt sich um ein Interlaced-Signal mit 15,5 kHz Horizontalfrequenz, 50 Hz Bildwechselfrequenz, 625 Zeilen, davon 576 sichtbar, Farbdarstellung mit YUV. Die Chrominanz-Information wird (wie bei NTSC) mit QAM (Quadratur Amplituden Modulation) im Frequenzbereich des Luminanzsignals übertragen (FBAS-Signal). Im Unterschied zu NTSC wird aber die Polarität der Chrominanz-V-Komponente (R-Y) zeilenweise umgeschaltet. Auf diese Weise werden Phasenfehler (und dadurch bedingte Farbfehler) weitgehend eliminiert. Der PAL Standard ist definiert nicht mit NTSC kompatibel.
RGB Farbraum - Im RGB Farbraum setzt sich jedes sichtbare Pixel aus den drei Komponenten R(ot), G(rün) und B(lau) zusammen. Will man eine naturgetreue Farbwiedergabe am Computer erreichen, so muss jede dieser Komponenten mindestens 256 Ausprägungen haben. Dies entspricht genau einem Byte Speicherplatz pro Farbkomponente. Für ein einziges vollständiges Videobild benötigt man daher 768 Pixel x 576 Pixel x 3 Byte = 1327104 Byte. Dies entspricht ungefähr 1,2 MB pro Bild. Will man also eine Sekunde Video im RGB Farbraum darstellen, benötigt man ca 31,6 MB Speicherplatz. Eine 2 Gigabyte Festplatte hätte bei diesem Verfahren eine Videokapazität von ungefähr einer Minute. Abgesehen davon, dass es (noch) keine Festplatte gibt, die diese Datenmengen in Echtzeit übertragen könnte, gibt es Möglichkeiten die Datenmenge des Videosignals durch Transformation in einen anderen Farbraum (meist

YUV) und durch Komprimierung (meist MJPEG) stark zu reduzieren.
SD - Abkürzung für Standard Definition
SDTV – Abkürzung für Standard Definition Television. Mit der Abkürzung wird auch das heutige PAL-Fernsehen bezeichnet.
SMPTE – Abkürzung für Society of Motion Picture and Television Engineers. Internationale berufsständische Organisation (gegründet in den USA), die Arbeits- und Normenvorschläge erarbeitet.
Transcoder - Transcoder, Normenwandler, Umkodierer.
Verlustbehaftete Kompression - Eine Methode zur Reduzierung von Bild-Dateigrößen, bei der Bildpunkte in einem Feld zusammengefasst und einander angeglichen werden. Dadurch gehen Detailinformationen unwiederbringlich verloren. Die Qualität eines Bildes nach der Dekompression ist schlechter wie das einer nicht komprimierten Datei. JPEG verwendet eine solche Kompressionsmethode. Das Ausmaß der Veränderung hängt von der gewählten Kompression ab, je höher die Kompression, desto stärker die Veränderungen.
Trimediale Produktion – bei der Produktion wird bereits auf die unterschiedlichen Ausspielwege Radio, Fernsehen, Internet geachtet
Verlustfreie Kompression - Reduzierung von Bild-Dateigrößen durch Zusammenfassung gleicher Pixel. Die Dekompression liefert das qualitativ gleiche Ergebnis wie eine nicht komprimierte Datei. Es gehen hierbei keine Detailinformationen verloren. Nach der Dekompression sind die Daten völlig wiederhergestellt. Verlustfreie Kompression ist sehr viel uneffektiver als verlustbehaftete Kompression, d.h. es werden nur sehr geringe Kompressionsraten erreicht (Faktor 2 bis 3). Beispiele: GIF, TIFF.
Wavelet-Codierung - Die Waveletcodierung ist ein entwickeltes Kompressions-/Dekompressionsverfahren zur Reduzierung der Datenmengen bei digitalen Fotos und Video, ein so genannter Codec, bei dem bestimmte Wellenmuster genutzt werden und das der zunehmenden digitalen Datenflut Rechnung trägt und daher Kompressionsverfahren wie JPEG überlegen ist. Bei der Waveletcodierung wird das Signal in Teilbänder zerlegt, die komprimiert, gespeichert bzw. übertragen und am Schluss wieder zusammengesetzt werden. Die Codierung erfolgt in drei Schritten, wobei der erste in einer diskreten Wavelet-Transformation besteht (ähnlich wie bei JPEG, wo allerdings eine diskrete Cosinustransformation benutzt wird; außerdem besteht die JPEG-Codierung aus fünf Schritten).
XDCAM - von Sony entwickeltes Aufzeichnungsformat
XML - Abkürzung für Extensible Markup Language. Neue Standardsprache für Web-Publishing und Dokumenten-Management in Client/Server-Umgebungen, welche es dem Entwickler ermöglicht, dynamische und animierte, also

sich verändernde Webseiten zu erstellen. Das alte HTML wird in XML integriert. Mit XML wird das Web damit flexibler und einfacher: Umlaute brauchen nicht mehr als ue oder ü geschrieben zu werden. Aktive Elemente benötigen nicht mehr extra Javascript oder Java Applets. Die Strukturierung einer Website wird übersichtlicher, die Suche nach einzelnen Begriffen im XML-Web schneller und effizienter.

Y/R-Y/B-Y - Abkürzung für Componenten Signale. Statt den Farbdaten (Rot, Grün, Blau) verwendet man in der Videotechnik Schwarzweiß (Luminanz) und Farbsignal (Chroma). Das Auge ist für Chrominaz Daten weniger empfindlich, sie lassen sich durch diese Aufspaltung gezielt schlechter (und damit sparsamer) übertragen - ein wesentlicher Trick in Fernsehnormen wie PAL und NTSC. Typische Chrominanzsignale sind einerseits U (Rot-Cyan Balance) und V (Gelb-Blau Balance) sowie andererseits I (Cyan-Orange-Balance) und Q (Magenty-Grün Balance). Durch Hinzunahme des Luminanzsignals Y enstehen die YIQ und YUV Farbsysteme. Zur Digitalisierung dient normalerweise (CCIR-601) Y R-Y B-Y, was bis auf Skalierung identisch mit YUV ist. Die Chrominanzsignale bestehen hier aus der Differenz von Y und Rot beziehungsweise Blau.

YCrCb - YCrCb bezeichnet einerseits die analogen Quellensignale (Komponenten) für die Übertragung nach einem MAC-Verfahren oder für die Abtastung nach CCIR 601-Empfehlung (Digitalisierung), andererseits auch die bereits digitalisierten Komponenten. Y Luminanz, Cr R-Y, Cb B-Y. (Anm.: Oftmals werden mit YCrCb auch nur die digitalen Komponentensignale bezeichnet.)

Zeilensprungverfahren - Darstellungsart bei Monitoren, bei der der Elektronenstrahl pro Bilddurchlauf nur jede zweite Zeile beschreibt. Abwechselnd werden so einmal die geraden und danach die ungeraden Zeilen abgetastet Dieser Modus wird verwendet, wenn die Grafikkarte bei hohen Auflösungen nicht mehr in der Lage ist, eine ausreichende Bildwiederholrate zu erzeugen. Bei modernen Monitoren mit kurz oder mittellang nachleuchtendem Phosphor ist diese Darstellungsart nicht flimmerfrei und somit auch nicht ergonomisch einsetzbar.

17.6 Audio

Winfried Bergmeyer

Einerseits werden Tondokumente auf alten und gefährdeten Speichertechnologien, wie Tonwalzen oder Tonbändern, digitalisiert um sie langfristig zu erhalten, andererseits werden durch die neuen Möglichkeiten der Informationstechnologie große Mengen digitaler Audiodateien von Privatpersonen über das Internet verbreitet. Neben der technischen Aufgabe archivierte Digitalaufnahmen in Bibliotheken, Archiven und Museen zu erhalten, zwingt der Umfang der aktuellen Produktion von Tondokumenten aller Art Konzepte zur Auswahl der zu erhaltenen Dokumente zu entwickeln.

Die Langzeitarchivierung von digitalen Audiodaten ist eine Herausforderung für Bibliotheken, Archive und Museen. Ob Sprachaufnahmen, Konzerte, Tierstimmen oder Geräusche, die Variabilität der Inhalte ist groß. Das Ziel der Langzeitarchivierung ist der Erhalt der akustischen Inhalte in der vorhandenen Qualität, die Sicherung der Nutzbarkeit und die Bewahrung der zugehörenden Informationen.

Die für die Speicherung auditiven Contents verwendeten Medien unterlagen in den letzten 100 Jahren einem permanenten Wandel und tun dies weiterhin. Ersten Aufzeichnungen auf Tonwalzen folgten Schellack- und Vinyl-Platten, daneben entwickelten sich die wieder beschreibbaren Medien wie Tonbänder und Kassetten unterschiedlicher Formate. Die Revolution der digitalen Aufzeichnung und ihrer Wiedergabe bediente sich ebenfalls unterschiedlicher Speichermedien wie Kassetten, CDs, Minidiscs und DVDs. Im Gegensatz zu analogen Technologien sind allerdings digitale Informationen nicht an ein bestimmtes Speichermedium gebunden. Spätestens seit der Verbreitung von Musik und Hörbuchern durch Internetportale ist diese Abhängigkeit verschwunden.

Mit diesem Medien- und Formatspektrum sowie den z. T. umfangreichen Datenmengen wird die Langzeitarchivierung zu einer technologischen Herausforderung. Stehen wir bei den analogen und digitalen Speichermedien vor dem Problem der physischen Zerstörung und der selten werdenden medienspezifischen Abspielgeräte, so muss man bei digitalen Daten zusätzlich den Dateiformaten eine besondere Beachtung schenken.

Digitalisierung analoger Aufnahmen für eine dauerhafte Bewahrung

Eine Speicherung auf einem Medium gleichen Typs ist bei vielen Technologien heute kaum mehr möglich, da die Medien und die Aufnahme- und Abspielgeräte kaum noch zur Verfügung stehen werden. Audio-Material auf älteren Tonträgern wie Walzen oder Schellackplatten wurden daher vor dem digitalen Zeitalter zur Archivierung auf Tonbänder aufgenommen. Diese für die dauerhafte Konservierung gedachten Tonbänder sind aber mehreren Verfallsmechanismen ausgeliefert (Entmagnetisierung, Ablösung der Trägerschichten, Sprödigkeit, Feuchtigkeitsbefall etc.) und damit stark gefährdet. Zudem gibt es weltweit zur Zeit (2009) nur noch zwei Produzenten dieser Bänder und nur noch wenige Hersteller von Abspielgeräten. Die Zukunft der Konservierung von Audio-Objekten ist die Übertragung in digitale Computerdaten. Digitale audiovisuelle Archive, wie sie von Rundfunk- und Fernsehanstalten geführt werden, sind heute so organisiert, dass sie das gesicherte Material in definierten Zeitabständen in einem neuen und damit aktuellen Format sichern. Sogenannte *DMSS* (Digital-Mass-Storage-Systems) beinhalten Sicherheitsmechanismen, die die Datenintegrität bei der Migration sicherstellen.

Zur Digitalisierung analogen Materials benötigt man einen Analog-to-Digital-Converter (ADC), der in einfachster Form bereits in jedem handelsüblichen PC in Form der Soundkarte vorhanden ist. Professionelle Anbieter von Digitalisierungsmassnahmen verfügen allerdings über technisch anspruchsvollere Anlagen, so dass hier ein besseres Ergebnis zu erwarten ist. Es gibt mittlerweile zahlreiche Anbieter, die auch spezielle Aufnahmegeräte für die einzelnen Technologien bereitstellen, so z.B. für die Digitalisierung von Tonwalzen-Aufnahmen.

Die Qualität der Digitalisierung vorhandener analoger Objekte ist neben der Qualität des technischen Equipments von der Abtastrate und der Abtasttiefe abhängig. Erstere bestimmt die Wiederholungsfrequenz, in der ein analoges Signal abgetastet wird, letztere die Detailliertheit der aufgezeichneten Informationen. Wurde lange Zeit CD-Qualität (Red Book, 44.1 kHz, 16 bit) als adäquate Archivqualität angesehen, so ist mit der technischen Entwicklung heute Audio DVD-Qualität (bis zu 192 kHz und 24 bit) im Gebrauch. Hier sind zukünftige Weiterentwicklungen zu erwarten und bei der Langzeitarchivierung zu berücksichtigen. Auf Datenkompression, die von vielen Dateiformaten unterstützt wird, sollte verzichtet werden, da es um das möglichst originäre Klangbild geht. PCM (Pulse-Code-Modulation) hat sich als Standardformat für den unkomprimierten Datenstrom etabliert. Eine Nachbearbeitung (Denoising und andere Verfahren) zur klanglichen Verbesserung des Originals ist nicht vorzunehmen,

da sonst das originäre Klangbild verändert würde. Eine Fehlerkorrektur ist hingegen zulässig, da es bestimmte, durch die Aufnahmetechnik bedingte, Fehlerpotentiale gibt, deren Korrektur dem Erhalt des originären Klangs dient. Bei „Born digital"-Audiodaten ist allerdings abzuwägen, ob das originale Dateiformat erhalten werden kann oder ob auf Grund der drohenden Obsoleszenz eine Format- und Medienmigration vorzunehmen ist.

Maßnahmen zur Langzeitarchivierung

Die permanente Weiterentwicklung von Aufnahme- und Abspielgeräten sowie die Entwicklung der für die Verfügbarkeit, vor allem über das Internet oder für Mobilgeräte, verwendeten Dateiformate erfordert eine dauerhafte Überwachung des Technologiemarktes. Datenmigration in neue Datenformate und Speichermedien wird deshalb zum grundlegenden Konzept der Langzeitarchivierung gehören müssen. Musikarchive, die sich die Archivierung von kommerziell vertriebenen Audio-CDs zur Aufgabe gemacht haben, stellen mittlerweile bereits erste Verluste durch Zersetzung der Trägerschichten fest. Auch hier wird ein Wechsel der Speichermedien und die Migration der Daten in Zukunft nicht zu vermeiden sein.

In den letzten Jahren wurde die Archivierung von Tondokumenten in Form von digitalen Audiodateien zur gängigen Praxis. Als Containerfomat hat sich das WAVE[37]-Format als de-facto-Standard durchgesetzt. Zudem findet das AIFF-Format[38] des MacOS -Betriebssystems breite Anwendung. Beide können als stabile und langfristig nutzbare Formate gelten. Als Sonderformat für den Rundfunkbereich wurde das BWF-Format (Broadcast-Wave-Format) von der European Broadcasting Union erarbeitet. Dieses Format wird vom Technischen Komitee der *International Association of Sound and Audiovisual Archives* offiziell empfohlen (vgl. IASA-TC 04, 6.1.1.1 und 6.6.2.2).[39] Das Format ist WAVE-kompatibel, beinhaltet aber zusätzliche Felder für Metadaten. Ein ambitioniertes Beschreibungsformat ist MPEG-21 der Moving Pictures Expert Group, das ein Framework für den Austausch und die Weitergabe multimedialer Objekte durch unterschiedliche Benutzer bildet. Es findet bereits in einigen audiovisuellen Archiven Anwendung.[40]

---

37 Zum Wave-Format siehe: http://www.it.fht-esslingen.de/~schmidt/vorlesungen/mm/seminar/ss00/HTML/node107.html und http://ccrma.stanford.edu/courses/422/projects/WaveFormat/

38 http://www.digitalpreservation.gov/formats/fdd/fdd000005.shtml

39 Weitere Informationen zu diesem Format unter: http://www.sr.se/utveckling/tu/bwf/ und http://www.iasa-online.de/

40 Beispielsweise in der Los Alamos National Laboratory Digital Library: http://www.dlib.

Die Bereitstellung des digitalen Materials für den Zugriff kann auch über Formate mit verlustbehafteter Datenkompression erfolgen, wie dies bei der Nutzung über das Internet in Form von Streaming-Formaten (z.B. Real Audio) oder bei MP3-Format der Fall ist. Diese Formate eignen sich jedoch nicht für die Langzeitarchivierung.[41]

Neben der Sicherung der Nutzung des Audio-Datenstromes erfordert eine effektive und erfolgreiche Langzeitarchivierung die Verfügbarkeit von Metadaten aus unterschiedlichen Bereichen.[42] *Inhaltliche* Metadaten betreffen die Beschreibung des Inhaltes, beispielsweise des Genres oder den Namen des Komponisten.[43] Für die Erhebung *technischer* Metadaten stehen Programme zur Verfügung, die diese aus den Dateien auslesen können. Der Digitalisierungsvorgang sollte ebenfalls in den technischen Metadaten abgelegt werden und Informationen zum originalen Trägermedium, seinem Format und dem Erhaltungszustand sowie zu den für seine Wiedergabe notwendigen Geräten und Einstellungs-Parametern beinhalten. Zusätzlich sind die Parameter des Digitalisierungsvorganges und die verwendeten Geräte zu dokumentieren. *Administrative* Metadaten beinhalten die rechtlichen Informationen, *strukturelle* Metadaten Zeitstempel, SMIL-Dokumente[44] u.a. Informationen zur Struktur des Tondokumentes. Für die Kontrolle der Integrität des Datenstroms sind Prüfsummen zu sichern.

Ausblick

Eine neue Herausforderung für die langfristige Sicherung unseres Kulturgutes ist das sich immer stärker als Distributionsweg etablierende Internet. Der Verkauf, aber auch der Tausch und die kostenlose Bereitstellung von digitalen Tondokumenten über das Web erreichen explosionsartig zunehmende Ausmaße. Die neuesten Songs werden über Internetportale und Internetshops per

org/dlib/november03/bekaert/11bekaert.html

41 Umfangreiche Beispiele von Konzepten zur Langzeitarchivierung digitaler Tondokumente findet man bei Schüller, Dietrich (2008) oder bei Casey, Mike/Gordon, Bruce (Hrsg.) (2007).

42 Das Probado-Projekt der Bayerischen Nationalbibliothek ist ein Beispiel für die Definition von Metadatenschemata im Rahmen der Langzeitarchivierung. Diet, Jürgen/Kurth, Frank (2007): The Probado Music Repository at the Bavarian State Library. In: 8th International Conference on Music Information Retrieval, September 23rd-27th 2007. 8th International Conference on Music Information Retrieval, September 23rd-27th 2007

43 Casey, Mike/Gordon, Bruce (2007) empfehlen hierfür die Metadatenschemata MARC oder MODS. Casey, Mike/Gordon, Bruce (Hrsg.) (2007), S. 62.

44 SMIL (Synchronized Multimedia Integration Language)

Download erworben oder im Internetradio mitgeschnitten. Selbstproduzierte Aufnahmen der Nutzer werden getauscht oder einer breiten Öffentlichkeit in entsprechenden Portalen angeboten. Podcasts werben für Politiker, Fahrzeuge, aber auch für den Besuch von Museumsaustellungen. Die Möglichkeiten der neuen Produktions- und Informationsmedien verändern unseren Umgang mit der auditiven Ware.

Für die kulturbewahrenden Institutionen bedeuten diese neuen Produktions- und Verteilungswege, dass bewährte Zuliefer- und Ingestverfahren überarbeitet und den neuen Anforderungen angepasst werden müssen. Neben der Notwendigkeit neue Auswahlkriterien für die zu archivierenden Daten zu definieren, gibt es zusätzliche Hindernisse in Form der zunehmenden Verbreitung technischer Schutzmaßnahmen. Durch immer neue Kopierschutzmechanismen versuchen die Musikverlage ihre Rechte zu sichern. Die daraus erwachsenden technischen wie auch rechtlichen Auswirkungen müssen bei der Langzeitarchivierung berücksichtigt werden. Leider gibt es keine generelle Sonderregelung für Institutionen, die für den Erhalt unseres kulturellen Erbe zuständig sind. Sogenannte „Schrankenregelungen" im Urheberrechtsgesetz ermöglichen allerdings Institutionen aus kulturellen oder wissenschaftlichen Bereichen individuelle Regelungen mit den Branchenvertretern zu vereinbaren. Hier könnten auch die besonderen Aspekte in Bezug auf die Langzeitarchivierung geregelt werden.

Literatur

Block, Carsen et. al. (2006): *Digital Audio Best Practice*, Version 2.1. CDP Digital Audio Working Group. In: http://www.bcr.org/cdp/best/digital-audio-bp.pdf

Breen, Majella (2004): *Task Force to establish selection criteria of analogue and digital audio contents for transfer to data formats for preservation purposes*. International Association of Sound and Audiovisual Archives (IASA). In: http://www.iasa-web.org/downloads/publications/taskforce.pdf

Schüller, Dietrich (2008): *Audiovisual research collections and their preservation*. TAPE (Training for Audiovisual Preservation in Europe). http://www.tape-online.net/docs/audiovisual_research_collections.pdf

Casey, Mike/Gordon, Bruce (Hrsg.) (2007): *Sound Directions. Best Practices for Audio Preservation*. In: http://www.dlib.indiana.edu/projects/sounddirections/bestpractices2007/

17.7 Langzeitarchivierung und -bereitstellung im E-Learning-Kontext

Tobias Möller-Walsdorf

In der elektronisch unterstützten Lehre hat sich in den letzten zehn Jahren ein breites Spektrum unterschiedlicher Technologien, E-Learning-Werkzeuge und didaktischer Szenarien entwickelt. Unter dem Aspekt der Archivierung kann das Themenfeld E-Learning in zwei Bereiche unterteilt werden, zum einen in E-Learning-Kurse bzw. -Kursangebote, zum anderen in Lehr- oder Lernmaterialien (E-Learning-Content). Liegt bei den E-Learning-Kursen der Fokus mehr auf der formalen oder rechtlichen Notwendigkeit einer Archivierung, so kommt bei E-Learning-Content die Nachnutzbarkeit und Weiterverwendung der Materialen hinzu. E-Learning-Kursbestandteile sind technisch sehr eng mit dem jeweiligen E-Learning-System verbunden und damit in der Langzeitarchivierung komplex zu handhaben. E-Learning-Content kann, in Form unterschiedlichster multimedialer oder auch dynamischer Objekte, in einer Vielzahl technischer Formate vorliegen. Gerade dieses breite Spektrum macht die Langzeitarchivierung schwierig. Metadaten und Standard-Formate sind daher eine wichtige Voraussetzung für die Langzeitarchivierung von E-Learning-Kursinformationen und E-Learning-Content.

Einführung

Möchte man sich der Frage der Archivierung und Langzeitarchivierung im Kontext des E-Learnings nähern, so ist zuerst eine Differenzierung und Definition des Themenfeldes nötig, denn was konkret unter dem Begriff E-Learning verstanden wird, hat sich in den letzten Jahren stark gewandelt. Bezeichnete der Begriff bei seiner Etablierung in den 1990er Jahren besonders eigenständige Lern-Anwendungen, sog. Computer Based Trainings bzw. später mit der Etablierung des Internets sog. Web Based Trainings, so wird der Begriff heute allgemein weiter gefasst.

Beispielsweise definiert Michael Kerres E-Learning wie folgt: „Unter E-Learning (englisch electronic learning – elektronisch unterstütztes Lernen), auch E-Lernen genannt, werden alle Formen von Lernen verstanden, bei denen digitale Medien für die Präsentation und Distribution von Lernmaterialien und/oder zur Unterstützung zwischenmenschlicher Kommunikation zum Einsatz kommen.“[45]

45 http://de.wikipedia.org/wiki/E-learning

Es geht somit im E-Learning heute neben dem technisch gestützten Selbstlernen mehr auch um die Unterstützung von Präsenzlehre. Unter dem Begriff E-Learning werden daher mittlerweile eine Vielzahl unterschiedlicher Technologien zusammengefasst, deren Spektrum technisch von Autorensystemen, Simulationen, Videokonferenzen und Teleteaching, Audiomitschnitten und Podcasts, Lernmanagementsystemen bis zu Lernspielen und Web-3D-Plattformen reicht.[46] Diese Technologien können in vielen unterschiedlichen didaktischen Szenarien mit unterschiedlichstem Umfang und unterschiedlichster Ausprägung eingesetzt werden. Galt in den Anfängen E-Learning noch als Alternative zu klassischen Lernformen, so wird es heute vor allem als sinnvolle Unterstützung und Ergänzung in der Lehre und im Lernprozess (dem sogenannten „Blended Lerarning") eingesetzt. Das niedersächsische (Open-Source-)„Erfolgsprodukt" Stud.IP ist ein gutes Beispiel für diese Entwicklung.[47] Traditionelle Lehre und E-Learning werden so gemeinsame Bestandteile eines hybriden Lernarrangements.

Dies hat zur Folge, dass bei der Betrachtung der Bereitstellung und besonders bei der Archivierung und Langzeitarchivierung das Themenfeld E-Learning in zwei Bereiche geteilt werden sollte, die differenziert betrachtet werden müssen: Gemeint ist die Unterscheidung zwischen a) E-Learning-Kursen bzw. Kursangeboten und b) E-Learning-Content. Also dem E-Learning-Kurs als organisatorischer Veranstaltungsform oder virtuellem Ort der Lernorganisation und Kommunikation und E-Learning-Content als die elektronischen Materialien, die bei der Lehre und dem Lernen Einsatz finden. Hierbei kann E-Learning-Content Teil eines E-Learning-Kurses sein, es kann aber auch selbständig unabhängig von einem Kurs nutzbar sein. Ein E-Learning-Kursangebot ist auch gänzlich ohne E-Learning-Materialien möglich, beispielsweise wenn E-Learning-Komponenten wie Foren, Wikis oder elektronische Semesterapparate in einem Lernmanagementsystem eingesetzt werden.

E-Learning-Kurse

Ein großer Teil des E-Learning hat heute mit dem Einsatz neuer Medien und Technologien zur Organisation, Durchführung und Effizienzsteigerung der Lehre zu tun. Hierbei stellt sich die Frage, was von den dabei anfallenden Daten auf den Servern der Bildungseinrichtungen archiviert werden sollte. Welchen Sinn macht es E-Learning-Kurse zu archivieren bzw. welche Bestandteile

46 Vgl. http://www.elan-niedersachsen.de/index.php?id=134
47 http://www.studip.de

eines E-Learning-Kurses sollten bzw. müssten archiviert werden: Veranstaltungsdaten, Teilnehmerlisten, Foreneinträge und Chats, Umfragen, Test- und Prüfungsergebnisse?

Da diese Informationen zu E-Learning-Kursen sehr stark personenbezogen sind, hat eine Archivierung dieser Daten eher einen reinen Archivierungscharakter und nur wenig Aspekte einer Nachnutzbarkeit und Weiterverwertung; der Zugriff auf diese Daten wäre aus Datenschutzgründen stark eingeschränkt.

Die genannten Bestandteile der E-Learning-Kurse sind technisch sehr eng mit dem System zur Kursorganisation (beispielsweise dem Lernmanagement-System) oder einem E-Learning-Tool (z.B. für Foren und Wikis) verbunden, so dass für die Archivierung zukünftig eine Emulationsumgebung des gesamten Systems (inkl. beispielsweise der Datenbank) notwendig wäre. Alternativ könnte nur ein Export einzelner, losgelöster Bestandteile des Kurses (beispielsweise der Foreneinträge in Textform oder von Lerneinheiten nach dem SCORM-Standard) erfolgen.

E-Learning-Content

E-Learning-Content bezeichnet in dieser Aufteilung im Gegensatz zu den E-Learning-Kursen die elektronischen Lehr- und Lernmaterialien, die im E-Learning eingesetzt werden. Die Art dieses E-Learning-Contents ist sehr heterogen und vom technischen System und didaktischen Szenario abhängig. Es kann sich u.a. um reine Textdateien, Bilddateien, Power-Point-Präsentationen, Audio- und Videodateien, Simulationen und Animationen (Flash-Dateien), HTML-Projekte und komplexe Multimedia-Programme handeln.

Oftmals sind dies unterschiedlichste multimediale und dynamische Objekte, die zusätzlich durch Interaktionen mit dem Nutzer gesteuert werden, also einer komplexen Programmierung folgen. Eine Vielzahl technischer Formate, unzureichende Normierung und besonders ein sehr hoher Innovationszyklus bei den Dateiformaten der multimedialen Objekte, machen das Thema der Archivierung von E-Learning-Content zu einem der Komplexesten, vergleichbar vielleicht mit der Archivierung von Multimedia-Anwendungen oder Computerspielen.

Werden die Dateien archiviert, besteht zudem die Gefahr, dass sie – losgelöst vom Kontext und ohne den Kurszusammenhang – didaktisch unbrauchbar oder für den Lehrenden und Lernenden inhaltlich unverständlich werden. Zusätzlich können rechtliche Aspekte den zukünftigen Zugriff auf diese Archivmaterialien erschweren, da für den Einsatz im Kurs-Zusammenhang des

E-Learning-Kurses andere rechtliche Rahmenbedingungen für den E-Learning-Content bestehen, als bei frei zugänglichen Materialien (§52a UrhG).

E-Learning-Content ist oftmals in einem technischen, proprietären System erstellt bzw. bedarf eines speziellen E-Learning-Systems, um ihn anzuzeigen, beispielsweise bei Kurs-Wikis, Contentmanagement-Systemen oder speziellen Authoring-Tools wie z.B. ILIAS. Ist ein Export der Materialien in ein Standardformat möglich bzw. wurden die Materialien bereits in einem gebräuchlichen Format erstellt, so ist die Archivierung einfacher. Die möglichen Formate, die im E-Learning zum Einsatz kommen, entsprechen zum größten Teil den gebräuchlichen Multimedia-Formaten, also beispielsweise PDF, Power-Point, Flash, AV-Formate, HTML-Projekte. Dazu aber auch noch Spezialformate wie z.B. Dateien des weit verbreiteten Aufzeichnungstools Lecturnity.[48]

Um die Lesbarkeit digitaler Materialien möglichst lange zu gewährleisten, sollten allgemein Dateiformate verwendet werden, deren Spezifikation offen gelegt ist (z.B. ODF, RTF, TIFF, OGG). Proprietäre Formate, die an die Produkte bestimmter Hersteller gebunden sind, wie z.B. DOC oder PPT, sind zu vermeiden. Der Grund hierfür liegt darin, dass langfristig zumindest die Wahrscheinlichkeit hoch sein sollte, dass eine Interpretationsumgebung (Hardware, Betriebssystem, Anwendungsprogramm) für das archivierte Objekt in der künftigen Nutzergemeinde vorhanden sein wird.[49] Diese Forderung ist für den Bereich E-Learning allerdings heute nur schwer umsetzbar. Auf jeden Fall sollten aber für die Erstellung von E-Learning-Content die auch in anderen Bereichen üblichen Multimediaformate eingesetzt werden. Die Archivierung ist dann zumindest analog zu anderen multimedialen Objekten zu sehen, natürlich mit allen dort auftretenden Schwierigkeiten der Emulierung oder Migration.

Archivierungskriterien

Betrachtet man beispielsweise den im Rahmen des Projektes ELAN in Niedersachsen entstandenen E-Learning-Content (www.elan-niedersachsen.de), so zeigt sich, dass nicht alle entstehenden E-Learning-Materialien auch langfristig relevant sind und nicht immer eine Archivierung und Bereitstellung mit dem Zweck der Nachnutzung und Weiterverwendung sinnvoll ist. Oftmals wandeln sich Kurse pro Semester so stark, dass von der Seite der Dozenten kein Interesse an der Archivierung und späteren Bereitstellung besteht. Eine Selektion des Materials, besonders unter dem Aspekt der Nachnutzbarkeit, ist daher angebracht. Allerdings sollte bei der Archivierung die Meinung des Autors bezüg-

48 http://www.lecturnity.de/

49 siehe hierzu auch nestor Handbuch 17.8 „Interaktive Applikationen“

lich der Relevanz der Archivierung nicht immer ausschlaggebend sein, denn für viele Materialien ist es derzeit nur sehr schwer vorhersehbar, welcher Wert ihnen in Zukunft beigemessen wird. Dass heute beispielsweise sehr frühe (Magnetophon-)Aufzeichnungen der Vorlesungen von Max Planck als großer Glücksfall angesehen werden, war zum Zeitpunkt ihrer Erstellung in vollem Umfang sicher noch nicht abschätzbar.[50] Das „absehbare historische Interesse" ist somit besonders für Bibliothekare und Archivare, die mit diesen Materialien zu tun haben, eine der wichtigen und auch schwierigen Fragen bei der Archivierung.

Auch für die Dozenten interessant ist bei der Archivierung die Wiederverwendbarkeit und Nachnutzung von Lehrmaterial. Hier sind beispielsweise Unterlagen für Grundlagenvorlesungen zu nennen. Material also, dass in dergleichen Form regelmäßig verwendet wird und sich ggf. nur in seiner jeweiligen Zusammenstellung unterscheidet. Solche Materialien könnten zudem über die Universität hinaus im Umfeld von Weiterbildung und Erwachsenenbildung (Lifelong Learning) eingesetzt werden. Auch Kostenreduktion bei zum Teil sehr kostenintensiven E-Learning-Produktionen, wie z.B. Videoaufzeichnungen oder komplexen Multimedia-Anwendungen, könnte bei der Archivierung eine Rolle spielen (vgl. z.B. die IWF Campusmedien[51]).

Ein weiterer Grund für die Archivierung von erstellten Lehr-, Lern- und besonders Prüfungsmaterialien können zukünftig rechtliche Anforderungen sein, nämlich zur späteren Kontrolle von Prüfungsergebnissen. Derzeit besteht allerdings noch keine konkrete rechtliche Verpflichtung, solche E-Learning-Dokumente längerfristig zu archivieren. Bei weitergehender Etablierung von E-Learning-Bestandteilen, besonders durch den Anstieg der nötigen Prüfungsleistungen beispielsweise bei den Bachelor-Master-Studiengängen, wird sich diese Situation aller Voraussicht nach zukünftig ändern.

Metadaten für E-Learning-Kurse und E-Learning Content

Um die Bereitstellung von E-Learning-Archivobjekten, also E-Learning-Kursen und E-Learning-Content oder Bestandteilen daraus, zu gewährleisten, werden neben technischen Metadaten inhaltsbeschreibende Metadaten und nachhaltig gültige Identifikatoren (Persistent Identifier) für die zu archivierenden Objekte benötigt. Nur anhand dieser Metadaten ist eine Suche in den Datenbeständen möglich. Im Bereich der Metadaten erfolgt u.a. im Rahmen von ELAN eine rege Forschungsaktivität mit Fokus auf der Entwicklung von Standards für solche Metadaten. Welche inhaltsbeschreibenden Metadaten für E-Learning-

50 http://webdoc.sub.gwdg.de/ebook/a/2002/nobelcd/html/fs_planck.htm
51 http://www.iwf.de/campusmedien/

Objekte geeignet sind und an welchen bestehenden Standard (z.B. Dublin Core, LOM[52]) sie orientiert werden, wurde im Rahmen des ELAN-Projektes in Niedersachsen ausgearbeitet, auf die Ergebnisse des „ELAN Application Profile" sei hier verwiesen.[53] Daneben ist das vom Bundesministerium für Wirtschaft und Technologie (BMWi) 2004 bis 2006 geförderte Projekt Q.E.D. (http://www.qed-info.de) zu nennen, welches das Ziel verfolgte, die Etablierung von innovativen Lernszenarien und eben auch internationalen Qualitätsstandards und Normen im E-Learning in Deutschland weiterzuentwickeln. Projektpartner war unter anderem das Deutsche Institut für Normung e.V. (DIN).

Bei allen diesen Bemühungen der Erfassung von Metadaten und Standardisierung mit dem Ziel der strukturierten Bereitstellung, Archivierung und Langzeitarchivierung sollten die Bibliotheken und Archive mehr als bisher in die Entwicklungsprozesse eingebunden werden. E-Learning-Content sollte, wie andere elektronische Materialien auch, in den regulären Geschäftsgang besonders der Bibliotheken einfließen und damit auch unabhängig von Projekten und temporären Initiativen Berücksichtigung finden. Nur so ist eine langfristige Bereitstellung und Archivierung dieses Teils unseres kulturellen Erbes möglich.

52 Vgl. http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf, Januar 2009.

53 DINI Schriften 6: ELAN Application Profile: Metadaten für elektronische Lehr- und Lernmaterialien [Version 1.0, Oktober 2005]. http://nbn-resolving.de/urn:nbn:de:kobv:11-10050226

17.8 Interaktive digitale Objekte

Dirk von Suchodoletz

Interaktive Applikationen sind spezielle digitale Objekte, die sich aufgrund ihres dynamischen, nichtlinearen Charakters schlecht mit traditionellen Archivierungsstrategien wie der Migration bewahren lassen. Sie treten dem Archivbetreiber typischerweise in zwei Ausprägungen entgegen. Entweder sie sind aus sich heraus von Bedeutung und primärem Interesse: In diese Klasse zählen Datenbankprogramme oder Computerspiele oder auch besondere technische Umgebungen, die als solche bewahrt werden sollen. Diese Objekte kann man auch als Primärobjekte fassen. Zusätzlich benötigt werden weitere dynamische Objekte, wie Applikationen zum Anzeigen oder Abspielen bestimmter Datenformate.
Sie werden als Hilfsmittel für die eigentlich interessierenden Objekte gebraucht und könnten daher als Sekundärobjekte bezeichnet werden. Allen dynamischen Objekten ist gemein, dass sie sich nur durch die Rekonstruktion ihrer Nutzungsumgebungen, einer bestimmten Zusammenstellung aus Software und / oder Hardware bewahren lassen. Diese Umgebungen lassen sich mithilfe der Emulationsstrategie langzeitbewahren. Oft genügt das interaktive Objekt alleine nicht: Je nach Primärobjekt oder Komplexität sind weitere Komponenten wie Schriftarten, Codecs oder Hilfsprogramme erforderlich, die in einem Softwarearchiv zusätzlich aufgehoben werden müssen.

Einführung

Mit der Durchdringung fast aller Bereiche des täglichen Lebens mit Computern änderten sich die Verfahren zur Speicherung, Verbreitung und Vervielfältigung von Informationen. Die elektronische Universalmaschine übernimmt eine dominierende Rolle: Eine zunehmende Zahl traditioneller Objekte wie Texte, Bilder, Musik und Film sind nicht mehr an analoge Medien gebunden, sondern können effizient digital bearbeitet, kopiert und verbreitet werden. Wissenschaftler fast aller Disziplinen erheben ihre Daten immer seltener ohne elektronische Maschinen, erfasste Informationen nehmen ohne Umweg über Papier und Stift den Weg zur Verarbeitung, Nutzung, Auswertung und Archivierung.

Die erzeugten digitalen Objekte können, anders als klassische Medien wie Papier oder Leinwände, nicht aus sich alleine heraus betrachtet werden. Die Erstellung und der Zugriff auf sie ist nur mithilfe eines nicht unerheblichen technischen Apparates möglich und gerade dieser Apparat unterliegt einer rasanten Fortentwicklung. Digitale Objekte erfordern eine bestimmte technische Umgebung, die sich aus Soft- und Hardwarekomponenten zusammensetzt. Diese Umgebung, hier als Erstellungs- oder Nutzungsumgebung bezeichnet, sorgt

dafür, dass ein Benutzer ein digitales Objekt je nach Typ betrachten, anhören oder ausführen kann.

Bei Computerprogrammen und Betriebssystemen, allgemein unter dem Begriff Software zusammengefasst, handelt es sich um dynamische, interaktive digitale Objekte. Vielfach wurden sie für die direkte Benutzerinteraktion mit dem Computer geschrieben, damit Menschen überhaupt erst sinnvoll Rechnerhardware nutzen können. Interaktive Objekte zeichnen sich durch einen nicht-linearen Aufbau und Ablauf aus. Erst die Interaktion mit dem Computeranwender bestimmt, wie sich das Objekt verhält. Jede Sitzung verläuft anders, deshalb ist die Zahl der möglichen Handlungswege typischerweise unbeschränkt.

Solche interaktiven Objekte treten dem Archivbetreiber und -nutzer in zwei Ausprägungen gegenüber. Einerseits handelt es sich um *Primärobjekte*. Diese sind als archivwürdig eingestufte Objekte, an denen ein direktes Interesse besteht. Hierzu zählen beispielsweise Computerspiele, interaktive Medien, Unterhaltung oder digitale Kunst, deren Archivierung in eigenen Abschnitten dargestellt wird. *Sekundärobjekte* meint alle digitalen Objekte, die zur Darstellung oder zum Ablaufenlassen von Primärobjekten erforderlich sind.[54]

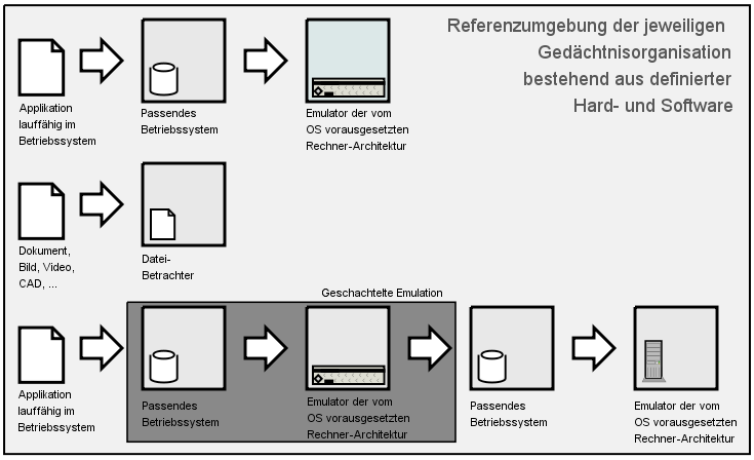


Abbildung 1: Je nach Art des Primärobjekts sind unterschiedliche Schritte zu seiner Darstellung oder Ausführung notwendig.

54 Sekundärobjekte werden je nach Vorhandensein bestimmter Primärobjekte nachgefragt und besitzen keine eigene Archivwürdigkeit aus sich heraus.

Diese Zusammenhänge lassen sich durch View-Paths (Kapitel 9.3) formalisieren. Darunter versteht man Darstellungswege, die vom darzustellenden oder auszuführenden digitalen Primärobjekt starten und bis in die tatsächliche Arbeitsumgebung des Archivnutzers reichen (Abbildung 1). Sie unterscheiden sich in ihrer Länge je nach Objekttyp und eingesetzter Archivierungsstrategie. Ein klassisches statisches Objekt wie ein Textdokument, Bild oder eine Videosequenz (linearer Ablauf ohne beliebige Verzweigungen) benötigen eine Darstellungsapplikation. Diese kann im Fall einer Migration direkt in der Arbeitsumgebung des Archivnutzers (Abbildung 1, Mitte) ablaufen. Damit fällt der View-Path kurz aus. Ist diese Applikation nicht mehr auf aktuellen Systemen installierbar und lauffähig, so sind weitergehende Maßnahmen erforderlich. Dann muss die Nutzungsumgebung auf geeignete Weise nachgebildet werden, was durch Emulation realisiert werden kann. Damit verlängert sich der View-Path und die Menge der benötigten Sekundärobjekte vervielfacht sich mindestens um den Emulator und ein zur Applikation passendes Betriebssystem.

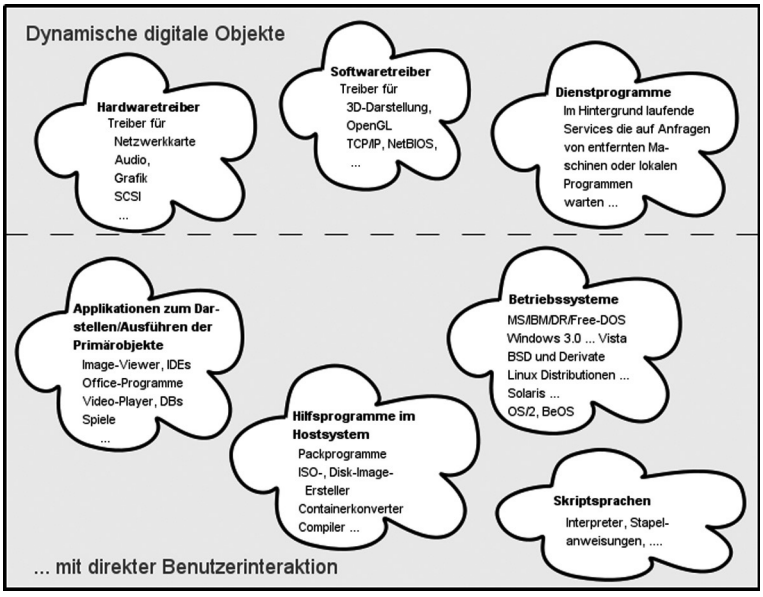


Abbildung 2: Die Klasse der dynamischen digitalen Objekte umfasst die interaktiven Objekte, die für eine direkte Benutzerinteraktion erstellt wurden. Ihre Einordnung muss nicht immer eindeutig ausfallen.

Typen interaktiver Objekte und ihre Bedeutung

Die Ausdifferenzierung und technologische Entwicklung auf dem Gebiet der elektronischen Datenverarbeitung hat inzwischen einen hohen Grad der Differenzierung erreicht, der sich in einer Reihe spezialisierter Komponenten niederschlägt. Erst das Zusammenspiel von ausführbaren Programmen, Betriebssystemen, Softwarebibliotheken und weiteren Komponenten (Abbildung 2) erlaubt die Erschaffung der typischen Objekte, mit denen sich die digitale Langzeitarchivierung befasst.

Die Liste wichtiger Sekundärobjekte umfasst:

- Betriebssysteme - sind die Grundkomponenten eines jeden Rechners neben seiner (physikalischen) Hardware (Abbildung 3). Sie übernehmen die Steuerung der Hardware, kümmern sich um Ressourcenverwaltung und -zuteilung und erlauben die Interaktion mit dem Endanwender. Sie sind im kompilierten Zustand[55] deshalb nur auf einer bestimmten Architektur ablauffähig und damit angepasst an bestimmte Prozessoren, die Art der Speicheraufteilung und bestimmte Peripheriegeräte zur Ein- und Ausgabe. Da eine Reihe von Funktionen von verschiedenen Programmen benötigt werden, sind diese oft in sogenannte Bibliotheken ausgelagert. Programme, die nicht alle Funktionen enthalten, laden benötigte Komponenten aus den Bibliotheken zur Laufzeit nach. Bibliotheken und Programme hängen dementsprechend eng miteinander zusammen.

- Anwendungsprogramme - oberhalb der Betriebssystemebene[56] befinden sich die Anwendungen (Abbildung 3). Dieses sind Programme, die für bestimmte, spezialisierte Aufgaben erstellt wurden. Mit diesen Programmen generierten und bearbeiteten Endanwender Daten der verschiedensten Formate. Der Programmcode wird im Kontext des Betriebssystems ausgeführt. Er kümmert sich um die Darstellung gegenüber dem Benutzer und legt fest, wie beispielsweise die Speicherung von

55 Computerprogramme werden in verschiedenen Programmiersprachen erstellt. Dieses sind typischerweise sogenannte Hochsprachen, die nicht direkt von einem Prozessor interpretiert werden können und erst in Maschinensprache übersetzt werden müssen. Während die Hochsprache relativ abstrakt von der konkreten Hardware ist, kann Maschinensprache immer nur auf einer bestimmten Computerarchitektur ausgeführt werden.

56 Für die schematische Darstellung der Arbeit eines Computers wird oft ein Schichtenmodell gewählt.

Objekten in einer Datei organisiert ist und wie der Anwender mit dem Computer interagiert. Das Betriebssystem übernimmt die Speicherung von Dateien auf Datenträgern üblicherweise angeordnet in Verzeichnissen. Zur Ausführung auf einer bestimmten Rechnerarchitektur werden Betriebssysteme und Applikationen aus dem sogenannten Quellcode in den passenden Binärcode übersetzt. Deshalb können Programme und Bibliotheken nicht beliebig zwischen verschiedenen Betriebssystemen verschoben werden.

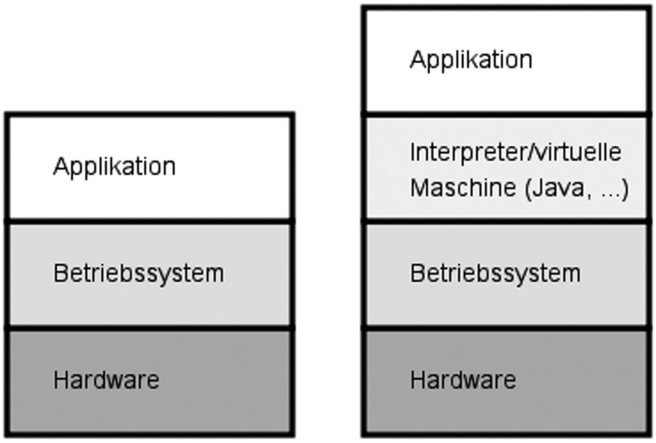


Abbildung 3: Eine typische Ablaufumgebung neuerer Rechnerarchitekturen für digitale Objekte bestehend aus Hard- und Software als Schichtenmodell. Interpreter und abstrakte Programmiersprachen wie Java können eine weitere Schicht einführen (rechts).

Ein wichtiges Beispiel digitaler Objekte primären Interesses sind Datenbanken. Die Bewegung, Durchsuchung und Verknüpfung großer Datenbestände gehört zu den großen Stärken von Computern. Zur Klasse der datenbankbasierten Anwendungen zählen Planungs- und Buchhaltungssysteme, wie SAP, elektronische Fahrpläne diverser Verkehrsträger bis hin zu Content Management Systemen (CMS) heutiger Internet-Auftritte von Firmen und Organisationen. Wenn von einer Datenbank sehr verschiedene Ansichten ad-hoc erzeugt werden können, ist sehr schwer abzusehen, welche dieser Ansichten zu einem späteren Zeitpunkt noch einmal benötigt werden könnten. Unter Umständen hat man sich dann auf Teilmengen festgelegt, die von nachfolgenden Betrachtern als unzureichend oder irrelevant eingestuft werden könnten. Gerade bei Datensammlungen wichtiger langlebiger Erzeugnisse wie Flugzeugen besteht großes allgemeines Interesse eines zeitlich unbeschränkten Zugriffs.

Alle genannten dynamischen Objekttypen zeichnen sich dadurch aus, dass sie außerhalb ihres festgelegten digitalen Kontextes heraus nicht sinnvoll interpretiert und genutzt werden können. Zudem ist ihre Migration nicht trivial:[57] Es fehlt typischerweise die gesamte Entwicklungsumgebung und der Zugriff auf den Quellcode, die eine Anpassung und Übersetzung auf aktuelle Systeme erlauben würden. Zudem wäre der personelle Aufwand immens und stünde häufig nicht in einem sinnvollen Verhältnis zum Wert des Objekts. Ebenso scheidet eine Überführung in ein analoges Medium in den meisten Fällen aus: Der Ausdruck des Programmcodes auf Papier oder Mikrofilm oder die Aufnahme einer Programmsitzung auf Video sind derart „verlustbehaftete" Speicherverfahren, dass sie im Sinne der Archivierung vielfach die gestellten Anforderungen nicht erfüllen.

Emulation – Erhalt von Nutzungsumgebungen

Emulation heißt zuerst einfach erstmal nur die Schaffung einer virtuellen Umgebung in einer gegebenen Nutzungsumgebung, üblicherweise dem zum Zeitpunkt des Aufrufs üblichen Computersystem. Das kann bedeuten, dass Software durch eine andere Software nachgebildet wird, ebenso wie Hardware in Software.

Emulation setzt dabei nicht am digitalen Objekt selbst an, sondern beschäftigt sich mit der Umgebung, die zur Erstellung dieses Objektes vorlag. Das bedeutet beispielsweise die Nachbildung von Software durch andere Software, so dass es für ein betrachtetes digitales Objekt im besten Fall keinen Unterschied macht, ob es durch die emulierte oder durch die Originalumgebung behandelt wird. Dem Prinzip folgend kann Computerhardware durch Software nachgebildet werden, auch wenn dieses erstmal deutlich komplexer erscheint. Einen ausführlichen Überblick zur Emulation als Langzeitarchivierungsstrategie gibt Kapitel 8.4, das entsprechende Kapitel in Borghoff (2003) oder auch Holdsworth (2001). Das Grundlagenwerk zur Emulation in der Langzeitarchivierung stammt von Jeff Rothenberg (1999) und (2000). Eine erste praktische Einbindung von Emulationsansätzen erfolgt derzeit im EU-geförderten PLANETS Project.[58]

---

57 Maschinencode kann nicht trivial von einer Rechnerarchitektur auf eine andere übersetzt werden so wie dieses für wohldefinierte Datenformate statischer Objekte möglich ist.

58 Es wird unter dem „Information Society Technologies (IST) Programme" des Framework 6 anteilig finanziert (Project IST-033789) und beschäftigt sich mit der prototypischen Erstellung von Tools

Emulatoren sind spezielle Software-Applikationen, die in der Lage sind Nutzungsumgebungen, wie Hardware-Plattformen, in derart geeigneter Weise in Software nachzubilden, dass ursprünglich für diese Nutzungsumgebung erstellte Applikationen weitgehend so arbeiten wie auf der Originalplattform. Emulatoren bilden damit die Schnittstelle, eine Art Brückenfunktion, zwischen dem jeweils aktuellen Stand der Technik und einer längst nicht mehr verfügbaren Technologie. Dabei müssen sich Emulatoren um die geeignete Umsetzung der Ein- und Ausgabesteuerung und der Peripherienachbildung bemühen.

Die Hardwareemulation setzt auf einer weit unten liegenden Schicht an (Abbildung 3). Das bedeutet auf der einen Seite zwar einen sehr allgemeinen Ansatz, erfordert umgekehrt jedoch eine ganze Reihe weiterer Komponenten: Um ein gegebenes statisches digitales Objekt tatsächlich betrachten zu können oder ein dynamisches Objekt ablaufen zu sehen, müssen je nach Architektur die Ebenen zwischen der emulierten Hardware und dem Objekt selbst „überbrückt" werden. So kann ein Betrachter nicht auf einer nackten X86-Maschine ein PDF-Dokument öffnen (Abbildung 4). Er braucht hierfür mindestens ein Programm zur Betrachtung, welches seinerseits nicht direkt auf der Hardware ausgeführt wird und deren Schnittstellen direkt programmiert. Dieses Programm setzt seinerseits ein Betriebssystem als intermediär voraus, welches sich um die Ansteuerung der Ein- und Ausgabeschnittstellen der Hardware kümmert.

Die Auswahl der inzwischen kommerziell erhältlichen oder als Open-Source-Software verfügbaren Emulatoren oder Virtualisierer (Abbildung 4) ist inzwischen recht umfangreich geworden, so dass häufig sogar mehr als ein Emulator für eine bestimmte Rechnerarchitektur zur Verfügung steht. Der überwiegende Anteil von Emulatoren und Virtualisierern wurde oftmals aus ganz anderen als Langzeitarchivierungsgründen erstellt. Sie sind heutzutage Standardwerkzeuge in der Software-Entwicklung. Nichtsdestotrotz eignen sich viele der im folgenden vorgestellten Werkzeuge für eine Teilmenge möglicher Langzeitarchivierungsaufgaben. Institutionen und privaten Nutzern reichen in vielen Fällen derzeitig verfügbare Programme aus. Jedoch eignet sich nicht jeder Emulator gleichermaßen für die Zwecke des Langzeitzugriffs, weil sich die nachgebildete Hardware ebenso wie die reale weiterentwickelt. Wird alte Hardware nicht mehr unterstützt, kann es passieren, dass ein bestimmtes Betriebssystem nicht mehr auf das Netzwerk zugreifen oder Audio abspielen kann.

Das Schichtenmodell (Abbildung 3) lässt sich nicht auf alle Rechnerarchitekturen anwenden. So existiert für frühe Architekturen keine deutliche Unterscheidung zwischen Betriebssystem und Applikation. Frühe Modelle von Home-Computern verfügten über eine jeweils recht fest definierte Hardware,

die zusammen mit einer Art Firmware ausgeliefert wurde. Diese Firmware enthält typischerweise eine einfache Kommandozeile und einen Basic-Interpreter. Nicht alle für den Betrieb von Emulatoren benötigten Komponenten, wie beispielsweise die genannte Home-Computer-Firmware ist frei verfügbar und muss deshalb mit geeigneten Rechten abgesichert sein. Ohne die Archivierung dieser Bestandteile ist ein Emulator für die Plattform wertlos und das Ausführen entsprechender archivierter Primärobjekte unmöglich.

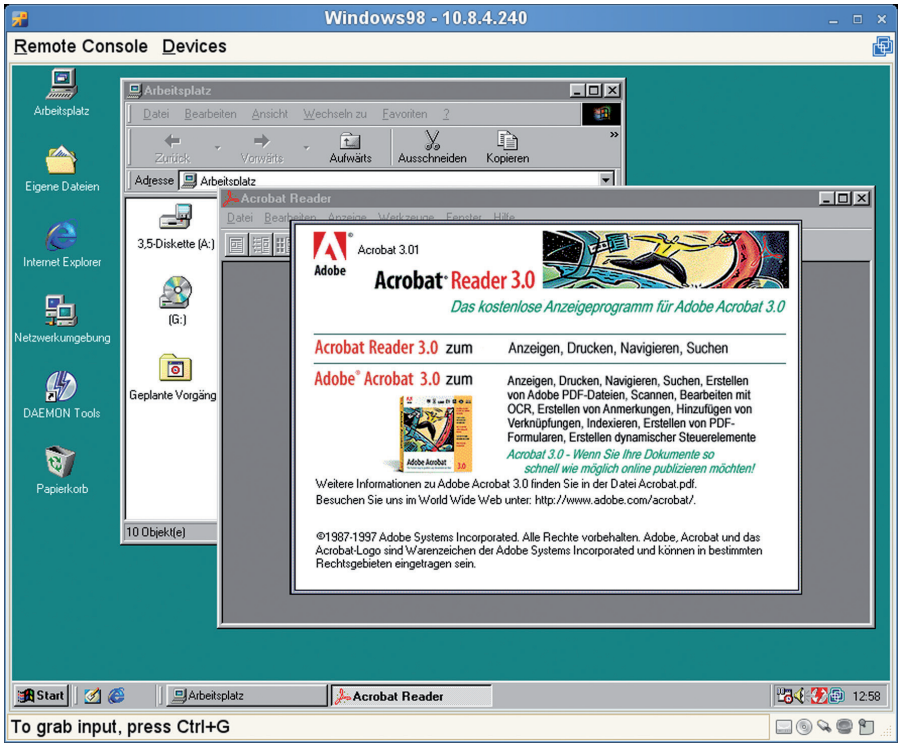


Abbildung 4: Der VMware-Server 2 ist ein X86 Virtualisierer, der einen Zugriff auf alte Nutzungsumgebungen, wie Windows98 über das Netz erlaubt.

Softwarearchiv

Nach den eher theoretisch angelegten Vorbetrachtungen zum Ansatzpunkt der Emulation und erforderlicher Zusatzkomponenten steht nun das Softwarearchiv als ein zentrales Hilfsmittel der Emulationsstrategie im Mittelpunkt. Zu den Erfolgsbedingungen für den Erhalt möglichst originalgetreuer Nutzungsumgebungen für die verschiedensten Typen digitaler Objekte zählen nicht nur

die Primärwerkzeuge – die Emulatoren. Das Archiv muss eine ganze Reihe verschiedener Softwarekomponenten umfassen:

- Geeignete Emulatoren sind zu speichern, so dass mit ihrer Hilfe die Wiederherstellung einer Rechnerarchitektur für bestimmte Nutzungsumgebungen erfolgen kann. Hierzu kann es nötig sein Firmware-Komponenten,[59] wie Home-Computer-ROMs oder X86-BIOS, ebenfalls zu speichern.
- Betriebssysteme, die je nach Rechnerplattform einen Teil der Nutzungsumgebung ausmachen, sind im Softwarearchiv abzulegen.
- Treiber der Betriebssysteme müssen zusätzlich gespeichert werden, da sie den Betriebssystemen erst ermöglichen mit einer bestimmten Hardware umzugehen.
- Alle Applikationen, mit denen die verschiedenen digitalen Objekte erstellt wurden, sind zu archivieren. Diese Anwendungsprogramme sind ebenfalls Bestandteil der Nutzungsumgebung des Objektes. Sie sind in vielen Fällen auf die vorgenannten Betriebssysteme angewiesen.
- Unter Umständen notwendige Erweiterungen einer Applikationsumgebung, wie bestimmte Funktionsbibliotheken, Codecs[60] oder Schriftartenpakete zur Darstellung, sind aufzubewahren.
- Hilfsprogramme, welche den Betrieb der Emulatoren vereinfachen oder überhaupt erst ermöglichen, sind zu sammeln. Hierzu zählen beispielsweise Programme, die direkt mit dem jeweiligen Containerformat eines Emulators umgehen können.
- Je nach Primärobjekt oder gewünschter Nutzungsumgebung sind mehrere Varianten derselben Software zu archivieren, um beispielsweise die Anpassung an den deutschen oder englischsprachigen Raum zu erreichen. Das betrifft einerseits die Verfügbarkeit verschiedensprachiger Menüs in den Applikationen aber auch geeignete Schriftarten für die Darstellung von Sonderzeichen oder Umlauten.

59 Basissoftware, die direkt mit der Hardware verknüpft vom Hersteller ausgeliefert wird.

60 Codecs sind in Software gegossene Verfahren zur Digitalisierung und Komprimierung analoger Medien, wie Audio, Filme.

Mittels View-Paths lässt sich der Vorgang zur Bestimmung der benötigten Softwarekomponenten formalisieren. Ausgehend von den Metadaten des Primärobjekts über die Metadaten der benötigten Applikation zur Betrachtung oder zum Abspielen bis hin zu den Metadaten des Betriebssystems werden die Komponenten ermittelt. Hierzu müssten bereits bestehende Format-Registries, wie beispielsweise PRONOM[61] erweitert werden.

Datenaustausch von Objekten

Ein weiteres nicht zu unterschätzendes Problem liegt im Datentransport zwischen der im Emulator ablaufenden Software und der Software auf dem Computersystem des Archivnutzers. Diese Fragestellung unterscheidet sich nicht wesentlich vom Problem des Datenaustauschs zwischen verschiedenen Rechnern und Plattformen. Mit fortschreitender technischer Entwicklung ergibt sich unter Umständen ein größer werdender Spalt zwischen dem technologischen Stand des stehenbleibenden emulierten Systems und dem des Host-Systems , das die Emulation ausführt. Zum Teil halten die verfügbaren Emulatoren bereits Werkzeuge oder Konzepte vor, um die Brücke zu schlagen.

Nach der Rekonstruktion einer bestimmten Nutzungsumgebung möchte man in dieser die gewünschten Daten ansehen, ablaufen lassen oder in selteneren Fällen bearbeiten. In der Zwischenzeit haben sich mit einiger Wahrscheinlichkeit die Konzepte des Datenaustausches verändert. Hier ist nun dafür zu sorgen, dass die interessierenden Objekte geeignet in die (emulierte) Nutzungsumgebung gebracht werden können, dass die Betrachtung für den Archivnutzer in sinnvoller Form möglich ist und dass eventuell Bearbeitungsergebnisse aus der Nutzungsumgebung in die aktuelle Umgebung transportiert werden können. Vielfach wird sich je nach Erstellungsdatum des Objektes die damalige Erstellungs- oder Nutzungsumgebung dramatisch von der jeweils aktuellen unterscheiden.

Diese Unterschiede überbrückt der Emulator, indem er statt mit physischen Komponenten mit virtuellen arbeitet. Während früher ein Computerspiel von einer Datasette[62] oder eine Datenbank von einer 8“ Diskette geladen wurde, stehen diese Varianten nicht mehr zur Verfügung. Die Daten sind im Augenblick der Archivaufnahme in Abbilder der früheren Medien umgewandelt worden, die dem logischen Aufbau des originalen Datenträgers entsprechen. Die

61 „The technical registry PRONOM“, http://www.nationalarchives.gov.uk/pronom des britischen Nationalarchivs. Diese Formatregistratur kennt eine große Zahl verschiedener Datenformate. Gleichzeitig steht mit DROID ein Programm zur Ermittlung bereit.

62 Spezielles Compact-Kassenten-Gerät für die Datenspeicherung von Home-Computern, wie dem C64.

virtuellen Datenträger können sodann vom Emulator gelesen und dem nachgebildeten System wie die originalen angeboten werden.

In vielen Fällen liegen die interessierenden Primärdaten als Dateien im Host-System und noch nicht auf einem passenden virtuellen Datenträger vor. Da Emulatoren selten direkt auf Dateien im Host-System zugreifen können, müssen geeignete Wege geschaffen werden. Typischerweise lesen Computer Daten von einem Peripheriegerät wie Festplatten, optischen oder Diskettenlaufwerken. Emulierte Maschinen verwenden hierzu virtuelle Hardware. Über diese muss für einen geeigneten Transport von digitalen Objekten gesorgt werden, da der Benutzer diese normalerweise über das Host-System in die gewünschte Nutzungsumgebung einbringen wird. In einigen Fällen wird auch der umgekehrte Weg benötigt: Der Transport von Primärobjekten aus ihrer Nutzungsumgebung in die jeweils gültige Referenzumgebung.

Generell stehen eine Reihe von Varianten zur Verfügung (Abbildung 5), die jedoch von der jeweiligen Referenzplattform und ihrer technischen Ausstattung abhängen.

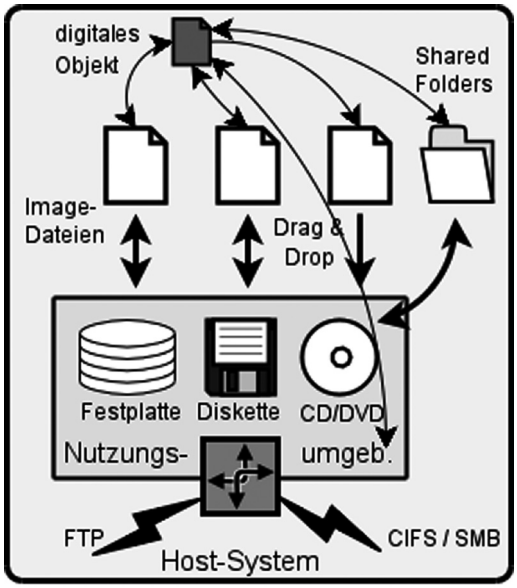


Abbildung 5: Digitale Objekte können auf verschiedene Weise zwischen der im Emulator laufenden Nutzungsumgebung und dem Host-System ausgetauscht werden.

Es lassen sich drei grundlegende Klassen von Austauschverfahren unterscheiden:

- Der klassische Weg, ohne Eingriffe in die Nutzungsumgebung und daher am universellsten einsetzbar, ist wohl der Einsatz von Datenträgerabbildern (auch als Disk-Images oder Containerdateien bezeichnet). Bei diesen handelt es sich um spezielle Dateien, welche die komplette Struktur und den vollständigen Inhalt von Datenspeichern der verschiedensten Art, wie Disketten, CD-ROMs, DVDs oder Festplatten enthalten. Der wesentliche Vorteil von Images ist der Erhalt der logischen Struktur des vormaligen Datenträgers bei Überführung in eine gut archivierbare Repräsentation.
- Im Laufe der technologischen Entwicklung begannen sich Computernetzwerke durchzusetzen. Betriebssysteme, die solche unterstützen, erlauben mittels virtueller Netzwerkverbindungen digitale Objekte zwischen Referenz- und Nutzungsumgebung auszutauschen.
- Neben diesen Verfahren bieten einige Virtualisierer und Emulatoren spezielle Wege des Datenaustauschs wie sogenannte „Shared Folders". Dieses sind spezielle Verzeichnisse im Host-System, die direkt in bestimmte Nutzungsumgebung eingeblendet werden können. Eine weitere Alternative besteht in der Nutzung der Zwischenablage[63] zum Datenaustausch.

Die notwendigen Arbeitsabläufe zum Datenaustausch können vom Erfahrungshorizont der jeweiligen Nutzer abweichen und sind deshalb auf geeignete Weise (in den Metadaten) zu dokumentieren oder zu automatisieren.

Disketten-, ISO und Festplatten-Images

Diskettenlaufwerke gehören zur Gruppe der ältesten Datenträger populärer Computersysteme seit den 1970er Jahren. Der überwiegende Anteil der Emulatoren und Virtualisierer ist geeignet, mit virtuellen Diskettenlaufwerken unterschiedlicher Kapazitäten umzugehen. Die virtuellen Laufwerke entsprechen, wie virtuelle Festplatten auch, einer Datei eines bestimmten Typs im Host-System. Das Dateiformat der virtuellen Disketten ist beispielsweise für alle virtuellen X86er identisch, so dass nicht für jeden Emulator ein eigenes Image ab-

63 Dieses entspricht einem Ausschneiden-Einfügen zwischen Host- und Nutzungsumgebung. Diese Möglichkeit muss vom Emulator unterstützt werden, da er die Brückenfunktion übernimmt. Die Art der austauschbaren (Teil-)Objekte hängen dabei vom Emulator und beiden Umgebungen ab.

gelegt werden muss und ein einfacher Austausch erfolgen kann. Die dahinter stehende einfache Technik sollte es zudem erlauben, sie auch in fernerer Zukunft zu emulieren.
Die physikalische Struktur der Diskette wird durch eine logische Blockstruktur in einer Datei ersetzt (Abbildung 6), indem die Steuerhardware des Laufwerkes und der physische Datenträger weggelassen werden. Aus Sicht des Betriebssystems im Emulator ändert sich nichts. Diskettenzugriffe werden einfach durch Zugriffe auf die Abbilddatei im Host-System umgesetzt.

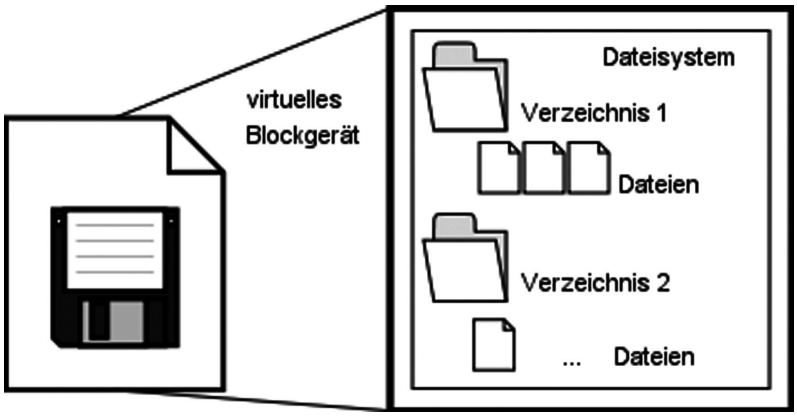


Abbildung 6: Der logische Aufbau einer Datei, die ein Disketten-Image repräsentiert. Die innere Blockstruktur kann mit einem Dateisystem versehen sein, welches Verzeichnisse und Dateien enthält.

Wegen des einfachen Aufbaus einer solchen Datei gibt es in den meisten Host-Systemen die Möglichkeit den Inhalt als virtuelles Blockgerät mit Dateisystem einzuhängen. Dadurch kann der Container so einfach wie eine normale Festplatte gelesen und beschrieben werden, wenn das Host-System das Dateisystem im Container unterstützt. Beispielsweise gelingt das Einbinden eines DOS-, VFAT- oder HFS-formatierten Disketten-Images auf Linux-Hosts ohne Schwierigkeiten, solange diese Dateisysteme und das spezielle Loop Blockdevice[64] vom Kernel[65] unterstützt werden. In Zukunft könnten langzeitrelevante

64 Das sogenannte Loop Device erlaubt Dateien innerhalb eines Dateisystems dem Betriebssystem als virtuellen Datenträger anzubieten. Auf diese Weise wird es möglich auf einzelne Komponenten innerhalb einer Containerdatei zuzugreifen.

65 Oder Kern, ist das eigentliche Betriebssystem, die zentrale Software zur Hardware- und Prozesssteuerung.

Dateisysteme mittels FUSE[66] realisiert und gepflegt werden. Für den Zugriff auf Disketten-Images in einer Referenzumgebung sind unter Umständen spezielle Hilfsprogramme notwendig, die zusätzlich archiviert werden sollten.

Disketten-Images können entweder beim Start des Emulators bereits verbunden sein. Fast alle Emulatoren unterstützen zudem das Ein- und Aushängen zur Laufzeit, welches normalerweise vom Benutzer getriggert[67] werden muss. Dieser sollte zudem darauf achten, dass eine Image-Datei nicht mehrfach eingebunden ist, da innerhalb des Containers Blockoperationen ausgeführt werden und kein Schutz vor konkurrierenden und potenziell zerstörenden Zugriffen stattfindet.

Hardwareemulatoren von Rechnerarchitekturen, die mit CD- oder DVD-Laufwerken umgehen können, verfügen über geeignete Implementierungen, um dem Gastsystem angeschlossene optische Wechseldatenträger anbieten zu können. Bei CDs und DVDs handelt es sich wie bei Disketten und Festplatten um blockorientierte Datenträger. Es besteht eine wesentliche Beschränkung: Der Datenaustausch kann nur in eine Richtung erfolgen, da einmal erstellte Datenträger ein typischerweise nur lesbares Dateisystem (ISO 9660 mit eventuellen Erweiterungen) enthalten.

CD-ROMs als Datenträger für den Einsatz im Computer gibt es erst seit Mitte der 1990er Jahre. Während Disketten quasi nativ in den meisten Plattformen unterstützt sind, muss für die Verwendung von CD-ROMs üblicherweise noch ein Treiber geladen werden, der ebenfalls im Softwareachiv abgelegt sein sollte.

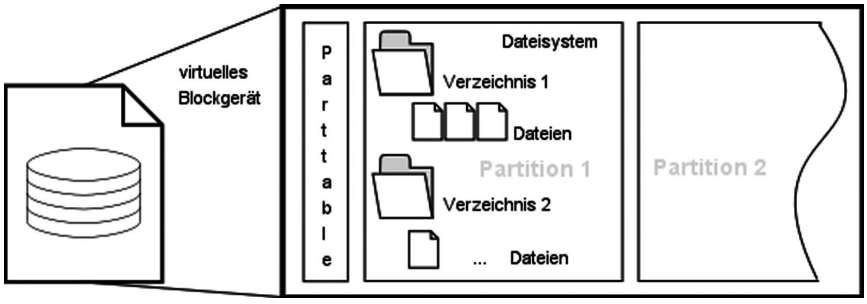


Abbildung 7: Der logische Aufbau einer Datei, die ein Festplatten-Image enthält. Dieses kann durch eine Partitionstabelle und mehrere Partitionen mit eventuell unterschiedlichen Dateisystemen strukturiert sein.

66 Filesystem im Userspace sind Dateisystemtreiber, die nicht im Betriebssystemkern implementiert sind.

67 Dieses entspricht dem traditionellen Einlegen und Auswerfen von Disketten mittels Softwarefunktion des Emulators.

Genauso wie bei Disketten, optischen Datenträgern wie CD-ROM, DVD oder Blu-ray-Disk, handelt es sich bei Festplatten um standardisierte blockorientierte Geräte. Dabei sorgen gerätespezifische Firmware und die jeweilige Hardwareplattform dafür, dass ein Betriebssystem in abstrakter Form auf das Laufwerk zugreifen kann. Spezielle Treiber, wie für die Einbindung optischer Wechseldatenträger, sind primär nicht notwendig.

Seit ungefähr 30 Jahren haben sich blockorientierte Festspeichergeräte als dauerhafter Speicher durchgesetzt. Der Inhalt dieser Speichermedien wird als virtuelles Blockgerät in Form von Dateien im Host-System repräsentiert. Ist der Aufbau dieser Dateien bekannt, so können diese ohne laufenden Emulator erzeugt und verändert werden.

Gegenüber Disketten ist bei Festplatten eine Partitionierung, eine logische Aufteilung der Gesamtkapazität, durchaus üblich.
Wegen des komplexeren inneren Aufbaus von Containerdateien wird ein einfaches direktes Einbinden im Referenzsystem fehlschlagen, wie es für Disketten- und ISO-Images vorgenommen werden kann. Stattdessen braucht man spezielle Programme, die mit Partitionen geeignet umgeben können.

Netzwerke zum Datenaustausch

Neuere Betriebssysteme verfügen über die Fähigkeit, Daten über Netzwerke auszutauschen. Viele Emulatoren enthalten deshalb virtuelle Netzwerkhardware und können zwischen Host- und Nutzungsumgebung virtuelle Datenverbindungen einrichten. Während Ende der 1980er Jahre noch proprietäre Protokolle wie Appletalk, NetBIOS und IPX/SPX neben TCP/IP existierten, verdrängte letzteres inzwischen weitestgehend die anderen. Hierbei ist bemerkenswert, dass der jetzige IPv4-Standard, welcher in seiner ersten Fassung am Ende der 1970er Jahre festgelegt wurde, im Laufe der Zeit erstaunlich stabil geblieben ist. Gleichzeitig offenbart sich der Vorteil freier, offener und damit langlebiger Standards.

Für den Austausch von digitalen Objekten bieten sich in erster Linie Protokolle an, die bereits seit vielen Jahren standardisiert sind. An prominenter Stelle stehen das seit den 1980er Jahren definierte File Transfer Protocol (FTP) und das Server Message Block Protocol (SMB). Handelt es sich bei diesen Softwarekomponenten nicht bereits um einen Teil des Betriebssystems, müssen sie bei Bedarf separat im Softwarearchiv abgelegt werden.

Bewahrung interaktiver, dynamischer Objekte

Unabhängig vom Charakter des Objekts, ob Primär- oder Sekundärobjekt, muss seine Nutzungsumgebung wiederherstellbar sein und damit kommt nur die Emulation als Strategie in Betracht. Eine zentrale Erfolgsbedingung für den Einsatz von Emulationsstrategien besteht deshalb im Aufbau und Betriebs eines Softwarearchivs. Je nach gewähltem Ansatzpunkt der Emulation wird eine Reihe zusätzlicher Softwarekomponenten benötigt (Abbildung 8)[68].

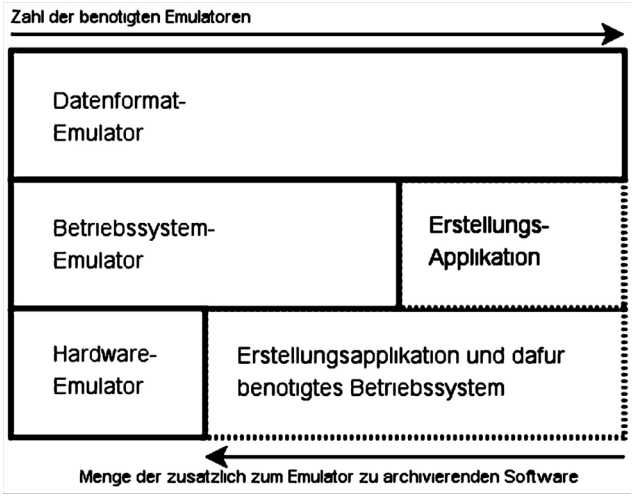


Abbildung 8: Je nach Ansatzpunkt der Emulation können eine Reihe zusätzlicher Sekundärobjekte benötigt werden.

Im Idealfall laufen die emulierten Applikationen[69] auf einer aktuellen Plattform und erlauben aus dieser heraus den direkten Zugriff auf die digitalen Objekte des entsprechenden Formates (Abbildung 1, Mitte). Solange es gelingt die entsprechende Applikation bei Plattformwechseln zu migrieren oder bei Bedarf neu zu erstellen, ist dieser Weg für die Langzeitarchivierung bestimmter Dateitypen durchaus attraktiv. Einsetzbar wäre dieses Verfahren durchaus für statische Dateitypen wie die verschiedenen offenen und dabei gut dokumentierten Bildformate.

68 Reichherzer (2006)

69 Beispielsweise ist OpenOffice ein Emulator für diverse Microsoft-Office Formate.

Die Emulation eines Betriebssystems oder dessen Schnittstellen erlaubt theoretisch alle Applikationen für dieses Betriebssystem ablaufen zu lassen. Dann müssen neben dem Emulator für das entsprechende Betriebssystem auch sämtliche auszuführende Applikationen in einem Softwarearchiv aufbewahrt werden (Abbildung 9). Bei der Portierung des Betriebssystememulators muss darauf geachtet werden, dass sämtliche in einem Softwarearchiv eingestellten Applikationen weiterhin ausgeführt werden können.

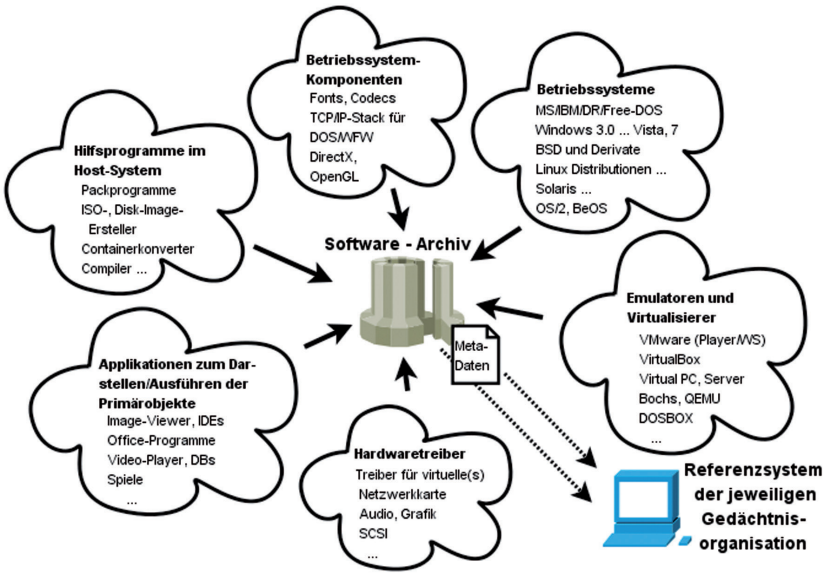


Abbildung 9: Auswahl der in einem Softwarearchiv für die X86-Architektur zu berücksichtigenden Sekundärobjekte.

Die Nachbildung einer kompletten Hardware in Software verspricht die besten Ergebnisse und verfolgt den allgemeinsten Ansatz. Dann benötigt man mindestens eines oder je nach darauf aufsetzenden Applikationen mehrere Betriebssysteme. Das bedeutet für ein Softwarearchiv, dass neben dem Emulator für eine Plattform auch die darauf ablauffähigen Betriebssysteme aufbewahrt werden müssen. Ebenso gilt dieses für die Liste der Applikationen, die zur Darstellung der verschiedenen archivierten Datentypen erforderlich sind. Erfolgt eine Portierung, also Migration des Hardwareemulators, muss anschließend überprüft werden, dass die gewünschten Betriebssysteme weiterhin ablauffähig bleiben. Da die meisten Applikationen typischerweise nicht direkt die Hardware

nutzen, sondern dafür auf die Schnittstellen der Betriebssysteme zugreifen, sollte deren Funktionsfähigkeit direkt aus der der Betriebssysteme folgen.

Virtualisierer entwickeln sich wie die nachgebildete Hardware weiter, um den technischen Entwicklungen und Marktbedürfnissen zu folgen. Damit kann es passieren, dass auch die nachgebildete Hardware für ein altes Betriebssystem zu neu ist. Dieses Problem löst beispielsweise VMware derzeit noch durch die Pflege und Bereitstellung geeigneter Treiber für jedes offiziell unterstützte Betriebssystem. Derzeit ist die Liste noch sehr lang und im Sinne der Archivierung fast vollständig. In Zukunft könnten jedoch am Ende des Feldes jedoch Betriebssysteme schrittweise herausfallen. Ein weiteres Problem von Virtualisierern ist die Art der Prozessornutzung. Viele virtuelle Maschinen reichen CPU-Befehle des Gastes direkt an die Host-CPU weiter (Beispiel VMware, Abbildung 4). Das setzt voraus, dass diese mit diesen Befehlen umgehen kann.

Geeigneter im Sinne einer wirklichen Langzeitarchivierung sind quelloffene Implementierungen.[70] Sie erlauben zum einen die Übersetzung für die jeweilige Plattform und zum anderen auch langfristige Anpassungen an völlig neue Architekturen. Zudem kann sichergestellt werden, dass auch alte Peripherie dauerhaft virtualisiert wird und nicht einem Produktzyklus zum Opfer fällt.

Ein weiterer Punkt ist die unter Umständen notwendige Anpassung der Containerdateien, in denen die Gastsysteme installiert sind. Ändert der Emulator das Datenformat, sind diese Dateien genauso wie andere digitale Objekte in ihrer Les- und Interpretierbarkeit gefährdet. Üblicherweise stellen jedoch die kommerziellen Anbieter Importfunktionen für Vorgängerversionen zur Verfügung. Bei freien, quelloffenen Emulatoren kann alternativ zur Weiterentwicklung dafür gesorgt werden, dass ein bestimmtes Dateiformat eingefroren wird. Ändert sich das Containerformat eines Emulators oder Virtualisierers müssen alle virtuellen Festplatten im Laufe der Zeit migriert werden, da sonst zu einem bestimmten Zeitpunkt keine Appliktion mehr existiert, die mit diesem umgehen kann. Damit wiederholt sich das Problem der Langzeitarchivierung – enthaltene Objekte sind nicht mehr zugreifbar und damit verloren.

---

70 Open Source Emulatoren können jederzeit von jedem Anwender mit geeigneten Programmierkenntnissen angepasst und damit für neue Plattformen übersetzt werden. Steht der Quellcode nicht zur Verfügung ist man der Produktstrategie des jeweiligen Unternehmens ausgeliefert, die häufig in deutlich kürzen Zyklen als denen der Langzeitarchivierung denkt.

Literatur

Borghoff, Uwe M. et al. (2003): *Langzeitarchivierung: Methoden zur Erhaltung digitaler Dokumente*. Heidelberg, dpunkt.verlag. ISBN 3-89864-245-3

Holdsworth, David / Wheatley, Paul (2001): *Emulation, Preservation and Abstraction*. In: RLG DigiNews, Nr. 4. Vol. 5

Rothenberg, Jeff (2000): *An Experiment in Using Emulation to Preserve Digital Publications*. http://nedlib.kb.nl/results/emulationpreservationreport.pdf

Rothenberg, Jeff (1999): *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. In: The State of Digital Preservation: An International Perspective. Conference Proceedings, Documentation Abstracts, Inc., Institutes for Information Science, Washington, D.C., April 24-25

Reichherzer, Thomas / Brown, Geoffrey (2006): *Quantifying software requirements for supporting archived office documents using emulation*. In: JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries

17.9 Web-Archivierung zur Langzeiterhaltung von Internet-Dokumenten

Andreas Rauber und Hans Liegmann (†)

Das World Wide Web hat sich in den letzten Jahren zu einem essentiellen Kommunikations- und Publikationsmedium entwickelt. Aus diesem Grund hat sich die Archivierung des Web auch zu einer wichtigen Aufgabe entwickelt, die international vor allem von Nationalbibliotheken, Staatsarchiven bzw. Institutionen mit fokussierten Sammlungsgebieten übernommen werden. Während die ersten Initiativen in diesem Bereich hochgradig experimentellen Projektcharakter hatten, existiert mittlerweile eine stabile Basis an Softwaretools und Erfahrungen zur Durchführung derartiger Projekte. In diesem Kapitel wird einerseits kurz die Geschichte der wichtigsten Webarchivierungs-Initiativen beleuchtet, sowie in der Folge detailliert auf die unterschiedlichen Sammlungsstrategien eingegangen, die zum Aufbau eines Webarchivs verwendet werden. Weiters werden Werkzeuge und Standards vorgestellt, die beim Aufsetzen einer solchen Intitiative hilfreich sind. Zum Abschluss werden offene Fragen sowie ein Ausblick auf die nächsten Herausforderungen in diesem Bereich gegeben.

Einführung

Das Web hat sich zu einem integralen Bestandteil unserer Publikations- und Kommunikationskultur entwickelt. Als solches bietet es uns einen sehr reichhaltigen Schatz an wertvollen Informationen, die teilweise ausschließlich in elektronischer Form verfügbar sind, wie z.B. Informationsportale wie Wikipedia, Informationen zu zahlreichen Projekten und Bürgerinitiativen, Diskussionsforen und Ähnlichem. Weiters beeinflussen die technischen Möglichkeiten sowohl die Art der Gestaltung von Webseiten als auch die Art, wie wir mit Information umgehen, wie unsere Gesellschaft vernetzt ist, wie sich Information ausbreitet bzw. wie sie genutzt wird. All dies stellt einen immens wertvollen Datenbestand dar, dessen Bedeutung uns erst bewusst werden mag, wenn dieser nicht mehr verfügbar ist.

Nun ist aber just diese (fehlende langfristige) Verfügbarkeit eine der entscheidenden Schwachstellen des World Wide Web. Unterschiedlichen Studien zufolge beträgt die durchschnittliche Lebensdauer eine Webressource zwischen wenigen Tagen und Wochen. So können schon binnen kürzester Zeit wertvolle Informationen nicht mehr über eine angegebene URL bezogen werden, bzw. stehen Forschern in naher und ferner Zukunft de-fakto keine Materialien zur Verfügung um diese unsere Kommunikationskultur zu analysieren. Auch Firmen haben zunehmend Probleme, Informationen über ihre eigenen Projekte,

die vielfach nicht über zentrale Dokumentmanagementsysteme sondern Webbasiert und zunehmen kollaborativ in wikiartigen Systemen abgewickelt werden, verfügbar zu halten.

Aus diesem Grund haben in den letzten Jahren vor allem Bibliotheken und Archive zunehmend die Aufgabe übernommen, neben konventionellen Publikationen auch Seiten aus dem World Wide Web zu sammeln, um so diesen wertvollen Teil unseres kulturellen Erbes zu bewahren und wichtige Informationen langfristig verfügbar zu halten. Diese massiven Datensammlungen bieten faszinierende Möglichkeiten, rasch Zugriff auf wichtige Informationen zu bekommen, die im Live-Web bereits verloren gegangen sind. Sie stellen auch eine unentbehrliche Quelle für Wissenschafter dar, die in der Zukunft die gesellschaftliche und technologische Entwicklung unserer Zeit nachvollziehen wollen.

Dieser Artikel gibt einen Überblick über die wichtigsten Fragestellungen zum Thema der Webarchivierung. Nach einer kurzen Vorstellung der wichtigsten Webarchivierungsinitiativen seit Beginn der Aktivitäten in diesem Bereich in Abschnitt 2 folgt in Abschnitt 3 eine detaillierte Darstellung der einzelnen Sammlungsstrategien und technischen Ansätzen zu ihrer Umsetzung. Abschnitt 4 fasst die einzelnen Themenbereiche, die beim Aufbau eines Webarchivs zu berücksichtigen sind, zusammen, während in Abschnitt 5 eine Reihe von Tools vorgestellt werden, die derzeit beim Aufbau von Webarchiven verwendet werden. Abschnitt 6 fasst zum Abschluss die wichtigsten Punkte nochmals kurz zusammen und bietet weiters einen Ausblick auf offene Fragestellungen, die weiterer Forschung und Bearbeitung bedürfen.

Überblick über Webarchivierungs-Projekte

Die Anfänge der Webarchivierung gehen zurück bis ins Jahr 1996, als das *Internet Archive*[71] in den USA durch Brewster Khale gegründet wurde (Brewster 1997). Ziel war es, eine „Bibliothek des Internet" aufzubauen. Ursprünglich wurden dazu die von der Suchmaschine Alexa indizierten HTML-Seiten archiviert. In weiterer Folge wurden auch andere Dateiformate wie Bilder etc. hinzugenommen, da nur so eine zuverlässige Rekonstruktion der jeweiligen Webseiten gewährleistet werden konnte – ein Hinweis auf die Tatsache, dass nicht ausschließlich die Bewahrung des textlichen Inhaltes des WWW relevant ist. Erfasst wurden dabei anfänglich nur Webseiten bis zu einer geringen Tiefe innerhalb einer Website, dafür aber für das gesamte weltweite Internet – auch dies wurde über die Jahre hinweg zunehmend ausgebaut, um die jeweiligen Websites vollständiger zu erfassen.

71 http://www.archive.org

Auf die gleiche Zeit geht das erste nationale Webarchiv zurück, das von der Royal Library in Schweden seit 1996 aufgebaut wird (*KulturarW3*) (Mannerheim et al. 2000). Dabei handelt es sich um das erste nationale Webarchiv, d.h. ein Webarchiv, welches dezidiert die Aufgabe hatte, in regelmäßigen Abständen eine Kopie des nationalen Webspace zu erstellen. Hier wird ein Crawler (ursprünglich *Combine*[72]) verwendet, um alle Seiten des nationalen Webspace in regelmäßigen Abständen zu sammeln. Erfasst werden dabei alle Dateitypen, die mit Hilfe eines Bandroboters gespeichert werden.

Neben Combine wurde im Rahmen des EU-Projekts *Nedlib* ein eigener Crawler entwickelt, der speziell für Webarchivierung bestimmt war. Dieser kam vor allem in Pilotstudien in Finnland (Hakala, 2001), Norwegen und Island zum Einsatz, wird mittlerweile jedoch nicht mehr weiterentwickelt.

Ebenfalls seit 1996 aktiv ist das Projekt *Pandora* (Webb (2001), Gatenby (2002)) der australischen Nationalbibliothek. Im Unterschied zu den bisher angeführten Projekten setzte Australien auf eine manuelle, selektive Sammlung wichtiger Dokumente. (Die Vor- und Nachteile der unterschiedlichen Sammlungsstrategien werden im folgenden Abschnitt detaillierter erläutert.)

Diese beiden Crawler (Nedlib, Combine) waren auch die Basis des an der Österreichischen Nationalbibliothek durchgeführten Pilotprojekts *AOLA – Austrian On-Line Archive*[73] (Aschenbrenner, 2005), wobei die Entscheidung letztendlich zugunsten von Combine ausfiel. Im Rahmen dieser Pilotstudie wurde eine unvollständige Sammlung des österreichischen Web erfasst. Dabei wurden sowohl Server innerhalb der nationalen Domäne .at erfasst, als auch ausgewählte Server in anderen Domänen, die sich in Österreich befanden (.com, .org, .cc). Weiters wurden explizit „Austriaca" wie z.B. das Österreichische Kulturinstitut in New York mit aufgenommen. Seit 2008 ist nunmehr eine permanente Initiative zur Webarchivierung an der österreichischen Nationalbibliothek eingerichtet.[74]

In Deutschland gibt es eine Reihe unabhängiger Webarchivierungsinitiativen. So gibt es einig Institutionen, die themenspezifische Crawls durchführen. Diese umfassen u.a. das Parlamentsarchiv des deutschen Bundestages[75] (siehe auch Kapitel 18.X), das Baden-Württembergische Online-Archiv[76], edoweb Reinland Pfalz[77], DACHS - Digital Archive for Chinese Studies[78] in Heidelberg,

72 http://combine.it.lth.se
73 http://www.ifs.tuwien.ac.at/~aola/
74 http://www.onb.ac.at/about/webarchivierung.htm
75 http://webarchiv.bundestag.de
76 http://www.boa-bw.de
77 http://www.rlb.de/edoweb.html
78 http://www.sino.uni-heidelberg.de/dachs/

und andere. Die Deutsche Nationalbibliothek hat in den vergangenen Jahren vor allem auf die individuelle Bearbeitung von Netzpublikationen und das damit erreichbare hohe Qualitätsniveau im Hinblick auf Erschließung und Archivierung gesetzt. Eine interaktive Anmeldeschnittstelle kann seit 2001 zur freiwilligen Übermittlung von Netzpublikationen an den Archivserver infodeposit.d-nb.de[79] genutzt werden. Im Herbst 2005 wurde zum Zeitpunkt der Wahlen zum Deutschen Bundestag in Kooperation mit dem European Archive[80] ein Experiment durchgeführt, um Qualitätsaussagen über die Ergebnisse aus fokussiertem Harvesting zu erhalten.

Ein drastischer Wechsel in der Landschaft der Webarchivierungs-Projekte erfolgte mit der Gründung der *International Internet Preservation Coalition (IIPC)*[81] im Jahr 2003. Im Rahmen dieses Zusammenschlusses erfolgte die Schaffung einer gemeinsamen Software-Basis für die Durchführung von Webarchivierungsprojekten. Insbesondere wurde ein neuer Crawler (HERITRIX) entwickelt, der speziell auf Archivierungszwecke zugeschnitten war – im Gegensatz zu den bisher zum Einsatz kommenden Tools, welche primär für Suchmaschinen entwickelt waren. Dieser Crawler wird mittlerweile von der Mehrzahl der Webarchivierungsprojekte erfolgreich eingesetzt. Weitere Tools, die im Rahmen des IIPC entwickelt werden, sind Nutch/Wax als Indexing-/Suchmaschine, sowie Tools für das Data Management und Zugriff auf das Webarchiv. Weiters wurde im Rahmen dieser Initiative das ARC-Format als de-facto Standard für Webarchiv-Dateien etabliert und mittlerweile als WARC[82] an die neuen Anforderungen angepasst. (Eine detailliertere Beschreibung dieser Tools findet sich in Abschnitt 5 diese Kapitels).

Inzwischen werden weltweit zahlreiche Webarchivierungsprojekte durchgeführt (USA, Australien, Singapur, …). Auch die Mehrzahl der europäischen Länder hat eigene Webarchivierungsprojekte eingerichtet. Entsprechende Aktivitäten werden z.B. von der Isländischen Nationalbibliothek, Königlichen Bibliothek in Norwegen, Nationalbibliotheken in Schweden, Dänemark und Frankreich als Teil des IIPC durchgeführt. In Großbritannien existieren zwei parallele Initiativen: einerseits das UK Webarchive Consortiums, sowie für die Regierungs-Webseiten eine Initiative des Nationalarchivs. Italien hat das European Webarchive mit der Erstellung eines nationalen Snapshot beauftragt. Eigenständige Aktivitäten existieren weiters in Tschechien (Nationalbibliothek

79 http://www.d-nb.de/netzpub/index.htm
80 http://europarchive.org
81 http://netpreserve.org
82 http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

in Kooperation mit der Bibliothek in Brno) sowie Slowenien, ebenfalls an der Nationalbibliothek angesiedelt.

Ein guter Überblick zu den Problemstellungen im Bereich Web Archivierung, Erfahrungsberichte einzelner Initiativen, sowie eine detaillierte Auflistung der Schritte zum Aufbau von Webarchiven finden sich in (Brown (2006), Masanes (2006)). Ein Forum zum internationalen Erfahrungsaustausch ist der jährlich stattfindende Internationale Workshop on Web Archiving (IWAW[83]). Im Rahmen dieses Workshops werden sowohl wissenschaftliche Beiträge präsentiert, als auch insbesondere eine Reihe von Best-Practice Modellen bzw. Erfahrungsberichte aus den einzelnen Projekten diskutiert. Die Beiträge sind als on-line Proceedings auf der Website der Workshopserie frei verfügbar.

Sammlung von Webinhalten

Grundsätzlich können vier verschiedene Arten der Datensammlung zum Aufbau eines Webarchivs, sowie einige Sonderformen unterschieden werden:

Snapshot Crawls:

Hierbei wird versucht, ausgehend von einer Sammlung von Startseiten (sog. Seed-URLs) den gesamten nationalen Webspace zu sammeln. Jede gefundene Seite wird auf weiterführende Links analysiert, diese werden zur Liste der zu sammelnden Seiten hinzugefügt. Unter der Annahme, dass alle Webseiten in irgendeiner Weise miteinander verlinkt sind, kann so der gesamte nationale Webspace prinzipiell erfasst werden – wobei natürlich keine Garantie dafür abgegeben werden kann, dass alle Websites entsprechend verlinkt sind. Üblicherweise kann mit Hilfe dieser Verfahren ein sehr großer Teil, jedoch keinesfalls der vollständige Webspace erfasst werden. Irreführend ist weiters die für diese Art der Datensammlung übliche Bezeichnung „Snapshot", da es sich dabei keineswegs – wie die Übersetzung vermuten ließe – um eine „Momentaufnahme" des nationalen Webspace handelt, sondern eher – um bei der Metapher zu bleiben – um eine Langzeitbelichtung, deren Erstellung mehrere Monate in Anspruch nimmt.

Im Rahmen dieser Snapshot-Erstellung muss definiert werden, was als „nationaler Webspace„ erfasst werden soll. Dieser umfasst primär alle Websites, die in der entsprechenden nationalen Top-Level Domäne (z.B. „*.at*", „*.de*" oder „*.ch*" für Österreich, Deutschland und die Schweiz) angesiedelt sind, sowie Websites, die in anderen Top-level Domänen (z.B. .com, .org, .net, .cc, etc.)

83 http://www.iwaw.net

gelistet sind, jedoch geographisch in den jeweiligen Ländern beheimatet sind. Diese können von den entsprechenden Domain Name Registries in Erfahrung gebracht werden. Weiters werden zur Erstellung eines Archivs des nationalen Webspace auch Sites erfasst, die weder unter der jeweiligen nationalen Domäne firmieren, noch im jeweiligen Land angesiedelt sind, sich jedoch mit Themen mit Länder-Bezug befassen. Diese müssen manuell ermittelt und in den Sammlungsbereich aufgenommen werden. Üblicherweise werden solche Snapshot-Archivierungen 1-4 mal pro Jahr durchgeführt, wobei jeder dieser Crawls mehrere TB an Daten umfasst.

Event Harvesting / Focused Crawls

Da die Erstellung eines Snapshots längere Zeiträume in Anspruch nimmt, eignet er sich nicht zur ausreichenden Dokumentation eines bestimmten Ereignisses. Zu diesem Zweck werden zusätzlich zu den „normalen" Snapshot-Archivierungen auch so genannte Focused Crawls durchgeführt. Bei diesen wird eine kleine Anzahl von Websites zu einem bestimmten Thema zusammengestellt und diese mit erhöhter Frequenz (täglich, wöchentlich) durch einen Crawler gesammelt. Typische Beispiele für solche Focused Crawls bzw. Event Harvests sind üblicherweise Wahlen, sportliche Großereignisse, oder Katastrophen (vgl. Library of Congress / Internet Archive: Sammlungen zu den Presidential Elections, zu 9/11; Netarchive.dk: Sondersammlung zum dänischen Mohammed-Karikaturen-Streit, etc.) Diese Sondersammlungen werden üblicherweise durch Kuratoren initiiert, wobei bestimmte Aktivitäten bereits für das jeweilige Jahr im Voraus geplant werden, andere tagesaktuell bei Bedarf initiiert werden.

Selective Harvesting

Dies ist eine Sonderform des Focused Crawls, der sich auf spezifische Websites konzentriert. Dieser Ansatz wird für Websites angewandt, die in regelmäßigen Abständen in das Archiv aufgenommen werden sollen, um eine vollständige Abdeckung des Inhalts zu gewährleisten. Üblicherweise wird dieser Ansatz vor allem bei Periodika angewandt, die z.B. täglich, wöchentlich etc. in das Archiv kopiert werden. Hierbei kann zusätzlich der Crawling-Prozess auf die jeweilige Website optimiert werden, um nur relevante Information in hoher Frequenz zu übernehmen. So werden z.B. oft nur die entsprechenden Nachrichtenartikel unter Ausblendung von Diskussionsforen, Werbung, oder on-line Aktionen, die laut entsprechender Sammlungsstrategie nicht ins Archiv Eingang finden sollen, regelmäßig mit hoher Frequenz kopiert.

Manual Collection / Submission

Manuelle Sammlung wird einerseits für Websites verwendet, die nicht durch Crawler automatisch erfassbar sind. Dabei handelt es sich meist um Websites, die aus Datenbanken (Content Management Systemen) generiert werden, die nicht durch Linkstrukturen navigierbar sind, sondern z.B. nur ein Abfrage-Interface zur Verfügung stellen (Deep Web, siehe „Sonderformen" unten). In anderen Fällen kann eine Kopie von Netzpublikationen über ein spezielles Web-Formular vom Eigentümer selbst abgeliefert werden. Weiters können bestimmte einzelne Webseiten oder wichtige Dokumente aus dem Netz selektiv in ein manuell verwaltetes und gepflegtes Archiv übernommen werden. Diese werden allerdings üblicherweise nicht in das „normale" Webarchiv übernommen, sondern gesondert in einen Datenbestand (z.B. OPAC) eingepflegt.

Sonderformen

Eine Sonderform stellt die Archivierung des Deep Web" dar. Dabei handelt es sich um Webseiten, die nicht statisch vorliegen, sondern die basierend auf Anfragen dynamisch aus einer Datenbank generiert werden. (z.B. Telefonbuch, Kataloge, geographische Informationssysteme, etc.) In diesen Fällen wird meist die Datenbank direkt nach Absprache mit dem Provider kopiert und für Archivzwecke umgewandelt, um die Information zu bewahren.

Ein anderer Ansatz, der die interaktive Komponente des Internet stärker betont, ist Session Filming. Dabei werden die Aktivitäten am Bildschirm mittels Screen-Grabbern „gefilmt", während BenutzerInnen bestimmte Aufgaben im Internet erledigen, und somit die Eigenschaft der Interaktion dokumentiert (z.B. Dokumentation, wie eine Internet-Banking Applikation im Jahr 2002 abgelaufen ist – inklusive Antwortzeiten, Arbeitsabläufe, Ablauf von Chat-Sessions, Netz-Spiele, etc.).

Zusätzlich werden weitere Sondersammlungen angelegt, die spezifische Quellen aus dem Internet ins Archiv übernehmen, wie zum Beispiel ausgewählte Videos der Plattform YouTube[84] (Shah 2007). Diese Ansätze werden meist ergänzend durchgeführt – sie stellen jedoch üblicherweise Sondersammlungen innerhalb eines Webarchivs dar.

Kombinationsstrategien

Die meisten Initiativen zum Aufbau eines Webarchivs verwenden derzeit eine Kombination der oben angeführten Strategien, d.h. regelmäßige Snapshots (1-2

84 http://www.youtube.com

mal pro Jahr), kombiniert mit fokussierten Sammlungen und Selective Crawling. Auf jeden Fall herrscht mittlerweile fast einstimmig die Meinung, dass ein rein selektiver Ansatz, d.h. die ausschließliche Erfassung manuell ausgewählter „wichtiger" Websites keine akzeptable Strategie darstellt, da auf diese Weise kein repräsentativer Eindruck des jeweiligen nationalen Webspace gegeben werden kann. Aus diesem Grund sind mittlerweile beinahe alle Initiativen, die ursprünglich auf rein manuelle Datensammlung gesetzt haben (z.B. Australien), dazu übergegangen, auch breites Snapshot Crawling in ihre Sammlungsstrategie aufzunehmen.

Sammlungsstrategien

Nationalbibliotheken fassen grundsätzlich alle der im World Wide Web erreichbaren Dokumente als Veröffentlichungen auf und beabsichtigen, ihre Sammelaufträge entsprechend zu erweitern, soweit dies noch nicht geschehen ist. Eine Anzahl von Typologien von Online-Publikationen wurde als Arbeitsgrundlage geschaffen, um Prioritäten bei der Aufgabenbewältigung setzen zu können und der Nutzererwartung mit Transparenz in der Aufgabenwahrnehmung begegnen zu können. So ist z.B. eine Klassenbildung, die mit den Begriffen „druckbildähnlich" und „webspezifisch" operiert, in Deutschland entstanden (Wiesenmüller 2004). In allen Nationalbibliotheken hat die Aufnahme von Online-Publikationen zu einer Diskussion von Sammel-, Erschließungs- und Archivierungsverfahren geführt, da konventionelle Geschäftsgänge der Buch- und Zeitschriftenbearbeitung durch neue Zugangsverfahren, die Masse des zu bearbeitenden Materials und neue Methoden zur Nachnutzung von technischen und beschreibenden Metadaten nicht anwendbar waren. Die neue Aufgabe von Gedächtnisorganisationen, die langfristige Verfügbarkeit digitaler Ressourcen zu gewährleisten, hat zu neuen Formen der Kooperation und Verabredungen zur Arbeitsteilung geführt.

Ein „Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications"[85] (CENL/FEP 2005) der Conference of European National Librarians (CENL) und der Federation of European Publishers (FEP) hat folgende Prinzipien im Umgang zwischen Verlagen und nationalen Archivbibliotheken empfohlen (unabhängig davon, ob sie gesetzlich geregelt werden oder nicht):

85 http://www.nlib.ee/cenl/docs/05-11CENLFEP_Draft_Statement050822_02.pdf

- Ablieferung digitaler Verlagspublikationen an die zuständigen Bibliotheken mit nationaler Archivierungsfunktion
- Geltung des Ursprungsland-Prinzip für die Bestimmung der Depotbibliothek, ggf. ergänzt durch den Stellenwert für das kulturelle Erbe einer europäischen Nation
- Einschluss von Publikationen, die kontinuierlich verändert werden (Websites) in die Aufbewahrungspflicht
- nicht im Geltungsbereich der Vereinbarung sind: Unterhaltungsprodukte (z.B. Computerspiele) und identische Inhalte in unterschiedlichen Medienformen (z.B. Online-Zeitschriften zusätzlich zur gedruckten Ausgabe).

Das Statement empfiehlt, technische Maßnahmen zum Schutz des Urheberrechts (z.B. Kopierschutzverfahren) vor der Übergabe an die Archivbibliotheken zu deaktivieren, um die Langzeitverfügbarkeit zu gewährleisten.
Zur Definition einer Sammlungsstrategie für ein Webarchiv müssen eine Reihe von Entscheidungen getroffen und dokumentiert werden. Dies betrifft einerseits die Definition des jeweiligen Webspace, der erfasst werden soll (z.B. in wie weit Links auf Webseiten im Archiv, die auf externe Seiten außerhalb des nationalen Webspace zeigen, auch erfasst werden sollen). Weiters ist zu regeln (und rechtlich zu klären), ob Robot Exclusion Protokolle (siehe unten) respektiert werden, oder ob Passwörter für geschützte Seiten angefordert werden sollen. Weitere Entscheidungen betreffend die Art und Größe der Dokumente, die erfasst werden sollen – insbesondere für Multimedia-Streams (z.B. bei Ausstrahlung eines Radioprogramms über das Internet); ebenso müssen Richtlinien festgelegt werden, welche Arten von Webseiten häufiger und mit welcher Frequenz gesammelt werden sollen (Tageszeitungen, Wochenmagazine, Seiten öffentlicher Institutionen, Universitäten, …) bzw. unter welchen Bedingungen ein bestimmtes Ereignis im Rahmen einer Sondersammlung erhoben werden soll. Diese Sondersammlungen können dann weiters auch in einem zentralen Katalogsystem erfasst und somit auch direkt über dieses zugänglich gemacht werden. Üblicherweise werden in der Folge von geschulten Fachkräften, die insbesondere diese Sondersammlungen verwalten, entsprechende Crawls gestartet und von diesen auch auf Qualität geprüft.

In diesem Zusammenhang soll nicht unerwähnt bleiben, dass die technischen Instrumentarien zur Durchführung zurzeit noch mit einigen Defiziten behaftet sind:

- Inhalte des so genannten „deep web“ sind durch Crawler nicht erreichbar. Dies schließt z.B. Informationen ein, die in Datenbanken oder Content Management Systemen gehalten werden. Crawler sind noch nicht in

der Lage, auf Daten zuzugreifen, die erst auf spezifische ad-hoc-Anfragen zusammengestellt werden und nicht durch Verknüpfungen statischer Dokumente repräsentiert sind.
- Inhalte, die erst nach einer Authentisierung zugänglich sind, entziehen sich verständlicherweise dem Crawling-Prozess.
- dynamische Elemente als Teile von Webseiten (z.B. in Script-Sprachen) können Endlosschleifen (Crawler traps) verursachen, in denen sich der Crawler verfängt.
- Hyperlinks in Web-Dokumenten können so gut verborgen sein (deep links), dass der Crawler nicht alle Verknüpfungen (rechtzeitig) verfolgen kann und im Ergebnis inkonsistente Dokumente archiviert werden.

Vor allem bei der Ausführung großen Snapshot Crawls führen die genannten Schwächen häufig zu Unsicherheiten über die Qualität der erzielten Ergebnisse, da eine Qualitätskontrolle aufgrund der erzeugten Datenmengen nur in Form von Stichproben erfolgen kann. Nationalbibliotheken verfolgen deshalb zunehmend Sammelstrategien, die das Web-Harvesting als eine von mehreren Zugangswegen für Online-Publikationen etablieren.

Aufbau von Webarchiven

Durchführung von Crawls

Zur automatischen Datensammlung im großen Stil wird in laufenden Projekten als Crawler meist HERITRIX eingesetzt. Durch den Zusammenschluss wichtiger Initiativen innerhalb des IIPC stellen die innerhalb dieses Konsortiums entwickelten Komponenten eine stabile, offene und gemeinsame Basis für die Durchführung von Webarchivierungsaktivitäten dar. Als Crawler, der explizit für Archivierungszwecke entwickelt wurde, vermeidet er einige der Probleme, die bei zuvor entwickelten Systemen für Suchmaschinen bestanden.
Um eine möglichst gute Erfassung des nationalen Webspace zu erreichen, sind einige Konfigurationen vorzunehmen. Dieses „Crawl Engineering“ ist eine der Kernaufgaben im Betrieb eines Webcrawling-Projekts und erfordert eine entsprechende Expertise, um vor allem für große Snapshot-Crawls effizient einen qualitativ hochwertigen Datenbestand zu erhalten.

Robot Exclusion Protokolle erlauben den Betreibern von Websites zu spezifizieren, inwieweit sie Crawlern erlauben, ihre Webseite automatisch zu durchsuchen. Auf diese Weise können zum Beispiel gewisse Bereiche des Webspace für automatische Crawler-Programme gesperrt werden oder nur bestimmte Crawler zugelassen werden (z.B. von einer bevorzugten Suchmaschine). Üblicherwei-

se sollten diese Robot Exclusion Protokolle (robots.txt) befolgt werden. Andererseits haben Studien in Dänemark ergeben, dass just Websites von großem öffentlichen Interesse (Medien, Politische Parteien) sehr restriktive Einstellungen betreffend Robot Exclusion hatten. Aus diesem Grund sieht die gesetzliche Regelung in manchen Ländern vor, dass für den Aufbau des Webarchivs diese Robot Exclusion Protokolle nicht gelten und nicht befolgt werden müssen. Zu bedenken ist, dass manche Informationsanbieter Gebühren entsprechend dem anfallenden Datentransfervolumen bezahlen. Sie schließen daher oftmals große Bereiche ihrer Websites mittels robots.txt vom Zugriff durch Webcrawler aus – womit ein Crawler, der dieses Konzept ignoriert, unter Umständen hohe Kosten verursacht.

Speicherung

Für die Speicherung der vom Crawler gesammelten Dateien hat sich das ARC bzw. WARC Format als de-facto Standard durchgesetzt. Diese Dateien sind XML-basierte Container, in denen die einzelnen Webdateien zusammengefasst und als solche in einem Speichersystem abgelegt werden. Üblicherweise werden in diesen Containern jeweils Dateien bis zu einer Größe von 100 MB zusammengefasst. Über dieses werden verschiedene Indexstrukturen gelegt, um auf die Daten zugreifen zu können. Betreffend Speicherung ist generell ein Trend zur Verwendung hochperformanter Speichersysteme, meist in Form von RAID-Systemen, zu erkennen.

Zugriff

Mit Ausnahme des Internet Archive in den USA bietet derzeit keines der über großflächiges Crawling aufgebauten Webarchive freien, öffentlichen Zugriff auf die gesammelten Dateien an. Dies liegt einerseits an ungenügenden rechtlichen Regelungen betreffend *Copyright*, andererseits bestehen auch Bedenken bezüglich des Schutzes der *Privatsphäre*. Dies liegt darin begründet, dass das World Wide Web nicht nur eine Publikationsplattform, sondern auch eine Kommunikationsplattform ist. Somit fallen viele der Webseiten eher in den Bereich eines „schwarzen Bretts" bzw. werden Postings auf Blogs oder Kommentarseiten von vielen BenutzerInnen nicht als „Publikation" gesehen. Durch die Sammlung personenbezogener Informationen über lange Zeiträume bestehen Bedenken hinsichtlich einer missbräuchlichen Verwendung der Informationen (Rauber, 2008) (Beispiel: Personalabteilung, die Informationen über BewerberInnen bis ins Kindesalter zurückverfolgt). Aus diesen Gründen gewähren viele Archive

derzeit noch keinen oder nur eingeschränkten Zugriff und warten rechtliche sowie technologische Lösungen ab, um diesen Problemen zu begegnen.

Andererseits bietet das Internet Archiv von Beginn an öffentlichen Zugriff auf seine Daten und entfernt Webseiten auf Anforderung, bzw. nimmt keine Daten in das Archiv auf, die durch das Robot Exclusion Protokoll geschützt sind. Bisher kam es zu keinen nennenswerten Klagen oder Beschwerdefluten. Andererseits sind einzelne Klagen aus den skandinavischen Ländern bekannt, in denen es primär um das Recht der Sammlung der Daten ging, die jedoch zugunsten des Sammlungsauftrags der Nationalbibliotheken entschieden wurden. Dennoch sollten diese Bedenken zum Schutz der Privatsphäre ernst genommen werden.

Langzeitarchivierung

Abgesehen von der redundanten Speicherung werden derzeit von den einzelnen Webarchivierungsprojekten kaum Schritte betreffend einer dezidierten Langzeit-Archivierung gesetzt. Insbesondere werden keine Migrationsschritte etc. durchgeführt. Dies kann teilweise damit begründet werden, dass ein Webarchiv inhärent unvollständig ist, und somit ein höheres Risiko hinsichtlich des Verlusts einzelner weniger Seiten eingegangen werden kann. Andererseits stellt ein Webarchiv durch die Heterogenität des Datenmaterials eine der größten Herausforderungen für die Langzeitarchivierung dar.

Werkzeuge zum Aufbau von Webarchiven

Es gibt mittlerweile eine Reihe von Werkzeugen, die als Open Source Komponenten zur Verfügung stehen. Erwähnenswert sind insbesondere folgende Softwarepakete:

HERITRIX

Heritrix[86] ist ein vom Internet Archive in den USA speziell für Webarchivierungszwecke entwickelter Crawler, der unter der GNU Public License verfügbar ist. Dieser Crawler wird von einer großen Anzahl von Webarchivierungsprojekten eingesetzt, und ist somit ausgiebig getestet. Er hat mittlerweile eine Stabilität erreicht, die einen laufenden Betrieb und die Durchführung großer Crawls ermöglicht. Aktuelle Verbesserungen betreffen vor allem eine höhere Intelligenz des Crawlers z.B. zur automatischen Vermeidung von Duplikaten,

86 http://crawler.archive.org

sowie eine flexibere Gestaltung des Crawling-Prozesses. Daten werden in ARC-files gespeichert.

HTTRACK

HTTRACK[87] ist ebenfalls ein Crawler, der jedoch für selektives Harvesting einzelner Domänen eingesetzt wird. Er ist sowohl über ein graphisches Interface als auch als Command-line Tool steuerbar und legt die Dateien in einer lokalen Kopie entsprechend der vorgefundenen Struktur am Webserver ab.

NetarchiveSuite

Die NetarchiveSuite[88] wurde seit dem Jahr 2004 im Rahmen des Netarchive Projekts in Dänemark entwickelt und eingesetzt. Sie dient zur Planung und Durchführung von Harvestingaktivitäten mit Hilfe des Heritrix Crawlers. Die Software unterstützt bit-level preservation, das heisst redundante Speicherung und Prüfung der Objekte. Die Software kann auf mehreren Rechnern verteilt ausgeführt werden.

NutchWAX

Nutchwax[89] ist eine in Kooperation zwischen dem Nordic Web Archive, dem Internet Archive und dem IIPC entwickelte Suchmaschine für Daten in einem Webachiv. Konkret baut NutchWAX auf ARC-Daten auf und erstellt Index-Strukturen, die eine Volltextsuche ermöglichen.

WERA

WERA[90] ist ein php-basiertes Interface, das auf den Tools des Nordic Web Archive, bzw. nunmehr auch NutchWAX aufbaut und eine Navigation im Webarchiv ermöglicht. Die Funktionalität ist vergleichbar mit jener der WayBack-Machine des Internet Archive, erweitert um Volltextsuche in den Archivdaten.

WayBack Machine

Die WayBack Machine[91] erlaubt - ähnlich wie WERA – den Zugriff auf das Webarchiv. Sie wird vom Internet Archive entwickelt, basiert rein auf Java, und unterstützt

87 http://www.httrack.com
88 http://netarchive.dk/suite
89 http://archive-access.sourceforge.net/projects/nutch
90 http://archive-access.sourceforge.net/projects/wera
91 http://archive-access.sourceforge.net/projects/wayback

zusätzlich zur Funktionalität von WERA einen Proxy-basierten Zugriff, d.h. alle Requests, alle Anfragen, die vom Webbrowser ausgehend von Archivdaten abgesetzt werden, können direkt wieder in das Archiv umgeleitet werden. (Tofel, 2007)

WCT - Web Curator Tool

Das Web Curator Tool[92], in Kooperation mit der British Library und der Nationalbibliothek von Neuseeland von Sytec Resources entwickelt, ist unter der Apache License als Open Source verfügbar. Es bietet ein Web-basiertes User Interface für den HERITRIX Crawler zur Steuerung von Selective Harvesting Crawls bzw. Event Harvesting. Ziel ist es, mit Hilfe dieses Interfaces die Durchführung von Crawls ohne spezielle IT-Unterstützung zu ermöglichen. Mit diesem Tool können BibliothekarInnen thematische Listen von Websites zusammenstellen und diese als Sondersammlungen in das Webarchiv integrieren.

DeepArc

DeepArc[93] ist ein Tool, das von der französischen Nationalbibliothek gemeinsam mit XQuark entwickelt wurde. Es dient zur Archivierung von Datenbanken, indem relationale Strukturen in ein XML-Format umgewandelt werden. Im Rahmen von Webarchivierungsprojekten wird es vorallem für den sogenannten „Deep-Web"-Bereich eingesetzt.

Zusammenfassung und Ausblick

Die Archivierung der Inhalte des Web ist von essentieller Bedeutung, um diese Informationen für zukünftige Nutzung retten zu können. Dies betrifft die gesamte Bandbreite an Webdaten, angefangen von wissenschaftlichen (Zwischen) ergebnissen, online Publikationen, Wissensportalen, elektronischer Kunst bis hin zu Diskussionsforen und sozialen Netzwerken. Nur so können wertvolle Informationen verfügbar gehalten werden, die es zukünftigen Generationen ermöglichen werden, unsere Zeit und Gesellschaft zu verstehen.

Andererseits wirft die Sammlung derartig enormer Datenbestände in Kombination mit den zunehmend umfassenderen technischen Möglichkeiten ihrer Analyse berechtigte ethische Fragestellungen auf. Welche Daten dürfen gesammelt und zugänglich gemacht werden? Gibt es Bereiche, die nicht gesammelt werden sollen, oder die zwar zugreifbar, aber von der automatischen Analyse ausgeschlossen sein sollten. Können Modelle entwickelt werden, die sowohl

92 http://webcurator.sourceforge.net
93 http://deeparc.sourceforge.net

eine umfassende Webachivierung erlauben, andererseits aber auch ethisch unbedenklich umfassenden Zugang zu (Teilen) ihrer Sammlung gewähren dürfen? Denn nur durch möglichst umfangreichen Zugriff können Webarchive ihr Nutzpotential entfalten. Die mit Webarchivierung befassten Institutionen sind sich ihrer Verantwortung in diesem Bereich sehr wohl bewusst. Aus diesem Grund sind daher derzeit fast alle derartigen Sammlungen nicht frei zugänglich bzw. sehen Maßnahmen vor um dem Nutzer Kontrolle über seine Daten zu geben. Nichtsdestotrotz sind weitere Anstrengungen notwendig, um hier eine bessere Nutzung unter Wahrung der Interessen der Betroffenen zu ermöglichen. (Rauber, 2008)

Allerdings sind diese ethischen Fragestellungen bei weitem nicht die einzigen Herausforderungen, mit denen Webarchivierungsinitiativen derzeit zu kämpfen haben. Die Größe, Komplexität des Web sowie der rasche technologische Wandel bieten eine Unzahl an enormen technischen Herausforderungen, deren Behandlung die zuvor aufgeführten Probleme oftmals verdrängt. So stellt alleine die Aufgabe, diese Daten auch in ferner Zukunft nutzbar zu haben, enorme Herausforderungen an die digitale Langzeitarchivierung – ein Thema, das schon in viel kontrollierbareren, konsistenteren Themenbereichen erheblichen Forschungs- und Entwicklungsaufwand erfordert. Die Problematik der digitalen Langzeitarchivierung stellt somit eine der größten technologischen Herausforderungen dar, der sich Webarchive mittelfristig stellen müssen, wenn sie ihre Inhalte auch in mittlerer bis ferner Zukunft ihren Nutzern zur Verfügung stellen wollen.

Weiters erfordern die enormen Datenmengen, die in solchen Archiven über die Zeit anfallen, völlig neue Ansätze zur Verwaltung, und letztendlich auch zur Analyse und Suche in diesen Datenbeständen – bieten doch diese Archive kombiniert nicht nur den Datenbestand diverser populärer Websuchmaschinen, sondern deren kumulativen Datenbestand über die Zeit an.

Somit stellt die Archivierung der Inhalte des World Wide Web einen extrem wichtigen, aber auch einen der schwierigsten Bereiche der Langzeitarchivierung Digitaler Inhalte, sowohl hinsichtlich der technischen, aber auch der organisatorischen Herausforderungen dar.

Bibliographie

Brown, Adrian (2006): *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publisching.

CENL/FEP Committee (2005): *Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications*. In: Proceedings Annual Conference of European National Libraries, Luxembourg.

Gatenby, Pam (2002) : *Legal Deposit, Electronic Publications and Digital Archiving. The National Library of Australia's Experience*. In: 68th IFLA General Conference and Council, Glasgow.

Hakala, Juha (2001): *Collecting and Preserving the Web: Developing and Testing the NEDLIB Harvester*. In: RLG DigiNews 5, Nr. 2.

Kahle, Brewster (1997): *Preserving the Internet. Scientific American,* March 1997.

Mannerheim, Johan, Arvidson, Allan und Persson, Krister (2000): *The Kulturarw3 project – The Royal Swedish Web Archiw3e. An Example of »Complete« Collection of Web Pages*. In: Proceedings of the 66th IFLA Council and General Conference, Jerusalem, Israel.

Masanes, Julien (Hrsg.) (2006): *Web Archiving*. Springer.

Aschenbrenner, Andreas und Rauber, Andreas (2005): *Die Bewahrung unserer Online-Kultur. Vorschläge und Strategien zur Webarchivierung*. In: Sichtungen, 6/7, Turia + Kant. 99-115.

Rauber, Andreas, Kaiser, Max und Wachter, Bernhard (2008): *Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda*. In: Proceedings of the 8th International Web Archiving Workshop, Aalborg, Dänemark

Shah, Chirag, Marchionini, Gary (2007): *Preserving 2008 US Presdential Election Videos*. In: Proceedings of the 7th International Web Archiving Workshop, Vancouver, Kanada.

Tofel, Brad (2007): *"Wayback" for Accessing Web Archives*. In: Proceedings of the 7th International Web Archiving Workshop, Vancouver, Kanada.

Webb, Colin und Preiss, Lydia (2001): *Who will Save the Olympics? The Pandora Archive and other Digital Preservation Case Studies at the National Library of Australia*. In: Digital Past, Digital Future – An Introduction to Digital Preservation. OCLC / Preservation Resources Symposium.

Wiesenmüller, Heidrun et al. (2004): *Auswahlkriterien für das Sammeln von Netzpublikatio-nen im Rahmen des elektronischen Pflichtexemplars*. In: Bibliotheksdienst 38, H. 11, 1423-1444.

17.10 Digitale Forschungsdaten

Jens Klump

Einführung

Durch eine Reihe von Aufsehen erregenden Wissenschaftsskandalen in den neunziger Jahren des 20. Jahrhunderts sah sich die Deutsche Forschungsgemeinschaft (DFG) gezwungen, „Empfehlungen für eine gute wissenschaftliche Praxis" auszusprechen, die in vergleichbarer Form auch von anderen Wissenschaftsorganisationen übernommen wurden. In ihren Empfehlungen bezieht sich die DFG auf Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren. Sie verlangt von ihren Zuwendungsempfängern, dass diese Daten für mindestens zehn Jahre auf geeigneten Datenträgern sicher aufbewahrt werden müssen.[94] Für die einzelnen Disziplinen ist der Umgang mit Daten im einzelnen zu klären, um eine angemessene Lösung zu finden.[95] Diese Policy dient jedoch in erster Linie einer Art Beweissicherung; über Zugang zu den Daten und ihre Nachnutzbarkeit sagen die Empfehlungen nichts aus. Zudem ist bisher noch kein Fall bekannt geworden, in dem die DFG Sanktionen verhängt hätte, allein weil der Pflicht zur Archivierung von Daten nicht nachgekommen wurde.

Auf Einladung der DFG wurde im Januar 2008 in einem Rundgespräch der Umgang mit Forschungsdaten und deren Bereitstellung diskutiert. Auch die Auswirkungen der „Empfehlungen" von 1998 wurden einem kritischen Review unterzogen.[96] Als Ergebnis dieser Konsultationen und der darauf folgenden Entscheidungen der Fachausschüsse veröffentlichte die DFG im Januar 2009 ihre „Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten".[97]

Welche Daten sollen und müssen archiviert werden? Die Empfehlungen der DFG sprechen von Forschungsprimärdaten. Dieser Begriff „Primärdaten" sorgt immer wieder für Diskussion, denn die Definition des Begriffs ist sehr von der eigenen Rolle in der wissenschaftlichen Wertschöpfungskette bestimmt. Für den einen sind „Primärdaten" der Datenstrom aus einem Gerät, z.B. einem Satelliten. In der Fernerkundung werden diese Daten „Level 0"

94 DFG (1998), Empfehlung 7
95 DFG (1998), Empfehlung 1
96 Kluttig (2008)
97 DFG (2009)

Produkte genannt. Für andere sind „Primärdaten“ zur Nachnutzung aufbereitete Daten, ohne weiterführende Prozessierungsschritte. Wieder andere differenzieren nicht nach Grad der Verarbeitung sondern betrachten alle Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren, als Primärdaten. Der begrifflichen Klarheit wegen sollte daher das Präfix „Primär-“ nicht mehr verwendet werden und statt dessen nur noch von wissenschaftlichen Daten oder Forschungsdaten gesprochen werden.

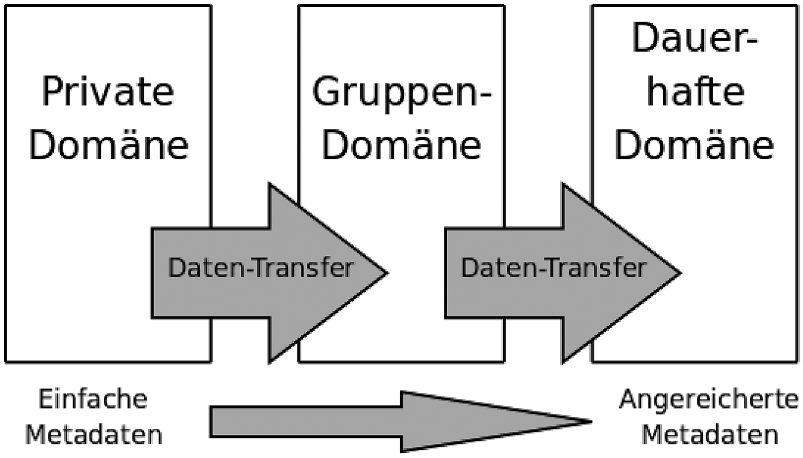
Welche Definition des Begriffs man auch wählt, wissenschaftliche Daten sind geprägt durch ihre Herkunft aus experimentellem Vorgehen, d.h. anders als Daten aus Arbeitsabläufen der Industrie oder Verwaltung stammen Forschungsdaten überwiegend aus informellen Arbeitsabläufen, die immer wieder ad hoc an die untersuchte Fragestellung angepasst werden.[98] Da in den meisten Fällen keine Formatvorgaben vorhanden sind, werden Forschungsdaten in einer Vielfalt von Dateiformaten hergestellt, die semantisch selten einheitlich strukturiert und nur lückenhaft mit Metadaten beschrieben sind. Diese Faktoren stellen für die digitale Langzeitarchivierung von Forschungsdaten eine größere Herausforderung dar, als die Datenmenge, auch wenn diese in einzelnen Fällen sehr groß sein kann.[99]

Im Laufe des digitalen Lebenszyklus von Forschungsdaten werden zudem in den verschiedenen Phasen sehr unterschiedliche Anforderungen an die Persistenz der Daten und der Werkzeuge zum Umgang mit Forschungsdaten gestellt. Zwischen dem Entstehen der Daten in wissenschaftlichen Arbeitsprozessen und der sicheren, nachnutzbaren Archivierung der Daten besteht ein breites Spektrum von teilweise gegensätzlichen Anforderungen, auch *Digital Curation Continuum* genannt. Organisatorisch ist ein Kontinuum allerdings nicht handhabbar, weswegen es notwendig ist, innerhalb einer Organisation zu bestimmen, wer in welcher Phase des Lebenszyklus von Forschungsdaten für deren Pflege verantwortlich ist. Auf Grund des vorhandenen Kontextwissens reicht in den Phasen vor der Speicherung in der dauerhaften Domäne ein eingeschränktes Metadatenprofil aus, das bei der Überführung in die nächste Domäne (teil-)automatisch angereichert werden kann.[100]

98 Barga & Gannon (2007)

99 Klump (2008)

100 Treloar & Harboe-Ree (2008); Treloar et al. (2007)



Für den Forscher liegt es nicht im Fokus seines wissenschaftlichen Arbeitens, Daten zu archivieren und zugänglich zu machen, denn bisher bestehen keine Anreize an Wissenschaftler, zumindest Daten, die Grundlage einer Veröffentlichung waren, für andere zugänglich zu machen.[101] Nur an sehr wenigen Stellen besteht heute im wissenschaftlichen Veröffentlichungssystem oder in der Forschungsförderung die Pflicht, Forschungsdaten für andere zugänglich zu machen. Darüber hinaus ist nicht geklärt, wer für die Langzeitarchivierung von Forschungsdaten verantwortlich ist und wie diese Leistung finanziert wird.[102] Dies führt zu Defiziten im Management und in der Archivierung wissenschaftlicher Daten mit möglichen negativen Folgen für die Qualität der Forschung.[103] Diese Entwicklung ist inzwischen auch den Herausgebern wissenschaftlicher Fachzeitschriften bewusst geworden. Als Konsequenz daraus verlangen bereits einige Zeitschriften, dass die Daten, die Grundlage der eingereichten Arbeit sind, für die Gutachter und später auch für die Leser zugänglich sind.

Trotz der Empfehlungen für eine gute wissenschaftliche Praxis sind kohärente Datenmanagementstrategien, Archivierung von Forschungsdaten und, soweit möglich, Zugang zu Daten meist nur in größeren Forschungsverbünden zu finden, die für Erfolge in der Forschung auf enge Zusammenarbeit angewiesen sind, oder in Fällen, in denen es gesetzliche Vorgaben für den Umgang mit Daten gibt. Wie schon in der Diskussion um den Offenen Zugang zu wissen-

101 Klump et al. (2006)
102 Lyon (2007)
103 Nature Redaktion (2006)

schaftlichem Wissen (Open Access) zeigt sich hier, dass eine Policy nur wirksam ist, wenn sie eine Verpflichtung mit sich bringt und gleichzeitig Anreize zur Zusammenarbeit bietet.[104] Um das Ziel einer nachhaltigen digitalen Langzeitarchivierung von Forschungsdaten zu erreichen, muss sowohl eine organisatorische Strategie verfolgt werden, die Langzeitarchivierung von Daten zu einem anerkannten Beitrag zur wissenschaftlichen Kultur macht und die gleichzeitig von einer technischen Strategie unterstützt wird, die den Akteuren für die digitalen Langzeitarchivierung von wissenschaftlichen Forschungsdaten geeignete Werkzeuge in die Hand gibt. Mit dazu gehören eine Professionalisierung des Datenmanagements und der digitalen Langzeitarchivierung von Forschungsdaten auf Seiten der Projekte und Archive.

Organisatorische Strategien

Auf Grund der enormen Summen, die jährlich für die Erhebung wissenschaftlicher Daten ausgegeben werden, beschäftigt sich die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) bereits seit einigen Jahren mit der Frage, wie mit Daten aus öffentlich geförderter Forschung umgegangen werden sollte. Auf dem Treffen der Forschungsminister im Januar 2004 wurde beschlossen, dass der Zugang zu Daten aus öffentlich geförderter Forschung verbessert werden muss.[105] Mit diesem Mandat im Hintergrund befragte die OECD die Wissenschaftsorganisationen ihrer Mitgliedsländer zu deren Umgang mit Forschungsdaten. Basierend auf den Ergebnissen der Befragung wurde eine Studie verfasst und im Dezember 2006 verabschiedete der Rat der OECD eine „Empfehlung betreffend den Zugang zu Forschungsdaten aus öffentlicher Förderung".[106] Diese Empfehlung ist bindend und muss von den Mitgliedsstaaten der OECD in nationale Gesetzgebung umgesetzt werden, die Umsetzung wird von der OECD beobachtet. In Abschnitt M der Empfehlung wird vorgeschlagen, dass schon bei der Planung von Projekten eine nachhaltige, langfristige Archivierung der Daten berücksichtigt wird.

Parallel dazu, und mit mehr Aufsehen in der Öffentlichkeit, wurde im Oktober 2003 von den Wissenschaftsorganisationen die „Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen" veröffentlicht, deren Schwerpunkt auf dem Zugang zu wissenschaftlicher Literatur für Forschung und Lehre liegt.[107] In ihre Definition des offenen Zugangs bezieht die „Berliner

104 Bates et al. (2006); Spittler (1967)
105 OECD (2004)
106 OECD (2007)
107 Berliner Erklärung (2003)

Erklärung“ auch Daten und Metadaten mit ein. Die Langzeitarchivierung ist hier ein Mittel zum Zweck, das den offenen Zugang zu wissenschaftlichem Wissen über das Internet auf Dauer ermöglichen soll. Aufrufe dieser Art wurden stets begrüßt, aber blieben leider ohne Folge.[108] Dieses Problem betrifft die Institutional Repositories des Open Access genauso wie die Datenarchive. Es sollte daher geprüft werden, inwiefern die Strategien, die bei der Umsetzung von Open Access angewandt werden, sich auch auf den offenen Zugang zu Daten anwenden lassen.[109]

Wenngleich es einige Policies gibt, die den Zugang zu Daten ermöglichen sollen, so hat sich erst recht spät die Erkenntnis durchgesetzt, dass die digitale Langzeitarchivierung von Forschungsdaten eine Grundvoraussetzung des offenen Zugangs ist. Eine umfangreiche Studie wurde dazu bereits in der ersten Förderphase des Projekts nestor erstellt.[110] Eine ähnliche Studie wurde auch für das britische Joint Information Systems Committee (JISC) veröffentlicht[111] und das Thema in einer weiteren Studie vertieft[112]. Einzelne Systeme, die als Best-Practice Beispiele gelten dürfen, da sie die Voraussetzungen von Offenem Zugang und vertrauenswürdiger digitaler Langzeitarchivierung erfüllen, existieren bereits.

Die Finanzierung der digitalen Langzeitarchivierung von Forschungsdaten ist eine offene Frage, denn bislang gab es für Datenmanagement jenseits des Projektendes weder die notwendigen finanziellen Mittel, noch waren Antragsteller verpflichtet einen entsprechenden Plan vorzulegen. Hier tritt bei den Förderorganisationen inzwischen ein Umdenken ein. Durch die Umsetzung der „Empfehlung betreffend den Zugang zu Forschungsdaten aus öffentlicher Förderung“[113] kann damit gerechnet werden, dass hier neue Möglichkeiten für den Aufbau von wissenschaftlichen Datenzentren und -archiven entstehen werden.

Technische Strategien

Voraussetzung für die digitale Langzeitarchivierung wissenschaftlicher Forschungsdaten ist, dass es vertrauenswürdige Archive gibt, die diese Aufgabe übernehmen können. Diese Aufgabe wird bereits in einigen Disziplinen von

108 Zerhouni (2006)
109 Bates et al. (2006); Sale (2006)
110 Severiens & Hilf (2006)
111 Lord & Macdonald (2003)
112 Lyon (2007)
113 OECD (2007)

Datenzentren übernommen und auch die Weltdatenzentren des International Council of Scientific Unions (ICSU WDCs) haben sich dieser Aufgabe verpflichtet. In den vielen Fällen, in denen es kein disziplinspezifisches Datenzentrum und –archiv gibt, fehlen Konzepte für eine digitale Langzeitarchivierung von wissenschaftlichen Forschungsdaten. Eine mögliche Lösung wäre, in Analogie zur Open Archive Initiative, für diese Daten lokale Institutional Repositories aufzubauen.[114] Die Herausforderungen liegen dabei weniger bei den Archivsystemen, wo sie oft vermutet werden, sondern häufiger im Zusammenspiel der Prozesse des Managements von Forschungsdaten und der digitalen Langzeitarchivierung. So beziehen sich nur wenige Datenarchive in der Organisation ihrer Archivprozesse auf das OAIS-Referenzmodell (ISO 14721:2003)[115], das die Prozesse der digitalen Langzeitarchivierung beschreibt.[116]

Besondere Herausforderungen an die digitale Langzeitarchivierung von Forschungsdaten erwachsen aus Grid- und eScience-Projekten, die sich auf den ersten Blick in vielen Aspekten nicht wesentlich von anderen Datenproduzierenden Forschungsprojekten unterscheiden. Die enorm großen Datenmengen, die in Grid-Projekten erzeugt und verarbeitet werden und die hohe Komplexität von Daten aus eScience-Projekten lassen jedoch vermuten, dass aus diesen Projekttypen neuartige Anforderungen an die digitale Langzeitarchivierung erwachsen.[117] Gerade wegen dieser extremen Anforderungen an Prozessierungs- und Speicherressourcen und zusätzlichen Managementvorkehrungen durch Virtualisierung der Ressourcen sind Communities, die große Datenmengen erzeugen oder verarbeiten, in der Anwendung von Grid-Technologien vergleichsweise weit fortgeschritten. Astrophysik, Klimaforschumg, biomedizinische Forschung, und andere Communities mit rechenintensiven Verfahren der Datenverarbeitung wenden bereits seit einiger Zeit Grid-Technologien an.

Die enorm großen Datenmengen erfordern von den Grid-Projekten konsistente Richtlinien für die Auswahl der Daten, die für lange Zeiträume archiviert werden sollen. Ähnlich wie in den Richtlinien des British Atmospheric Data Centre wird in den Projekten evaluiert, ob die Daten grundsätzlich und mit vertretbarem Aufwand neu generiert werden können, und ob die Daten in der vorliegenden Form nachnutzbar sind. Allerdings liegen die Herausforderungen an die Langzeitarchivierung von Forschungsdaten in Grid-Projekten weniger in den Datenmengen, sondern eher im neuartigen technologischen und organisa-

114 Lyon (2007)
115 OAIS (2002)
116 Lyon (2007)
117 Hey & Trefethen (2003)

torischen Rahmen (z.B. Virtuelle Organisationen, Authentifizierung und Autorisierung, semantische Bezüge, Metadaten).[118]

Langzeitarchive für Forschungsdaten und organisatorische Rahmenbedingungen in den Instituten und bei der Forschungsförderung sind notwendige Voraussetzungen für die digitale Langzeitarchivierung von wissenschaftlichen Forschungsdaten. Sie müssen aber auch durch technische Lösungen unterstützt werden, die die Mitwirkung durch die Wissenschaftler an der digitalen Langzeitarchivierung von wissenschaftlichen Forschungsdaten so einfach wie möglich gestalten, so dass sie sich möglichst nahtlos in die wissenschaftlichen Arbeitsabläufe einfügt. Ein Beispiel dafür ist die Beschreibung der Forschungsdaten durch Metadaten. Erstellen und Pflege von Metadaten stellt eine enorme Hürde dar, denn die notwendigen Metadatenprofile sind meist komplex, sie manuell zu erstellen ist aufwändig.[119] In der Praxis hat sich gezeigt, dass das Management von Daten und Metadaten eine bessere Chance zum Erfolg hat, wenn das Erstellen und Pflegen von Metadaten weitgehend automatisiert ist. Ein hoher Grad an Technisierung des Datenmanagements erlaubt den Wissenschaftlern, sich ihrem eigentlichen Tätigkeitsschwerpunkt, der Forschung, zu widmen. In den vom Bundesministerium für Bildung und Forschung geförderten Projekten C3-Grid und TextGrid sind sowohl für die Naturwissenschaften, als auch für die Geisteswissenschaften vorbildliche Verfahren für die Erzeugung und Verwaltung von Metadaten entwickelt worden.[120]

Während bereits die inhaltliche Beschreibung der zu archivierenden Daten durch Metadaten eine Hürde darstellt, kommen für eine spätere Nachnutzung weitere Probleme hinzu. Vielfach trifft man auf das Missverständnis, dass die Angabe des MIME-Type eine ausreichende Beschreibung des Dateiformats und seiner Nutzung sei. Ein Archivsystem müsste jedoch nicht nur den MIME-Type der archivierten Daten kennen, sondern auch deren semantische Struktur und ihr technisches Format. Die semantische Struktur maschinenlesbar zu dokumentieren ist eine Grundvoraussetzung für die in Zukunft geforderte Interoperabilität der Archivsysteme.[121] Zusätzlich müssen sich die Archivbetreiber und ihre Nutzer darüber verständigen, welche Dateiformate archiviert werden, denn nicht jedes bei den Nutzern populäre Format ist für eine verlustfreie Langzeitarchivierung geeignet.[122]

---

118 Klump (2008)
119 Robertson (2006)
120 Neuroth et al. (2007)
121 Klump (2008)
122 Lormant et al. (2005)

Ungeachtet des in der „Berliner Erklärung“ durch die Universitäten, Wissenschafts- und Forschungsförderungsorganisationen geleisteten Bekenntnisses zum offenen Zugang gibt es Gründe, warum manche Daten nicht offen zugänglich sein können. Aus diesem Grund sind Zugriffsbeschränkungen in der digitalen Langzeitarchivierung von Forschungsdaten ein wichtiges Thema. Die Zugriffsbeschränkungen dienen hierbei nicht primär der Sicherung von Verwertungsrechten, sondern sie sind entweder gesetzlich vorgeschrieben (Datenschutz) oder dienen dem Schutz von Personen oder Objekten, die durch eine Veröffentlichung der Daten Gefährdungen ausgesetzt würden. Für geschützte Datenobjekte müssen Verfahren und Policies entwickelt werden, die auch über lange Zeiträume hinweg zuverlässig die Zugriffsrechte regeln und schützen können.[123] Auch der Umgang mit „verwaisten“ Datenbeständen muss geregelt werden.

Zum Schutz der intellektuellen Leistung der Wissenschaftler sollten Daten in wissenschaftlichen Langzeitarchiven mit Lizenzen versehen sein, die die Bedingungen einer Nachnutzung regeln, ohne dadurch den wissenschaftlichen Erkenntnisgewinn zu behindern. Entsprechende Vorarbeiten sind bereits in den Projekten Creative Commons (CC) und Science Commons (SC) geleistet worden. Zusätzlich zur erwiesenen Praxistauglichkeit können die hier entwickelten Lizenzen auch maschinenlesbar hinterlegt werden, was eine künftige Nachnutzung deutlich vereinfacht. Die Diskussion, welche Lizenzen für Daten empfohlen werden sollten, ist noch offen.[124] Sie wird zusätzlich erschwert durch die rechtliche Auffassung im Urheberrecht, die die Schutzwürdigkeit von Daten stark einschränkt.

Nachnutzung von Daten

Keine der Infrastrukturen für eine digitale Langzeitarchivierung lässt sich dauerhaft betreiben, wenn es keine Nutzer gibt, denn erst wenn eine Nachfrage der Wissenschaft nach einer digitalen Langzeitarchivierung besteht, können dauerhafte Strukturen entstehen. Für die meisten Forschungsdaten gilt heute noch, dass die Nachfrage schon in den ersten Jahren stark abnimmt.[125] Dies gilt jedoch nicht für unwiederholbare Messungen wie z.B. Umweltmessdaten.[126] Im heutigen Wissenschaftsbetrieb sind der Gewinn an Distinktion und Reputation wichtige Motivationskräfte. Digitale Langzeitarchivierung muss als Praxis

123 Choi et al. (2006); Simmel (2004)
124 Uhlir & Schröder (2007)
125 Severiens & Hilf (2006)
126 Pfeiffenberger (2007)

in der Wissenschaft verankert sein und im selbst verstandenen Eigeninteresse der Wissenschaftler liegen. Die wissenschaftliche Publikation ist dabei ein entscheidendes Medium.[127] Ein möglicher Anreiz, Daten zu veröffentlichen und dauerhaft zugänglich zu machen, ist es daher, die Veröffentlichung von Daten zu formalisieren und als Bestandteil des wissenschaftlichen Arbeitens zu institutionalisieren. Dazu ist nötig, dass die veröffentlichten Daten findbar, eindeutig referenzierbar und auf lange Zeit zugänglich sind. Allerdings werden Datenveröffentlichungen nur dann auch nachgenutzt und zitiert, wenn ihre Existenz den potenziellen Nutzern auch bekannt ist. Ein geeigneter Weg, Daten recherchierbar und zugänglich zu machen, ist ihre Integration in Fachportale und Bibliothekskataloge. Eine entscheidende Voraussetzung für die Zitierbarkeit von Daten ist, dass sie eindeutig und langfristig referenzierbar sind.[128]

Da in der Praxis URLs nur kurzlebig sind, werden sie nicht als zuverlässige Referenzen angesehen. Persistente, global auflösbare Identifier, wie z.B. Digital Object Identifier (DOI) oder Universal Resource Names (URN) schließen diese Lücke.[129] In der bisherigen Praxis fehlten bisher wichtige Bestandteile, die eine nachhaltige Publikation von Daten möglich machen. Diese Defizite wurden im DFG-Projekt „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten" (STD-DOI) analysiert. Mit der Einführung von persistenten Identifikatoren für wissenschaftliche Datensätze wurden die Voraussetzungen für eine Publikation und Zitierbarkeit wissenschaftlicher Daten geschaffen.[130]

Zusammenfassung

In der Einführung zum OAIS-Referenzmodell zur Langzeitarchivierung digitaler Objekte ist treffend formuliert worden, dass ein Archivsystem für digitale Objekte mehr ist als nur ein technisches System. Das OAIS-Referenzmodell beschreibt es als das Zusammenwirken von Menschen und Systemen mit dem Ziel der Langzeiterhaltung von digitalen Objekten für eine definierte Nutzergruppe.[131] Die digitale Langzeitarchivierung von Forschungsdaten ist daher nicht allein eine technische Herausforderung, sondern muss in einen entsprechenden organisatorischen Rahmen eingebettet sein, der im Dialog mit der Wissenschaft gestaltet wird. Der wissenschaftliche Wert, Forschungsdaten für lange Zeit zu archivieren und zugänglich zu machen, ist erkannt worden. In dem Maße, wie

127 Pfeiffenberger & Klump (2006)
128 Klump et al. (2006)
129 Hilse & Kothe (2006); Klump et al. (2006)
130 Brase & Klump (2007)
131 OAIS (2002)

die Auswertung von Daten für die Forschung an Bedeutung zunimmt, wird sich auch der Umgang mit Daten in der Forschungspraxis und in der Langzeitarchivierung verändern.

Quellenangaben

Barga, Roger, und Dennis B Gannon (2007): Scientific versus business workflows. In *Workflows for e-Science*, I. J. Taylor et al. (Hrsg.), S. 9-16, Springer-Verlag, London, Großbritannien.

Bates, Melanie et al. (2006): *Rights and Rewards Project - Academic Survey Final Report*, JISC. Bath, Großbritannien.

Berliner Erklärung (2003): Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

Brase, Jan, und Jens Klump (2007): Zitierfähige Datensätze: Primärdaten-Management durch DOIs, In *WissKom 2007 : Wissenschaftskommunikation der Zukunft ; 4. Konferenz der Zentralbibliothek, Forschungszentrum Jülich, 6. - 8. November 2007*, Bd. 18, R. Ball (Hrsg.), S. 159-167, Forschungszentrum Jülich, Jülich.

Choi, Hee-Chul, et al. (2006): Trust Models for Community Aware Identity Management, in *WWW2006*, Edinburgh, Großbritannien.

DFG (1998): *Sicherung guter wissenschaftlicher Praxis* , Deutsche Forschungsgemeinschaft, Bonn. http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

DFG (2009): *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*, Deutsche Forschungsgemeinschaft, Bonn. http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/veroeffentlichungen/dokumentationen/download/ua_inf_empfehlungen_200901.pdf

Hey, Tony, und Anne Trefethen (2003): e-Science and its implications, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 361(1809): 1809-1825, doi:10.1098/rsta.2003.1224.

Hilse, Hans-Werner, und Jochen Kothe (2006): *Implementing Persistent Identifiers*, Consortium of European Research Libraries, London, Großbritannien.

Klump, Jens (2008): *Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten*, Expertise, Kompetenznetzwerk Langzeitarchivierung (nestor), Frankfurt (Main). http://nbn-resolving.de/urn:nbn:de:0008-2008040103

Klump, Jens, et al. (2006): Data publication in the Open Access Initiative, *Data Science Journal*, 5, 79-83, doi:doi:10.2481/dsj.5.79.

Kluttig, Thekla (2008), Bericht über das DFG-Rundgespräch „Forschungsprimärdaten“ am 17.01.2008, Bonn, Germany. http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/forschungsprimaerdaten_0108.pdf

Lord, Philipp, und Alison Macdonald (2003): *e-Science Curation Report - Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*, JISC. Bath, Großbritannien.

Lormant, Nicolas, et al. (2005): How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?, In: *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005)*, S. 11, Edinburgh, Großbritannien. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf

Lyon, Liz (2007): *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, consultancy report, UKOLN, Bath, Großbritannien. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Nature Redaktion (2006): A fair share, *Nature*, 444(7120), 653-654, doi:10.1038/444653b.

Neuroth, Heike, et al. (2007): *Die D-Grid Initiative*, Universitätsverlag Göttingen, Göttingen. http://resolver.sub.uni-goettingen.de/purl/?webdoc-1533

OAIS (2002): *Reference Model for an Open Archival Information System (OAIS). Blue Book*, Consultative Committee for Space Data Systems, Greenbelt, MD, USA.

OECD (2004): *Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué*, Communiqué, Organisation for Economic Co-operation and Development, Paris, Frankreich.

OECD (2007): *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Organisation for Economic Co-operation and Development, Paris, Frankreich.

Pfeiffenberger, Hans (2007): Offener Zugang zu wissenschaftlichen Primärdaten, *Zeitschrift für Bibliothekswesen und Bibliographie*, 54(4-5), 207-210.

Pfeiffenberger, Hans, und Jens Klump (2006): Offener Zugang zu Daten - Quantensprung in der Kooperation, *Wissenschaftsmanagement*, (Special 1), 12-13. http://oa.helmholtz.de/fileadmin/Links/Artikel/Wissenschafts_Management_Open_Access/Daten.pdf

Robertson, R. John (2006): *Evaluation of metadata workflows for the Glasgow ePrints and DSpace services*, University of Strathclyde, Glasgow, Großbritannien.

Sale, Arthur (2006): The acquisition of Open Access research articles, *First Monday*, 11(10).

Severiens, Thomas, und Eberhard R Hilf (2006): *Langzeitarchivierung von Rohdaten*, nestor - Kompetenznetzwerk Langzeitarchivierung, Frankurt (Main). http://nbn-resolving.de/urn:nbn:de:0008-20051114018

Simmel, Derek (2004): *TeraGrid Certificate Management and Authorization Policy*, policy, Pittsburgh Supercomputing Center, Carnegie Mellon University, University of Pittsburgh, Pittsburg, PA, USA. http://www.teragrid.org/policy/TGCertPolicy-TG-5.pdf

Spittler, Gerd (1967), *Norm und Sanktion. Untersuchungen zum Sanktionsmechanismus*, Walter Verlag, Olten, Schweiz.

Treloar, Andrew, et al. (2007): The Data Curation Continuum - Managing Data Objects in Institutional Repositories, *D-Lib Magazine*, 13(9/10), 13, doi:10.1045/september2007-treloar.

Treloar, Andrew, und Cathrine Harboe-Ree (2008): Data management and the curation continuum: how the Monash experience is informing repository relationships. In: *VALA2008*, Melbourne, Australien. http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Uhlir, Paul F, und Peter Schröder (2007): Open Data for Global Science, *Data Science Journal*, 6(Open Data Issue), OD36-53, doi:10.2481/dsj.6.OD36.

Zerhouni, Elias A (2006): *Report on the NIH public access policy*, National Institute of Health, Bethesda, MD, USA. http://www.mlanet.org/government/gov_pdf/2006_nihrpt_pubaccessplcy.pdf

17.11 Computerspiele

Karsten Huth

Die langfristige Erhaltung von Software wird in Gedächtnisinstitutionen bislang kaum betrieben. Während man sich über die Notwendigkeit der Erhaltung eines Teils des digitalen kulturellen Erbes im Klaren ist, wird die Software als ein wesentliches Kulturgut der digitalen Welt beinahe vollständig ignoriert. Computerspiele bilden dahingehend eine Ausnahme, weil sie Liebhaber in aller Welt gewonnen haben, die sich zumindest dem kurzfristigen Erhalt gewidmet haben. Diese Gruppe hat technische Lösungen entwickelt, von denen klassische Einrichtungen nun profitieren könnten.

Einführung

Das Computerspiel ist, neben den frühen Datenbanken, eines der ältesten digitalen Artefakte, das von seiner Natur her als „born digital" zu betrachten ist. Sieht man von dem ersten Vorläufer des Videospiels, einem Ausstellungsstück auf einem „Tag der offenen Tür" der Atomforschung, und dem ersten Wohnzimmergerät ab, beide Geräte beruhten noch auf analoger Technik, so sind alle Video- und Computerspiele technisch betrachtet Computerprogramme. Das IBM Dictionary of Computing ordnet sie der „application software", also der „Anwendungssoftware" zu, zu der auch Textverarbeitungsprogramme, Tabellenkalkulation und andere Office-Programme gezählt werden. Computerspiele bilden dennoch eine Sondergruppe innerhalb der Anwendungssoftware. Mit ihnen wird kein Problem gelöst oder die täglich anfallende Büroarbeit bewältigt. Computerspiele dienen einzig der Unterhaltung und dem Vergnügen des Nutzers. Ihre unterhaltende Funktion hat technische Konsequenzen. Computerspiele müssen sich auf einem wachsenden Markt behaupten und die Aufmerksamkeit der Käufer erregen. Sie operieren deshalb oft am oberen technischen Limit der jeweiligen aktuellen Hardwaregeneration. Überlieferte Beispiele aus den siebziger oder achtziger Jahren mögen gegen die Leistungsfähigkeiten eines aktuellen PCs rührend anmuten, für den Nutzer vergangener Tage waren sie ein Beispiel für rasenden technischen Fortschritt, das nicht selten Begeisterung auslöste. Diese Begeisterung machte den Einzug des Computers in den privaten Haushalt möglich. Sie legte einen Grundstein für unseren alltäglichen Umgang mit der digitalen Medienwelt.

Video- und Computerspiele werden häufig nach ihren Hardware/Software Plattformen klassifiziert. Man unterscheidet[132]:

- die Arcade-Spiele: Automaten, die in Spielhallen stehen und gegen den Einwurf von Geld benutzt werden können. Die Software befindet sich meistens auf austauschbaren Platinen im sogenannten Jamma-Standard.
- die Computerspiele: Spiele, die auf Computern gespielt werden, welche nicht ausschließlich zum Spielen gedacht sind. Ein aktuelles Beispiel sind die PCs. In den achtziger Jahren waren die Homecomputer sehr populär. Das früheste Beispiel ist das Spiel „Spacewar" aus dem Jahr 1962, geschrieben für den ersten Minicomputer der Welt, den PDP-1. Die Datenträger für Computerspiele reichen von üblichen Musikkassetten über die ersten Floppydisks bis hin zu den heute gebräuchlichen DVDs. Die Darstellung des Spiels erfolgte damals über den Fernseher, heute über den PC-Monitor.
- die Videospiele: Plattform ist hierbei die sogenannte „Konsole". Die Konsole ist ein Computer, der einzig zum Spielen dient. Seine Hardware ist deshalb für eine gute grafische Darstellung und eine gute Audio-Ausgabe optimiert. Die Datenträger sind ebenso wie die Software an einen bestimmten Konsolentyp gebunden.
- die tragbaren Videospiele: Die sogenannten Handhelds vereinigen den Computer, den Monitor und das Steuerungsgerät in einem kompakten Taschenformat. Neu hinzugekommen sind die Spiele für Mobiltelefone. Bei manchen Geräten sind die Spiele fest implementiert, bei anderen sind sie über spezielle Datenträger austauschbar.

Gründe für die Archivierung

Folgende Gründe sprechen für eine nachhaltig betriebene Langzeitarchivierung von Computerspielen:

Wissenschaftliche Forschung: Computer- und Videospiele sind zum interdisziplinären Untersuchungsgegenstand für die Wissenschaft geworden, vor allem in den Bereichen der Pädagogik, Psychologie, Kultur- und Medienwissenschaften. Das „Handbuch Medien Computerspiele", herausgegeben von der Bundeszentrale für politische Bildung verzeichnet im Anhang ca. 400 Titel zum Thema Computerspiele. Diese Zahl der größtenteils deutschen Titel aus dem Jahr 1997

132 Fritz, Jürgen : Was sind Computerspiele? In: Handbuch Medien: Computerspiele: Theorie, Forschung, Praxis/ hrsg. Jürgen Fritz und Wolfgang Fehr – Bonn: Bundeszentrale für politische Bildung Koordinierungsstelle Medienpädagogik; 1997. (S. 81-86)

zeigt, dass die wissenschaftliche Untersuchung von Computerspielen keine Randerscheinung ist. Die Artikel des Handbuchs beziehen sich oft auf konkrete Spielsoftware. Während das Zitieren der Literatur in diesen Artikeln nach wissenschaftlichen Regeln abläuft, werden Angaben zu den verwendeten Spielen oft gar nicht oder nur in unzureichender Weise gemacht. Man kann somit die wissenschaftlichen Hypothesen eines Artikels, der spezielle Computerspiele als Gegenstand behandelt, nicht überprüfen. Neben dem Problem des wissenschaftlichen Zitierens besteht natürlich auch das Problem des gesicherten legalen Zugriffs auf ein zitiertes Computerspiel. Streng genommen, ist ohne eine vertrauenswürdige Langzeitsicherung von Computerspielen die Wissenschaftlichkeit der Forschung in diesem Bereich gefährdet.

Kulturelle Aspekte: Die Anfänge des Computerspiels reichen zurück bis in das Jahr 1958.[133] Seitdem hat sich das Computerspiel als eigenständiges Medium etabliert. Zum ersten Mal in der Geschichte könnten wir die Entwicklung einer Medienform, von den ersten zaghaften Versuchen bis zur heutigen Zeit, beinahe lückenlos erhalten und damit erforschen. Es wird allgemein bedauert, dass aus der frühen Stummfilmzeit nur noch ca. 10% des einst verfügbaren Materials erhalten geblieben sind. Der Bestand an Computerspielen wäre noch zu einem ökonomisch vertretbaren Preis zu erhalten und könnte auch der übrigen Medienforschung dienen.

Als Zeugnis der technischen Entwicklung: Wie bereits erwähnt, testen Computerspiele, wie keine zweite Software, die technischen Fähigkeiten der jeweiligen Hardwaregeneration aus. Sie eignen sich dadurch für eine anschauliche Demonstration des Mooreschen Gesetzes. Zudem wurden bei alter Software Programmiertechniken verwendet, die auf einen sparsamen und ökonomischen Einsatz von Hardware-Ressourcen (Speicherplatz und Rechenzeit) ausgerichtet waren. Diese Techniken wurden im Zuge der Hardwareverbesserungen aufgegeben und vergessen. Niemand kann jedoch sagen, ob sie nicht irgendwann einmal wieder von Nutzen sein werden.[134]

---

133 o.V. (2004) : Video Games-Did They Begin at Brookhaven, http://www.osti.gov/accomplishments/videogame.html

134 Dooijes, Edo Hans: Old computers, now and in the future – 2000: Im Internet: http://www.science.uva.nl/museum/pdfs/oldcomputers_dec2000.pdf

Computerspiele und Gedächtnisinstitutionen

Die Integration von Video- und Computerspielen in die Medienarchive, Bibliotheken und Museen steht noch aus. Die Erhaltung der frühen Spiele ist der Verdienst von privaten Sammlern und Initiativen, die sich über das Internet gefunden und gebildet haben. Beinahe jede obsolete Spielplattform hat ihre Gemeinde, die mit großem technischen Expertentum die notwendigen Grundlagen für eine langfristige Archivierung schafft. Den wichtigsten Beitrag schaffen die Autoren von Emulatoren, die oft zur freien Verfügung ins Netz gestellt werden. Aber auch das Sammeln von Metadaten, welches oft in umfangreiche Softwareverzeichnisse mündet, die aufwendige Migration der Spielsoftware von ihren angestammten Datenträgern auf moderne PCs sowie das Sammeln des Verpackungsdesigns und der Gebrauchsanleitungen sind notwendige Arbeiten, die unentgeltlich von den Sammlern erbracht werden. Leider bewegen sich die privaten Initiativen oft in einer rechtlichen Grauzone. Die Software unterliegt dem Urheberrecht. Ihre Verbreitung über das Internet, auch ohne kommerzielles Interesse, stellt einen Rechtsbruch dar, selbst wenn die betroffenen Produktionsfirmen schon längst nicht mehr existieren. Besonders die Autoren von Emulatoren werden von der Industrie in eine Ecke mit den aus Eigennutz handelnden Softwarepiraten gestellt. Es soll hier nicht verschwiegen werden, dass es auch Emulatoren gibt, die aktuelle Spielplattformen emulieren und dadurch die Softwarepiraterie fördern. Die Motivation dieser Autoren ist deutlich anders gelagert. Die Sammler von historischen Systemen nutzen die Emulation zur Erhaltung ihrer Sammlungen. Die obsoleten Systeme sind im Handel in dieser Form nicht mehr erhältlich. Zudem hat die Industrie bislang kaum Interesse an der Bewahrung ihrer eigenen Historie gezeigt. Zumindest gibt es innerhalb der International Game Developers Association (IGDA) eine Special Interest Group (SIG), die sich mit dem Problem der digitalen Langzeitarchivierung befassen will.

Beispiele für die Langzeitarchivierung von Computerspielen in den klassischen Institutionen sind rar. Die Universitätsbibliothek in Stanford besitzt wohl die größte Sammlung innerhalb einer Bibliothek. Die Sammlung trägt den Namen des verstorbenen Besitzers: Stephen M. Cabrinety. Sie besteht aus kommerziellen Videospielen, sowie den Originalverpackungen, Gebrauchsanleitungen, gedruckten Materialien und dokumentiert somit einen großen Teil der Geschichte der Computerspiele in der Zeitspanne von 1970-1995. Neben den 6.300 Programmen verfügt die Sammlung über 400 original Hardwareobjekte von Motherboards, Monitoren bis hin zu CPUs. Die Sammlung wird verwaltet

von Henry Lowood und ist Teil des "Department. of Special Collections" der Stanford University Library.[135]

Das Computerspielemuseum in Berlin wurde im Februar 1997 eröffnet. Getragen wird das Museum vom Förderverein für Jugend- und Sozialarbeit e.V. Das Museum besitzt rund 8.000 Spiele und ist auf der Suche nach einem neuen Ort für eine permanente Ausstellung seiner Exponate. Das Museum ist der Ausrichterder Ausstellung Pong-Mythos[136] mit Stationen in Stuttgart, Leipzig,Bern und Frankfurt a.M..

Der Verein „Digital Game Archive"[137] hat sich den Aufbau eines legalen Medienarchivs für Computerspiele zum Ziel gesetzt. Der Nutzer kann die archivierten Spiele über die Internetseite des Archivs beziehen. Alle angebotenen Spiele wurden von den Rechteinhabern zur allgemeinen Verwendung freigegeben. Neben der Erhaltung der Software sammelt das Digital Game Archive auch Informationen zum Thema Computerspielarchivierung und versucht die Geschichte des Computerspiels zu dokumentieren. Die Mitglieder sind Fachleute aus verschiedenen wissenschaftlichen Disziplinen. Sie vertreten den Verein auch auf Fachkonferenzen. Das Digital Game Archive arbeitet eng mit dem Computerspielemuseum Berlin zusammen.

Das Internet Archive hat eine kleine Sektion, die sich der Sammlung von historischen Computerspielen widmet. Diese hat das Classic Software Preservation Project[138] im Januar 2004 ins Leben gerufen. Ziel des Projekts ist die Migration gefährdeter Software von ihren originalen Datenträgern auf aktuelle, nicht obsolete Medien. Nach der Migration werden die Programme solange unter Verschluss gehalten, bis die Rechtslage eine legale Vermittlung der Inhalte erlaubt. Um dieses Vorhaben rechtlich möglich zu machen, erwirkte das Internet Archive eine Ausnahmeregelung vor dem Digital Millenium Copyright Act[139]. Das Copyright Office entsprach den Vorschlägen des Internet Archives und erlaubte die Umgehung eines Kopierschutzes sowie die Migration von obsoleter

135 Lowood, Henry: Playing History with Games : Steps Towards Historical Archives of Computer Gaming - American Institute for Conservation of Historic and Artistic Works. Electronic Media Group: 2004 Im Internet: http://aic.stanford.edu/sg/emg/library/pdf/lowood/Lowood-EMG2004.pdf

136 Fotos der Ausstellung in Frankfurt a.M.: Im Internet: http://www.computerspielemuseum.de/index.php?lg=de&main=Ausstellungen&site=03:00:00&id=164

137 The Digital Game Archive (DiGA): http://www.digitalgamearchive.org/home.php

138 Internet Archive: Software Archive: Im Internet: http://www.archive.org/details/clasp

139 Rulemaking on Exemptions from Prohibition on Circumvention of Technological Measures that Control Access to Copyrighted Works: Im Internet: http://www.copyright.gov/1201/2003/index.html

Software auf aktuelle Datenträger zum Zwecke der Archivierung in Gedächtnisorganisationen. Diese Ausnahmeregelung wird 2006 erneut vom Copyright Office geprüft werden.

Emulation und andere Strategien

Wenn man sich zum Ziel gesetzt hat, die Spielbarkeit der Programme zu erhalten, gibt es zwei mögliche digitale Erhaltungsstrategien (Hardware Preservation und Emulation) für die Langzeitarchivierung von Computerspielen,. Die Möglichkeit das Spiel nur durch Bilder (Screenshots) und eine ausreichende Spielbeschreibung zu dokumentieren und einzig diese Dokumentation zu bewahren, soll hier nicht weiter betrachtet werden. Langzeitarchivierung eines Computerspiels in diesem Kapitel heißt: „*Der originale Bitstream des Computerspiels muss erhalten bleiben und das Programm soll auch in Zukunft noch lauffähig und benutzbar sein.*"

Diese Vorgabe schränkt die möglichen Erhaltungsstrategien von vornherein ein. Migration scheidet als langfristige Strategie aus, da sie bei einer Anpassung an eine neue Softwareplattform den Bitstream des Programms verändert. Solche Portierungen von Programmen auf neue Plattformen sind sehr viel aufwendiger als die vergleichbare Konvertierung von Dateien in ein anderes Dateiformat. Bei einer Dateikonvertierung kann ein einzelnes Konverterprogramm unbegrenzt viele Dateien bearbeiten. Bei einer Software-Portierung muss jedes einzelne Programm von Hand umgeschrieben und angepasst werden. Zudem bräuchte man ein hohes technisches Wissen über die obsoleten Programmiersprachen, welches oft nicht mehr verfügbar ist. Die Kosten und der Aufwand für eine langfristige Migrationsstrategie wären somit immens hoch.

Praktiziert werden zurzeit zwei Erhaltungsstrategien. Zum einen die der Hardware Preservation (Computermuseum) und die der Emulation[140]. Beide Strategien erhalten den originalen Bitstream eines Programms. Diese Zweigleisigkeit findet man sowohl in privaten Sammlerkreisen, als auch bei den Computerspiele bewahrenden Institutionen wieder. Befürworter der Hardware Preservation Strategie bemängeln den Verlust des sogenannten „Look and Feel" bei der Emulation. Diese Kritik ist nicht ganz unberechtigt. Ältere Spiele der 8-Bit Hardwaregeneration wurden beispielsweise für die Ausgabe auf einem NTSC oder PAL Fernsehbildschirm konzipiert. Die Betrachtung mittels eines

140 Rothenberg, Jeff: Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources – Washington D.C.: Council on Library and Information Resources, 1998: S. 18: Im Internet: http://www.clir.org/pubs/reports/rothenberg/criteria.html; vgl. auch nestor Handbuch 2.0, Kap. 8.4

Emulators über einen PC-Monitor gibt nicht zu einhundert Prozent den ursprünglichen Eindruck wieder. Die Farben wirken, je nach Einstellung, auf jedem PC etwas anders. Teilweise ist die Emulation auch nicht vollständig, sodass z.B. die Tonwiedergabe nicht bei allen Sound-Effekten glückt. Manche Emulatoren bieten zusätzlich eine Anpassung des Bildes an die alten NTSC- oder PAL-Verhältnisse, um Abweichung des „Look and Feel" zu kompensieren. Jenseits von Bild und Ton bleibt aber noch das Problem der Steuerung. Die originalen Steuerungsgeräte (Joystick, Paddle usw.) werden bei einer Emulation auf dem PC durch die dort vorhandenen Steuerungsgeräte Tastatur und Maus ersetzt. Dies kann zu einem abweichenden Spielerlebnis und Ergebnis führen. Manche Spiele sind mit PC-Tastatur oder Maus nur sehr schwer oder auch gar nicht zu bedienen. Wir werden später beim Thema „notwendige Metadaten" näher auf dieses Problem eingehen.

Bei der Hardware Preservation muss man sich hingegen fragen, ob es sich hierbei überhaupt um eine Langzeitarchivierungsstrategie handelt. Es dürfte auf lange Sicht hin unmöglich sein, die originale Hardware und die dazugehörigen Datenträger lauffähig zu halten. Einige Datenträger, z.B. EPROMS haben sich als sehr haltbar erwiesen, andere Datenträger z.B. Floppy-Disks halten bestenfalls 10 Jahre. Regelmäßiges Überspielen der Programme auf frische Datenträger des gleichen Typs als Strategie zur Lebensverlängerung scheidet aus, da die betreffenden Datenträgertypen obsolet geworden sind und somit nicht mehr nachproduziert werden. Somit bleibt nur die Emulation als erfolgversprechende Langzeitstrategie.

Computerspielarchiv nach ISO 14721:2003

Zur Zeit gibt es noch kein funktionierendes Langzeitarchiv für Computerspiele, das den kompletten Anforderungen des Open Archival Information System-Funktionsmodells[141] entspricht. Im folgenden Abschnitt wird in einfachen Schritten ein OAIS-konformes Modell für ein Computerspielarchiv entworfen. Die einzelnen Abschnitte sind dementsprechend in Ingest (Accession/Erfassung), Data Management/Archival Storage (Erschließung/Magazin), Access (Benutzung) unterteilt. Wenn möglich, werden zu den einzelnen Abschnitten Beispiele angeführt. Dies können bestimmte Organisationen sein, die in diesem Bereich arbeiten und ihre Ergebnisse publizieren oder konkrete Hinweise auf nutzbare Werkzeuge z.B. Emulatoren oder Metadaten Schemata usw. sein. Das

141 Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1: Blue Book – Consultative Committee for Space Data Systems; 2002: Im Internet http://public.ccsds.org/publications/archive/650x0b1.pdf

entworfene Archiv stellt eine erste Annäherung an ein mögliches Archiv dar. Das OAIS-Funktionsmodell wurde wegen seines hohen Bekanntheitsgrades und seines Status als ISO-Standard (ISO 14721:2003) gewählt.

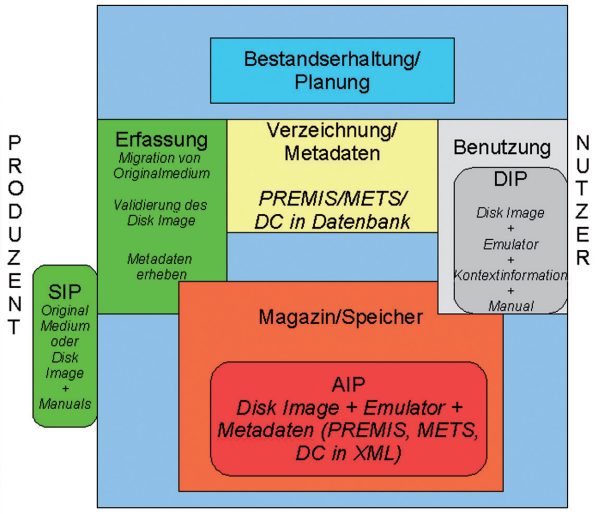


Abbildung 1: OAIS-Funktionsmodell

Das hier angenommene Archiv für Computerspiele nutzt die Emulation als digitale Erhaltungsmaßnahme. Es wird angenommen, dass das Archiv alle rechtlichen Fragen geklärt hat und die Benutzung der Computerspiele durch die Archivbesucher legal ist.

Ingest/Produzent/Erfassung:
Bevor ein Spiel in das Magazin des Archivs eingestellt werden kann, muss es von seinem originalen Datenträger auf einen für das Archiv nutzbaren Datenträger überspielt werden. Dieser Vorgang ist mit einem hohen Aufwand verbunden, da die obsoleten Systeme nicht ohne weiteres mit den aktuellen Systemen über kompatible Schnittstellen verbunden werden können. Insbesondere das Auslesen von ROM-Chips erfordert ein hohes Maß an technischer Kenntnis. Teilweise muss auch erst ein Kopierschutz umgangen werden. Da sich fast alle obsoleten Systeme technisch unterscheiden, ist für jede Plattform ein anderes Expertenwissen gefragt. Glücklicherweise wurden diese Arbeiten schon zu weiten Teilen erbracht. Teilweise könnten nahezu komplette Sammlungen fast aller damals gebräuchlichen Systeme aus dem Internet bezogen werden. Ein

Nachteil dieser Methode ist allerdings, dass einem über die Herkunft der bereits migrierten Programme vertrauenswürdige Informationen fehlen. Dies kann zu Problemen führen, wenn die Programme beim Umgehen des Kopierschutzes verändert oder beschädigt wurden. Viele Spiele des C64 Homecomputers, die heute über das Internet im Umlauf sind, sind Produkte der damaligen Softwarepiraterie. Ihr Programmcode wurde von den sogenannten „Crackern", den Knackern des Kopierschutzes, abgeändert. Teilweise wurden die Programme dadurch zerstört. Ein Archiv muss deshalb innerhalb seiner Sammelrichtlinien festlegen, ob es veränderte Programme von unbestimmter Herkunft in seinen Bestand aufnehmen möchte oder nicht.

Die Software Preservation Society[142], eine Gruppe von Technikexperten für die Migration von Disk Images, akzeptiert nur originale, unveränderte Programme, die mitsamt ihrem Kopierschutz auf neue Datenträger überspielt wurden. Dazu wurde das Interchangeable Preservation Format entwickelt, mit dem sich die Disk Images mit der Hilfe eines Emulators auf einer aktuellen Plattform nutzen lassen. Die Sammlung der SPS umfasst weite Teile der Amiga Spiele.

Eine weitere Frage des Ingests ist: Welche weiteren Informationen werden neben dem Programm noch benötigt, um es später zu archivieren und zu nutzen? Diese Informationen sollten ein Bestandteil des Submission Information Packages (SIP nach der OAIS-Terminologie) sein.

Manche Computerspiele, wie z.B. „Pong", erklären sich von selber. In der Regel benötigt man aber eine Bedienungsanleitung, um ein Spiel zu verstehen. Teilweise enthalten die Anleitungen auch Passwörter, die ein Spiel erst in Gang setzen. Dies war eine häufige Form des Kopierschutzes. Die Bedienungsanleitung ist somit ein fester Teil des Data Objects, das vom Archiv bewahrt werden muss. Genauso wichtig sind Informationen darüber, welcher Emulator verwendet werden soll. Es wäre auch denkbar, dass der Emulator ein Bestandteil des SIPs ist, wenn das Archiv noch nicht über ihn verfügt. Zur Vollständigkeit trägt auch eine technische Dokumentation der obsoleten Plattform bei, auf der das Spiel ursprünglich betrieben wurde. Außerdem werden Informationen über den Kopiervorgang, die Herkunft des Spiels und die rechtlichen Bestimmungen benötigt. Um das Bild abzurunden, sollten digitalisierte Bilder der Verpackung, des obsoleten Datenträgers und der Hardware dem Data Object beigefügt werden. Beispiele für solche Scans findet man auf der Web-Seite von ATARI Age[143] oder lemon64.com[144]. Informationen über Langzeitarchivierungsformate für

142 Software Preservation Society (SPS): Im Internet: http://www.softpres.org/

143 AtariAge: Im Internet: http://www.atariage.com/

144 Lemon64: Im Internet: http://www.lemon64.com/

Bilder und Text finden sich in den betreffenden Kapiteln 17.3 bzw. 17.2 dieses Handbuchs.

Es wäre günstig für ein Computerspielarchiv, wenn die Zeitspanne zwischen der Veröffentlichung eines Spiels und seiner Aufnahme in das Archiv möglichst kurz wäre. Nur solange das Spiel auf seiner originalen Plattform läuft, kann das authentische Verhalten und Look and Feel des Programms durch das Archiv dokumentiert werden. Diese Dokumentation wird später zur Beurteilung des Emulatorprogramms benötigt. Ohne ausreichende Angaben kann später niemand sagen, wie authentisch die Wiedergabe des Spiels mittels des Emulators ist.

Es ist sehr wahrscheinlich, dass sich der Bestand eines Computerspielarchivs nicht allein auf die Spiele als Archivobjekte beschränken kann. Zur technischen Unterstützung müssen, neben den Emulatorprogrammen auch obsolete Betriebssysteme, Treiberprogramme, Mediaplayer usw. archiviert werden.

Archival Storage/Magazin

Die Haltbarkeit der Datenträger hängt von der Nutzung und den klimatischen Lagerungsbedingungen ab. Hohe Temperaturen und hohe Luftfeuchtigkeit können die Lebensdauer eines Datenträgers, ob optisch oder magnetisch, extrem verkürzen. Die Wahl des Datenträgers hängt auch mit der Art des Archivs, seinen finanziellen und räumlichen Möglichkeiten ,sowie den Erwartungen der Nutzer ab.

Sicher ist, dass die Bestände in regelmäßigen Abständen auf neue Datenträger überspielt werden müssen. Dabei sollten die Bestände auf Datenträger des gleichen oder eines ähnlichen Typs überspielt werden, wenn sich das angegebene Verfallsdatum des alten Trägers nähert, oder die Datenträger besonderen Strapazen ausgesetzt waren. Die Bestände sollten auf einen Datenträger eines neuen Typs überspielt werden, wenn der alte Datenträger technisch zu veralten droht. Es ist unwahrscheinlich, dass ein Langzeitarchiv ohne diese beiden Typen von Migration auskommt. Informationen zu den möglichen digitalen Speichermedien finden Sie im Kapitel 10.3 dieses Handbuchs.

Genauso wie die Datenträger ständig überprüft und erneuert werden, müssen auch die Emulatorprogramme an die sich wandelnden technischen Bedingungen angepasst werden. Die möglichen Strategien zur Nutzung von Emulatoren entnehmen Sie bitte den Kapiteln 8.4 (Emulation) bzw. 17.4.4 (Interaktive Applikationen) dieses Handbuchs.

Die Wahl des Emulatorprogramms ist abhängig vom Spiel, das emuliert werden soll. Ein Spiel, das von einer Commodore64 Plattform stammt, kann

nicht mit einem Emulator verwendet werden, der eine ATARI VCS Plattform emuliert. Zudem sollten Archive bei der Auswahl ihrer Emulatoren weitere Faktoren, wie Benutzungsbedingungen, technische Weiterentwicklung und Hilfestellung durch die Entwicklergemeinde, Leistungsfähigkeit und Authentizität der Darstellung, Einfachheit der Bedienung und Installation, Verbreitung auf verschiedenen Hardware/Software Plattformen usw. bedenken. Es gibt Emulatorprogramme, die von einer internationalen Entwicklergemeinde ständig verbessert und an neue Plattformen angepasst werden. Die weltweit größte Gemeinde hat bisweilen der Multiple Arcade Machine Emulator, der für Arcade-Spiele verwendet wird. Der MAME Emulator[145] unterstützt zurzeit ca. 3.000 Spiele. Ein Ableger von MAME ist das Multiple Emulator Super System (MESS)[146], der Konsolen, Homecomputer, Handhelds und sogar Taschenrechner emuliert. Zur Zeit kann MESS für 442 unterschiedliche Plattformen genutzt werden. Es ist davon auszugehen, dass für nahezu jedes obsolete Spielsystem ein Emulator existiert.

Data Management/Ordnung/Verzeichnis

Es gibt keine moderne Bibliothek ohne Katalog und kein Archiv ohne Findmittel. Auch ein Archiv für Computerspiele braucht ein Verzeichnis. Benötigt werden Metadaten zur inhaltlichen und formalen Erschließung des Bestandes. Bibliotheken nutzen für die formale Erschließung von Computerspielen die Regeln für die alphabetische Katalogisierung für elektronische Ressourcen. Für ein digitales Archiv wäre der Metadatensatz des Dublin Core möglicherweise besser geeignet und unkomplizierter in der Anwendung. Die SPS hat für ihren Katalog einen kleinen Metadatensatz mit den wichtigsten formalen Daten entwickelt.

Die inhaltliche Erschließung erfolgt in der klassischen Bibliothek über Klassifikationen und Systematiken. Einige öffentliche Bibliotheken, die auch Computerspiele in ihrem Bestand führen, haben die verschiedenen Genre, nach denen sich die Computerspiele klassifizieren lassen, in ihre Systematiken eingebaut. Diese Klassifikationen sind aber nicht für ein Spezialarchiv geeignet, das ausschließlich Computerspiele sammelt. Die Klassifikation nach Genres und Subgenres scheint für die inhaltliche Erschließung zumindest der richtige Ansatz zu sein. Es sollte von diesem Punkt aus möglich sein, Spezialsystematiken mit einer höheren Indexierungsspezifität zu entwickeln, die für ein Computerspielarchiv angemessen sind.

---

145 Multiple Arcade Machine Emulator: Im Internet: http://mamedev.org/

146 Multiple Emulator Super System: Im Internet: http://www.mess.org/

Die inhaltliche und formale Erschließung eines Bestandes findet man auch in der traditionellen Bibliothek. Neu hinzukommen alle Metadaten, die wichtig für den langfristigen Erhalt eines digitalen Objektes sind. Seit neuestem gibt es Metadatenschemata, die diese Informationen erfassen und strukturieren. Bisher werden diese Schemata vor allem für die Langzeitarchivierung von digitalen Texten und Bildern verwendet. Erfahrungen mit der Erfassung von Computerspielen stehen noch aus. Das Metadatenschema PREMIS[147] scheint jedoch ein vielversprechender Kandidat für die Verzeichnung von Langzeitarchivierungsdaten und die Abbildung der Struktur von komplexen digitalen Objekten zu sein.

Ausgehend vom OAIS sollten die Metadaten und das Data Object gemeinsam in ein Archival Information Package (AIP) integriert werden.

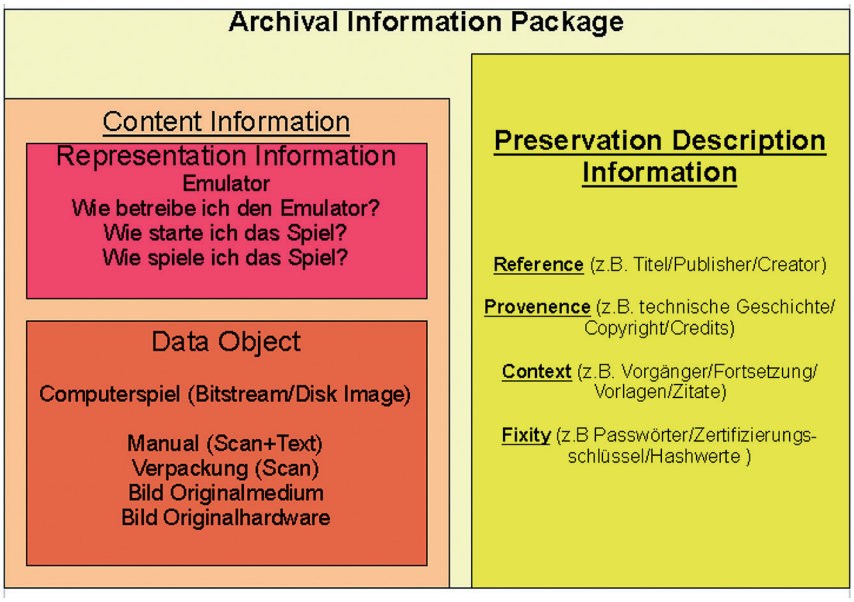


Abbildung 2: AIP für ein Computerspiel [148]

147 Data Dictionary for Preservation Metadata, Version 2.0, März 2008 Im Internet: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

148 Huth, Karsten; Lange, Andreas: Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive; In: ICHIM Berlin 04 – Proceedings: 2004; Im Internet: http://www.archimuse.com/publishing/ichim04/2758_HuthLange.pdf

Alle Informationen des SIP sollen auch im AIP enthalten sein. Als wichtigster Teil des AIP wird die sogenannte Representation Information angesehen. Sie umfasst alle Informationen, die nötig sind, um das Data Object, in unserem Fall das Computerspiel, zu nutzen und zu verstehen. Es wäre demnach ratsam, entweder den entsprechenden Emulator mit Gebrauchsanleitung dort abzulegen oder an dieser Stelle auf den benötigten Emulator zu verweisen. Einige Emulatoren sind schwer zu bedienen. Oft braucht man auch Kenntnisse über die emulierte Plattform, da man sonst nicht weiß, wie das Programm gestartet werden kann. Es ist deshalb ratsam, die nötigen Anweisungen zum Starten des Spiels mittels eines Emulator in einfachsten Schritten, der Representation Information beizufügen.

Die Representation Information ist nicht statisch. Es ist anzunehmen, dass auch die aktuellen Hardware/Software Konfigurationen in absehbarer Zeit veralten. Ebenso wie das Emulatorprogramm muss dann auch die Representation Information an die neuen technischen Bedingungen angepasst werden. Wie bereits erwähnt, scheint PREMIS für diese Aufgabe der beste Kandidat zu sein. Für Archive, die eine größere Freiheit bei der Auswahl ihrer Metadaten benötigen, scheint METS[149] eine gute Alternative zu sein. Beide Metadatenschemata sind in XML-Schemas umgesetzt worden und beanspruchen für sich, OAIS-konform zu sein. Näheres zu PREMIS und METS sowie über Langzeitarchivierungsmetadaten finden Sie im Kapitel 6 des Handbuchs.

Benutzung

Je besser und genauer die Angaben der Representation Information sind, umso einfacher wird die Benutzung des archivierten Computerspiels. Die Benutzung und die Übermittlung des Spiels hängen hauptsächlich von den Möglichkeiten des Archivs ab. Die Benutzung könnte Online, innerhalb der Räume des Archivs oder durch den Versand eines Datenträgers erfolgen. Neben dem Spiel muss auch der Emulator und die entsprechende Representation Information übermittelt werden. Alle genannten Teile zusammen ergeben das Dissemination Information Package (DIP in OAIS-Terminologie). Ein Beispiel für eine benutzerfreundliche Vermittlung wird zurzeit an der Universität Freiburg im Rahmen einer Dissertation entwickelt[150]. Der Nutzer kann einen Emulator und ein Computerspiel über ein Web-Applet in seinem Browserfenster laden und

149 Metadata Encoding and Transmission Standard: Official Website: Im Internet: http://www.loc.gov/standards/mets/

150 Suchodoletz, Dirk von; Welte, Randolph: Emulation: Bridging the Past to the Future: Im Internet: http://www.wepreserve.eu/events/nice-2008/programme/presentations/dirk_von_suchodoletz.pdf

starten. Das Spiel läuft ausschließlich auf seinem Bildschirm, es wird nicht auf die Festplatte des Archivnutzers heruntergeladen.

Zusammenfassung

Eine nachhaltige Archivierung von Computerspielen in einem größeren, öffentlichen, institutionellen Rahmen steht noch aus. Kleinere Organisationen mit dem nötigen technischen Know-how stehen bereit. Technische Arbeitsmittel wie Emulatoren oder Metadatenschemata im XML-Format sind bereits verfügbar. Eine Langzeitarchivierung von Computerspielen ist technisch möglich. Benötigt werden die entsprechenden Mittel, geeignete rechtliche Vorgaben und ein noch zu etablierender Wissenstransfer zwischen den klassischen Institutionen (Bibliotheken, Medienarchive, Museen) und den engagierten kleineren Organisationen mit den technischen Spezialkenntnissen. Auch die Europäische Union hat das Potential erkannt und 2009 das Projekt KEEP (Keep Emulator Environments Portable) ins Leben gerufen.

Die bisherige Arbeit des Computerspielemuseums Berlin und des Digital Game Archives zeigt, dass ein vielfältiger Bedarf (kultureller, wissenschaftlicher Art) auf der Nutzerseite existiert.

Literatur

AtariAge: Im Internet: http://www.atariage.com/

Data Dictionary for Preservation Metadata - Version 2.0, März 2008 Im Internet: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

Digital Game Archive (DiGA): Im Internet: http://www.digitalgamearchive.org/home.php

Dooijes, Edo Hans: *Old computers, now and in the future*, 2000: Im Internet: http://www.science.uva.nl/museum/pdfs/oldcomputers_dec2000.pdf

Fritz, Jürgen (1997): *Was sind Computerspiele?* In: Handbuch Medien: Computerspiele: Theorie, Forschung, Praxis/ hrsg. Jürgen Fritz und Wolfgang Fehr – Bonn: Bundeszentrale für politische Bildung Koordinierungsstelle Medienpädagogik. (S. 81-86)

Huth, Karsten / Lange, Andreas (2004): *Die Entwicklung neuer Strategien zur Bewahrung und Archivierung von digitalen Artefakten für das Computerspiele-Museum Berlin und das Digital Game Archive*; In: ICHIM Berlin 04 – Proceedings: 2004; Im Internet: http://www.archimuse.com/publishing/ichim04/2758_HuthLange.pdf

Internet Archive: Software Archive: Im Internet: http://www.archive.org/details/clasp

Lemon64: Im Internet: http://www.lemon64.com/

Lowood, Henry (2004): *Playing History with Games*: Steps Towards Historical Archives of Computer Gaming - American Institute for Conservation of Historic and Artistic Works. Electronic Media Grou, 2004 Im Internet: http://aic.stanford.edu/sg/emg/library/pdf/lowood/Lowood-EMG2004.pdf

Metadata Encoding and Transmission Standard: Official Website: Im Internet: http://www.loc.gov/standards/mets/

Multiple Arcade Machine Emulator: Im Internet: http://mamedev.org/

Multiple Emulator Super System: Im Internet: http://www.mess.org/ *Reference Model for an Open Archival Information System* (OAIS): CCSDS 650.0-B-1: Blue Book – Consultative Committee for Space Data Systems; 2002: Im Internet http://public.ccsds.org/publications/archive/650x0b1.pdf

Rothenberg, Jeff (1998): *Avoiding Technological Quicksand*: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources – Washington D.C.: Council on Library and Information Resources, 1998: Im Internet: http://www.clir.org/pubs/abstract/pub77.html

Rulemaking on Exemptions from Prohibition on Circumvention of Technological Measures that Control Access to Copyrighted Works: Im Internet: http://www.copyright.gov/1201/2003/index.html

Software Preservation Society (SPS): Im Internet: http://www.softpres.org/

Suchodoletz, Dirk von / Welte, Randolph: *Emulation: Bridging the Past to the Future*: Im Internet: http://www.wepreserve.eu/events/nice-2008/programme/presentations/dirk_von_suchodoletz.pdf *Video Games-Did They Begin at Brookhaven*, 2004 Im Internet: http://www.osti.gov/accomplishments/videogame.html

17.12 E-Mail-Archivierung

Karin Schwarz

Aufgrund der gestiegenen Verwendung von E-Mails in Verwaltung und Unternehmen und der gesetzlichen Anforderungen an die digitale Aufbewahrung digital erstellter, steuerrelevanter Dokumente ist die E-Mail-Archivierung[151] zu einem der bedeutendsten Themen im Bereich des Records Management avanciert. Hierbei ist die Beweisfähigkeit der Dokumente durch eine qualifizierte Langzeitarchivierung zu gewährleisten. Die gesetzlichen Bestimmungen für die Aufbewahrung von steuerrelevanten E-Mails und deren Anhänge stellen spezielle und teils sehr konkrete Anforderungen an die digitale Langzeitarchivierung. Die Ablage und Archivierung von E-Mails berücksichtigt deren Dokumentenstruktur und die Einbindung in E-Mail-Systeme. Für die Aufbewahrung von E-Mails bis zum Ablauf der Aufbewahrungsfrist werden Enterprise-Content-Management-Systeme empfohlen, in die die E-Mails importiert werden. Die praktizierte Aufbewahrung in Mail-Systemen und E-Mail-Archivierungssystemen wird ebenfalls beschrieben. Die dauerhafte Archivierung ähnelt derjenigen von elektronischen Dokumenten, muss aber ebenfalls den Export aus den Mail-Systemen berücksichtigen.

E-Mails haben im Geschäftsverkehr eine immer größere Bedeutung gewonnen. Auch wenn Unternehmen und Verwaltungen noch nicht vollständig auf die Führung elektronischer Unterlagen umgestiegen sind, fallen mittlerweile große Mengen an elektronischen Dokumenten in Form von E-Mails an. Darunter befinden sich immer mehr steuerrelevante oder beweiskräftige E-Mails mit Anhängen, die aus gesetzlichen Gründen oder wegen Erhaltung formatabhängiger Anwendungsmöglichkeiten (z.B. Tabellenkalkulation) auf lange Zeit bzw. auch auf Dauer elektronisch aufbewahrt werden. Sie sind ebenso wie die im Geschäftsumfeld erzeugten Text- und Bilddateien elektronische Dokumente und den papiergebundenen Unterlagen zunehmend gleichgestellt. In fast allen Gesetzen und Verordnungen in Bezug auf Schriftgut finden sich entsprechende Bestimmungen, insbesondere im Zusammenhang mit der qualifizierten elektronischen Signatur.[152] Die meisten Unternehmen und Verwaltungen werden wohl

151 Die Verwendung des Begriffs folgt hier dem mittlerweile im Bereich Records Management gebräuchlichen Sprachgebrauch, worunter eigentlich die geordnete Ablage von E-Mails zu verstehen ist. Aus archivfachlicher Sicht erfolgt eine Archivierung jedoch erst dann, wenn nach Feststellung des bleibenden Wertes eine E-Mail dauerhaft aufbewahrt wird.

152 Im Zivilrecht § 126 (1) BGB (Bürgerliches Gesetzbuch): „Soll die gesetzlich vorgeschriebene schriftliche Form durch die elektronische Form ersetzt werden, so muss der Aussteller der Erklärung dieser seinen Namen hinzufügen und das elektronische Dokument

bei der Aufbewahrung ihrer E-Mails erstmals mit der Umsetzung der digitalen Langzeitarchivierung konfrontiert. Bei der langfristigen Aufbewahrung und Archivierung von E-Mails kumulieren daher auch die allgemeinen Anforderungen der digitalen Langzeitarchivierung (DLZA). Daher kann dieser Anwendungsfall als ein Paradebeispiel der DLZA gelten.

Dabei haben E-Mails inhaltlich gesehen keine herausragende Rolle gegenüber anderen Textdokumenten wie Handelsbriefen, Verträgen, Weihnachtsgrüßen und Einladungen, Weisungen und Vermerken. Von den Papierunterlagen unterscheiden sie sich durch ihren digitalen Charakter („digital born“) und durch die Möglichkeiten der schnellen elektronischen Übermittlung einer Nachricht mit oder ohne anhängendes Dokument. Im Stellenwert der menschlichen Kommunikation sind sie zwischen Telefonanrufen und Papierbriefen einzuordnen. Durch den speziellen Charakter dieser Dokumententypen vermischen sich geschäftliche Angelegenheiten leicht mit persönlichem Vokabular oder privaten Themen. Ähnlichen Charakter haben auch die E-Mail-Accounts der Mitarbeiter: Mittlerweile wird es problematisch, private und geschäftliche E-Mails auseinanderzuhalten. Dies ist insofern von Bedeutung, als Unternehmen und Verwaltungen bei privaten Inhalten verpflichtet sind das Post- und Fernmeldegeheimnis zu wahren.[153] Als Pendant hierzu treten auch geschäftliche E-Mails in privaten Accounts auf, die außerhalb des Zugriffs der Unternehmen und Verwaltungen liegen.[154]

Unternehmen und Verwaltungen stehen vor der Aufgabe, die aufbewahrungswürdigen E-Mails herauszufinden, geordnet zu verwalten, sie gesetzeskonform aufzubewahren und zu archivieren. E-Mail-Management und E-Mail-Archivierung sind somit eng miteinander verknüpft, denn ein gutes E-Mail-Management erleichtert es die gesetzlichen, technischen und archivischen Anforderungen der digitalen Langzeitarchivierung zu gewährleisten. Dabei führen E-Mails – so scheint es – Unternehmen und Verwaltungen zu einst etablierten

mit einer qualifizierten elektronischen Signatur nach dem Signaturgesetz versehen.“ Sowie §§ 130a und 371a ZPO (Zivilprozessordnung) betreffend Beweiskraft von elektronischen Dokumenten. Im öffentlichen Recht § 3 (2) VwVfG (Verwaltungsverfahrensgesetz): „Eine durch Rechtsvorschrift angeordnete Schriftform kann, soweit nicht durch Rechtsvorschrift etwas anderes bestimmt ist, durch die elektronische Form ersetzt werden. In diesem Fall ist das elektronische Dokument mit einer qualifizierten elektronischen Signatur nach dem Signaturgesetz zu versehen....“

153 Unternehmen, die die private Nutzung von E-Mail-Systemen zulassen, werden zu einem Diensteanbieter nach §3(6) Telekommunikationsgesetz und müssen nach §85 des Telekommunikationsgesetzes das Post- und Fernmeldegeheimnis nach Art. 10 Grundgesetz wahren.

154 Knolmayer/ Disterer (2007), S.15.

Maximen des Records Management zurück: zu den unternehmens- oder behördenweit gültigen Policies und der Beachtung ihrer Einhaltung.[155] Denn die in den vergangenen Jahrzehnten sich immer weiter entwickelnde von Bearbeiter zu Bearbeiter variierende Ordnung der Unterlagen führt zu einer schwer handhabbaren Masse an Informationen, aus der man nicht unmittelbar das benötigte Expertenwissen adäquat herausziehen kann. Die Automatisierung in elektronischen Systemen, die z. T. für die E-Mail-Archivierung benutzt werden, setzt jedoch Eindeutigkeit in der Handhabung der Unterlagen voraus. Records Management hat über die E-Mail-Archivierung daher zunehmend an Bedeutung in Unternehmen und Verwaltungen gewonnen.[156]

Was sind E-Mails?

E-Mails sind über ein Computernetzwerk ausgetauschte Briefe. Zum Verständnis soll daher die Vorstellung eines herkömmlichen Papier-Briefes herangezogen werden: Ebenso wie dieser besteht die E-Mail aus den Übermittlungsdaten (quasi dem Briefumschlag), einem Header (dem Briefkopf) und einem Body (dem Text).

Die Übermittlung erfolgt im SMTP-Protocol (Simple Mail Transfer Protocol) als Dialog zwischen den Mailservern. Hier sind die IP (Internetprotokoll)-Adresse sowie Absender- und Empfängerdaten enthalten.

Der Header beinhaltet die Metadaten zu der E-Mail, die teilweise aus den Übermittlungsdaten entnommen werden: Daten zum Absender und Empfänger zur Adressierung, der Betreff zur Kodierung des Textes, Daten zum Übermittlungsweg (auch Angaben zu den genutzten Mail-Programmen) sowie nicht standardisierte Angaben (gekennzeichnet durch ein „X“ am Anfang der Headerzeile) zur Kontrolle der korrekten Übermittlung.[157] Die meisten E-Mail-Programme zeigen in der Grundeinstellung nicht den vollständigen Header an, was aber durch entsprechende Programmeinstellungen änderbar ist.

Der Body kann sowohl Texte als auch Bilder und aktive Elemente enthalten. Zum Body gehört auch die elektronische Signatur nach dem Signaturgesetz[158] zur Sicherung der Beweiskraft der E-Mail.

---

155 Vgl. hierzu: Kahn/ Blair (2004).

156 Hierzu insbesondere: Stettler (2006).

157 Für die Interpretation der komplizierten E-Mail-Header werden verschiedene Tools angeboten, insbesondere zur Recherche nach dem Besitzer einer IP-Adresse.

158 Signaturgesetz vom 16. Mai 2001 (BGBl. I S. 876), zuletzt geändert durch Artikel 4 des Gesetzes vom 26. Februar 2007 (BGBl. I S. 179).

Die E-Mail wird im Internet Message Format übertragen, die durch die Network Working Group als reines Textformat ursprünglich mit 7-Bit-ASCII-Zeichen festgelegt war. Da der 7-Bit-ASCII-Code nur über einen geringen Zeichensatz verfügt - bspw. sind die deutschen Umlaute nicht enthalten – ist eine Kodierung nach MIME notwendig, einem Kodierstandard, der es ermöglicht, Daten in beliebigen Formaten zu übertragen. MIME kodiert die Daten auf der Seite des Senders und dekodiert sie auf Seite des Empfängers.[159] Erweiterungen des Standards erlauben auch das Verschlüsseln und digitale Signieren der E-Mails (S/MIME).

An die E-Mails angehängt sind oftmals Dateien in den unterschiedlichsten Formaten, die für die Langzeitarchivierung jeweils spezifisch behandelt werden müssen.

Das Empfangen, Lesen, Schreiben und Senden von E-Mails übernehmen herstellerabhängige E-Mail-Programme. E-Mails entstehen also in einer proprietären Umgebung, die sich nur bedingt für die langfristige und in keiner Weise für die dauerhafte Archivierung eignet.

Auswahl aufbewahrungswürdiger E-Mails

Für den Zweck der Nachvollziehbarkeit des unternehmerischen oder Verwaltungshandelns sind allein records, d.h. Unterlagen, die innerhalb eines Vorgangs bearbeitet werden, aufbewahrungswürdig. Grüße, Notizen, Werbung, persönliche Post etc. werden ebenso wie in Papierform gehandhabt: entweder vernichtet oder in persönlichen Ordnern aufbewahrt. Im Zuge der E-Mail-Archivierung bedienen sich die Anwender der automatisierten Löschung nicht aufbewahrungswürdiger Dokumente, die daher genau definiert werden müssen und zwar noch vor ihrer eigentlichen Entstehung. Diese prospektive Festlegung führt jedoch zu einer Verunsicherung in der Praxis weswegen die Bearbeiter und Verwaltungen mitunter dazu neigen, jede E-Mail auf lange Zeit aufzubewahren. Die National Archives und Records Administration (NARA) in den USA hat hierzu eine Liste erstellt, die entsprechende nicht aufbewahrungswürdige Dokumente definiert.[160] Sie kann als Anhaltspunkt für die Vernichtung von Dokumenten dienen.

---

159 Daher müssen die für die verschiedenen Formate verwendeten MIME-Typen beim Mailsystem von Sender und Empfänger bekannt sein.

160 National Archives and Records Administration (Hrsg.) (2004), S.3. Genauer heißt es "Records...which have minimal or no documentary or evidential value."

Daneben sind für Unternehmen und Verwaltungen auch andere elektronische Dokumente aufbewahrungswürdig, bspw. Bilder zu Dokumentationszwecken oder Aufsätze, Artikel etc. als Teil eines Informationspools.

Von großer Tragweite und Bedeutung ist das Definieren der von Gesetz wegen aufbewahrungspflichtigen Dokumente, denn erst der wachsende Druck auf Unternehmen durch gesetzliche und regulatorische Anforderungen zur Erhöhung der Transparenz und Kontrollierbarkeit von Unternehmen und deren Mitarbeitern (Compliance) zwingt sie letztlich zur digitalen Langzeitarchivierung. So definiert der Gesetzgeber steuerrelevante digital-born-Dokumente,[161] die mit einer qualifizierten elektronischen Signatur versehen sind, zum rechtsverbindlichen Original, welches weiterhin in digitaler Form aufbewahrt werden muss.[162]

Die Unternehmen müssen hierbei verschiedene allgemeine und spezielle rechtliche Bestimmungen beachten. Grundlegend gelten hier v. a. die Bestimmungen der Abgabenordnung für die ordnungsgemäße Aufbewahrung von Unterlagen (§147) sowie für die Buchführung und weitere Aufzeichnungen (§§145, 146), die für digitale Unterlagen in den „Grundsätzen zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen“ (GdPdU) sowie für die Buchführungssysteme in den „Grundsätzen ordnungsmäßiger DV-gestützter Buchführung“ (GoBs) näher spezifiziert sind. Dies betrifft die steuerrelevanten Unterlagen, die in jedem Unternehmen anfallen und daher den Hauptteil der E-Mail-Archivierungsthematik ausmachen. Die Anbieter von Programmen zur E-Mail-Archivierung gehen daher überwiegend von einer sechs bis zehn Jahren währenden Aufbewahrung von Dokumenten aus.

Branchenspezifische Bestimmungen schreiben mitunter längere Aufbewahrungsfristen vor, bspw. bei Patientenakten oder Strahlenschutzunterlagen (30 Jahre). Ob sich unter den anfallenden E-Mails aufbewahrungsrelevante Dokumente befinden und ob diese weiterhin digital aufbewahrt werden müssen, muss von Fall zu Fall geprüft werden. Entsprechend müssten die Produkte zur E-Mail-Archivierung die Möglichkeit einer Aufbewahrung über einen längeren Zeitraum anbieten.

---

161 Eine Auflistung findet sich in: Brand, Thorsten/ Groß, Stefan/ Zöller, Bernhard (2003), S.7-8.

162 Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GdPdU) (2001), Abschnitt III, Punkt 1.

In den Unternehmen werden zur besseren Übersichtlichkeit Richtlinien für die Aufbewahrung von E-Mails erstellt. Auch für die öffentliche Verwaltung existieren entsprechende Policies. Zu nennen wären beispielsweise die E-Mail-Regularien der NARA.[163]

Anforderungen an die langfristige Aufbewahrung

Die Anforderungen an die langfristige Aufbewahrung von E-Mails ergeben sich aus den genannten gesetzlichen Bestimmungen.
Zu diesen Anforderungen gehören nach der GDPdU und der GoBs:

- Die Unveränderbarkeit und Erhaltung der E-Mails. Eine E-Mail darf nicht mehr nachträglich verändert[164] oder gelöscht werden oder gar bei der Übermittlung in ein Archivierungssystem verloren gehen.[165] Eine Komprimierung sowie eine Migration in andere Formate, Systeme, Software und auf andere Medien muss entsprechend verlustfrei erfolgen.
- Integrität der E-Mails und des Systems. Änderungen am Dokument, aber auch in der Organisation und der Struktur der Dokumentenmenge müssen so protokolliert werden, dass der ursprüngliche Zustand wiederherstellbar wäre.
- Der zeitnahe Zugriff auf E-Mails. Die Ablage, Klassifizierung und Indexierung erfolgt so, dass mit geeigneten Retrievaltechniken das Dokument zeitnah[166] aufgerufen werden kann. Ebenso muss auch genau das Dokument gefunden werden, das gesucht worden ist.
- Die Authentizität der E-Mails. Das System muss die Dokumente in der Form anzeigen, in der es erfasst wurde.
- Datensicherheit und Datenschutz der E-Mails. Nach den gesetzlichen Bestimmungen müssen über die Lebensdauer des verwendeten Systems hinaus Datensicherheit und –schutz gewährleistet werden können.

163 National Archives and Records Administration Regulation on E-mail (2006) §1234.24 Standards for managing electronic mail records.

164 § 146 (4) AO (Abgabenordnung).

165 § 146 (1) AO (Abgabenordnung).

166 Die Finanzverwaltung unterscheidet beim Zugriff auf steuerrelevante Unterlagen zwischen unmittelbarem, mittelbarem Zugriff und Datenträgerüberlassung, vgl. dazu: Brand, Thorsten/ Groß, Stefan/ Zöller, Bernhard (2003). S.14-15.

Die GdPdU fordert zusätzlich die Aufbewahrung in einem Format, das eine maschinelle Auswertbarkeit ermöglicht.[167]

Die Verantwortlichkeit für die Aufbewahrung von E-Mails liegt ab dem Zeitpunkt des Übergangs in die Verfügungsgewalt des Empfängers bzw. eines empfangsberechtigten Dritten beim Empfänger. Dies gilt unabhängig davon, wie beim Eingang mit E-Mails verfahren wird. Die Folgen einer nicht dem Bearbeiter zugestellten E-Mail trägt der Empfänger, auch wenn eine automatische Viren- oder Spam-Prüfung positiv ausgefallen sein sollte.

Benötigte Metadaten für die Aufbewahrung von E-Mails

Zur Aufbewahrung von E-Mails werden Metadaten aus dem Header der E-Mail sowie Metadaten über die Anhänge und die Verknüpfung zwischen E-Mail und Anhang benötigt, um die oben beschriebenen Anforderungen umsetzen zu können. Die Metadaten müssen wegen des Nachweises der Vollständigkeit und der Authentifizierung einheitlich benannt und in einer strukturierten Form, fest mit dem Dokument verbunden, aufbewahrt werden.[168]

Die Auswahl der Metadaten kann die für elektronische Dokumente verwendete Metadatenstruktur übernehmen, so dass keine besonderen E-Mail-Metadatenstruktur formuliert werden muss. Sofern für elektronische Dokumente kein Metadatensatz definiert ist, können Anregungen anderer Unternehmen oder der in der Verwaltung verwendete Metadatensatz XDOMEA für den Austausch von Dokumenten als Vorlage genutzt werden.[169] Dies empfiehlt sich zumal bei XDOMEA für den Bereich „Adresse“ eine entsprechende Komponente existiert, die auch die Übermittlungsart berücksichtigt.[170]

Zu den Metadaten sollten sogenannte Protokolldaten bzw. Bearbeitungsdaten, die die Dokumentenverwaltung betreffen und für jedes andere Dokument auch anfallen, hinzugefügt werden: Angaben zum Kontext, in welchem die E-Mail bearbeitet wurde, und zum Verlauf der Bearbeitung.

---

167 Darunter fallen beispielsweise PDF-Dokumente nicht. Eine Liste möglicher Formate ist abgedruckt in: Brand, Thorsten/ Groß, Stefan/ Zöller, Bernhard (2003), S.20.

168 Eine virtuelle Verbindung über eine Datenbank bedürfte der langfristigen Verfügbarkeit und Lesbarkeit der Datenbank und hängt von der Langlebigkeit des Produkts etc. ab.

169 Die Felder in XDOMEA sind optional. Derzeit wird XDOMEA 1.0 verwendet, das jedoch momentan erweitert wird zu XDOMEA 2.0. Verschiedene Dokumente zum XML-Schema sowie eine tabellarische Übersicht sind abrufbar unter: http://www.koopa.de/produkte/xdomea.html

170 Für die weitere dauerhafte Archivierung empfiehlt sich die Berücksichtigung von Metadatensätzen für die Übermittlung in ein Archiv und deren Aufbewahrung im Archiv (bspw. XArchiv und XBarch).

Erst die Beziehung zwischen E-Mail, Bearbeiter und Geschäftsprozess zeigt deren Bedeutung und Funktion im Unternehmen bzw. der Verwaltung auf.[171] Auch diese Anforderungen werden in XDOMEA unterstützt.

Bei der Auswahl der festen Metadaten ist es sinnvoll, sich auf die standardisierten Metadaten des E-Mail-Headers zu stützen und die Mail-System bezogenen – mit X gekennzeichneten – Metadaten nur optional aufzunehmen, ansonsten kann aufgrund der Unterschiedlichkeit der Header-Informationen eine fehlerfreie automatische Extraktion der Metadaten nicht gewährleistet werden.[172]

Formen der langfristigen Aufbewahrung von E-Mails

Im Mail-System

Mail-Systeme stehen entweder über einen eigenen Mailserver (v. a. in Unternehmen und Verwaltungen) oder einen Online-Mailserver (v. a. im privaten Bereich) zur Verfügung, bei letzteren spricht man statt von E-Mails auch von Webmails. Vom Mailserver werden die E-Mails mittels eines Protokolls übertragen oder abgerufen. Hierfür werden meistens die Protokolle POP 3 (Post Office Protocol Version 3) und IMAP (Internet Message Access Protocol) verwendet. Bei IMAP verbleiben die E-Mails auf dem Mailserver und werden dort verwaltet während POP 3 die E-Mails auf den PC überträgt. Eine langfristige Aufbewahrung von E-Mails ist für Nutzer der Online-Mailserver entsprechend auch von dem verwendeten Protokoll abhängig: bei IMAP *scheinen* sich die E-Mails nur auf dem eigenen Rechner zu befinden, werden jedoch von fremden Stellen aufbewahrt. Daher ist eine Speicherung von E-Mails aus Online-Mailservern auf den eigenen Server oder Rechner entsprechend ratsam. Weil die Verwendung von IMAP und sein Nachfolger[173] wegen der Möglichkeiten der implizierten E-Mail-Verwaltung zunimmt, sollte beständig überprüft werden, ob E-Mails online oder auf dem eigenen Rechner vom Mail-System abgelegt werden. Da

171 Dies ist eine der zentralen Forderungen des Records Management. Die Kontextinformationen stellen sicher, dass zu einem Vorgang sämtliche Dokumente vorliegen. Dies bildet die Basis für die Nachvollziehbarkeit von *dokumentenbasierten* Prozessen und deren Optimierung im Qualitätsmanagement. Die kontextgebundene Aufbewahrung von *werkbezogenen* Unterlagen bzw. Daten (z.B. bei literarischen Texten und Forschungsdaten) wird seltener im Kontext eines Prozesses vorgenommen.

172 Model Requirements for the Management of electronic records. MoReq 2 Specification (2008). S.73-76.

173 IMAP wird derzeit durch SMAP (Simple Mail Access Protocol) weiterentwickelt, das verschiedene Änderungen beim Verwalten der E-Mails vorsieht.

Online-Mailserver insbesondere im privaten Bereich genutzt werden, sollten auch Privatleute steuerrelevante oder beweiskräftige Dokumente nicht auf diesen Online-Mailservern belassen.[174]

Der tägliche Speicherbedarf der E-Mails liegt bei etwa 10 MB pro Mitarbeiter, was einem Durchschnitt von täglich 34 versendeten und 84 empfangenen E-Mails entspricht.[175] Die langfristige Speicherung von E-Mails auf dem Mailserver ist mit hohen Kosten verbunden. Durch die Bereitstellung schneller Zugriffsmöglichkeiten wird teurer Speicherplatz verbraucht, die für ältere E-Mails nicht nötig wären. Die Nutzung der Speicherhierarchie[176] zur Ablage der E-Mails nach der Zugriffshäufigkeit bietet hier eine Lösung.

Unter der Belastung großer Speichermengen von E-Mails werden die Mailsysteme instabiler, so dass die für die Datensicherung bereitstehende Zeit nicht mehr ausreicht, um die Datenmengen zuverlässig zu sichern. Störend wirkt sich hierbei die Anhäufung redundanter Daten aus durch die Sicherung der gesamten E-Mail-Datenbank, auch wenn sich nur ein kleiner Teil verändert hat. Zur Entlastung der Systeme löschen viele E-Mail-Programme die Daten nach Ablauf einer bestimmten Frist. Erfolgt dies jedoch unkontrolliert und ohne verwaltungs- oder unternehmensinterne Richtlinien können wichtige oder auch rechtsrelevante E-Mails und Dokumente verloren gehen. Die Wiederherstellung der Daten aus Backup-Datenbeständen erfordert zusätzlichen Aufwand durch IT-Spezialisten.

Die Klassifizierung von E-Mails im Mail-System wird über das Anlegen von Ordnern geregelt. Da meistens jeder Benutzer selbstständig eine Ordnerstruktur festlegen kann, führt dies zu individuellen Ablagesystemen mit der bekannten Unübersichtlichkeit für andere Benutzer. Zudem ist durch die Verwendung von Passwörtern der Zugriff auf die E-Mails an die Anwesenheit des Bearbeiters gebunden. Für die Ordnerstrukturen nutzen Die Mail-Programme komprimierte Dateien oder eigene (proprietäre) kleinere Datenbankanwendungen, die auf längere Sicht ebenfalls zu den bekannten Problemen bei der Langzeitarchivierung führen können.

---

174 Die E-Mail-Archivierung im privaten Bereich hat das Projekt PARADIGMA (Personal Archives Accessible in Digital Media) thematisiert und Empfehlungen für die Errichtung eines privaten (digitalen) Archivs herausgegeben: (o. V.) Guidlines for creators of personal archives (o.J.).

175 Knolmayer/ Disterer (2007), S.20.

176 Die Speicherhierarchie ordnet die Arbeits- und Datenspeicher nach dem Kriterium der Zugriffsgeschwindigkeit und der Speicherkapazität. Danach sind Speicher mit einer hohen Kapazität langsamer im Zugriff als solche mit geringerer Speicherkapazität. Entsprechend können nicht häufig benötigt digitale Unterlagen auf langsamen, aber mit großem Platz ausgestatteten Speichern abgelegt werden.

In E-Mail-Archivierungssystemen

Es handelt sich dabei um speziell für die E-Mail-Aufbewahrung entwickelte Produkte vor dem Hintergrund der Speicher- und Verwaltungsproblematik. Unternehmen greifen auf diese sogenannten E-Mail-Archivierungssysteme zurück, die für die langfristige, nicht jedoch für die dauerhafte Archivierung brauchbar sind. Zu berücksichtigen bleibt auch hier, dass es sich um hochproprietäre Systeme handelt und auf die Ansprüche der digitalen Langzeitarchivierung ebenso zu achten ist wie auch auf eine mögliche Exportfunktion in eine nicht-proprietäre Archivierungslösung zur dauerhaften Aufbewahrung.

Die Systeme zeichnen sich durch verschiedene Lösungen der Speicherung und der Verwaltung von E-Mails aus. Hinsichtlich der Verwaltung bieten die Hersteller zum größten Teil Systeme an, die die Redundanz von Anhängen oder auch E-Mails vermeiden helfen, die E-Mails automatisch indexieren und auch klassifizieren. Da diese Funktionen kostenintensiv sind, erfolgt mitunter auch keine Auswahl der aufbewahrungswürdigen Daten, sondern eine komplette Aufbewahrung der E-Mail-Bestände. Vor- und Nachteile dieser Lösung müssen die Unternehmen entsprechend abwägen.

Bei der Speicherung werden die Daten in der Regel komprimiert. Was aus speichertechnischer Sicht günstig sein kann, bleibt jedoch ein Risiko für die Anforderung der Unveränderbarkeit von E-Mails, da komprimierte Daten zu Datenverlusten führen können, insbesondere bei Bildern. Eine verlustfreie Komprimierung in Archivierungsformaten sollte daher zum Anspruch eines Archivierungssystems gehören, zumal viele Daten über den ersten Migrationsschritt hinaus aufbewahrt werden müssen.

Im Hinblick auf die Anforderungen der Lesbarkeit, Zugriffsmöglichkeit und die Verwaltung von E-Mails hat eine belgische Studie 2006 zu dem Ergebnis geführt, dass die angebotenen E-Mail-Archivierungssysteme keinen Mehrwert gegenüber den E-Mail-Systemen haben. Ziel sei v. a. die Verringerung der Ladezeiten durch Datenkomprimierung mit den genannten Gefahren.[177]

Bei der Anschaffung eines E-Mail-Archivierungssystem muss auch beachtet werden, dass diese Lösung mittelfristig zu Informationsinseln führen wird. Andere elektronische Dokumente zum selben Geschäftsprozess oder Vorgang liegen in anderen elektronischen Systemen, eventuell kommt sogar noch eine papiergebundene Aufbewahrung hinzu. Eine zumindest virtuelle Zusammenführung der Dokumente ist nicht möglich, die Bildung von elektronischen Akten wird behindert. E-Mails unterscheiden sich inhaltlich nicht von anderen

177 Boudrez, Filip (2006). S.14.

elektronischen Dokumenten, die Tatsache der elektronischen Übermittlung kann daher nicht ausschlaggebend für einen eigenen Aufbewahrungsort sein.

In Enterprise-Content-Management (ECM)-System abhängigen Produkten und Middleware-Produkten

Beide Produkte integrieren E-Mails in ein ECM-System[178] wobei die Middleware-Produkte[179] nur die Verwaltung der E-Mails bis zur Speicherung in ein ECM-System abdecken.

Besonderes Augenmerk gilt hier der Schnittstelle zwischen Mail-System und ECM-System. Für den Export stellen die Mail-Systeme Formate zur Verfügung, die eine Formatkonvertierung der E-Mail, nicht aber des Anhangs bedeuten. Unter den Formaten befinden sich solche, die sich wegen ihrer Proprietät nur für eine mittelfristige Ablage, nicht jedoch für eine Langzeitarchivierung eignen würden (darunter bspw. das msg-Format von Microsoft). Bei der Auswahl eines Export-Formats ist zu beachten, dass die Übertragung aller nötigen Metadaten gewährleistet ist und – aus bearbeitungstechnischen Gründen – dass eine Rückübertragung in das System möglich ist, um die E-Mail zu beantworten oder weiterzuleiten.[180] Die Formatkonvertierung der E-Mail-Anhänge muss gesondert erfolgen und die gleichen Anforderungen wie andere elektronische Dokumente erfüllen. Bei steuerrelevanten Dokumenten ist die Anforderung der maschinellen Auswertbarkeit von Dokumenten, die zur Weiterverarbeitung in Datenbanken verwendet werden, ein wichtiges Kriterium für die Langzeitarchivierung. Nach dieser Anforderung können bspw. Excel-Tabellen nicht im eigentlich archivfähigen Format PDF oder PDF/A aufbewahrt werden.[181]

Das anforderungsgerechte Exportieren ist auf „händischem" Weg möglich; dafür muss das Mail-System über die Einstellungsfunktionen angepasst werden.[182] Dies ist jedoch auch sehr fehleranfällig: so muss etwa bei jeder Aktion das Exportformat gewählt werden und die Anhänge gesondert in ein anderes System überführt werden.

---

178 Andere Systeme wie Dokumenten-Management-Systeme oder Vorgangsbearbeitungssysteme sind hier ebenfalls gemeint.

179 Programme, die zwischen zwei Applikationen vermitteln, hier zwischen Mail- und ECM-System.

180 Ohne letztere Möglichkeit würden die Bearbeiter die E-Mails so lange wie möglich im Mail-System halten und die Aktualität des ECM-Systems verhindern.

181 Vgl. hier insbesondere Abschnitt 3 der GDPdU.

182 Auf die Angaben hierzu wird an dieser Stelle verzichtet. Genaue Informationen finden sich hierzu bei Boudrez, Filip (2006). S.21-27.

Der Export durch die ECM-Systeme oder die Middleware-Produkte verläuft zwar automatisch, bedarf aber der erwähnten Regularien bzw. Nennung der Anforderungen gegenüber dem Hersteller. Auch wenn diese vielfach vordefinierte Schnittstellen anbieten, sollten sie auf die oben genannten Anforderungen hin geprüft werden. Das betrifft nicht nur den Export und Import der E-Mails, sondern auch deren Weiterbehandlung im ECM: Das Zusammenspiel von E-Mail, Metadaten und Anhängen muss auch nach den serienmäßig implementierten Migrationsmöglichkeiten[183] gewährleistet sein. So darf bei einer Umwandlung in ein pdf-Format der Anhang nicht verloren gehen. Manche Systeme fügen zwar das Symbol für den Anhang in das Dokument ein, jedoch keinen Link, so dass der Anhang entsprechend schwer zu finden oder bei Löschung der ursprünglichen Datei gar nicht mehr auffindbar ist.[184]

Die Überführung in ein ECM-System sollte so schnell wie möglich erfolgen, um durch die Bearbeitung von Dokumenten in nur einem statt zwei verschiedenen Systemen Arbeitsabläufe zu vereinfachen. Zum anderen sind die ECM-Speicher wesentlich sicherer als die Mail-System-Speicher.

In einem Digitalen Archiv

Die dauerhafte Aufbewahrung von E-Mails bedeutet den Export aus einer proprietären Umgebung in eine möglichst nicht-proprietäre Archivierungslösung. Hierbei sind E-Mails entweder aus E-Mail-Systemen oder, bei Verwendung von ECM-Systemen, aus diesen herauszulösen. Relevant ist hier der Export aus den E-Mail-Systemen, da in EMC-Systemen E-Mails wie die übrigen elektronischen Dokumente behandelt werden. Ebenso wie bei anderen elektronischen Dokumenten ist hier die Überführung in ein Standard-Austauschformat zusammen mit den Metadaten erforderlich.[185]

---

183 So wird bspw. nach Import einer E-Mail mit Anhang in ein ECM-System die E-Mail im Word-Format gespeichert und mit dem Anhang als selbständigem Dokument im ursprünglichen Format verlinkt. Zur längerfristigen Sicherung bieten diese Systeme oftmals die Umwandlung in das pdf-Format an.

184 Zu den Anforderungen an ein ECM hinsichtlich der E-Mail-Aufbewahrung sind die Angaben in Moreq 2 aufschlussreich. Moreq hat sich zu einem europäischen Standard mit spezifischen Anforderungen an das Records Management entwickelt. Dem E-Mail-Management wird ein eigenes Kapitel gewidmet. Model Requirements for the Management of electronic records. MoReq 2 Specification (2008). S.73 – 76. Vgl. auch die Bestimmungen der National Archives and Records Administration Regulation on E-mail (2006), §1234.24 (b) – (3)-(ii).

185 National Archives and Records Administration Regulation on E-mail (2006), §1234.22 (a) - (3) und (4), §1234.24 (a) - (1) und (2) sowie (b) – (1) und (2).

Als Archivierungsformate für E-Mails können die Standardformate XML oder PDF/A herangezogen werden. XML hat hierbei den Vorteil, dass die Daten strukturiert aufbewahrt werden und gegebenenfalls für eine firmeninterne oder –externe Auswertung in eine Datenbank überführt werden könnten. Wie oben erwähnt ist eine direkte Migration in ein Archivierungsformat zur revisionssicheren Langzeitarchivierung serienmäßig nicht möglich. Dies würde die direkte Archivierung unter Umgehung proprietärer Systeme bedeuten.

Eben dies hat sich das niederländische Projekt „Testbed Digitale Bewaring" zur Aufgabe gemacht. Ziel war es eine Software zu entwickeln, die den Export von E-Mails aus den E-Mail-Systemen zur direkten dauerhaften Archivierung vollführt. Während des laufenden Betriebs soll dies ohne den Umweg über ein ECM-System geschehen und damit die Tatsache berücksichtigen, dass elektronische Dokumente auch ohne ECM-Systeme anfallen und dauerhaft archiviert werden sollen. Bei der Umsetzung wurde das Interface des E-Mail-Programms durch Buttons modifiziert, die Funktionen zur Konvertierung der E-Mails mit Text und Metadaten in ein XML-Format auslösen. Die Daten werden auf den Server des Archivs kopiert und in eine HTML-Datei umgewandelt, welche an das E-Mail-Programm zurückgeführt wird. Anhänge der E-Mail werden wie andere elektronische Dokumente an das Archiv übergeben, wobei jedoch die Verbindung zwischen E-Mail und Anhang durch einen Link in der XML-Datei gewährleistet bleibt.[186]

Literatur:

Boudrez, Filip (2006): *Filing and archiving e-mail*. http://www.expertisecentrumdavid.be/docs/filingArchiving_email.pdf (3.11.2008), in verkürzter Fassung: Boudrez, Filip (2005) Archiveren van E-Mail. http://www.expertisecentrumdavid.be/davidproject/teksten/Richtlijn1.pdf

Brand, Thorsten/ Groß, Stefan/ Zöller, Bernhard (2003): *Leitfaden für die Durchführung eines Projektes zur Abdeckung der Anforderungen der Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GDPdU)*. Version 1.0. November 2003. http://www.elektronische-steuerpruefung.de/checklist/voi_leitf.pdf

186 Der Aufbau der XML-DTD ist online verfügbar unter: www.digitaleduurzaamheid.nl/bibliotheek/images/E-mailplaatje.gif. Auf der Homepage steht auch eine Demonstration der Bearbeitungsschritte zur Verfügung.

Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen (GdPdU) (2001). BMF [Bundesministerium für Finanzen]-Schreiben vom 16. Juli 2001 – IV D 2 – S0316 – 136/01. http://www.bundesfinanzministerium.de/nn_314/DE/BMF__Startseite/Aktuelles/BMF__Schreiben/Veroffentlichungen__zu__Steuerarten/abgabenordnung/006,templateId=raw,property=publicationFile.pdf

Institut der Wirtschaftsprüfer (Hrsg.) (2006) : *IDW RS FAIT 3. Grundsätze ordnungsmäßiger Buchführung beim Einsatz elektronischer Archivierungsverfahren*. In: Die Wirtschaftsprüfung Nr.22. 2006. S. 1465ff.

Kahn, Randolph A./ Blair, Barclay T. (2004): *Information Nation. Seven keys to information management compliance*. Silver Spring: AIIM.

Knolmayer, Gerhard / Disterer, Georg (2007): *Anforderungsgerechte Dokumentation der E-Mail-Kommunikation. Rechtliche Vorschriften, technische Lösungen und betriebliche Regelungsbedarfe*. Arbeitsbericht Nr. 192 vom Institut für Wirtschaftsinformatik der Universität Bern. http://www.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-192.pdf

Model Requirements for the Management of electronic records. MoReq 2 Specification (2008). http://www.cornwell.co.uk/moreq2/MoReq2_typeset_version.pdf

National Archives and Records Administration (Hrsg.) (2004): General Records Schedule 23. Records Common to Most Offices Within Agencies. http://www.archives.gov/records-mgmt/ardor/grs23.html

National Archives and Records Administration Regulation on E-mail (2006). 36 Code of Federal Regulations (C. F. R.) Part 1234, *Electronic Records Management*. Subpart c – Standards for the Creation, Use, Preservation, and Disposition of Electronic Records. http://www.archives.gov/about/regulations/part-1234.html

Stettler, Niklaus et. al. (2006): *Synthesebericht Records Management Survey Schweiz in ausgewählten Sektoren der Privatwirtschaft (2005/2006)*, hg. v. Ausschuss eArchiv des Vereins Schweizer Archivarinnen und Archivare und der HTW Chur. http://isb.admin.ch/themen/architektur/00078/00201/00202/index.html?download=NHzLpZeg7t,Zn4Z2qZpnO2Yuq2Z6gpJCDdoF3hGym162epYbg2c_JjKbNoKSn6A--&lang=fr

(o. V., 2002): *E-mail-XML-Demonstrator. Technical description*. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/email-demo-en.pdf

(o. V., o. J.) *Guidlines for creators of personal archives*. http://www.paradigm.ac.uk/workbook/appendices/guidelines-tips.html

18 Praxisbeispiele

18.1 Einführung

Regine Scheffel

nestor hat als Kompetenzzentrum in Sachen Langzeitarchivierung digitaler Objekte ganze Arbeit geleistet: Fachleute aus dem Kulturerbebereich haben das neue Arbeitsfeld angenommen, Projekte und erste Erfahrungsberichte zeigen, dass auf unterschiedlichen Gebieten Lösungen für die Herausforderung, unser digitales Erbe zu erhalten, erarbeitet werden. Neben Archiven, Bibliotheken und Museen arbeiten wissenschaftliche Einrichtungen, Firmen, Organisationen und Behörden daran die digitale Langzeitarchivierung in Organisationsstruktur, Aufgabenbereichen und Arbeitsprozessen zu verankern. Noch sind viele auf dem Weg und betrachten ihre Lösungen nicht als endgültig abgeschlossen. Das ist bei diesem Arbeitsfeld wohl auch gar nicht anders möglich, zu stark sind die zu erhaltenden Objekte, aber auch Konzepte und technische Lösungen zur digitalen Langzeitarchivierung Entwicklungsprozessen unterworfen.

Dennoch freut sich das Herausgeberteam die ersten Praxisbeispiele vorstellen zu können:

Die Deutsche Nationalbibliothek (DNB) hat in Deutschland die Vorreiterrolle gespielt bei der Einrichtung eines Archivsystems. Mit kopal liegt nun ein System vor, das nicht nur die große Menge digitaler Publikationen für die DNB archiviert, sondern auch anderen Einrichtungen offen steht.

Das Bibliotheksservicezentrum (BSZ) in Konstanz hat die Bildarchivierung für die Staatsgalerie Stuttgart auf Dauer übernommen. Noch ist das Archivsystem nicht abschließend realisiert, doch sind wesentliche OAIS-konforme Festlegungen und Workflows definiert und stehen weiteren Einrichtungen zur Nachnutzung zur Verfügung.

Der Blick über den Tellerrand zeigt, dass auch außerhalb von Landes- und Bundeseinrichtungen der Funke übergesprungen ist: Der Bundestag hat forsch damit begonnen seine Internetangebote in transparenter Auswahl in ein langzeitarchivierungstaugliches Webarchiv zu überführen und arbeitet nun an der Optimierung des Systems.

Man darf gespannt darauf sein, welche Einrichtungen oder Firmen ihre Praxis der digitalen Langzeitarchivierung in den nächsten Ausgaben dieses nestor Handbuchs vorstellen werden.

18.2 Langzeitarchivierung von elektronischen Publikationen durch die Deutsche Nationalbibliothek

Maren Brodersen und Sabine Schrimpf [1]

Seit das Gesetz über die Deutsche Nationalbibliothek vom 22. Juni 2006 in Kraft getreten ist, erstreckt sich der Sammelauftrag der Deutschen Nationalbibliothek auch auf „Medienwerke in unkörperlicher Form", d.h. auf Netzpublikationen. Konkret hat sie den Auftrag, alle in Deutschland veröffentlichten und deutschsprachigen Medienwerke „zu sammeln, zu inventarisieren, zu erschließen und bibliografisch zu verzeichnen, auf Dauer zu sichern und für die Allgemeinheit nutzbar zu machen".[2] Dabei gelten als Medienwerke alle Darstellungen in Schrift, Bild und Ton, die in körperlicher Form verbreitet werden (d.h. auf Papier, elektronischen oder anderen Datenträgern) oder in unkörperlicher Form über öffentliche Netze, in der Regel das Internet, zugänglich gemacht werden.

Für diese Situation werden Verfahren für die Sammlung und Langzeitarchivierung von Netzpublikationen an der Deutschen Nationalbibliothek permanent weiter entwickelt und implementiert. Dieser Artikel kann daher lediglich einen Überblick über den aktuellen Stand der Sammlung und Langzeitarchivierung von Netzpublikationen an der Deutschen Nationalbibliothek geben. Es wird eingegangen auf den in Pflichtablieferungsverordnung und Sammelrichtlinien näher bestimmten Sammelauftrag der Deutschen Nationalbibliothek, auf speziell für Netzpublikationen entwickelte Ablieferungs- und Erschließungsverfahren und die Langzeitarchivierung von Netzpublikationen.

Sammelgebiet

Nicht jede deutschsprachige Webseite im Internet gehört automatisch zum Sammelauftrag der Deutschen Nationalbibliothek. Zwei Dokumente regeln die Einzelheiten zum Sammelgebiet Netzpublikationen und schränken die Ablieferungspflicht nach bestimmten Selektionskriterien ein: die Pflichtablieferungsverordnung und die Sammelrichtlinien.

---

1 Mit Unterstützung von Sarah Hartmann, Susanne Puls und Tobias Steinke.

2 Gesetz über die Deutsche Nationalbibliothek (DNBG) vom 22. Juni 2006, hier § 2 Abs. 1, veröffentlicht im Bundesgesetzblatt Jahrgang 2006 Teil I Nr. 29, ausgegeben zu Bonn am 28. Juni 2006, verfügbar unter http://www.d-nb.de/wir/pdf/dnbg.pdf
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

In der Pflichtablieferungsverordnung vom 17. Oktober 2008 (PflAV)[3] werden eine Reihe von Netzpublikationen von der Ablieferungspflicht ausgenommen, darunter Publikationen, die nicht von besonderem öffentlichen Interesse sind, wie Netzpublikationen, die lediglich privaten oder gewerblichen Zwecken dienen, Netzpublikationen, die lediglich einer privaten Nutzergruppe zugänglich sind oder Netzpublikationen von Kreisen, Gemeinden und Gemeindeverbänden, die ausschließlich amtlichen Inhalt enthalten. Auch Vorabveröffentlichungen, reine Software- oder Anwendungstools, Fernseh- und Hörfunkproduktionen und Spiele fallen nicht unter den Sammelauftrag der Deutschen Nationalbibliothek. Ebenfalls nicht sammelpflichtig sind E-Mail-Newsletter, sofern sie kein Webarchiv haben und Kommunikations-, Diskussions- oder Informationsinstrumente ohne sachliche oder personenbezogene Zusammenhänge.

Ist ein Werk parallel als Printausgabe und Netzpublikation erschienen, so sind sowohl das gedruckte Werk als auch die Netzpublikation sammelpflichtig und müssen an die Deutsche Nationalbibliothek abgeliefert bzw. von ihr eingesammelt werden. Bei unterschiedlichen technologischen, sonst aber inhaltsgleichen Ausführungen von Netzpublikationen genügt die Ablieferung bzw. Sammlung einer Version.

Näher ausgeführt werden die Auswahlkriterien in den Sammelrichtlinien. Die Sammelrichtlinien haben Handreichungscharakter für die Bibliothekare und enthalten klare Anweisungen, wie beispielsweise: „Zu sammeln sind Netzpublikationen mit Themen- oder Personenbezug, wie z.B. Netzpublikationen von und über Persönlichkeiten des öffentlichen Lebens; dazu gehören insbesondere Politiker, Schauspieler, Musiker, Schriftsteller, Maler, Wissenschaftler, Publizisten, Journalisten usw.“ Die Sammelrichtlinien waren zur Drucklegung dieses Werkes noch nicht veröffentlicht, können nach ihrer Veröffentlichung aber auf der Website der Deutschen Nationalbibliothek eingesehen werden.[4]

Selektion und Praxis der Ablieferung

Neben den Sammelrichtlinien spielen die Auswirkungen auf die etablierten Geschäftsgänge eine große Rolle, denn mit der Verbreitung des digitalen Publizierens ist auch ein Wandel der bisherigen Vertriebs- und Verarbeitungswege

3 Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek vom 17. Oktober 2008, veröffentlicht im Bundesgesetzblatt Jahrgang 2008 Teil I Nr. 47, ausgegeben zu Bonn am 22. Oktober 2008, verfügbar unter http://www.bgblportal.de/BGBL/bgbl1f/bgbl108s2013.pdf

4 Die Sammelrichtlinien werden unter http://www.d-nb.de/netzpub/index.htm veröffentlicht werden.

verbunden. Für die Deutsche Nationalbibliothek verändert sich damit die Selektion der Ablieferer von Netzpublikationen. In der Printwelt sind die Ablieferer in der Regel Verlage, wirtschaftliche und wissenschaftliche Institutionen und Organisationen sowie ein kleiner Kreis von Privatpersonen; die Vertriebsstrukturen sind seit Jahrzehnten unverändert. Die Verlage sind bekannt und ein großer Teil der traditionellen Publikationen wird über das VLB (Verzeichnis Lieferbarer Bücher)[5] gemeldet. Teilweise sind die Ablieferer von Netzpublikationen identisch mit den bereits bekannten Ablieferern von Printpublikationen. Das Internet erweitert jedoch den Kreis der zur Ablieferung verpflichteten Produzenten um ein vielfaches und führt in gewisser Hinsicht zu einer Anonymisierung. De facto kann jeder zum Autor und damit zum Produzent von Netzpublikationen werden.

Wie spricht die Deutsche Nationalbibliothek diesen neuen Typ von Produzenten an? Eine Möglichkeit ist die Anmeldung als Ablieferer von Netzpublikationen über ein Webformular, das auf der Website der Deutschen Nationalbibliothek bereitgestellt wird. Die Anmeldung ist offen für jedermann. Nach der Übermittlung der Adressdaten überprüfen Bibliotheksmitarbeiter die Angaben und schalten die Produzenten für die Ablieferung frei. Im Rahmen von Veröffentlichungen zum Thema Netzpublikationen, wie beispielsweise zur Ankündigung der PflAV oder in Workshops, die zum Thema organisiert werden, wird dieses Verfahren erläutert.

Dynamische Entwicklung von Ausgabeformaten

Auch was die Ausgabenformate oder –formen betrifft, ist der traditionelle Publikationsmarkt im Umbruch. Große Verlage wie z.B. Springer arbeiten seit Jahren an der Optimierung ihrer Netzpublikationen und der entsprechenden Anpassung ihrer Geschäftsgänge. Wurde hier bis vor kurzem noch die Printpublikation zuerst auf dem Markt angeboten und erst danach die Netzpublikation über die Verlagsplattform, dann ist dies inzwischen umgekehrt der Fall.

Die dynamischen Entwicklungen des Internets und der damit verbundenen Technologien stellen große Herausforderungen für die Selektion, Sammlung und Langzeitarchivierung von Netzpublikationen dar, weil sich alle Planungen auf ein bewegtes Ziel richten. Galt beispielsweise lange Zeit das eBook[6] als die klassische Form der Netzpublikation, so werden die Endgeräte vielfältiger und

5 Informationen zum Verzeichnis Lieferbarer Bücher: http://www.vlb.de

6 Als eBook werden einerseits Netzpublikationen bezeichnet, die ein spezielles Lesegerät benötigen, häufig aber auch nur PDF-Dateien, die als Onlineversion die Printpublikationen abbilden.

mobiler und damit wächst die Suche nach Formaten, die multifunktional einsetzbar sind, wie beispielsweise XML oder auch das eigens für diesen Zweck entwickelte epub-Format.[7] Das hat für die Verlage zur Folge, dass auch die gewohnten Vertriebswege erweitert werden müssen, da sich die Zielgruppen verändern und damit die unterschiedlichsten Ansprüche und Erwartungen haben. Zunehmend wird nach Lösungen gesucht und in Form eigener Portalentwicklungen mit individueller Hard- und Software gefunden. Die Deutsche Nationalbibliothek steht vor der Herausforderung, Ablieferungsverfahren zu entwickeln, die mit diesen unterschiedlichen Portalsystemen harmonieren.

Auffällig ist, dass sich auf anderer Ebene ein Trend zur Standardisierung abzeichnet und zwar im Bereich der Metadaten. So wird beispielsweise im Verlagswesen seit 2000 der Metadatenstandard ONIX[8] entwickelt, um über dieses Datenformate verschiedene Geschäftsgänge zu bedienen: Informationen auf der Website, Daten für die Meldung an das VLB, ggf. auch Daten für die Presse bzw. die Verkaufskataloge.

Geschäftsgänge für die Sammlung von Netzpublikationen

Betrachtet man die Erfahrungen bei der Sammlung von Netzpublikationen auf freiwilliger Basis und die Entwicklung in diesem Bereich über die vergangenen sieben Jahre, dann wird deutlich, dass Geschäftsgänge, die einmal entwickelt wurden, um Netzpublikationen einzusammeln, steten Veränderungen unterliegen. Berücksichtigt man dann noch die unterschiedlichen Mengen, die produziert werden, dann zeigt sich auch hier, dass unterschiedliche Ablieferungsverfahren und damit Geschäftsgänge für die Verarbeitung erforderlich sind.

Aus den Erfahrungen mit den unterschiedlichen Dateiformaten und den verschiedenen Ablieferungsverfahren wurden deshalb neue Anforderungen abgeleitet und spezielle Geschäftsgänge entwickelt. Im Vordergrund stand ein pragmatischer Ansatz mit Konzentration auf das einzeln zu adressierende Objekt, d.h. die Netzpublikation mit Entsprechung in der Printwelt, die sog. druckbildähnliche Netzpublikation, die in der Regel im PDF-Format erscheint. Die nach wie vor verbreitete Trennung in Monografien und Zeitschriften ermöglicht die Orientierung an der Printwelt. Ein einheitliches Dateiformat erleichtert zudem die Ablieferung und ermöglicht automatisierte Prüfroutinen. Zentrales Ziel war in erster Linie die Automatisierung der verschiedenen Geschäftsgänge auf Bibliotheksseite: den automatisierten Import von Netzpublikationen in ein Archivsystem, die automatisierte Vergabe einer URN als Persistent Identifier

7 Informationen zum epub-Format: http://www.idpf.org

8 Informationen zu ONIX als Metadatenstandard: http://www.editeur.org/onix.html

(wenn die Netzpublikation keinen besitzt) und den Import von Metadaten in das eigene Katalogsystem, um hier einen Datensatz zu erstellen. Für Zeitschriften bedeutet dies die Ablieferung auf Heft- oder Artikelebene. Im Formular erfolgt die Verknüpfung über den in einer Auswahlliste angezeigten Zeitschriftentitel. Damit können Zeitschriften auf Heft- oder Artikelebene recherchiert und anzeigt werden.

Der automatisierte Import der Netzpublikation erfolgt über eine sog. Transfer-URL. Hier kann direkt auf das PDF zugegriffen und die Netzpublikation „abgeholt" werden. Wurde im Formular kein eindeutiger Identifier angegeben, dann wird an dieser Stelle auch eine URN der Deutschen Nationalbibliothek automatisch vergeben. Der Identifier ist das Bindeglied zwischen dem Katalogdatensatz und der Netzpublikation auf dem Archivsystem.

In einem Metadaten-Kernset[9] ist festgelegt, welche Metadaten erforderlich sind, um einen Kerndatensatz im Katalogsystem zu erstellen. Darüber hinaus sind weitere Metadaten festgelegt, deren Lieferung wünschenswert ist. Im Webformular sind die Kerndaten als Pflichtfelder festgelegt. Die angegebenen Metadaten können unmittelbar geprüft und ggf. korrigiert werden. Die über das Formular erfassten Metadaten werden dann in das Katalogsystem importiert und die Anzeige im Katalog erfolgt umgehend.

Eine Anzeige in der Deutschen Nationalbibliografie[10] erfolgt allerdings erst nach der Formal- und Sacherschließung; d.h. nach einer intellektuellen Erschließung anhand der geltenden Regelwerke und der Verknüpfung mit den Normdateien PND (Personennamendatei)[11] und GKD (Gemeinsame Körperschaftsdatei)[12] sowie der inhaltlichen Erschließung nach RSWK (Regeln für den Schlagwortkatalog) und/oder DDC. Aufgrund der Masse der Netzpublikationen ist dies aber auf Dauer nicht mehr zu leisten. Die Deutsche Nationalbibliothek entwickelt derzeit aber ein neues, stärker automatisiertes Erschließungskonzept.

---

9 Informationen zum Metadaten-Kernset: http://www.d-nb.de/netzpub/info/pdf/metadaten_kernset_extern.pdf

10 Informationen zur Deutschen Nationalbibliografie: http://www.d-nb.de/service/zd/dnb.htm

11 Informationen zur Personennamendatei: http://www.d-nb.de/standardisierung/normdateien/pnd.htm

12 Informationen zur Gemeinsamen Körperschaftsdatei: http://www.d-nb.de/standardisierung/normdateien/gkd.htm

Weitere Automatisierung der Geschäftsgänge

In einem nächsten Schritt ist die Automatisierung der Geschäftsgänge auf Seiten der Ablieferer geplant. Für die Ablieferung kleiner Mengen von Netzpublikationen ist das Webformular eine komfortable Lösung. Nach dem Einloggen in das Portal kann die Ablieferung erfolgen und sie dauert in der Regel auch nicht lange. Für die Massenablieferung wird derzeit gerade ein automatisiertes Harvestingverfahren[13] mit einem Verlag getestet. Auch hier ist die Voraussetzung, dass die Erstellung von Datensätzen automatisiert erfolgt ebenso wie die Verknüpfung mit der Netzpublikation. Es hat sich bereits im Umgang mit Online-Dissertationen gezeigt, welchen Vorteil einheitliche Metadatenstandards und Dateiformate bieten, insbesondere dann, wenn die Nachbearbeitung auf der Basis intellektueller Erschließungsinstrumente erfolgt. Deshalb sind standartisierte Metadaten erforderlich. Das Metadaten-Kernset bietet eine Konkordanz für den Import im ONIX-Format an. Weitere Einlieferformatstandards werden folgen, beispielsweise MARC21[14] und XMetaDiss(Plus).[15] Daneben sind individuelle Absprachen mit den Ablieferern erforderlich ebenso wie Testphasen, auch diese mit dem Ziel, die Automatisierung zu verbessern, mögliche Fehlerquellen bereits im Vorfeld auszuschalten und das Ausmaß notwendiger Datenprüfungen möglichst gering zu halten. Das Metadaten-Kernset für die Ablieferung von Zeitschriftenlieferungen ist in Vorbereitung. Hier wird beispielsweise in den Metadaten die Festlegung auf einen Identifier verlangt, über den die Heft- oder Artikellieferung mit dem Zeitschriftentitel verknüpft werden können.

Netzpublikationen stellen auch ein Mengenproblem dar. Die beschriebenen Verfahren sind ein erster Schritt hin zur notwendigen Automatisierung, zusätzliche Erweiterungen in diesem Bereich sind erforderlich, insbesondere in Bezug auf andere Dateiformate, aber auch auf weitere Ablieferungsverfahren wie z.B. das Webharvesting.

---

13 Informationen zum Verfahren der automatisierten Ablieferung über Harvestingverfahren finden sich auf der Website der Deutschen Nationalbibliothek unter http://www.d-nb.de/netzpub/ablief/pdf/automatisierte_ablieferung.pdf

14 Informationen zum Umstieg auf MARC21: http://www.d-nb.de/standardisierung/formate/marc21.htm

15 Informationen zum Metadatenstandard XMetaDiss: http://www.d-nb.de/standards/xmetadiss/xmetadiss.htm (09.02.2009). Allerdings ist hier eine Erweiterung auf weitere Hochschulschriften geplant. Bislang wurden nur die Online-Dissertationen gesammelt.

Prinzipien der Langzeitarchivierung an der Deutschen Nationalbibliothek

Für die schnell wachsende Speichermenge muss nicht nur eine geeignete Datenverarbeitungs-Infrastruktur bereitstehen und gepflegt, gewartet und weiterentwickelt werden. Um die Inhalte der gesammelten Netzpublikationen über wechselnde Hard- und Softwaregenerationen zu bewahren, müssen die Daten mit geeigneten Metadaten in einem Archivsystem verwaltet werden, das die gängigen Langzeitarchivierungsstrategien unterstützt: Migration, die Konvertierung in aktuell nutzbare Dateiformate, und Emulation, die Herstellung von früheren Systemumgebungen auf aktuellen Systemen mit Hilfe spezifischer Software.

Die Langzeitarchivierung von Netzpublikationen an der Deutschen Nationalbibliothek basiert auf folgenden Prinzipien:

Grundsätzlich nimmt die Deutsche Nationalbibliothek Netzpublikationen in jedem Format an, die Ablieferer werden aber auf die Präferenzregelung hingewiesen (derzeit: 1. PDF/A, 2. Andere PDF-Versionen, 3. HTML, 4. PS, 5. Weitere XML-basierte Formate, TXT, 6. Sonstige (DVI, RTF, etc.).[16] Die Deutsche Nationalbibliothek archiviert nur eine von ggf. mehreren vorliegenden inhaltsgleichen elektronischen Dokumentversionen, wobei die Auswahl der Präferenzregelung folgt.

Jedes Objekt wird vor der Langzeitarchivierung automatisch mit technischen Metadaten angereichert, die den gezielten Zugriff auf Archivobjekte und die Anwendung von Langzeitarchivierungsstrategien unterstützen. Zur Analyse von Dateiformaten und zur automatischen Generierung von technischen Metadaten setzt die Deutsche Nationalbibliothek das Open Source Tool Jhove ein. Jhove (JSTOR/Harvard Object Validation Environment) ist ein Gemeinschaftsprodukt von JSTOR und der Harvard University Library (HUL). Jhove wird von einer großen internationalen Gemeinschaft benutzt, gepflegt und weiterentwickelt. Die Deutsche Nationalbibliothek bringt sich hier aktiv ein und arbeitet mit internationalen Partnern wie Harvard und der Niederländischen Nationalbibliothek (Koninklijke Bibliotheek) zusammen, z.B. an der Entwicklung fehlender Module für zusätzliche Formate.

Aus den abgelieferten und mit Metadaten versehenen Objekten werden unter Nachnutzung vorhandener Standards (z.B. METS) Archivobjekte im offen definierten Paketformat Universellen Objektformat (UOF)[17] generiert. Dabei

16 Informationen zur Präferenzregelung: http://www.d-nb.de/netzpub/ablief/np_dateiformate.htm

17 Informationen zum Universellen Objektformat:

kann ein Archivobjekt mehrere Dateien umfassen, die gemeinsam ein logisches Objekt, d.h. eine Netzpublikation ausmachen.

Die so entstandenen Archivobjekte werden in eine sichere Umgebung, das Archivsystem, eingespielt, in dem das gespeicherte Material ständig routinemäßig überprüft wird. Jedes Objekt wird durch Bitstream Preservation mit regelmäßigen Maßnahmen wie Backups und Umkopieren zur Sicherstellung der Datenintegrität unverändert im Originalformat erhalten. Die wichtigste Langzeitarchivierungsstrategie der Deutschen Nationalbibliothek ist zurzeit die Migration, denn große Mengen der Objekte, die unter den Sammelauftrag fallen, können damit adressiert werden. Wenn sich abzeichnet, dass die Originalformate zu veralten drohen, werden die archivierten Objekte in aktuelle, zukunftsfähige Formate migriert. Die Zielformate werden auf Grundlage kontinuierlicher Marktbeobachtung (Technology Watch) bestimmt. Bei Migrationen wird das Ausgangsobjekt immer erhalten und zusammen mit dem migrierten Objekt weiter aufbewahrt. Alle Migrationsschritte werden dokumentiert und in den Metadaten des Objekts verzeichnet.

kopal-Archivsystem

Die technische Basis der Langzeitarchivierung an der Deutschen Nationalbibliothek bildet das Archivsystem, das im kopal-Projekt entwickelt wurde. Gefördert vom Bundesministerium für Bildung und Forschung (BMBF) hat die Deutsche Nationalbibliothek zwischen 2004 und 2007 in Partnerschaft mit der Niedersächsischen Staats- und Universitätsbibliothek Göttingen, der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) und der IBM Deutschland GmbH ein kooperativ nutzbares Langzeitarchiv aufgebaut.[18]

Das kopal-Archivsystems orientiert sich am OAIS-Referenzmodell (ISO 17421 „Open Archive Information System“) und setzt auf Standardsoftware auf. Für die Benutzung des kopal-Archivsystems entwickelten die Deutsche Nationalbibliothek und die SUB Göttingen die “kopal Library for Retrieval and Ingest” (koLibRI), die das Einspielen von Objekten in den Archivspeicher sowie den Zugriff auf die archivierten Objekte unterstützt.

Nach dem Abschluss der kopal-Projektphase 2007 erfolgt die Einbettung des Archivsystems in den Produktivbetrieb der Deutschen Nationalbibliothek. Die Einbettung erfordert einige konzeptionelle Anstrengung und Anpassungen in den vorhandenen technischen Systemen. Das Archivsystem wird so in die Geschäftsgänge eingebaut, dass die Netzpublikationen nahtlos von dem Zwi-

http://kopal.langzeitarchivierung.de/index_objektspezifikation.php.de

18 Informationen zu kopal: http://www.kopal.langzeitarchivierung.de/

schenspeicher, auf dem sie während des Erschließungsprozesses abgelegt sind, an das Archivsystem weitergegeben werden, wo sie langfristig und sicher aufbewahrt werden können. Um den Zugriff auf die Archivobjekte über Benutzerschnittstellen zu realisieren, müssen Schnittstellen angepasst werden. Weitere Schnittstellen müssen implementiert und Geschäftsgänge so umgestaltet werden, dass sie den Anforderungen der Archivobjekte gerecht werden. Zum Beispiel muss das Bereitstellungssystem darauf ausgerichtet werden, die archivierten Objekte im jeweils aktuellen Format (oder, alternativ, in dem vom Nutzer gewünschten Format) anzuzeigen. Auch die Anwendung von Langzeitarchivierungsstrategien im Praxisbetrieb, für die das Archivsystem ausgelegt ist, muss vorbereitet werden.

Weitere Herausforderungen

Doch selbst das Zusammenspiel von bewährten Tools und sicheren Archivsystemen kann nicht alle Herausforderungen der Langzeitarchivierung lösen. Neben technischen müssen vor allen Dingen organisatorische Vorkehrungen getroffen werden, hier illustriert am Beispiel von Konvertierungseinstellungen von Dateiformaten. Im Prinzip kann das alle möglichen Formate betreffen, hier wird dies aber am Beispiel PDF erläutert, weil große Mengen der Archivbestände der Deutschen Nationalbibliothek in PDF vorliegen. Das Format ist bei Verlagen und anderen Ablieferern akzeptiert und weit verbreitet. Doch viele Verlage und Ablieferer liefern passwortgeschützte oder verschlüsselte PDFs ab oder deaktivieren bestimmte Funktionen wie zum Beispiel Druck- und Kopiermöglichkeiten. Das bereitet einerseits in der Benutzung der Dateien Probleme, wirft aber auch essentielle Probleme für die Langzeitarchivierung auf: An solchen Dateien können nicht alle Langzeitarchivierungsmaßnahmen durchgeführt werden und es können Datenverluste entstehen. Die Deutsche Nationalbibliothek ist daher im Gespräch mit Verlegern, um auf diese Problematik aufmerksam zu machen und für einheitliche, offene Speichereinstellungen zu werben. Gleichzeitig gilt es aber auch, die internen technischen Prozesse auf dieses Problem hin anzupassen: Entsprechende Dateien müssen zunächst automatisch erkannt und – unter Beachtung urheberrechtlicher Rahmenbedingungen – in eine für die Langzeitarchivierung geeignete Struktur überführt werden.

Um weitere Entwicklungen auf dem Gebiet der Langzeitarchivierung voranzutreiben, arbeitet die Deutsche Nationalbibliothek intensiv mit zahlreichen nationalen und internationalen Partnern zusammen. Dabei geht es sowohl um die zukünftige Anwendung von nötigen Langzeitverfügbarkeitsstrategien wie Emulation und die kooperative Nutzung verschiedener Systeme (zum Beispiel

in den EU-Projekten SHAMAN[19] und KEEP[20]), als auch um die Weiterentwicklung von Formatregistries wie GDFR und Pronom oder die gezielte Unterstützung der Entwicklung und breiten Anwendung von archivierungsfreundlichen Standards wie PDF/A.

Ausblick

Die bestehenden Herausforderungen können insbesondere in Bezug auf die weiteren technischen Entwicklungen im Bereich der Netzpublikationen nur bewältigt werden, wenn die Verfahren verstärkt automatisiert werden. Dafür wurden die Grundlagen in der Systemarchitektur der Deutschen Nationalbibliothek gelegt. So können zumindest in Teilen die bereits entwickelten Verfahren für weitere Objekttypen nachgenutzt werden, aber es wird auch notwendig sein, neue Verfahren für Multimediaobjekte oder ablieferpflichtige Applikationen zu entwickeln.

Zum aktuellen Stand: Die Formulare zur Ablieferung von Zeitschriftenlieferungen (Hefte/Artikel) sind in der Testphase. Für die automatisierte Ablieferung über Harvestingverfahren sind Erweiterungen des Metadaten-Kernsets erforderlich, die u.a. auch für andere Objekttypen notwendig sein werden, wie beispielsweise für Audioobjekte. Weitere Entwicklungen betreffen die Metadatenformate, die in automatisierten Verfahren zur Anwendung kommen können, zum Beispiel auch die Anbindung weiterer Datenformate.

Eine Herausforderung stellt auch die Bereitstellung/Präsentation der Objekte dar. Erschwerend wirkt hier die rasche technologische Weiterentwicklung von Formaten und Abspielumgebungen. Übergeordnetes Ziel ist aber, die Bereitstellung für alle archivierten Objekte zu gewährleisten – auch für die Objekte, die auf einem Datenträger vorliegen. Angesichts der rund 700.000 Einheiten in den Sammlungen der Deutschen Nationalbibliothek ab ca. 1980 wird auch die historische Dimension dieses Problems offensichtlich.

19 Informationen zu SHAMAN: http://www.d-nb.de/wir/projekte/shaman.htm

20 Informationen zu KEEP: http://cordis.europa.eu/fetch?CALLER=PROJ_ICT&ACTION=D&DOC=13&CAT=PROJ&QUERY=011f22fab3a8:cf13:10a93ccb&RCN=89496

18.3 Langzeitarchivierung eines digitalen Bildarchivs – Projekt zum Aufbau eines Langzeitarchivs für hochaufgelöste digitale Bilddateien der Staatsgalerie Stuttgart am BSZ

Werner Schweibenz und Stefan Wolf [21]

Der Beitrag beschreibt das Projekt der Staatsgalerie Stuttgart mit dem Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) zum Aufbau eines Langzeitarchivs für hochaufgelöste digitale Bilddateien. Für die Archivierung wird das Langzeitarchiv SWBdepot des BSZ verwendet, die Metadaten für die Langzeitarchivierung werden mit dem Objektdokumentationssystem IMDAS-Pro erzeugt.

Die Situation in der Staatsgalerie Stuttgart

Die international bedeutsamen Sammlungen der Staatsgalerie Stuttgart (SGS)[22] reichen vom Tafelaltar bis zur modernen Medienkunst mit praktisch allen Formen z.B. der Malerei, Plastik, Graphik bis hin zu raumgreifenden Installationen. Sie werden fortlaufend erforscht, erschlossen und dokumentiert, gleichzeitig aber auch in verschiedenen Zusammenhängen eingebunden: beispielsweise in den eigenen Ausstellungen, im Leihverkehr zwischen Museen, in der Museumspädagogik, in Publikationen oder auch in Internetauftritten. Das Fotoatelier der SGS fertigt für diese Zwecke laufend eine große Zahl qualitativ hochwertiger Fotos an, dem Fortschritt der Technik folgend heute mit einer hochauflösenden Digitalkamera. Die anfallende Datenmenge wächst kontinuierlich. Für die Zukunft sucht die SGS nach einer kostengünstigen Lösung für Datenspeicherung und –sicherung bei gleichzeitiger Nutzung im Verbund. Dies gilt auch für die aus Sicherheitsgründen notwendige redundante Speicherung an einem zweiten Ort. Als Mitglied im MusIS-Verbund, dem landeseinheitlichen Verfahren für Museumsdokumentation der Staatlichen Museen in Baden-Württemberg wandte sich die SGS an das Bibliotheksservice-Zentrum Baden-Württemberg

21 Die Autoren danken herzlich den Kolleginnen und Kollegen der Staatsgalerie Stuttgart – allen voran Frau Dr. Elke Allgaier – für die kollegiale Unterstützung bei der Erstellung dieses Kapitels zum nestor Handbuch.

22 http://www.staatsgalerie.de/

BSZ),[23] das bereits für andere Institutionen ähnliche Dienstleistungen zu Dokumentmanagement und Langzeitarchivierung (LZA) anbietet.

Das Angebot des BSZ im Bereich Langzeitarchivierung

Das BSZ bietet auf dem Gebiet der Langzeitarchivierung eine umfangreiche Dienstleistungspalette an.[24] Der Fokus des BSZ richtet sich auf die Verbindung der Objekte mit einer qualitativ hochwertigen Dokumentation und der Schaffung eines Mehrwerts für seine Kunden. Der Gewinn liegt z.B. in der technischen Realisierung spezifischer Geschäftsgänge für verschiedene Häuser auf gemeinsamer Basis, der gemeinsamen Nutzung der gleichen Software oder in der automatisierten Erzeugung und Bereitstellung von Gebrauchsderivaten aus den archivierten Objekten – also auf klassischen Synergieeffekten. Mit diesem Ansatz vereint das BSZ scheinbar Gegensätzliches, indem Produktion, Dokumentation und Nutzung mit der Archivierung wertvoller Daten in einen durchgängigen Arbeits- und Archivierungsprozess gebündelt werden. Für die Kunden wird ein sicheres und kostensparendes Outsourcing großer Datenmengen in Kombination mit einer sicheren Datenhaltung der Originaldateien und einem flexiblen Zugriff auf Derivate geboten, also Mehrwerte, die eine Institution braucht, um sicher, rationell und ökonomisch arbeiten zu können. Für die Bildproduktion der Museen erschien die Durchführung eines Pilotprojekts als notwendig, das vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg gefördert wird.

Die Leitidee

Die Grundidee des Projektes zum „Aufbau eines Langzeitarchivs im BSZ für hochaufgelöste digitale Bilddateien der Staatsgalerie Stuttgart sowie die Entwicklung eines sicheren und auf Kontinuität basierenden Online-Daten-Transfers der Digitalisate“ (so der Name im Projektantrag) ist, unter strikter Beachtung des Primats einer sicheren und nachhaltigen Archivierung ein praktisches Verfahren zu entwickeln, bei dem möglichst nur ein Computersystem für die Dokumentation der zu archivierenden Bilddateien benötigt wird und bei dem

23 http://www.bsz-bw.de/

24 Wolf, Stefan ; Mainberger, Christof ; Schweibenz, Werner: Langzeitarchivierung am Bibliotheksservice-Zentrum Baden-Württemberg : Konzept, Aktivitäten und Perspektiven. – Preprint – Konstanz, BSZ, 2009.
URL: http://opus.bsz-bw.de/swop/volltexte/2009/465/
Erscheint voraussichtlich in: Bibliotheksdienst, Heft 43(2009), Heft 3.

die Mitarbeiter in allen beteiligten Arbeitsbereichen im Museum möglichst nur mit den ihnen bereits vertrauten Programmen und Werkzeugen arbeiten. Dieser Weg erhöht die Akzeptanz im Museum und hält den Einarbeitungs- und Schulungsaufwand gering. Gleichzeitig ist dies die Voraussetzung des zweiten Projektziels, das Verfahren nach Abschluss des Pilotprojekts den anderen Museen im MusIS-Verbund zur Verfügung zu stellen. Die Abläufe und Daten integrieren sich nahtlos in das Langzeitarchiv SWBdepot des BSZ. SWBdepot bezeichnet die am BSZ in Betrieb befindliche Speicherinfrastruktur, die nach Bedarf ausgebaut wird und auf der die üblichen Prozesse der Datensicherung wie z.B. Bandsicherung, redundante Speicherung, Konsistenzprüfung abgewickelt werden.

Der Produktionsablauf

Den Auftakt in der Produktion gibt die Bildbestellung, die in der Dokumentation festgehalten wird: nach ihrer Anweisung erstellt das Fotoatelier der SGS die Bildaufnahmen. Seit Januar 2005 arbeitet das Fotoatelier der SGS eingebettet in eine hausintern festgelegte Digitalisierungsstrategie mit einer Digitalkamera: Arbeitsprozesse, Geschäftsgänge und Dokumentation werden durchgängig elektronisch unterstützt und ausgeführt. Die professionellen Qualitätsansprüche an die Bildproduktion löst die digitale Fotografie mittlerweile ein. Je Museumsobjekt entsteht mindestens eine hochauflösende, unkomprimierte Masteraufnahme und ein farbkorrigierter, verlustfreier Submaster gleicher Auflösung im Tagged Image File Format (TIFF)[25] von durchschnittlich 50 MB bei 8 Bit Tiefe pro Farbkanal.

Das TIFF-Format bietet die Möglichkeit, IPTC-Metadaten[26] im Bild zu erfassen und zu speichern. Dieser Quasistandard der Pressefotografie erlaubt Angaben z.B. zu Bildrechten, Fotograf, Titel, Auflösung und Pixelzahl zu machen. Produkte wie z.B. Adobe Photoshop bieten dafür Erfassungsmasken, die im Fotoatelier von den Fotografen ausgefüllt werden. Teilweise handelt es sich um Angaben, die standardmäßig für jede Aufnahme aus der Staatsgalerie Stuttgart gemacht werden und in den Masken schon vorbelegt sind, teilweise aber auch um individuelle Merkmale, die zur einzelnen Fotografie eingetragen werden.

Sowohl Master als auch Submaster erhalten eindeutige Dateinamen, welche Hinweise auf Bildherkunft, Künstlername, Inventarnummer, Aufnahmegegebenheiten und Dateiformat enthalten. Zur Dateinamensgestaltung existiert eine formale und semantische Absprache zwischen SGS und BSZ. Die Konvention

25 http://de.wikipedia.org/wiki/Tagged_Image_File_Format
26 Vgl. http://de.wikipedia.org/wiki/IPTC-NAA-Standard

wird strikt eingehalten, da der Dateiname die Ablage im Archiv steuert und mit den IPTC-Daten zur Dokumentation herangezogen wird. Ideal ist es deshalb, wenn vor dem Transfer der Bilddatei das Objekt selbst schon als IMDAS-Pro-Museumsobjekt im Dokumentationssystem der SGS erfasst ist.

Vor dem sFTP-Transfer[27] der Dateien an das BSZ wird zur Sicherheit der Integrität die MD5-Prüfsumme[28] berechnet. Die Tagesproduktion wird gebündelt und an das BSZ durch einen Cron-Job nachts transferiert. Alle folgenden Arbeitsprozesse werden zur Kontrolle der Datenintegrität mittels Vergleich bzw. Erhebung der MD5-Prüfsumme begleitet.

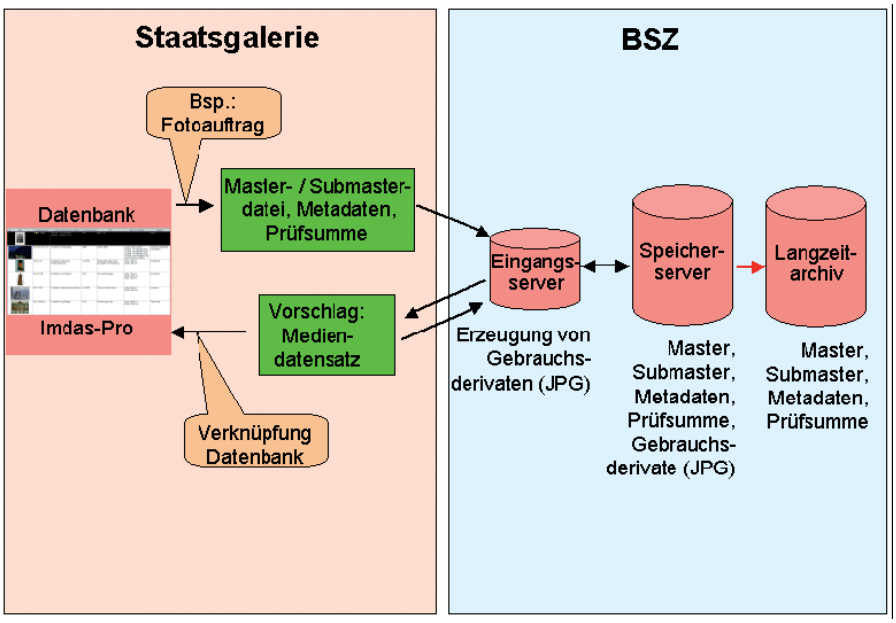


Abbildung 1[29]

27 http://de.wikipedia.org/wiki/SSH_File_Transfer_Protocol

28 http://de.wikipedia.org/wiki/Message-Digest_Algorithm_5

29 Abbildung 1: Allgaier, Elke (Staatsgalerie Stuttgart): Archivierung von digitalen Bilddaten. Beitrag des nestor-Seminars „Digitale Langzeitarchivierung in Museen und Archiven - Konzepte und Strategien“ Köln, 21.11.2008. In: URL: http://www.langzeitarchivierung.de/downloads/2008-11-21_allgaier.pdf

Aufgabe des BSZ ist die Übernahme, Archivierung und Bereitstellung der Bilddaten. Nach Prüfung der formalen Eingangsvoraussetzung (Einhaltung der Dateinamenskonvention, korrekte Benennung des Dateiformats, gleichzeitige Lieferung von Master, Submaster und Prüfsumme etc.) werden die Bilder vom offenen FTP-Bereich in einen sicheren Arbeitsbereich kopiert. Der Master befindet sich danach in einem geschützten Archivbereich. Er wird nicht weiter verwendet, sondern bildet die Grundlage, wenn später auf die Ursprungsdatei der Bildaufnahme zurückgegriffen werden muss. Der Submaster wird physisch dupliziert – eine Kopie wird mit dem Master im Langzeitarchiv abgelegt, die zweite Kopie bereitgestellt und zur Erzeugung der Gebrauchsderivate im JPG-Format herangezogen. Nötig sind bislang jeweils ein kleines Thumbnail für die Vorschau in IMDAS-Pro und für die Präsentation in BAM, dem gemeinsamen Portal zu Bibliotheken, Archiven und Museen[30], sowie eine größere Version zur differenzierten Betrachtung am Bildschirm oder für Restaurierung, Kunstvermittlung und entsprechende Zwecke.

Im gleichen Prozess werden Dateiname und IPTC-Daten gelesen und zusammen verarbeitet. Als Ergebnis entsteht eine Importdatei, die die nötigen Metadaten für eine Vorerfassung der Fotografie als Medienobjekt[31] in IMDAS-Pro bereitstellt. Sie enthält einen aus dem Dateinamen abgeleiteten Vorschlag zum Künstlernamen und zur Inventarnummer. Wegen der technischen Gegebenheiten in den Zeichensätzen kann die endgültige Ansetzung nicht transportiert werden. Die Importdatei enthält aber auch Angaben zu den technischen Daten der Fotografie und zu den Adressen bzw. Speicherorten der für die Nutzung bereitgestellten Derivate und Submaster, die aus den IPTC-Daten entnommen werden. Die Importdatei wird täglich in IMDAS-Pro eingelesen.

An dieser Stelle wird wieder die Staatsgalerie Stuttgart aktiv: die Dokumentation der Fotografie als IMDAS-Pro-Medienobjekt wird fertig gestellt und mit der Dokumentation des originalen Kunstwerks als IMDAS-Pro-Museumsobjekt verknüpft. Die Vorschläge zu Künstlername und Inventarnummer aus der importierten Vorerfassung werden in die endgültige, korrekte Ansetzungsform gebracht. Die Bilder werden geprüft und freigeben, so dass nun auch im Dokumentationswerkzeug eine Vorausschau auf das Bild verfügbar ist und weitere Derivate sowie der Submaster per Mausklick nach Berechtigung angefordert werden können.

---

30 http://www.bam-portal.de/

31 Das IMDAS-Pro-Medienobjekt ist eine programminterne Dokumentationsklasse, die der Aufnahme von Objekten dient, die mit der Dokumentationsklasse Museumsobjekt in IMDAS-Pro verknüpft werden kann.

Die Verarbeitungsprogramme des BSZ sind modular verkettet, konfigurierbar und parametrisierbar. Weitere Gebrauchsderivate können auf Anforderung erzeugt, weitere Inhalte aus dem IPTC-Header ausgelesen werden. Zur Bildverarbeitung im BSZ ist die OpenSource-Software ImageMagick[32] integriert. Gleichzeitig überwacht das Programm auch die Abläufe, so dass Verantwortlichen an BSZ und SGS nach Abschluss des täglichen Jobs die nötigen Rückmeldungen erhalten.

IMDAS-Pro erlaubt den Export von Daten im XML-Format. Sie beinhalten je nach Konfiguration sowohl Daten aus dem IMDAS-Pro-Medienobjekt als auch dem IMDAS-Pro-Museumsobjekt. Für die Langzeitarchivierung notwendig sind Metadaten in austauschfähigen, nicht proprietären Formaten – dafür bietet sich museumdat[33] an. Als Arbeitsergebnis der Fachgruppe Dokumentation des Deutschen Museumsbundes stellt es das Standardformat dar, mit dessen Hilfe Beschreibungsdaten aus Museumsbeständen ausgetauscht und gegenseitig nutzbar gemacht werden. Die aus IMDAS-Pro exportierten Daten werden in museumdat-Darstellung mit den Mastern und Submastern zu METS-Paketen verbunden und als Submission Information Packages an das Langzeitarchiv übertragen. In diese Pakete werden auch die notwendigen technischen Metadaten der Langzeitarchivierung eingebunden. Die automatische Erhebung dieser Daten geschieht mit der Open-Source-Software JHOVE, die am BSZ seit langem eingeführt ist und z.B. auch in kopal Anwendung findet. Im Archiv stehen diese Pakete den Prozessen der Langzeitarchivierung im engeren Sinn zur Verfügung.

Der Projektstand

Die erste Arbeitssitzung im Projekt lag im April 2008. Heute, im Februar 2009, sind große Teile des Projektes einsatzbereit: die Vorgaben für Dateinamen und IPTC-Daten sind definiert; auf ihrer Basis wurden die Arbeitsrichtlinien in Dokumentation und Fotoatelier der Staatsgalerie erarbeitet, die notwendigen Schulungen fanden statt. Die Verarbeitungsprogramme im BSZ sind fertig gestellt und getestet, die Import-Routinen für IMDAS-Pro stehen bereit. Die notwendigen Anpassungsarbeiten an IMDAS-Pro der Version 4.0 stehen vor dem Abschluss, so dass die Projektpartner damit rechnen, Mitte März die Produktion zu starten. Die Staatsgalerie Stuttgart hat einen großen Teil der seit 2005 hergestellten Fotografien mit mobiler Festplatte an das BSZ übertragen,

32 http://www.imagemagick.org/script/index.php

33 XML-Schema und Dokumentation finden sich unter der URL http://museum.zib.de/museumdat/

die dort gesichert werden. Mit ihnen wurde ein Massentest der Verarbeitungsprogramme durchgeführt, der zeigte, dass die Programme zuverlässig arbeiten. Gleichzeitig steht mit dem Festplattentransport ein zweiter Lieferweg neben dem sFTP-Transfer zur Verfügung. In Vorbereitung ist die Herstellung und Ablage der Submission Information Packages. Mit Produktionsbeginn wird ein internes Audit zur Einhaltung der BSI-Standards[34] und Schutzbedarfsfeststellung durchgeführt.

Die Perspektiven und Zusammenhänge

Nach gemeinsamer Einschätzung der Beteiligten hat das Projekt ein beträchtliches Potential im Hinblick auf Materialien, Anwendung und Ausbreitung. Das im Entstehen befindliche Verfahren entlastet die Staatsgalerie Stuttgart von der laufenden Server- und Softwarewartung für den digitalen Bildbestand und teilt die Verantwortlichkeit für die Langzeitarchivierung der Daten zwischen der SGS als Eigentümer der Daten und dem BSZ als Dienstleister. Neben positiven internen Organisationseffekten werden finanzielle Einsparungen erwartet. Das damit einhergehende fast vollständige Outsourcing des Datenbestandes ist für den Eigentümer der Originale und Originaldaten, die Staatsgalerie Stuttgart, nicht selbstverständlich, aber angesichts der Ergebnisse gewünscht, die Abläufe und Bedingungen zu erleichtern.

Gleichzeitig sind neben den hochauflösenden Fotografien eine Vielzahl weiterer, teilweise bereits historischer Bildbestände aus der Geschichte der Staatsgalerie, aus ihren Werkstätten, zu Veranstaltungen und Ausstellungen vorhanden, die für die Forschung immer relevanter werden. Auch wenn es sich vor allem um Aufnahmen im JPG-Format handelt, steht ihre langfristige Sicherung und Dokumentation an. Schon zur Vermeidung der Doppelerfassung bzw. paralleler Dokumentationswerkzeuge sind sie in die Langzeitarchivierung einzubeziehen.

Daneben entstehen auch Künstlervideos und weitere elektronische Inhalte aus der Museumsdokumentation, aus Ausstellungsunterlagen, Öffentlichkeitsarbeit in Form von Texten, Bildern und Tonaufnahmen, die besondere Anforderungen an die Speichertechnik stellen. Ihre Bearbeitung steht in künftigen Projekten an.

Die in Deutschland maßgeblichen Standards der Langzeitarchivierung werden konsequent beachtet, proprietäre Lösungen vermieden. Die Standardisierung im Museumsbereich ist längst nicht so weit fortgeschritten wie z.B. im

34 http://www.bsi.de/literat/bsi_standard/index.htm

Bibliotheksbereich. Mit museumdat steht jetzt ein Format zur Verfügung, das den Verbindlichkeitsanforderungen an inhaltlich beschreibende Metadaten genügt und sich hoffentlich durchsetzen wird. Seine Benutzung bei der Bildung der zur Langzeitarchivierung üblichen METS-Objekte ist die zukunftsweisende Lösung. Solche Objekte werden geeignet sein, z.B in eine kopal-Installation übertragen zu werden.

Speicherplatz wird am BSZ bedarfsgerecht bereit gestellt und kann laufend erweitert werden. Damit verbunden sind die üblichen Sicherungsverfahren in einem Rechenzentrum inklusive dislozierter, redundanter Speicherung an einem zweiten Aufbewahrungsort.

Das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg unterstützt das Projekt besonders auch aus einem weiteren Grund: nach Einführung des Verfahrens sollen die Projektergebnisse den anderen Museen des Landes Baden-Württemberg und des MusIS-Verbundes zur Verfügung stehen. Die Begleitung des Projekts durch das Badische Landesmuseum Karlsruhe (BLM) sichert genau diesen Sachverhalt: hausspezifische Lösungen werden vermieden, dafür werden Methoden und Wege gewählt, die sich mit geringem Aufwand verallgemeinern, übertragen und in weiteren Institutionen einführen lassen. Vereinbart ist schon heute, dass nach Inbetriebnahme der Produktion für die Staatsgalerie Stuttgart das BLM seinen Bestand an Fotografien im erarbeiteten Verfahren an das BSZ übertragen wird.

Es bestätigt sich, was in einem Gespräch zur Langzeitarchivierung in Baden-Württemberg festgehalten wurde, an dem Vertreter von Rechenzentren, Bibliotheken, Archiven und Museen beteiligt waren: der Bedarf an Lösungen für die Langzeitarchivierung in Museen besteht. Auch wenn die Digitalisierung der Bestände in den Museen später als z.B. in den Bibliotheken einsetzte und anderen Notwendigkeiten folgt, hat doch die Produktion von Daten, die verlässlich archiviert werden müssen, begonnen und wird einen beträchtlichen Aufschwung nehmen. Die Zusammenarbeit mit einem Dienstleister gewährleistet die gewünschte hohe Datensicherheit.

Literatur

Allgaier, Elke (Staatsgalerie Stuttgart): *Archivierung von digitalen Bilddaten*. Beitrag des nestor-Seminars „Digitale Langzeitarchivierung in Museen und Archiven - Konzepte und Strategien“ Köln, 21.11.2008. In: URL: http://www.langzeitarchivierung.de/downloads/2008-11-21_allgaier.pdf

Wolf, Stefan / Mainberger, Christof / Schweibenz, Werner: *Langzeitarchivierung am Bibliothekservice-Zentrum Baden-Württemberg : Konzept, Aktivitäten und Perspektiven*. – Preprint – Konstanz : BSZ, 2009. URL: http://opus.bsz-bw.de/swop/volltexte/2009/465/

18.4 ARNE – Archivierung von Netzressourcen des Deutschen Bundestages

Angela Ullmann

Der Deutsche Bundestag archiviert seit Januar 2005 seine Internetangebote und stellt die archivierten Snapshots wiederum über ein Webarchiv im Internet bereit. Ausgehend von archivischen Grundprinzipien und den speziellen Rahmenbedingungen beim Deutschen Bundestag wurden für die Archivierung sowohl ein Konzept als auch ein System entwickelt. Dabei standen insbesondere Fragen zur Wahrung von Authentizität und Kontext, der archivischen Bewertung von Netzressourcen, aber auch technische Aspekte wie Maßnahmen zur Langzeiterhaltung im Fokus. Seit einiger Zeit kommen neue Herausforderungen hinzu – so die Wahrung des Persönlichkeitsschutzes beim freien Zugang zu archivierten Netzressourcen über das Internet.

Fachliche Einordnung und organisatorische Rahmenbedingungen

Die Archivierung von Netzressourcen des Deutschen Bundestages stellt ein Anwendungsbeispiel für ein fokussiertes Web-Harvesting dar. Die Sicherungsaufgabe leitet sich von der archivischen Zuständigkeit des Parlamentsarchivs ab und umfasst somit ausschließlich Webangebote, die aus der Provenienz „Deutscher Bundestag" stammen. Konzeptionell einbezogen sind sowohl öffentlich zugängliche Angebote im Internet als auch nichtöffentliche wie das Intranet.[35] In den Wirkbetrieb überführt ist bislang nur die Archivierung einzelner Internetangebote, die zum Zeitpunkt ihrer Veröffentlichung bzw. des Downloads keinen Zugangsbeschränkungen unterlagen.

Die Archivierung wird in Kooperation zwischen zwei Organisationseinheiten der Bundestagsverwaltung realisiert: dem Referat „Parlamentsarchiv" und dem Referat „Online-Dienste / Parlamentsfernsehen". Die Arbeitsteilung beruht auf den verwaltungsmäßigen Zuständigkeiten und den daraus resultierenden Kompetenzen. Während das Parlamentsarchiv die archivische Bewertung und die anderen anfallenden (archiv)fachlichen Aufgaben wahrnimmt, wird die technische Abwicklung durch die Online-Dienste übernommen. Da

35 Das Intranetangebot des Deutschen Bundestages ist im Unterschied zum Internet nur für einen beschränkten Adressatenkreis zugänglich. Zu diesem gehören die Abgeordneten und deren Mitarbeiter sowie die Mitarbeiter der Fraktionen und der Bundestagsverwaltung.

auch die inhaltliche und technische Pflege der aktuellen Webangebote dort ressortiert, sind Informationsverluste oder Kommunikationslücken zu anstehenden Veränderungen der Webangebote weitgehend ausgeschlossen.

Entwicklung und Fortschreibung einer Archivierungslösung

Basierend auf Vorüberlegungen des Parlamentsarchivs zur Sicherung von Netzressourcen des Deutschen Bundestages aus den Jahren 2002 bis 2003 wurde im Jahre 2004 eine Übereinkunft mit den Online-Diensten zur Entwicklung einer Archivierungslösung für das Angebot „Bundestag im Internet“ (www.bundestag.de) getroffen. Der enge Fokus war die einzige Möglichkeit, sich dieser Aufgabe praxisnah zu nähern. Gleichzeitig liegt hier jedoch ein „Geburtsfehler“ des „Systems zur Archivierung von Netzressourcen des Deutschen Bundestages“ (ARNE), da es systemtechnisch eng an eine einzelne Ressource gebunden ist. Da diese Ressource jedoch das wichtigste Angebot des Bundestages im Netz ist, das unbedingt erhalten werden soll, war dieses Vorgehen gerechtfertigt.

Standards blieben in der Entwicklungsphase weitgehend unberücksichtigt. Zum einen wurde und wird der (Meta)Datenaustausch mit anderen Gedächtnisorganisationen nicht angestrebt, daher musste keine Evaluierung eventuell geeigneter Metadatenstandards erfolgen. Auch für die Zugrundelegung des OAIS-Modells gab es beim Start des Vorhabens kein Bedürfnis.

In Abweichung zur üblichen Verfahrensweise bei derartigen Projekten entstanden Konzept und System zeitgleich und in gegenseitiger Abhängigkeit. Die wenigen bislang existierenden Referenzprojekte waren für die beim Bundestag angestrebte Lösung nicht einschlägig – entweder hinsichtlich der archivfachlichen Prämissen oder der Einbettung in bestehende Systemlandschaften. Das weitaus bekannteste Referenzprojekt dürfte das Internet Archive (http://web.archive.org) sein. Eine Analyse der dort gespeicherten Snapshots (also Momentaufnahmen) von „Bundestag im Internet“ ergab, dass durch die Wayback-Machine Internetseiten verschiedener Zeitschnitte miteinander in einem Angebot verbunden werden, die vor der Archivierung nicht gleichzeitig online verfügbar waren. Der Benutzer erhält jedoch beim Laden der Seiten keine entsprechenden Hinweise. Dies verstößt eindeutig gegen den Grundsatz der Authentizität. Die Wahrung der Authentizität ist eine große Herausforderung, aber auch eine entscheidende Frage nicht nur bei der Webarchivierung.

Die während der Archivierungs- und Aufbereitungsvorgänge aufgetretenen Fehler und Fragen ermöglich(t)en das (Fort)Schreiben eines Konzeptes, das der Wirklichkeit entspricht. Diese induktive Methode empfiehlt sich zum jetzigen

Zeitpunkt insbesondere für kleinere Einrichtungen und einzellige Archive, deren Sicherungsauftrag nur wenige Webangebote umfasst.

Generell ist davon auszugehen, dass Systeme zur Webarchivierung einem unablässigen Wandel unterworfen sind, da sich auch die Technologien zur Erzeugung von Webangeboten rasant weiterentwickeln.

Grundsätze und Anforderungen

Die Bewahrung aller Webseiten in allen jemals veröffentlichten Versionen wurde bereits zu Beginn der internen Diskussion weder als realistisch noch als sinnvoll angesehen. Die Archivierung erfolgt daher auf zwei Wegen: einmal als Turnus- und darüber hinaus als Anlassarchivierung – auch bezeichnet als selective Harvesting und Eventharvesting[36], wobei diese Methoden nicht gleichberechtigt nebeneinander stehen, sondern das Eventharvesting eine Nebenform des zyklischen selektiven Harvesting darstellt.

Der Zyklus der Archivierung unterscheidet sich für jede Netzressource. Die dabei berücksichtigten Aspekte sind unter „Auswahlstrategie und Bewertung der Netzressourcen" ausgeführt.

Bewahrenswert sind aus archivischer Sicht nicht nur die Informationen, die auf einer Webseite publiziert sind, sondern auch das Aussehen, die Gestaltung, die Funktionalitäten, der Kontext und das Verhalten der Webseiten. Als Bezugspunkt dient dabei die Nutzersicht und nicht die Sicht des Systembetreuers, der auch auf Inhalte zugreifen kann, die aktuell nicht freigegeben sind. Somit finden bei der Archivierung nur die zu diesem Zeitpunkt veröffentlichten Seiten / Dateien Berücksichtigung. Diese Entscheidung gab den Ausschlag für die Wahl eines Crawlers und gegen FTP für den Downloadprozess.

Ein weiterer Grundsatz bestand im Bestreben, archivierte Netzressourcen möglichst zeitnah wiederum dem ursprünglichen Adressatenkreis auf einem gleichwertigen Zugangsweg wie vor der Übernahme in das Archiv bereitzustellen.

Damit unmittelbar verbunden war die Frage, ob alle im vorarchivischen Bereich angebotenen Funktionen (Druckversion erzeugen, mailto-Befehle etc.) im Webarchivsystem nachzubilden sind, ob dies vom Aufwand her vertretbar und auch sinnvoll ist. Aus unterschiedlichen Gründen wurde entschieden, diese Funktionen nicht anzubieten: Der mailto-Befehl bspw. soll nicht mehr ausführbar sein; das Erzeugen der Druckfunktion war technisch zu aufwändig. Letz-

36 Vgl. PoWR. *The Preservation of Web Resources Handbook. Digital preservation for the UK HE/FE web management community*. London 2008. URL http://www.jisc.ac.uk/media/documents/programmes/preservation/powrhandbookv1.pdf, S. 18 - 19

teres hat auch Kritik von Nutzern hervorgerufen, weil die Erwartungen an die Funktionalität archivierter Netzressourcen die gleichen sind wie die an Live-Angebote im Web.

Für die Erhaltung interaktiver Inhalte wie bspw. den virtuellen Adler auf „Bundestag im Internet" oder das Einreichen einer elektronischen Petition ist bislang noch keine Lösung entwickelt. Aus archivischer Sicht wäre eine Bewahrung dieser Angebote unstreitig wünschenswert, während andere interaktive Inhalte wie bspw. die Bestellung von Informationsmaterial zum Deutschen Bundestag als nicht archivwürdig gelten können. Bislang bleiben alle interaktiven Inhalte von der Archivierung ausgeschlossen.

Auswahlstrategie und Bewertung der Netzressourcen

Die Bewahrung von Webangeboten des Deutschen Bundestages erfolgt nach archivischen Prinzipien. Ein grundlegendes Prinzip ist die unter „Fachliche Einordnung und organisatorische Rahmenbedingungen" bereits erläuterte Provenienzbindung. Ein weiterer Grundsatz besteht in der Auswahl der zu archivierenden Unterlagen. Dieser findet auch auf Webangebote Anwendung. Die Webprojekte des Bundestages werden in ihrer Entstehung und Entwicklung beobachtet und bewertet. Die Bewertungsentscheidung ist zweistufig: zunächst wird entschieden, welche Netzressourcen grundsätzlich archivwürdig sind. Die Archivwürdigkeit wurde mit einer Ausnahme für alle aktuellen Webprojekte des Bundestages bejaht. Die Entscheidung, die Homepage „Das Parlament" (www.das.parlament.de) nicht zu archivieren, beruht auf dem fehlenden inhaltlichen Mehrwert, denn sie stellt eine nahezu identische Webaufbereitung der vom Deutschen Bundestag im Druck herausgegebenen Wochenzeitung „Das Parlament" dar. Die Darbietung im Netz ist lediglich ein alternativer Verbreitungsweg. Aus archivischer Sicht handelt es sich somit um eine Mehrfachüberlieferung.

Im positiven Falle ist darüber hinaus das Intervall einer Archivierung festzulegen. Diese archivische Bewertungsentscheidung orientiert sich an der Aussagekraft und dem Stellenwert der Netzressource, dem Aktualisierungsintervall des Live-Angebotes, den inhaltlichen Alleinstellungsmerkmalen des Webangebotes und dem Ziel der Archivierung. Das Webangebot „Mitmischen" (www.mitmischen.de) für Jugendliche behandelt allgemeine aktuelle politische Themen und soll Jugendliche an Politik heranführen. Der inhaltliche Bezug zum Deutschen Bundestag ist hier nicht unmittelbar gegeben. Bei diesem Angebot ist ein halbjährlicher Archivierungsturnus vorgesehen. Das Ziel der Archivierung besteht dabei nicht in der Nachvollziehbarkeit aller Informationen über politische Ereignisse, sondern einer auswahlweisen Veranschaulichung des Angebotes.

„Bundestag im Internet“ wird dagegen regelmäßig alle vier Wochen archiviert. Dieses Intervall kann in einer „heißen“ politischen Phase geändert werden. Die politischen Ereignisse im Jahre 2005 mit der Vertrauensfrage des Bundeskanzlers, der Verkürzung der Legislaturperiode und der vorzeitigen Neuwahl waren Anlass dafür, das Archivierungsintervall in dieser Zeit auf 14 Tage festzulegen.

Politische Ereignisse können ebenso wie technisch-inhaltliche Veränderungen (Redesign des Webangebotes o.ä.) eine zusätzliche Anlassarchivierung außerhalb des normalen Archivierungsturnus nach sich ziehen.

Die Grundsätze für die Auswahl und Bewertung sind durch die ständig aktualisierte Veröffentlichung des Konzeptes zur „Archivierung von Netzressourcen des Deutschen Bundestages“[37] für jedermann transparent.

Wahrung der Authentizität, Erschließung und Metadaten

Das oberste Gebot der Archivierung ist die Wahrung der Authentizität. Ein Dokument muss immer das sein, was es zu sein vorgibt. Die konsequente Anwendung dieses Prinzips gestattet durchaus Veränderungen an einer Netzressource im Zuge ihrer Archivierung – allerdings müssen alle Änderungen dokumentiert und jederzeit durch jedermann nachvollziehbar sein. Realisieren lässt sich dies vorrangig durch Metadaten.

Metadaten dienen nicht nur zur Dokumentation der technischen Ursprungsumgebung sowie der mit einer Archivierung verbundenen technischen Maßnahmen und deren Parameter. Sie geben auch inhaltliche Auskünfte und erschließen eine Netzressource – beispielsweise durch die Angabe der Provenienz, des Archivierungsdatums, des Archivierungsanlasses, der Domäne.

Für ARNE wurde eine Liste von Metadaten definiert, die sowohl den vorarchivischen Bereich als auch den Workflow der Archivierung beschreibt. Diese Liste wird im Rahmen der Fortentwicklung des Systems ebenfalls erweitert. Die Aufzählung aller Metadaten würde den Rahmen dieses Beitrages sprengen. Es soll daher an dieser Stelle abermals auf die Konzeption zur „Archivierung von Netzressourcen des Deutschen Bundestages“ verwiesen werden.

Weitere Gesichtspunkte für die Wahrung der Authentizität von Netzressourcen sind die Behandlung externer Links und die Art der Aufbereitung für eine Nutzung. Die Verlinkung von Webseiten gehört zur elementaren Charakteristik dieser Quellengattung. Links verbinden sowohl Inhalte innerhalb einer Netzressource, führen aber auch zu anderen Webangeboten entweder desselben Inhabers oder zu Angeboten Dritter. Bei der Archivierung von Netzressourcen des

37 zur URL siehe Ende des Beitrages

Deutschen Bundestages bleiben nur die Links innerhalb eines Webangebotes im archivierten Snapshot unmittelbar ausführbar. Absolute Links werden hierzu im Rahmen der Konvertierung in interne Links umgewandelt. Für alle anderen Links werden die Ziele und das Verhalten („öffne ein neues Fenster“ etc.) gesichert.

Auf die Sicherung der Authentizität im Rahmen der Aufbereitung für eine Nutzung wird unter „Bereitstellung und Nutzung“ eingegangen.

Workflow

Erst über die Jahre des Wirkbetriebs hinweg konnte der Workflow zu einer stimmigen Abfolge entwickelt werden. Aktuell besteht er aus den Arbeitsschritten:

- Archivische Bewertung aller Netzressourcen
- Registrierung technischer Metadaten zu potentiellen Dateiformaten, die in einer Netzressource des Bundestages enthalten sein können
- Technische Verankerung der Archivierungsoptionen
- Anlegen eines Snapshots in der Referenzdatenbank (damit automatisch verbunden Anlegen eines Verzeichnisses für die Ablage des Snapshots auf dem Webarchivserver)
- Start und Ablauf des Downloadvorganges
- Anlegen einer Kopie für den gesamten Snapshot
- Konvertierung (umfasst mehrere Arbeitsschritte wie Umwandlung der externen Links, Konvertierung der html-Seiten nach xhtml)
- Indexierung
- Qualitätssicherung durch die Prüfung von definierten Referenzseiten
- Freigabe für die Benutzung (damit automatisch verbunden Transfer auf den externen Webserver und Bereitstellung im Internet)
- Backup
- Weitere Erhaltungsmaßnahmen

In diesen Workflow eingebunden ist die Ermittlung der technischen Zusammensetzung einer Netzressource oder auch neu hinzugekommener Dateiformate.

Bereitstellung und Nutzung

Netzressourcen des Bundestages sollen so bald als möglich nach ihrer Archivierung im Internet verfügbar gemacht werden. Es existieren mehrere Zugangswege:

- Das Webarchiv ist eingebunden in „Bundestag im Internet".
- Von verschiedenen Seiten in „Bundestag im Internet" wird auf ältere Inhalte im Webarchiv direkt verlinkt.
- Das Webarchiv ist über Suchmaschinen zugänglich.
- Jeder Nutzer kann einen Link direkt auf eine archivierte Webseite erzeugen und diesen auf die übliche Art verwenden.
- Künftig soll es auch über die Archivdatenbank des Parlamentsarchivs erreichbar sein.

Archivierte Snapshots müssen jederzeit als Archivgut erkennbar sein. Der Nutzer muss sehen, dass er sich im Archiv befindet und um welchen Snapshot es sich handelt. Auch ein Wechsel in einen anderen Snapshot muss erkenn- und nachvollziehbar sein.

Alle archivierten Webseiten des Bundestages werden daher in einem roten Rahmen dargestellt. Die Kopf- und Fußzeile dieses Rahmens informieren über den Archivstatus und zeigen die wichtigsten Metadaten zur Identifizierung des Snapshots an.

Bei der Ausführung von Links, die aus dem Snapshot hinausführen („externe Links") unterscheidet das System verschiedene Arten und gibt in Abhängigkeit davon Hinweise für den Benutzer. Bei externen Links zu Datenbanken oder Angeboten Dritter erscheint ein Hinweis auf den „verlorenen" Kontext. Dem Benutzer wird erklärt, dass die referenzierte Datenbank bzw. das referenzierte Webangebot seit der Archivierung der genutzten Netzressource wahrscheinlich inhaltlich und gestalterisch verändert wurde. Damit kann nicht (mehr) von einem unmittelbaren Bezug der Netzressource auf die referenzierte Datenbank oder das referenzierte Webangebot ausgegangen werden („Verlorener Kontext"). Bei Links in andere Snapshots erhält der Benutzer eine Mitteilung darüber, dass er in einen anderen Snapshot wechselt.

Eine besondere Herausforderung stellen Inhalte dar, die nach ihrer Archivierung aufgrund datenschutzrechtlicher oder anderer Bestimmungen in der Benutzungsversion geändert werden müssen. In diesem Rahmen vorgenommene Änderungen sind zur Wahrung der Authentizität zu dokumentieren. Eine fachliche Lösung und deren technische Umsetzung hierfür werden im Rahmen von ARNE angestrebt und befinden sich derzeit in Entwicklung.

Maßnahmen zur langfristigen Erhaltung

Verschiedene Maßnahmen sollen zur langfristigen Erhaltung beitragen, auch wenn ein umfassendes Sicherungskonzept noch aussteht. Die technischen und

inhaltlichen Metadaten geben ein umfassendes Bild der Netzressource. Bestandteil dieser Metadaten sind auch Fehlerprotokolle und Logbücher der archivtechnischen Bearbeitung sowie Dateistatistiken.

Nach dem Download werden die Daten zunächst kopiert, um jederzeit auf die unbearbeitete Version des Snapshots zurückgreifen zu können. Konvertiert werden bislang lediglich html-Dateien nach xhtml.

Darüber hinaus existiert ein Datensicherungskonzept, in das interne und externe Datenträger an unterschiedlichen Standorten einbezogen sind.

Technische Eckpunkte

Beim System ARNE handelt es sich um eine Eigenentwicklung des Deutschen Bundestages. Es umfasst u.a. die Referenzdatenbank für die Metadaten und verbindet verschiedene am Markt erhältliche Tools bspw. für den Download (httrack) und die Indexierung.
Die Speicherung der Daten erfolgt auf einem gesonderten Webarchivserver, der sich bei den Online-Diensten befindet und dort auch technisch betreut wird.

Eine detaillierte technische Beschreibung ist dem Konzept zur Archivierung von Netzressourcen zu entnehmen.

Mehrwert der Webarchivierung für den Deutschen Bundestag

Was aber bedeutet die Archivierung von Netzressourcen für den Deutschen Bundestag? War es zunächst ausschließlich ein archivisches Anliegen, historisch wertvolle Quellen zu bewahren, so erbringt die Webarchivierung mittlerweile wichtige institutionelle Mehrwerte.

Bereits die Beschäftigung mit den Webangeboten im Rahmen der Vorbereitung der Webarchivierung brachte wichtige Erkenntnisse und Fragen auf die Tagesordnung. Das System zur Webarchivierung offenbarte technische Fehler im Live-Angebot, die so vorher nicht ersichtlich waren. Allein die aus nicht mehr zielführenden Links resultierenden „Fehlerseiten“ konnten in den letzten Jahren erheblich reduziert werden.

Das Webarchiv hat sich zu einem wichtigen institutionellen Gedächtnis entwickelt. So lässt sich nachweisen – wenn auch nicht lückenlos -, welche Informationen zu welcher Zeit online verfügbar waren. Es trägt damit nicht zuletzt zur Rechtssicherung bei.

Durch die regelmäßige Archivierung von Webseiten wird das vorarchivische Content-Management-System (CMS) entlastet. Ältere Inhalte, die nicht mehr fortgeschrieben werden, können im CMS gelöscht und aus dem aktuellen

Angebot ins Webarchiv verlinkt werden. Ältere Seiten müssen dann auch nicht mehr in Relaunches einbezogen werden.

Literatur und Quellenangaben

Ausführliche Darstellung in:

Ullmann, Angela / Rösler, Steven (2007): *Archivierung von Netzressourcen des Deutschen Bundestages*. Version 2.0. In: Online-Veröffentlichungen aus dem Parlamentsarchiv des Deutschen Bundestages. Dezember 2007. http://www.bundestag.de/wissen/archiv/oeffent/arch_netz_gross2.pdf

PoWR. The Preservation of Web Resources Handbook. Digital preservation for the UK HE/FE web management community. London 2008., S. 18 – 19. URL http://www.jisc.ac.uk/media/documents/programmes/preservation/powrhandbookv1.pdf

Webarchiv des Deutschen Bundestages im Internet:

http://webarchiv.bundestag.de

Rechercheanleitung für das Webarchiv:

http://webarchiv.bundestag.de/cgi/recherche_anleitung_webarchiv_bundestag.pdf

19 Qualifizierung im Themenbereich „Langzeitarchivierung digitaler Objekte“

Regine Scheffel, Achim Oßwald und Heike Neuroth

Qualifizierungsbestrebungen zur Langzeitarchivierung digitaler Objekte wurden im Projekt nestor II (2006-2009) auf der Basis der Erfahrungen aus nestor I fortgeführt und auf eine neue Ebene gehoben: Neben den Seminaren und Workshops, die z. T. von den nestor Arbeitsgruppen durchgeführt werden und dem nestor Handbuch, das eine ständig erweiterte Sicht auf das Themenfeld gestattet, entwickelte die AG „Kooperation mit Hochschulen im Bereich Aus-, Fort- und Weiterbildung“ ein Konzept für die Entwicklung curricularer Bausteine und Veranstaltungsformen. Richten sich die nestor schools an Praktiker und Studierende, so sind die e-Tutorials zu ausgewählten Themenbereichen der digitalen Langzeitarchivierung von Studenten für Studenten zum Selbststudium oder zur Unterstützung von Lehrveranstaltungen konzipiert. Die curricularen Angebote basieren auf einer Bedarfsanalyse und einer Vision kollaborativer und kooperativer Entwicklung eines LZA-Studienangebots, das aktuelle Ansätze in der Langzeitarchivierungs-Community ebenso aufgreift wie solche aus der Hochschuldidaktik.

19.1 Qualifizierung als Thema im Projekt nestor I

Aufklärung über die Gefahren eines drohenden Verlusts digital vorliegender Informationen sowie über Probleme und Lösungsansätze der Langzeitarchivierung digitaler Daten hatten im Projekt nestor I hohe Priorität: Die im Rahmen des nestor-Projektes entwickelte Website www.langzeitarchivierung.de dokumentiert den Sach- und Forschungsstand zum Thema Langzeitarchivierung erstmalig in deutscher Sprache in der gebotenen Breite und für jedermann zugänglich. Zugleich fördert sie die Vernetzung durch Informationen über Projekte und Fachleute in diesem noch relativ neuen Spezialgebiet. Damit ist sie eine wahre Fundgrube für Praktiker, für die interessierte Öffentlichkeit, aber auch für Lehrende. Dies hat nicht nur Auswirkungen auf die Wahrnehmung des Themas in Fachkreisen, die dank der parallelen Pressearbeit von nestor erkennbar gestiegen ist. Auch die Sensibilisierung der Öffentlichkeit für Fragen der Langzeitarchivierung ist durch Meldungen, Radio- und TV-Beiträge und die Darstellung von Problemfällen der Langzeitarchivierung wesentlich verbessert worden.

Die nestor-Projektbeteiligten haben in der ersten Phase des nestor-Förderzeitraums (2003-2006) mit einer Reihe von Seminarveranstaltungen die Grundproblematik sowie den aktuellen Stand der Problemlösungsangebote zielgruppenspezifisch thematisiert und dokumentiert.[1] Die Materialien der Seminarveranstaltungen sind auf den nestor-Seiten weiterhin zugänglich. Zudem wurden die ersten beiden dieser Veranstaltungen in Göttingen mittels Video aufgezeichnet und stehen auf einer 6 Stunden und 35 Minuten umfassenden DVD-ROM zur Verfügung (siehe nestor-Seminare (2006)). Dadurch sind auch andere als die klassischen Folien-Unterlagen von Vorträgen abrufbar und insgesamt deutlich bessere Voraussetzung für eine Multiplikation der Erkenntnisse gegeben, als dies sonst bei Forschungsprojekten der Fall ist.

Von all dem profitierten die Aktivitäten der Aus-, Fort- und Weiterbildung immens: Langzeitarchivierung, nestor und die Ergebnisse des Projektes sind dort Themen geworden. Konsequenterweise wurde am Ende der Förderphase I für nestor im sog. „Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland“ auch eine Aussage zum Thema Qualifizierung getroffen:

---

1 Siehe bspw. www.langzeitarchivierung.de , - Veranstaltungen, - nestor-Seminare:
1. Seminar „Einführung ...“ am 29.11.05 an der SUB Göttingen; 2. Seminar „Archivbereich ...“ am 13.01.06 ebenfalls an der SUB Göttingen; 3. Seminar „Museen ...“ am 13.06.06 in Nürnberg

„18. Mit der digitalen Langzeitarchivierung entstehen neue Aufgaben für die archivierenden Institutionen. Es muss PROFESSIONELLES PERSONAL *zum Einsatz kommen. Die Anforderungen und Aufgaben der digitalen Langzeitarchivierung sind als ein Schwerpunkt in die Aus- und Fortbildung einzubeziehen. Gezielte Fortbildungsangebote sollten sowohl themenspezifisch sensibilisierend wie auch konkret qualifizierend angelegt werden." nestor-Memorandum (2006)*

19.2 Fortentwicklung der Qualifizierungsanstrengungen im Projekt nestor II

Ziel der zweiten Phase des nestor-Projektes 2006-2009 war es daher, hierfür mittelfristig neue Angebote zu konzipieren. Dazu wurde ein Arbeitspaket „Einrichtung und Ausbau von Ausbildungs- und Fortbildungsangeboten" (AP 5) vorgesehen, das unter der Koordination und Leitung der SUB Göttingen vielseitige Aktivitäten entfaltete. Als besonders produktiv stellte sich die Initiierung einer Arbeitsgruppe heraus, der AG „Kooperation mit Hochschulen im Bereich Aus-, Fort- und Weiterbildung".

Bei den Qualifizierungsanstrengungen sind zwei Aktivitätsbereiche erkennbar:

- projektbasierte, weitgehend von der nestor AG initiierte und größtenteils auch realisierte Qualifizierungsangebote sowie
- hochschulbasierte Qualifizierungsangebote im Rahmen von einschlägigen Curricula.

Ziel der Aktivitäten der nestor AG „Kooperation mit Hochschulen im Bereich Aus-, Fort- und Weiterbildung" ist es, beide Aktivitätsbereiche strukturell zu stärken und zu verbinden, so dass im Sinne der Nachhaltigkeit projektbasierte Aktivitäten weitergeführt, auf jeden Fall aber das im Laufe der Projektzeit entwickelte Know-how für zukünftige Qualifizierungsaktivitäten dauerhaft produktiv gemacht werden kann.

Die wesentlichen Ergebnisse dieser Aktivitäten und die dabei deutlich gewordenen Bedarfe werden nachfolgend skizziert.

19.3 Langzeitarchivierung digitaler Objekte als neuer Gegenstand von Aus-, Fort- und Weiterbildungsangeboten

19.3.1 Die Breite des Qualifizierungsbedarfs

Im Bereich der Ausbildung von Informationsspezialisten sind in den diversen Fachhochschulen und Universitäten Fragen der Langzeitarchivierung von digitalen Medien in den vergangenen Jahren von verschiedenen Professorinnen und Professoren sowie Lehrbeauftragten als Thema aufgegriffen und in unterschiedlichen Lehrveranstaltungsformen thematisiert worden. Den aktuellen Stand haben Oßwald / Scheffel (2006) und Oßwald / Scheffel (2007) zusammenfassend dargestellt.

Solche Veränderungen der Lehrinhalte sind im weitesten Sinne eine Auswirkung der systematischen Umstellung des Publikationswesens auf digital basierte Prozesse und Produkte. Diese beeinflussen methodisch nahezu alle Tätigkeitssegmente im bibliothekarischen und weiteren informationswirtschaftlichen Kontext. Faktisch bedeutet dies für die meisten Institutionen parallele Aufgabenbereiche und Prozesse, weil nur in wenigen Anwendungsbereichen die völlige Umstellung auf digitale Bestände und Dienstleistungen realisierbar und gewünscht ist. Entsprechend hat schon im Jahr 2002 die Hochschulrektorenkonferenz empfohlen, das wissenschaftliche Informations- und Publikationswesen in den nächsten Jahren konsequent auf elektronische Kommunikations- und Informationsmöglichkeiten auszurichten (vgl. Hochschulrektorenkonferenz (2002)).

Dies gilt auch für das Thema Bestandssicherung, das in der Konsequenz in den Kulturerbeeinrichtungen zweigleisige Aktivitäten fordert: Die für die traditionellen Printprodukte einerseits wie auch jene für die in den letzten Jahren in Zahl und Relevanz einen wachsenden Stellenwert erfahrenden digitalen Objekte andererseits.

Ginge es um Bibliotheken und die anderen Kulturerbeeinrichtungen allein, würde dieser Umstand vermutlich wenig Beachtung erfahren. Weil jedoch das Bestandssicherungs-Know-how für digitale Objekte – in der öffentlichen (Fach-) Diskussion als Langzeitarchivierung digitaler Objekte bezeichnet – aus wichtigen anderen gesellschaftlichen Segmenten, speziell dem der Wirtschaft, aber auch aus dem privaten Bereich nachgefragt wird, ist der Erwartungsdruck auf die Experten in Bibliotheken, Archiven und Museen zur Bereitstellung spezifischer Problemlösungen sehr rasch und deutlich gestiegen – und damit

automatisch auch der Erwartungsdruck auf die Hochschuleinrichtungen, die für diese Arbeitsmarktsegmente qualifizieren.

Die dort entwickelten Qualifizierungskonzepte und -inhalte sind insofern auch für die deutlich weiter gefassten Zielgruppen von nestor gedacht, denn das Kompetenznetzwerk nestor verfolgt das Ziel, die digitalen Ressourcen in Deutschland zu sichern und verfügbar zu machen und dabei mit anderen Netzwerken und Entscheidungsträgern national und international zusammenzuarbeiten, um gemeinsam das kulturelle und wissenschaftliche Erbe Deutschlands langfristig zu bewahren. Zielgruppen sind demnach:

- Institutionen, zu deren Aufgaben die Archivierung und Langzeiterhaltung digitaler Ressourcen gehört
- Personen, die über Kompetenzen und Erfahrungen auf dem Gebiet verfügen
- Produzenten digitaler Ressourcen in Wissenschaft, Wirtschaft und Verwaltung
- Nutzer digitaler Ressourcen
- Förderinstitutionen mit deren Rahmenplanungen und Einzelaktivitäten
- ausländische Institutionen, Organisationen und Projekte, die auf dem Gebiet der Langzeitarchivierung digitaler Ressourcen aktiv sind
- Kommerzielle Dienstleister und Industriepartner, die Services oder Produkte zur Langzeitarchivierung anbieten

Deren Qualifizierungsbedarf muss durchaus differenziert gesehen werden. Perspektivisch sollte daher eine modular aufgebaute Qualifizierungsstrategie entwickelt werden, die sich an die folgenden drei prioritären Zielgruppen in den Berufsfeldern richtet:

- Entscheidungsträger (E)
- Allgemein Qualifizierte aus dem Kulturerbe-Bereich (Q)
- Mitarbeiterinnen und Mitarbeiter mit Langzeitarchivierungsaufgaben (M)

Die nachfolgende Tabelle konkretisiert, welche zu vermittelnden Inhalte für diese drei Zielgruppen aus Sicht der Autoren sinnvollerweise angeboten werden sollten. Je nach Interpretation der beschriebenen Inhalte sind hier vermutlich Modifikationen sinnvoll.

Handlungsorientierte Vermittlungsinhalte	**E**	**Q**	**M**
Sensibilisierung + grundlegende Kenntnisse der LZA	X	X	X
Vertiefte Kenntnisse theoretischer Konzepte der LZA (Strategien, Infrastruktur, Sammelrichtlinien, Policies)	X	X	X
Konzeption und Realisierung von Datensicherungs-, Datenrettungs- und Langzeitsicherungsstrategien		X	X
Vertiefte Kenntnisse der Realisierung von Datensicherungs-, -rettungs- und Langzeitsicherungsstrategien; Archivserverlösungen und deren Durchführung	X		X
Vertiefte Kenntnisse und Anwendungsfertigkeiten bezüglich der Standards, die bei der LZA zur Anwendung kommen			X
Kenntnisse, Fähigkeiten und Fertigkeiten des Daten- und Informations- bzw. Recordsmanagements		X	X
Vertiefte Kenntnis der Informatiklösungen für LZA und deren Anwendung			X
Kenntnis der rechtlichen Aspekte	X	X	X
Vertiefte Kenntnis der rechtlichen Aspekte und ihrer Anwendung			X
Kenntnis der Kostenaspekte	X	X	X

E: Entscheidungsträger
Q: Allgemein Qualifizierte aus dem Kulturerbe-Bereich
M: Mitarbeiterinnen und Mitarbeiter mit Langzeitarchivierungsaufgaben

Diese Zusammenstellung deckt sich auch mit der Erwartungshaltung aus der Branche jener Firmen, die sich mit Datenrettung nach Havariefällen befassen.[2]

Daraus leiteten die Autoren die folgenden Empfehlungen für die Entwicklung von Qualifizierungsangeboten ab:

- Kooperative bzw. kollaborative Entwicklungen von Lehreinheiten / Modulen zu den nestor-Forschungsfeldern in didaktisch und medial für verschiedene Zielgruppen aufbereiteter Form
- Konzeption der Module derart, dass sie für die Aus- und Weiterbildung sowie in der Fortbildung genutzt werden können, z.B. durch Einbettung in bzw. Umsetzung als e-Learning-Applikationen
- Vermittlung der gängigen nationalen und internationalen Normen / Standards an praktischen Beispielen.
- Vermittlung von best-practice-Lösungen der verschiedenen Konzepte auf internationaler, nationaler, regionaler und lokaler Ebene (inkl. der Verknüpfung mit entsprechenden Anwendungen zur Anschauung)

Die Angebote sollten – und dies wäre eine deutliche Veränderung zu früheren Angeboten an Aus- und Weiterbildungsangeboten im Bereich der Langzeitarchivierung – nicht vorwiegend theorielastig sein, sondern auch praktische Übungen einschließen.

19.3.2 Angebote einzelner Hochschulen

Über Jahre war es eher von der intrinsisch motivierten Innovationsoffenheit einzelner Lehrender abhängig, ob das Thema „Langzeitarchivierung digitaler Objekte“ an einer Hochschule aufgegriffen wurde.[3] Vor diesem Hintergrund

2 Beispielhaft für diese Branche wurde 2006 die Firma Ontrack (*http://www.ontrack.de/*) befragt.

3 Selbst wenn die fachliche Weitsicht und Innovationsorientierung bei den zuständigen Leitungsgremien oder -personen gegeben ist, so sind hier erst jüngst begrenzte Steuerungsmöglichkeiten (W-Besoldung mit Zielvereinbarungen) eröffnet worden. Die traditionellen Beschäftigungsverhältnisse (Beamtenverhältnis auf Lebenszeit und damit geringe Erneuerungszyklen des Dozentenstammes) erlaubten so gut wie keine

erfolgte eine curriculare Einbindung dieses Themas zum Beispiel mit den Schwerpunkten im Organisatorischen (Strategie, Sammlungsschwerpunkte etc.), Finanziellen (Geschäftsmodelle, Finanzierung etc.) oder Technologischen (Archivsysteme, Ingest-Prozesse etc.) bei der Qualifizierung von Bibliothekarinnen und Bibliothekaren, aber auch von Mitarbeiterinnen und Mitarbeitern anderer sog. Kulturerbeeinrichtungen wie z.B. Museen bislang nur sehr begrenzt.

Die curriculare Situation stellt sich zum Ende des Jahres 2008 wie folgt dar:

- In einer zunehmenden Anzahl von Hochschulen erfolgt eine thematische Einführung und Bestandsaufnahme in einzelnen Lehrveranstaltungen und insofern eine (auch weiterhin) eher punktuelle Qualifizierung (vgl. Oßwald / Scheffel 2007).
- Das in Deutschland bislang am deutlichsten auf den Bereich der Langzeitarchivierung digitaler Objekte ausgerichtete Qualifizierungsangebot, das MA-Studienangebot der Kunstakademie Stuttgart (vgl. AKB Stuttgart; http://www.mediaconservation.abk-stuttgart.de/) fokussiert auf einen speziellen Anwendungsbereich. Seine Weiterführung ist allerdings in Frage gestellt.[4]
- Als neuer Themenschwerpunkt und eventuelles zukünftiges Lehrgebiet einer darauf spezialisierten Professur könnte das Thema Langzeitarchivierung in den neuen Bachelor-Studiengangskonzepten nur bei großem Problemdruck und unter Verzicht auf ein anderes, traditionelles oder ggf. ebenfalls neues Thema in das Curriculum integriert werden – und dies auch frühestens beim nächsten anstehenden Studienreformzyklus (alle 5-7 Jahre).
- Hinzu kommt, dass das Thema Langzeitarchivierung digitaler Objekte ein typisches Vertiefungsthema und insofern vorzugsweise für einen Master-Studiengang geeignet ist.[5]
- Für die Realisierung eines MA-Angebotes fehlen sowohl die personellen Ressourcen (im Sinne einer kritischen Anzahl von Wissenschaftlern mit entsprechenden theoretischem und berufspraktischem Hintergrund, die

Steuerungsmöglichkeiten, sondern die Hochschulen mussten weitgehend auf die o. g. intrinsisch motivierte Innovationsoffenheit der Lehrenden vertrauen.

4 „Nach drei Durchgängen durchläuft der Studiengang momentan eine Evaluierung. Von ihrem Ergebnis wird abhängen, ob und wann neue Studierende aufgenommen werden können. Wir werden dies schnellstmöglich hier bekanntgeben, vermutlich jedoch nicht vor dem Frühjahr 2009.“ http://www.mediaconservation.abk-stuttgart.de/index.php?id=18

5 Aus kapazitären Gründen ist bei Bacherlorstudiengängen zumeist die Option auf Wahlpflichtangebote sehr reduziert. Inwieweit ein solches Wahlpflichtangebot von den Studierenden aufgegriffen würde, sei dahin gestellt.

die für ein solches Angebot notwendige breite fachliche Qualifikation und Lehrerfahrung mitbringen) als auch stabile arbeitsmarktbezogene Prognosen, die das Auslastungs- und damit auch Finanzierungsrisiko eines solchen Einzelangebotes für eine Hochschule allein rechtfertigen würden.[6]

Vor diesem Hintergrund eröffnet sich die Option − je nach fachlich verantwortlicher Sicht aber auch die Notwendigkeit − zu neuen Formen der Zusammenarbeit von Hochschuleinrichtungen und außerhochschulischen Qualifizierungseinrichtungen zu kommen, die die Relevanz des Themas für ihre jeweilige Klientel sehen – und das Aufgreifen des Themas auch als Indiz für ihre Innovationsoffenheit wahrgenommen wissen wollen.

19.4 Hochschulübergreifende Kooperationen

Zielsetzungen der hochschulübergreifenden Zusammenarbeit

Im Rahmen der von nestor projektbasiert erfolgten Zusammenarbeit verschiedener Hochschulen und anderer fachlich einschlägiger Qualifizierungseinrichtungen deutet sich ein neues Modell curricularer Kooperation an, bei dem aus der (kapazitären und spezialisierungsbezogenen) Not eine Tugend (mit kooperativen und kollaborativen Innovationspotenzialen) erwachsen könnte.
Die Kooperation zwischen nestor und den Hochschulen bzw. anderen Qualifizierungseinrichtungen wird im Projektzeitraum vom nestor-Partner Staats- und Universitätsbibliothek Göttingen als zentralem thematischen Ansprechpartner koordiniert. Modellhaft ist dabei nicht nur der Prozess und die Form der Zusammenarbeit, sondern auch die Zielsetzung, ein neues und hochrelevantes Thema unter den skizzierten hochschulrechtlichen Rahmenbedingungen im Interesse der Fachwelt nicht konkurrierend, sondern kooperativ aufzugreifen.
Schritte auf dem Weg zu diesem Ziel sind nachfolgend skizzierte Formen der Zusammenarbeit im Sinne einer gestuften Entwicklung von

- kooperativer Zusammenarbeit bei der Konzeption und Realisierung von Fortbildungsveranstaltungen;
- gemeinsam genutzten Lehr- und Lernmaterialien;

6 Das Angebot der Kunstakademie Stuttgart (http://www.mediaconservation.abk-stuttgart.de/) bestätigt diese Einschätzung prinzipiell, weil es nur mit ungewöhnlich hoher finanzieller Förderung aus Landesmitteln realisiert werden konnte und zudem durch seine inhaltliche Ausrichtung auf eine sehr spezielle Zielgruppe fokussiert.

- kooperativ und nach gemeinsam vereinbarten Standards entwickelten e-Tutorials;
- der gegenseitigen Anerkennung von fachlich einschlägigen Lehrmodulen (z.B. auf der Grundlage von ECTS Punkten) bis hin zu einem
- kooperativ konzipierten und realisierten Studienschwerpunkt bzw. Studienangebot.

Kooperationspartner bei der Entwicklung neuer Qualifizierungsangebote

Die Arbeitsgruppe „Kooperation mit Hochschulen im Bereich Aus-, Fort- und Weiterbildung" konnte neue Kooperationspartner für nestor gewinnen. Insgesamt acht Partner aus Hochschulen und Weiterbildungseinrichtungen aus Deutschland, Österreich und der Schweiz sind zurzeit in dieser Arbeitsgruppe engagiert:

- Niedersächsische Staats- und Universitätsbibliothek Göttingen (Leitung)
- Archivschule Marburg
- Fachhochschule Köln, Institut für Informationswissenschaft
- Humboldt University Berlin - Institut für Bibliotheks- und Informationswissenschaft (IBI)
- Hochschule für Technik, Wirtschaft und Kultur Leipzig, Fachbereich Medien
- Fachhochschule Potsdam, Fachbereich Informationswissenschaften
- Hochschule für Technik und Wirtschaft Chur (Schweiz), Informationswissenschaft
- Technische Universität Wien (Österreich), Information & Software Engineering Group

Ausgehend von Vorarbeiten im Rahmen von nestor I, insbesondere den Seminaren, die zum Teil als Videoaufzeichnung auf DVD vorliegen und auch für spätere Lehreinheiten nachgenutzt werden konnten, wurden die Arbeitsschwerpunkte für die 2. Laufzeit des Projektes konzipiert. Dabei war von Beginn an auch vorgesehen, im Bereich von Lern- und Lehreinheiten zusammenzuarbeiten.

Ein „Memorandum of Understanding" über die Zielsetzungen der Kooperationspartner

Die Ziele der seit 2006 aktiven Arbeitsgruppe sind in einem gemeinsamen, im Jahr 2007 von den verantwortlichen Vertretern der Hochschulen unterzeichneten „Memorandum of Understanding" (http://nestor.sub.uni-goettingen.de/education/mou.pdf) festgehalten:

- Die Partner vereinbaren den wechselseitigen Informationsaustausch zu Fragen der Qualifizierung im Bereich der langfristigen digitalen Archivierung.
- Die Partner streben an, mittelfristig den curricularen Anteil des Themas digitale Langzeitarchivierung in der Lehre auszubauen, soweit bei ihnen ein entsprechendes Lehrangebot besteht. Die Lehrenden streben an, sich bei der gemeinsamen Entwicklung der Module auf Aspekte der digitalen Langzeitarchivierung zu spezialisieren und sich gegenseitig zu ergänzen. Dadurch kann eine Profilbildung der jeweiligen Institution stattfinden.
- Soweit bei ihnen ein entsprechendes Lehrangebot besteht, treiben die Partner perspektivisch die Entwicklung eines gemeinsamen Curriculums voran, das zwischen den Hochschulen in verteilten, unterschiedlichen Schwerpunkten (wie z.B. Technik, Organisation, Standards etc.) angeboten und genutzt werden kann. Hierzu kann auch die Entsendung von Dozenten oder Studierenden zählen. Eine Möglichkeit der Realisierung könnte ein gemeinsames, modular angebotenes MA-Studium sein, soweit die entsprechenden Voraussetzungen an den Partnerinstitutionen gegeben sind.
- Alle beteiligten Partner werden Materialien in das curriculare Konzept einbringen, die von den hier genannten Partnern inhaltlich geprüft und ggf. mit einem Zertifikat versehen werden können.
- In Abhängigkeit von den institutionellen Rahmenbedingungen ist die gegenseitige Anerkennung von Lehrveranstaltungen und der damit erworbenen ECTS Punkte beabsichtigt.

Dieses „Memorandum of Understanding“ ist zunächst an die Projektlaufzeit von nestor II (Juni 2009) gekoppelt. Es werden aber Diskussionen darüber geführt, wie die Kooperation und Kollaboration nach Ende des BMBF-Projektes nestor aussehen kann. Zu diesen Überlegungen gehört auch die Frage, ob und welche weiteren Partner für diese Form der Zusammenarbeit gewonnen werden können.[7]

7 Beispielsweise entwickelt die Hochschule der Medien, Stuttgart, ab dem Sommersemester 2009 ein thematisch passendes Modul auf der Grundlage der im Projekt entwickelten Rahmenbedingungen, das in das vorhandene Konzept eingefügt werden soll.

19.5 Projektbasierte, vorcurriculare Qualifizierungsangebote im Themenbereich der digitalen Langzeitarchivierung

Im Projekt nestor wurde ein umfangreicher Qualifizierungsbedarf erkannt, der bis zur Verankerung des Themas in der Hochschulausbildung über Weiterbildungsangebote bedient werden muss.

Die bislang sehr erfolgreichen Veranstaltungen wie Workshops, Tagungen, Seminare etc. werden auch weiterhin angeboten und erfreuen sich regen Zuspruchs.

Ein weiterer Baustein der nestor Bemühungen um eine umfassende Qualifizierung ist das „nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung", das im Frühjahr 2007 in einer ersten Version der Öffentlichkeit vorgestellt wurde. Zukünftig wird dieses Handbuch in weiteren überarbeiteten und ergänzten Versionen veröffentlicht werden.

Um den Qualifizierungsbedarf jedoch bereits in der Ausbildung einschlägiger Studiengänge zu verankern, wurde damit begonnen, ein Konzept für ein Fort- und Weiterbildungsangebot in Zusammenarbeit mit bestehenden Qualifizierungseinrichtungen aus dem Bereich Bibliothek, Archiv und Museum zu entwerfen.

Diese Aktivitäten wurden konkret in nestor II mit der nestor Spring School 2007 eröffnet. Die Reihe der mehrtägigen Training Schools wurde und wird mit ähnlichen Veranstaltungen fortgesetzt – so z.B. durch die nestor Spring School 2009. Als weiteres Angebot erfolgte die Entwicklung von e-Learning-Modulen. Im Folgenden werden diese einzelnen Bausteine vorgestellt:

19.5.1 Seminare

Bisher wurden mehrere, meistens eintägige Seminare zu bestimmten Themen angeboten, die durchweg sehr gut besucht waren. Der Vorteil der Seminare ist, dass schnell und unkompliziert auf aktuelle Bedürfnisse der Langzeitarchivierungs-Community reagiert und in dieser Form relativ unaufwändig Unterstützung und Hilfe angeboten werden kann. Die Seminare sind in der Regel auf bestimmte Themenbereiche spezialisiert und haben nicht unbedingt immer nur einen einführenden und grundsätzlichen Charakter, zumal wenn eine der Arbeitsgruppen[8] innerhalb von nestor die Veranstaltung ausrichtet.

---

8 Vgl. http://nestor.cms.hu-berlin.de/moinwiki/
Alle hier aufgeführten URLs wurden im April 2009 auf Erreichbarkeit geprüft .

19.5.2 Das nestor Handbuch

Die Langzeitarchivierung digitaler Objekte gewinnt sowohl national als auch international zunehmend an Bedeutung. Das „nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung“ (http://nestor.sub.uni-goettingen.de/handbuch/) versucht, das derzeit vorhandene Wissen über das vielfältige und komplexe Thema und seine unterschiedlichen Teilaspekte zu sammeln und über eine „kleine Enzyklopädie“ in strukturierter Form einer deutschsprachigen Gemeinschaft zugänglich zu machen. Zielgruppen sind die breite Öffentlichkeit, Entscheidungsträger, Fachleute aus dem Kulturerbe-Bereich sowie Personen mit Langzeitarchivierungsaufgaben.

Einzelne, von verschiedenen Experten erstellte Fachaufsätze gestatten einen Einblick in die diversen Themengebiete der Langzeitarchivierung: von technischen und rechtlichen Aspekten bis hin zur Definition von Rahmenbedingungen.

Das nestor Handbuch wird als „living document“ verstanden, in dem bis zur Version 1.5 schon eine Reihe von Themen gesammelt werden konnten. Die Beiträge des Handbuchs werden im Laufe der Zeit ergänzt, vervollständigt und aktualisiert.

Damit verbunden ist auch, dass der Kreis der Experten laufend erweitert und ergänzt wird. Ziel ist dabei in naher Zukunft einen umfassenden Überblick über das anspruchsvolle und sich in stetiger Entwicklung befindliche Themengebiet zu erhalten.

Angeboten wird auch die Möglichkeit, die deutschsprachige Fachgemeinschaft in den Entstehungsprozess einzubeziehen, indem Kommentare zu den einzelnen Artikeln in das technische System eingestellt werden können.[9]

19.5.3 nestor Schools

Nach dem Vorbild der Delos Summer Schools[10] und den ab 2008 startenden Digital Preservation Europe (DPE) Schools (http://www.digitalpreservationeurope.eu/) sind im Jahr 2007 zwei nestor Schools veranstaltet worden, die nestor Spring School (http://nestor.sub.uni-goettingen.de/spring_school_2007/index.php) und die nestor Winter School (http://nestor.sub.uni-goettingen.de/winter_school_2007/index.php). Mit jeweils über 40 Teilnehmern und

9 Inwieweit das Handbuch z.B. mit neueren Web 2.0 Technologien insgesamt interaktiver gestaltet werden kann und ob dies zu einer Qualitätssteigerung führt, muss im Laufe der Zeit sicherlich geprüft werden.

10 Vgl. z.B. Delos Summer School 2007, http://www.dpc.delos.info/ss07/index.php

Referenten waren die Schools sehr erfolgreich. Erfreulicherweise konnte ein hoher Prozentsatz an Studierenden als Teilnehmer gewonnen werden.

Das Konzept der Schools besteht in der Zusammensetzung unterschiedlicher theoretischer und praktischer Blöcke. Jeweils eine 1,5-stündige Lektion führt in das Thema ein (z.B. Metadaten, Formate, Vertrauenswürdige Archive), danach folgt eine praktische Übung in mehreren kleinen Gruppen, die von den Referenten und Experten intensiv betreut wird. Die Teilnehmer stellen die Ergebnisse ihrer Übung dem Plenum vor und zum Abschluss wird das Themengebiet zusammenfassend diskutiert. Da die Schools jeweils für eine Woche angelegt sind, konnte eine umfassende Einführung in das Thema der digitalen Langzeitarchivierung (nestor Spring School) gegeben bzw. im Rahmen der Winter School der Fokus auf praktische Anwendungsfelder gelegt werden.

2008 wurde die nestor / DPE Summer School (http://nestor.sub.uni-goettingen.de/summer_school_2008/index.php) erstmals in einen Einführungsblock und einen fachlich vertiefenden Themenblock (Speichertechnologien und LZA-Strategien) aufgeteilt. In der nestor / DPE Spring School 2009 (http://nestor.sub.uni-goettingen.de/spring_school_2009/index.php) wurde wieder zum Modell einer durchgängigen Veranstaltung mit dem Fokus auf Archive für Forschungsdaten sowie auf Archivsysteme zurückgekehrt. Mit zum Konzept gehört, dass die Referentinnen und Referenten nicht nur möglichst während der gesamten Veranstaltung anwesend sind, sondern sich auch als Spezialisten bei Übungen und Gesprächen über die Praxiserfahrung der Teilnehmerinnen und Teilnehmer mit ihrer Expertise einbringen und so zusätzlich die Lernprozesse vertiefen.

19.5.4 e-Learning-Tutorials

Studierende aus den Fachhochschulen Köln, Potsdam und Leipzig sowie der Hochschule für Technik und Wirtschaft Chur in der Ostschweiz beteiligten sich im Wintersemester 2007/2008 an einem gemeinsamen Projekt zur Entwicklung von e-Learning-Tutorials zu verschiedenen Themenfeldern der Langzeitarchivierung digitaler Objekte. Die Tutorials bieten einführende und inhaltlich vertiefende Informationen, die unter Nutzung der auch international weit verbreiteten e-Learning-Plattform Moodle (http://www.moodle.de) entwickelt wurden. Diese Software-Anwendung wird von der Humboldt-Universität zu Berlin (HUB) technisch bereit gestellt und betreut. Außerdem wurde von der HUB ein gestalterisches und didaktisches Konzept für die e-Learning-Tutorials entworfen.

Auf der Grundlage von vier Seminar- bzw. Projekt-Veranstaltungen, die von Hochschullehrern[11] an den jeweiligen Standorten initiiert, koordiniert und betreut wurden, bereiteten die Studierenden z.B. folgende Themen in Form von e-Learning-Tutorials auf:

- Einführung in die Langzeitarchivierung digitaler Objekte
- Formate und Datenträger in der Langzeitarchivierung
- Langzeitarchivierung bestimmter Datentypen (CAD-Daten, GIS-Daten)
- Metadatenerzeugung für technische Abläufe in der Langzeitarchivierung (z.B. ingest)

Im Sommersemester 2008 wurden diese e-Tutorials an verschiedenen Hochschulstandorten in Lehrveranstaltungen evaluiert und optimiert. Nach einer Auswertung der ersten Erfahrungen wurde mit einem „Kick-off-Meeting" von Studierenden und Lehrenden am 10./11. Oktober 2008 die zweite Runde der e-Tutorial-Erstellung mit neuen Projektgruppen anderer Matrikeln in Chur, Köln und Leipzig gestartet. Neu hinzukommen sollen e-Tutorials aus Berlin und Wien, die nicht von studentischen Projektgruppen erstellt werden.

Im Sommersemester 2009 werden die „von Studenten für Studenten" erstellten e-Tutorials dann wiederum an den Hochschulstandorten getestet und evaluiert. Auch ist geplant, von Studierenden weitere Module im Rahmen hochschulübergreifender Seminare entwickeln zu lassen, die an den jeweiligen Hochschulen von gemeinsamen Projektveranstaltungen begleitet werden.

19.6 Curriculare Optionen für eine Integration in die hochschulbasierte Aus-, Fort- und Weiterbildung

Im „Memorandum of Understanding" wurde als Ziel formuliert:

> *„Soweit bei ihnen ein entsprechendes Lehrangebot besteht, treiben die Partner perspektivisch die Entwicklung eines gemeinsamen Curriculums voran, das zwischen den Hochschulen in verteilten, unterschiedlichen Schwerpunkten (wie z.B. Technik, Organisation, Standards etc.) angeboten und genutzt werden kann. Hierzu kann auch die Entsendung von Dozenten oder Studierenden zählen. Eine Möglichkeit der Realisierung könnte ein gemeinsames, modular angebotenes MA-Studium sein,*

11 Prof. Dr. N. Stettler, HTW Chur; Prof. R. Scheffel, HTWK Leipzig; Dr. K. Schwarz, FH Potsdam und Prof. Dr. A. Oßwald, FH Köln

> *soweit die entsprechenden Voraussetzungen an den Partnerinstitutionen gegeben sind."* (nestor "Memorandum of Understanding" 2007, http://nestor.sub.uni-goettingen.de/education/mou.pdf)

Auf Grundlage der bislang erfolgten Aktivitäten und Erfahrungen in der Zusammenarbeit steigen die Chancen, dass dieses Ziel der kooperativen Entwicklung eines gemeinsam und modular konzipierten MA-Studienangebotes realisierbar wird. Hierzu trägt die Entwicklung der e-Tutorials in starkem Maße bei.

Die aus den e-Tutorials ersichtliche thematische Schwerpunktsetzung bedeutet für die Beteiligten keine kompetenzbezogene Festlegung oder Einschränkung. Den einzelnen, jetzt schon beteiligten oder zukünftig hinzu kommenden Lehrenden und Hochschuleinrichtungen verbliebe weiterhin die Option, individuelle Kompetenzbereiche oder standortbezogene Forschungsschwerpunkte in kollegialer Abstimmung in dieses offene Konzept einzubringen. Damit ist auch weiterhin die Profilbildung jeder kooperierenden Hochschule sichergestellt.

Mit der stabilen Bereitstellung und Pflege der e-Learning-Tutorials bestünde für die beteiligten Hochschulen, ihre Lehrenden sowie die beteiligten kompetenten Praktiker, eine gemäß den Anforderungen der Kultusministerkonferenz (vgl. SAK (2005)) zertifikatsbasierte Zusatzqualifikation anzubieten, die solitär oder z.B. als Erweiterung der bereits angebotenen berufsbegleitenden Master-Fernstudiengänge in Berlin und Köln zertifiziert realisiert werden könnte. Hierfür bedarf es lediglich der Bereitschaft, die bislang schon begonnenen, von nestor initiierten Aktivitäten konsequent weiter zu führen und in ein qualitativ abgesichertes, hochschulübergreifendes Konzept einzubringen.

Damit wäre es den beteiligten Hochschulen und den anderen Qualifizierungseinrichtungen möglich, auf der Grundlage fachlicher Zusammenarbeit im Interesse der Fachcommunity und deren Entwicklung über ihren wettbewerbsorientierten „Schatten" zu springen und anstelle traditioneller, zu Zersplitterung der Ressourcen und Kompetenzen führenden Konkurrenz zukunftsorientierte Formen der Zusammenarbeit zu finden, die auch für andere Bereiche wegweisend sein könnten.

19.7 Kooperationsmöglichkeiten für die weiteren Entwicklungsschritte

Es gibt vielfältige Möglichkeiten der Kooperation, die in Zukunft – auch über das Projektende von nestor hinaus – mit Blick auf ein curricular eingebundenes Qualifizierungsangebot verfolgt werden können.

In den USA läuft zum Beispiel bis 2009 das für diese Perspektive hochinteressante Projekt, "DigCCurr - Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum" (http://www.ils.unc.edu/digccurr), an dem unter Federführung der School of Information and Library Science (SILS) der Universität von North Carolina / Chapel Hill ein „openly accessible graduate-level curriculum" (ebd.) entwickelt wird (vgl. Lee (2007)). Erste Ergebnisse zeigen, dass hier versucht wird, das Thema der digitalen Langzeitarchivierung umfassend auf allen Ebenen eines auf den Bereich digitaler Bibliotheksaufgaben orientierten Curriculums mit zu denken und entsprechend zu berücksichtigen. Digitale Langzeitarchivierung wird verstanden als ein Prozess, der den kompletten Lebenszyklus eines digitalen Objektes umfasst (digital curation). Dementsprechend sind in einer ersten, unveröffentlichten Version neun von zehn Kernthemen, die für die Qualifizierung im Bereich Digitale Bibliothek identifiziert wurden, langzeitarchivierungsrelevante Themenaspekte zugeordnet (z.B. für den Themenbereich „Digital Objects" u.a. der Aspekt „2-c File formats, transformation, migration"). Ein Themencluster nennt sich sogar explizit „Preservation". Interessant an diesem Ansatz ist, dass das Forschungs- und Lehrgebiet der digitalen Langzeitarchivierung komplett in das ganze Studienprogramm integriert werden soll. Das Projekt wird in den USA mit anderen Hochschulen gemeinsam durchgeführt, so dass die Chance besteht, die integrative Sicht des Digital Curation auch an anderen Studienstandorten einzubringen.

Die sich daran anschließende Frage in diesem Projektkontext ist, wie ein adäquater Studienabschluss aussehen könnte. Diskutiert werden, wie in Deutschland auch, verschiedene Ansätze wie zum Beispiel Informationsspezialist mit Schwerpunkt digitale Langzeitarchivierung (data steward oder data curator). Sicherlich wäre hilfreich, hier in Kooperation mit allen relevanten Partnern national, aber auch international, zu gemeinsamen Konzepten zu kommen.

Eine weitere Gestaltungsvariante im nestor Kooperationsrahmen besteht darin, mit Hilfe englischsprachiger internationaler Partner, die vorhandenen e-Tutorials in die englische Sprache zu übersetzen, sie damit zu „internationalisieren" und hierbei auch zu standardisieren. Gerade im Hinblick auf den Bologna Prozess könnte hiermit ein Grundstock gelegt werden, um sich zumindest auf europäischer Ebene enger zu vernetzen. Es besteht jedoch auch Interesse bei Partnern in außereuropäischen Ländern. Dies könnte in Zukunft zu einer Zusammenarbeit auf Hochschulebene führen, im Rahmen derer Studierende einen Teil ihres Studiums an einer deutschsprachigen und einen weiteren Teil an einer anderen europäischen Hochschule absolvieren. Voraussetzung hierfür wäre die gemeinsame Abstimmung über relevante Lehrinhalte sowie die

gegenseitige Anerkennung von ECTS Punkten. Dies wäre bei der gemeinsamen Entwicklung von e-Tutorials fast automatisch gegeben.

Darüber hinaus bietet eine umfangreiche, qualitätsgeprüfte, z.B. nestorzertifizierte Sammlung von e-Tutorials natürlich die Möglichkeit eine Qualifizierungsmaßnahme auf individuell oder institutionell abgestimmte Bedürfnisse auszurichten. Einzelne e-Tutorial Module könnten für bestimmte Anforderungen zusammengestellt werden, so dass externe Schulungs- und Fortbildungsveranstaltungen im Hochschulbereich, aber auch bei den Gedächnisorganisationen (Bibliotheken, Archive, Museen) oder Industrievertretern angeboten werden könnten. Der willkommene Nebeneffekt könnte darin bestehen, dass hierüber Einkünfte erzielt werden könnten, die wiederum z.B. in den Ausbau der technischen Plattform oder in den Erwerb von kostenpflichtigen Tools zur Erstellung von Lehr- und Lernmodulen, in Pflege und Weiterentwicklung des Angebots investiert werden könnten. Ein solches Angebotskonzept böte gleichzeitig die Chance, die durch verschiedene Förderprogramme des BMBF (vgl z.B. http://www.bmbf.de/foerderungen/12128.php) oder der Europäischen Kommission (vgl. EC (2007)) signalisierte Zielvorstellung lebenslangen Lernens zu bedienen, die für das Bestehen in unserer heutigen Wissensgesellschaft unabdingbar ist.

Diese Vorstellung scheint sich auch mit den derzeitigen Entwicklungen und Diskussionen in Deutschland zu decken, die der Lehre im Hochschulbereich einen höheren Stellenwert verschaffen möchten (vgl. Wiarda (2008)). Auch wenn im Gegensatz zu der Exzellenzinitiative für die Forschung, der insgesamt 1,9 Milliarden Euro zur Verfügung standen, für die zurzeit (in Stifterverband und KMK) diskutierte „Exzellenzinitiative für die Lehre“ nur ganze 5 Millionen Euro geplant sind, so könnte diese Initiative dennoch einem höheren Stellenwert der Lehre den Weg bereiten. Die Grundzüge der drei Förderlinien zeichnen sich nach dem derzeitigen Stand wie folgt ab:

- „Nachwuchsförderung“: Stipendien, Weiterbildungsangebote, Berufung von Gastprofessoren
- „Strukturbildung“: Kompetenzzentren für die Lehre inklusive Weiterbildung der Lehrenden
- „Strategieentwicklung“: Entwicklung von Zukunftskonzepten für die Angliederung an internationales Spitzenniveau im Bereich der Lehre.

Die Gründung einer "Deutschen Lehrgemeinschaft" (vgl. z.B. http://idw-online.de/pages/de/news243140) analog zur Deutschen Forschungsgemeinschaft erscheint hier nur ein nächster konsequenter Schritt zu sein. Eine

solche Entwicklung könnte für die Hochschullandschaft in Deutschland ein interessantes Entwicklungspotential bergen: Fachhochschulen und Universitäten könnten sich gemeinsam im Bereich der kooperativen Lehre engagieren. Ein hochschulpolitisch interessanter Nebeneffekt aus bundesdeutscher Sicht ist dabei das Erodieren der – außerhalb des deutschsprachigen Raums ohnehin kaum kommunizierbaren – Differenzierung zwischen Universitäten und Fachhochschulen.

Im Rahmen dieser sich abzeichnenden Entwicklung könnten die gemeinsamen Erfahrungen der nestor-Kooperation im Bereich der digitalen Langzeitarchivierung hilfreich, vielleicht sogar wegweisend sein (vgl. Neuroth / Oßwald (2008).

19.8 Zitierte Quellen und Literatur

Delos Summer School 2007: http://www.dpc.delos.info/ss07/index.php

Hochschulrektorenkonferenz 2002: Hochschulrektorenkonferenz: *Zur Neuausrichtung des Informations- und Publikationssystems der deutschen Hochschulen*; Empfehlung des 198. Plenums vom 5. November 2002. http://www.hrk.de/de/download/dateien/Empfehlung_Bibliothek.pdf

ABK Stuttgart 2009: *Staatlichen Akademie der Bildenden Künste Stuttgart: Konservierung neuer Medien und digitaler Information* [Studiengang an der Staatlichen Akademie der Bildenden Künste Stuttgart]: http://www.mediaconservation.abk-stuttgart.de/index.php?id=18

Lee 2007: Lee, Christopher: International Digital Curation Curriculum: DigCCurr Project. Folien des Vortrags bei *"iPRES 2007 – International Conference on Preservation of Digital Objects"*; Beijing, 11-12 October, 2007 http://ipres.las.ac.cn/pdf/ipres2007-digccurr.pdf

nestor Handbuch: http://nestor.sub.uni-goettingen.de/handbuch/index.php

nestor-memorandum 2006: nestor „*Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland*" 2006. http://www.langzeitarchivierung.de/downloads/memo2006.pdf

nestor „Memorandum of Understanding" 2007: *Kooperative Entwicklung curricularer Module zur digitalen Langzeitarchivierung im Rahmen des nestor II Arbeitspaketes 5*. http://nestor.sub.uni-goettingen.de/education/mou.pdf

nestor / DPE Spring School 2007: http://nestor.sub.uni-goettingen.de/spring_school_2007/index.php

nestor / DPE Winter School 2007: http://nestor.sub.uni-goettingen.de/winter_school_2007/index.php

nestor / DPE Summer School 2008:
http://nestor.sub.uni-goettingen.de/summer_school_2008/index.php

nestor-Seminare 2006: nestor-Seminare; Göttingen 2006, CD-ROM (ISBN 3-938616-41-5)

Neuroth / Oßwald 2008: Neuroth, Heike; Oßwald, Achim: *Curriculare Innovation im Spezialbereich*: Qualifizierung im Themenbereich „Langzeitarchivierung digitaler Objekte". In: ZfBB (3) 2008, S. 190-197

Oßwald / Scheffel 2006: Oßwald, Achim; Scheffel, Regine: *Lernen und weitergeben – Aus- und Weiterbildungsangebote zur Langzeitarchivierung.* Folien des Vortrags bei 3 Jahre nestor – Abschlussveranstaltung - Frankfurt - 19.6.2006
http://www.langzeitarchivierung.de/downloads/nestor_2006_06_19_osswald_scheffel.pdf

Oßwald / Scheffel 2007: Oßwald, Achim; Scheffel, Regine: *Lernen und Weitergeben – Aus- und Weiterbildungsangebote zur Langszeitarchivierung.* In: nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung - Version 0.1 [Elektronische Ressource] / Hrsg. Heike Neuroth [u.a.] – 2007.
http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_22.pdf

EC 2007: European Commission: *The Lifelong Learning Programme 2007-2013* / European Commission
http://ec.europa.eu/education/programmes/newprog/index_en.html

SAK 2005: *Verfahren und Standards zur Evaluierung und Akkreditierung von Weiterbildenden Studiengängen und Modulen* [Elektronische Ressource] / Hrsg. Ständige Akkreditierungskommission (SAK) - Arbeitsgruppe Weiterbildungsstudiengänge - 12.07.2005
http://www.zeva.org/service/akkred/Weiterbildung.pdf

Wiarda 2008: Wiarda, Jan-Martin: *Exzellenzinitiative light.* In: DIE ZEIT, 31. Januar 2008

Anhang

Herausgeberverzeichnis

Dr. Heike Neuroth - Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) & Max Planck Digital Library (MPDL)

neuroth@mail.sub.uni-goettingen.de

Dr. Heike Neuroth arbeitet seit Februar 2008 als Consultant für eHumanities an der Max Planck Digital Library (MPDL). Sie hat einen Doktortitel der Geologie und arbeitet seit 1997 an der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB). Dort leitet sie die Abteilung Forschung und Entwicklung (RDD). Als Expertin auf dem Gebiet der digitalen Langzeitarchivierung, wissenschaftlichen Forschungsdaten und digitalen Bibliotheksentwicklungen ist sie in diversen nationalen und internationalen Initiativen, Projekten und Arbeitsgruppen involviert. Seit 2004 ist sie darüber hinaus in nationale und internationale Projekte und Aktivitäten eingebunden, die die Entwicklung einer Grid-basierten Forschungsinfrastrutur für unterschiedliche Wissenschaftsdisziplinen vorantreiben. Ihr besonderes Interesse gilt dabei der Entwicklung einer Virtuellen Forschungsumgebung für die eHumanities.

Prof. Dr. Achim Oßwald - Fachhochschule Köln, Institut für Informationswissenschaft

achim.osswald@fh-koeln.de

Dr. rer. soc., Dipl.-Inf.wiss., M.A, geb. 1956, studierte Geschichte und Germanistik in Stuttgart und Freiburg i.Br., sowie Informationswissenschaft in Berlin und Konstanz, arbeitete mehr als 10 Jahre im Bereich Bibliothek, Information und Dokumentation - als Anwender, Vertriebsmitarbeiter eines Softwareanbieters, Dozent und Leiter einer Weiterbildungseinrichtung (Lehrinstitut für Dokumentation, Frankfurt) sowie freiberuflich als Consultant. Seit 1990 Lehraufträge an Fachhochschulen des Archiv-, Bibliotheks- und Dokumentationsbereiches.

Seit 1994 Professor an der Fachhochschule für Bibliotheks- und Dokumentationswesen (FHBD) in Köln, jetzt FH Köln, Fakultät für Informations- und

Kommunikationswissenschaft (Berufungsgebiet: Anwendung der Datenverarbeitung im Informationswesen).

Schwerpunkte seiner Lehre im Institut für Informationswissenschaft: Konzeption und Realisierung IT-basierter bibliothekarischer und dokumentarischer Arbeitsprozesse; Nutzung elektronischer Kommunikationsnetze mit spezieller Ausrichtung auf elektronisch gestützte Informationsdienstleistungen (Mehrwertdienste); Verfahren und Anwendungsbereiche des Digitalen Publizierens und der Elektronischen Dokumentlieferung; Langzeitarchivierung digitaler Objekte sowie kommerzielle und Open Source Software für bibliothekarische Geschäftsprozesse.

Leiter des Zentrums für Bibliothekarische und Informationswissenschaftliche Weiterbildung (ZBIW) der FH Köln sowie Studiengangsbeauftragter des berufsbegleitenden Masterstudiengangs Bibliotheks- und Informationswissenschaft.

Prof. Regine Scheffel M.A. - Hochschule für Technik, Wirtschaft und Kultur Leipzig, Fakultät Medien

scheffel@fbm.htwk-leipzig.de

Nach dem Studium der Romanistik, Germanistik und Volkskunde an der Georg-August-Universität Göttingen und der Weiterbildung zur Wissenschaftlichen Dokumentarin arbeitete Regine Scheffel am Bayerischen Nationalmuseum München. Seit 2000 vertritt sie als Professorin das Lehrgebiet „Computergestützte Informationssysteme in Museen und Bibliotheken“ an der Fakultät Medien der Hochschule für Technik, Wirtschaft und Kultur Leipzig.
Schwerpunkte der Arbeit sind Passgenauigkeit und Nachhaltigkeit beim Einsatz von IT-Systemen in Museen und Bibliotheken, Entwicklung und Verbreitung von Standards (Fachgruppe Dokumentation im Deutschen Museumsbund) sowie Informationsmanagement und Langzeitarchivierung digitaler Objekte (nestor).

Stefan Strathmann - Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)

strathmann@sub.uni-goettingen.de

Stefan Strathmann koordiniert in der Abteilung Forschung und Entwicklung (RDD) die Aktivitäten der Staats- und Universitätsbibliothek Göttingen zur digitalen Langzeitarchivierung. Er ist als wissenschaftlicher Mitarbeiter für die Projekte nestor und DPE tätig und darüber hinaus in eine Reihe von weiteren Projekten und Initiativen zur LZA involviert (Alliance for Permanent Access, IDEA Workshop etc.). Aktuell ist er insbesondere mit Fragen der Aus-, Fort- und Weiterbildung und der LZA von Forschungsdaten befaßt.

Dr. Mathias Jehn - Universitätsbibliothek Johann Christian Senckenberg in Frankfurt am Main

m.jehn@ub.uni-frankfurt.de

Mathias Jehn, geb. 1970, studierte Mittlere und Neuere Geschichte, Kulturanthropologie/Europäische Ethnologie und Ev. Theologie an der Goethe Universität Frankfurt am Main und der Università degli Studi di Firenze. Im Jahre 2002 promovierte er in San Marino über ein rechtshistorisches Thema des späten Mittelalters. Von 2002 bis 2005 war Jehn Archivreferendar im Landesarchiv Nordrhein-Westfalen und schloss die Assessorenprüfung an der Archivschule in Marburg ab. Nach einigen Erschließungsprojekten an den Staatsarchiven Düsseldorf und Marburg übte Jehn von 2006 bis 2008 Projektkoordinator des BMBF geförderten Projektes ‚nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Objekte in Deutschland' in der Deutschen Nationalbibliothek Frankfurt am Main aus. Seit März 2008 ist er an der Universitätsbibliothek Frankfurt am Main Abteilungsleiter für das Archivzentrum und der auf Frankfurt bezogenen Rara-Sammlung.
Schwerpunkte seiner Arbeit sind digitale und analoge Bestandserhaltung von Archivalien, Durchführung von Digitalisierungsprojekten sowie die Umsetzung von Verfahren und Anwendungsbereiche des Digitalen Publizierens.

Autorenverzeichnis

Altenhöner, Reinhard
Deutsche Nationalbibliothek
r.altenhoener@d-nb.de

Aschenbrenner, Andreas
Niedersächsische Staats- und Universitätsbibliothek Göttingen
aschenbrenner@sub.uni-goettingen.de

Becker, Christoph
Technische Universität Wien
becker@ifs.tuwien.ac.at

Bergmeyer, Dr. Winfried
Institut für Museumsforschung
w.bergmeyer@smb.spk-berlin.de

Brandt, Olaf
Behörde der Bundesbeauftragten für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik (BStU)
brandt.lib@gmail.com

Brase, Dr. Jan
Technische Informationsbibliothek Hannover
jan.brase@tib.uni-hannover.de

Brodersen, Maren
Deutsche Nationalbibliothek
m.brodersen@d-nb.de

Brübach, Nils
Sächsisches Staatsarchiv
nils.bruebach@sta.smi.sachsen.de

Dickmann, Frank
Georg-August-Universität Göttingen. Abteilung Medizinische Informatik
fdickmann@med.uni-goettingen.de

Dobratz, Susanne
Humboldt-Universität zu Berlin, Universitätsbibliothek
dobratz@cms.hu-berlin.de

Enders, Markus
The British Library
markus.enders@bl.uk

Funk, Stefan
Niedersächsische Staats- und Universitätsbibliothek Göttingen
funk@sub.uni-goettingen.de

Hackel, Dr. Siegfried, Dir. u. Prof.
Physikalisch-Technische Bundesanstalt
siegfried.hackel@ptb.de

Hänger, Dr. Andrea
Bundesarchiv
a.haenger@barch.bund.de

Heister, Carmen
Technische Universität Wien
heister@ifs.tuwien.ac.at

Hillegeist, Tobias
Akademie der Wissenschaften zu Göttingen
tobias.hillegeist@goettingerakademie.de

Huth, Karsten
Bundesarchiv
k.huth@barch.bund.de

Jehn, Dr. Mathias
Universitätsbibliothek J. C. Frankfurt
m.jehn@ub.uni-frankfurt.de

Keitel, Dr. Christian
Landesarchiv Baden Württemberg
christian.keitel@la-bw.de

Klump, Dr. Jens
Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum – GFZ
jens.klump@gfz-potsdam.de

Kulovits, Hannes
Technische Universität Wien
kulovits@ifs.tuwien.ac.at

Ludwig, Jens
Niedersächsische Staats- und Universitätsbibliothek Göttingen
ludwig@sub.uni-goettingen.de

Moeller-Walsdorf, Tobias
Niedersächsisches Ministerium für Wissenschaft und Kultur
tobias.moeller-walsdorf@mwk.niedersachsen.de

Neubauer, Matthias
Deutsche Nationalbibliothek
m.neubauer@d-nb.de

Neuroth, Dr. Heike
Niedersächsische Staats- und Universitätsbibliothek Göttingen
neuroth@sub.uni-goettingen.de

Oßwald, Prof. Dr. Achim
Fachhochschule Köln
achim.osswald@fh-koeln.de

Queitsch, Manuela
Sächsische Landesbibliothek-, Staats- und Universitätsbibliothek Dresden
queitsch@slub-dresden.de

Rauber, Prof. Dr. Andreas
Technische Universität Wien
rauber@ifs.tuwien.ac.at

Sauter, Prof. Dietrich
FKTG - Fernseh- und Kinotechnische Gesellschaft e.V.
sauter@beenen.de

Schäfer, Tobias
Physikalisch-Technische Bundesanstalt
tobias.schaefer@ptb.de

Scheffel, Prof. Regine
Hochschule für Technik, Wirtschaft und Kultur Leipzig
scheffel@fbm.htwk-leipzig.de

Schoger, Dr. Astrid
Bayerische Staatsbibliothek
astrid.schoger@bsb-muenchen.de

Schöning-Walter, Christa
Deutsche Nationalbibliothek
c.schoening@d-nb.de

Schrimpf, Sabine
Deutsche Nationalbibliothek
s.schrimpf@d-nb.de

Schroeder, Kathrin
Bundesarchiv
k.schroeder@barch.bund.de

Schumann, Natascha
Deutsche Nationalbibliothek
n.schumann@d-nb.de

Schwarz, Dr. Karin
Fachhochschule Potsdam
schwarz@fh-potsdam.de

Schweibenz, Dr. Werner
Bibliotheksservice-Zentrum Baden-Württemberg
werner.schweibenz@bsz-bw.de

Spindler, Prof. Dr. Gerald
Georg-August-Universität Göttingen, Lehrstuhl für Bürgerliches Recht, Handels- und Wirtschaftsrecht, Multimedia- und Telekommunikationsrecht
lehrstuhl.spindler@jura.uni-goettingen.de

Steinke, Tobias
Deutsche Nationalbibliothek
t.steinke@d-nb.de

Strathmann, Stefan
Niedersächsische Staats- und Universitätsbibliothek Göttingen
strathmann@sub.uni-goettingen.de

Suchodoletz, Dirk von
Rechenzentrum der Universität Freiburg / Institut für Informatik
dsuchod@rz.uni-freiburg.de

Ullmann, Angela
Parlamentsarchiv des Deutschen Bundestages
angela.ullmann@bundestag.de

Ullrich, Dagmar
Hochschulrechenzentrum der Universität Kassel
ullrichd@hrz.uni-kassel.de

Upmeier, Dr. Arne
Universitätsbibliothek Technische Universität Ilmenau
arne.upmeier@tu-ilmenau.de

Vlaeminck, Sven
Niedersächsische Staats- und Universitätsbibliothek Göttingen
vlaeminck@sub.uni-goettingen.de

Wiesenmüller, Prof. Heidrun
Hochschule der Medien Stuttgart
wiesenmueller@hdm-stuttgart.de

Wolf, Stefan
Bibliotheksservice-Zentrum Baden-Württemberg
stefan.wolf@bsz-bw.de

Wollschläger, Dr. Thomas
Deutsche Nationalbibliothek
t.wollschlaeger@d-nb.de

Zimmer, Dr. Wolf
CSC Deutschland Solutions GmbH
wzimmer2@csc.com

Akronym- und Abkürzungsverzeichnis

AAD	Access to Archival Databases
ADR	Advanced Digital Recording
AfP	Zeitschrift für Medien- und Kommunikationsrecht
AGLS	Australian Government Locator Service
AHDS	Arts and Humanities Data Service
AIP	Archival Information Package
AIT	Advanced Intelligent Tape
AJAX	Asynchronous JavaScript and XML
AKB	Staatliche Akademie der Bildenden Künste
AKEA	Arbeitskreis Elektronische Archivierung des Verbands der Wirtschaftsarchive
ANSI	American National Standards Institute
AOLA	Austrian On-Line Archive
AP	Arbeitspaket
ARNE	Archivierung von Netzressourcen des Deutschen Bundestages
ASCII	American Standard Code for Information Interchange
ATA	Advanced Technology Attachment
AVC	Advanced Video Coding
AVCHD	Advanced Video Codec High Definition
AVI	Audio Video Interleave
AWV	Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V.
B	Byte
BABS	Bibliothekarisches Archivierungs- und Bereitstellungssystem der Bayerischen Staatsbibliothek
BAM	Bibliotheken, Archive und Museen
BDSG	Bundesdatenschutzgesetz
BGB	Bürgerliches Gesetzbuch
BGBl	Bundesgesetzblatt
BGH	Bundesgerichtshof
bit	binary digit
BL	British Library

BMBF	Bundesministerium für Bildung und Forschung
BMF	Broadcast Metadata Exchange Format
BMWi	Bundesministerium für Wirtschaft und Technologie (BMWi)
BSB	Bayerische Staatsbibliothek
BSI	Bundesamt für Sicherheit in der Informationstechnik
BSZ	Bibliotheksservice-Zentrum Baden-Württemberg
BT-Drs	Bundestagsdrucksache
CAD	Computer Aided Design
CAS	Content Addressed Storage
CC	Creative Commons
CCD	Charge-coupled Device
CCSDS	The Consultative Committee for Space Data Systems
CD	Compact Disc
CD-ROM	Compact Disc Read-Only Memory
CDWA	Categories for the Description of Works of Art
CEN	Comité Européen de Normalisation
CENL	Conference of European National Librarians
CERN	Conseil Européen pour la Recherche Nucléaire
CF	Compact Flash
CIDOC-CRM	International Committee for Documentation - Conceptual Reference Model
CLOCKSS	Controlled Lots of Copies Keep Stuff Save
CMS	Cryptographic Message Syntax
CMS	Content Management System
CNRI	Corporation for National Research Initiatives
Codec	Compression - Decompression
COM	Computer Output on Microfilm/-fiche
CRiB	Conversion and Recommendation of Digital Object Formats
CRIG	Common Repository Interfaces Group
CRL	Center for Research Libraries
DACHS	Digital Archive for Chinese Studies

DAT	Digital Audio Tapes
DC	Dublin Core
DCC	Digital Curation Centre
DCT	Discrete Cosine Transform - Diskrete Kosinustransformation
DDC	Dewey Decimal Classification
DFG	Deutsche Forschungsgemeinschaft
DIAS	Digital Information Archiving System
DIDL	Digital Item Declaration Language
DigCCurr	Digital Curation Curriculum
DigiTool	Digital Asset Management Tool
DIMAG	Digitales Magazin (des Landesarchivs Baden-Württemberg)
DIN	Deutsches Institut für Normung
DINI	Deutsche Initiative für Netzwerkinformation
DIN-NABD	Deutsches Institut für Normung - Normenausschuss Bibliotheks- und Dokumentationswesen
DIP	Dissemination Information Package
DL	Dual-Layer
DLT	Digital Linear Tape
dLZA	Digitale Langzeitarchivierung
DMS	Document Management System
DMSS	Digital-Mass-Storage-Systems
DNB	Deutsche Nationalbibliothek
DNBG	Gesetz über die Deutsche Nationalbibliothek
DNG	Digital Negative
DOAR	Directory of Open Access Repositories
DOI	Digital Object Identifier
DOMEA	Dokumentenmanagement und elektronische Archivierung in der öffentlichen Verwaltung
DOS	Disk Operating System
DPC	Digital Preservation Coalition
DPE	Digital Preservation Europe
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
DRIVER	Digital Repository Infrastructure Vision for European Research
DRM	Digital Rights Management

DROID	Digital Record Object Identification
DTF	Digital Tape Format
DV	Digital Video
DVD	Digital Versatile Disc
DVI	Digital Visual Interface
EBU	European Broadcast Union
ECM	Enterprise-Content-Management
ECTS	European Credit Transfer System
EDV	Elektronische Datenverarbeitung
ELAK	Elektronischen Akt für die Verwaltung
ELAN	eLearning Academic Network Niedersachsen
e-Learning	electronic learning
EPK	Ereignisgesteuerte Prozessketten
EPROMS	Erasable Programmable Read Only Memory
EPS	Encapsulated Postscript
ERPANET	Electronic Resource Preservation and Access Network
EU	Europäische Union
EuGH	Europäische Gerichtshof
FBAS	Farb-Bild-Austast-Synchron(-Signal)
FC	Fibre Channel
FCLA	Florida Center for Library Automation
Fedora	Flexible Extensible Digital Object and Repository Architecture
FEP	Federation of European Publishers
FESAD	Fernseharchivdatenbank
FH	Fachhochschule
FTP	File Transfer Protocol
FUSE	Filesystem in Userspace
GB	Giga-Byte
GDFR	Global Digital Format Registry
GDPdU	Grundsätzen zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen
GenTAufzV	Gentechnikaufzeichnungsverordnung
GEVER	Geschäftsverwaltung
GG	Grundgesetz

GIF	Graphics Interchange Format
GIS	Geoinformationssystem
GNU	GNU is not Unix
GoBs	Grundsätzen ordnungsmäßiger DV-gestützter Buchführung
GOP	Group of Pictures
GPL	General Public License
GRATE	Global Remote Access To Emulation services
GRUR	Deutsche Vereinigung für gewerblichen Rechtsschutz und Urheberrecht
GWDG	Gesellschaft für wissenschaftliche Datenverarbeitung mbH
HD	High Definition
HDTV	High Definition TeleVision
HDV	High Definition Video
HFS	Hierarchical File System
HP	Hewlett Packard
HSM	Hierarchisches Speichermanagement
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
HTW	Hochschule für Technik und Wirtschaft
HTWK	Hochschule für Technik, Wirtschaft und Kultur
HUB	Humboldt-Universität zu Berlin
HUL	Harvard University Library
IBI	Institut für Bibliotheks- und Informationswissenschaft
IBM	International Business Machines Corporation
ICOM	International Council of Museums
ICSU	International Council of Scientific Unions
ID	Identity
IDF	International DOI Foundation
IETF	Internet Engineering Task Force
IFLA	International Federation of Library Associations and Institutions
IGDA	International Game Developers Association

IIPC	International Internet Preservation Coalition
IMAP	Internet Message Access Protocol
IMDAS	Integrated Museum Documentation and Administration Programme
IMX	Interoperability Material Exchange
IP	Internet Protocol
IPR	Intellectual Property Rights
IPX/SPX	Internetworking Packet Exchange/Sequence Packet Exchange
ISBN	International Standard Book Number
ISO	International Organisation for Standardization
IT	Informationstechnik
ITU	International Telecommunications Union
IWAW	International Web Archiving Workshop
J2EE	Java Platform, Enterprise Edition
JHOVE	JSTOR/Harvard Object Validation Environment
JISC	Joint Information Systems Committee
JPEG	Joint Photographic Experts Group
KB	Koninklijke Bibliotheek (Königliche Bibliothek der Niederlande)
KEEP	Keep Emulator Environments Portable
KMK	Kultusministerkonferenz
koLibRI	kopal Library for Retrieval and Ingest
kopal	Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen
KOST	Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen
LIFE	Lifecycle Information For E-Literature
LMER	Langzeitarchivierungsmetadaten für elektronische Ressourcen
LoC	Library of Congress
LOCKSS	Lots of Copies Keep Stuff Save
LRZ	Leibniz-Rechenzentrum
LTO	Linear Tape Open
LZA	Langzeitarchivierung

MA	Master
Mac OS	Macintosh Operating System
MAID	Massive Array of Idle Disks
MAME	Multiple Arcade Machine Emulator
MARC	Machine-Readable Catalog
MAZ	Magnetische AufZeichnung
MB	Megabyte
Mbit/s	Megabit pro Sekunde
MD5	Message-Digest algorithm 5
MDT	Medium Decay Time
MDZ	Münchner Digitalisierungszentrums
MEL	Medium Expected Lifetime
MESS	Multiple Emulator Super System
METS	Metadata Encoding and Transmission Standard
MIME	Multipurpose Internet Mail Extensions
MIT	Massachusetts Institute of Technology
MIX	Metadata for Images in XML
MODS	Metadata Object Description Schema
MoReq	Model Requirements for the Management of Electronic Documents and Records
MPEG	Moving Pictures Experts Group
MTBF	Mean Time Between Failures
MXF	Material eXchange Format
NARA	National Archives and Records Administration
NBN	National Bibliography Number
NDAD	National Digital Archive of Datasets
NDSG	Niedersächsisches Datenschutzgesetz
Nedlib	Networked European Deposit Library
nestor	Network of Expertise in long-term STOrage and availability of digital Resources in Germany
NetBIOS	Network Basic Input Output System
NISO	National Information Standards Organization
NLE	Non-Linear Editing
NOARK	Northwest Arkansas Human Resource Association

NoE	Network of Excellence
NTSC	National Television Systems Committee
NutchWAX	Nutch Web Archive eXtensions
OAI-ORE	Open Archives Initiative Protocol - Object Exchange and Reuse
OAIS	Open Archival Information System
OAK	Open Access to Knowledge
OCLC	Online Computer Library Center
ODF	Open Document Format
OECD	Organisation for Economic Co-operation and Development
OGF	Open Grid Forums
OLG	Oberlandesgericht
OMG	Object Management Group
ONIX	Online Information eXchange
OP	Operational Patterns
OSCI	Online Services Computer Interface
OSI	Open Society Institute
p	Pixel
PAL	Phase Alternating Line
PANDORA	Preserving and Accessing Networked Documentary Resources in Australia
PC	Personal Computer
PCB	Plychloriertes Biphenyl
PCM	Puls-Code-Modulation
PCMCIA	Personal Computer Memory Card International Association
PD	Professional Disc
PDF	Portable Document Format
PDF/A	Portable Document Format Level A
PDI	Preservation Description Information
PDP	Programmed Data Processor
PflAV	Pflichtablieferungsverordnung
PI	Persistent Identifier
PICA	Project of Integrated Catalogue Automation
PIN	Personal Identification Numbers

PKCS	Public Key Cryptography Standard
planets	Preservation and Long-term Access through Networked Services
Plato	Planets Preservation Planning Tool
PNG	Portable Network Graphics
POP3	Post Office Protocol Version 3
PREMIS	Preservation Metadata: Implementation Strategies
PRONOM	On-line information system about data file formats and their supporting software products
PUID	PRONOM Persistent Unique Identifier
Q.E.D	Qualitätsinitiative E-Learning in Deutschland
QAM	Quadratur Amplituden Modulation
RAID	Redundant Array of Independent Disks
RegBib	Regional Bibliothek
RFC	Requests for Comment
RGB	Rot, Grün, Blau
RLG	Research Libraries Group
RöntgV	Röntgenverordnung
RSWK	Regeln für den Schlagwortkatalog
RTF	Rich Text Format
S/MIME	Secure / Multipurpose Internet Mail Extensions
SAK	Ständige Akkreditierungskommission
SAM	Standard-Archivierungs-Moduls
SAML	Security Assertion Markup Language
SAP	SAP Business Suite
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment
SC	Science Commons
SCORM	Sharable Content Object Reference Model
SCSI	Small Computer System Interface
SD	Standard Definition
SDHC	Secure Digital High Capacity
SDLT	Super Digital Linear Tape

SDTV	Standard Definition Television
SFTP	Simple File Transfer Protocol
SGML	Standard Generalized Markup Language
SGS	Staatsgalerie Stuttgart
SHAMAN	Sustaining Heritage Access through Multivalent ArchiviNg
SIG	Special Interest Group
SigG	Signaturgesetz
SigV	Signaturverordnung
SILS	School of Information and Library Science
SIP	Submission Information Package
SMB	Server Message Block Protocol
SMIL	Synchronized Multimedia Integration Language
SMPTE	Society of Motion Picture and Television Engineers
SMTP	Simple Mail Transfer Protocol
SOA	Serviceorientierte Architektur
SOAP	Simple Object Access Protocol
SPS	Software Preservation Society
SRU	Search/Retrieve via URL
SRW	Search/Retrieve for the Web
StrlSchV	Strahlenschutzverordnung
Stud.IP	Studienbegleitender Internetsupport von Präsenzlehre
SUB	Niedersächsische Staats- und Universitätsbibliothek
TB	Terabyte
TCO	Total Cost of Ownership
TCP/IP	Transmission Control Protocol/Internet Protocol
TIB	Technische Informationsbibliothek
TIFF	Tagged Image File Format
TMG	Telemediengesetz
TRAC	Trustworthy Repositories Audit & Certification
UML	Unified Modelling Language
UOF	Universelles Objektformat

UrhG	Urheberrechtsgesetz
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
USB	Universal Serial Bus
UTF	Unicode Transformation Format
UV	Ultraviolett
UVC	Universal Virtual Computer
VFAT	Virtual File Allocation Table
VLB	Verzeichnis Lieferbarer Bücher
W3C	World Wide Web Consortium
WARC	Web ARChive file format
WDC	World Data Center
WORM	Write Once Read Many
WWW	World Wide Web
XCDL	Extensible Characterisation Definition Language
XENA	XML Electronic Normalising of Archives
XML	Extensible Markup Language
XMLDSig	XML Signatur Spezifikation
XMP	Extensible Metadata Platform
XSLT	Extensible Stylesheet Language Transformation