H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 17

Vorgehensweise für ausgewählte Objekttypen

nestor

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth
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

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.

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

Die digitale Version 2.3 steht unter folgender Creative-Commons-Lizenz:
„Namensnennung-Keine kommerzielle Nutzung-Weitergabe unter gleichen Bedingungen 3.0 Deutschland“
http://creativecommons.org/licenses/by-nc-sa/3.0/de/

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.

URL für Kapitel 17 „Vorgehensweise für ausgewählte Objekttypen“ (Version 2.3):
urn:nbn:de:0008-20100624163
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100624163

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

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 Veranwortungsbereich 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
Alle hier aufgeführten URLs wurden im Mai 2010 auf Erreichbarkeit geprüft.

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://igitur-archive.library.uu.nl/DARLIN/2005-0526-201749/VandeSompelDLib2003UsingMPEG.htm

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, File-sharing 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. Das Zielformat muss progressiv sein, da die Darstellung auf dem Endgerät auch progressiv erfolgt. Der Zeilensprung hat ausgedient (interlaced format) (Messerschmid 2006). Auch die Langzeitarchivierung muss in einem progressiven Format erfolgen, da die De-Interlacer (Hard- und Software) alle nicht zufrieden stellend arbeiten. Die Auflösung ist bei Bewegtbildern nur halb so groß.

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 Ar-

chive 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 in-

dividuellen 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),

- MPEG4/H264
- 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, Wür-

mern etc., siehe Oermann, A. (2007).

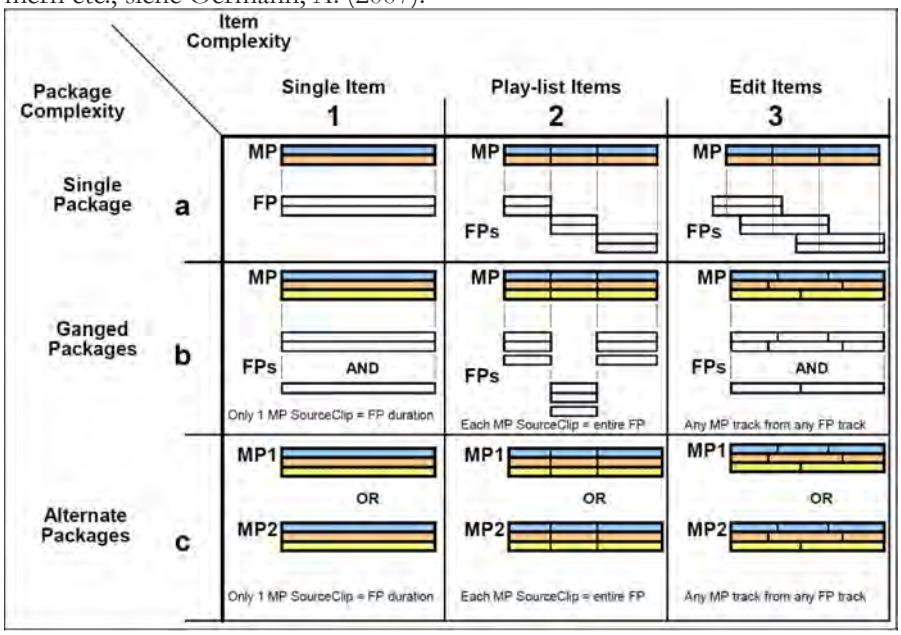


Abbildung 2: Achsen der operational pattems 1-c, 2a-c und 3a-c (FP = file package)(Sohst, Lennart (2006))

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.

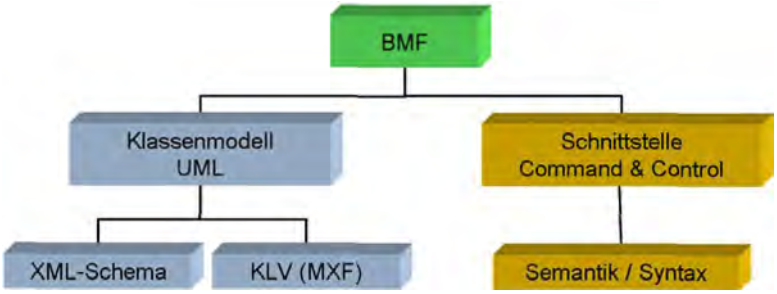


Abbildung. 3: Bestandteile des Broadcast Metadata Exchange Format – BMF

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

Für die hochwertigen Produktionen sind keine Festlegungen getroffen worden, hier werden 10 bit Auflösung und Datenraten > 200 Mbit/sec erwartet.

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 und dass keine Reduzierung der Auflösung z.B. 8-bit anstelle von 10-bit erfolgt. *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

Messerschmid, Ulrich (2006) Abschied vom Zeilensprung. In: Fernseh- und Kinotechnik FKT 11/2006

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
- DVCPRO50/100 DV-based-Kompression mit 50/100 Mbit/s (SMPTE 314M)
- IMX MPEG-2 422P@ML Kompression (EBU D94 und SMPTE 356M)

- HDCAM SR MPEG4 SMPTE 409M
- ProRes Apple eigene Kompression
- AVCIntra SMPTE RP 2027
- DNxHD SMPTE AVC-3
- 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 (50 Mbit/sec) 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 (SMPTE VC3).

ProRes

Das Apple eigene Kompressionsverfahren arbeitet nach dem Waveletverfahren und hat eine Datenrate bis zu 220 Mbit/sec. Sowohl 8 bit als auch 10 bit werden unterstützt. Das Abtast-verhältnis ist 4:2:2.

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 Maximalkapazität 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 ein-

zelne 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/dps/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]

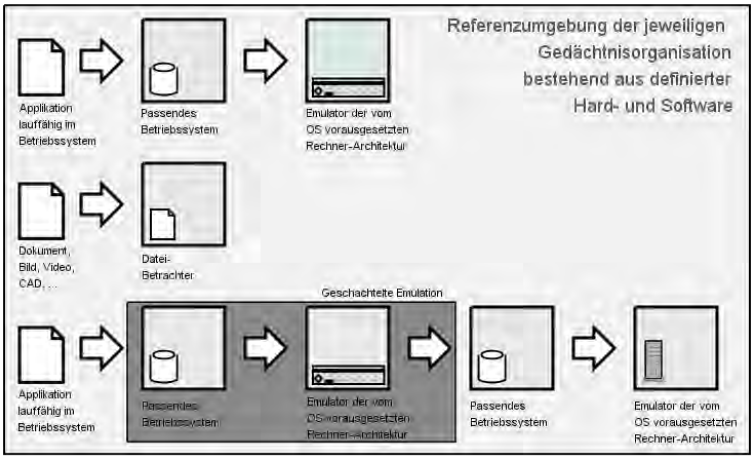


Abbildung 1: Je nach Art des Primärobjekts sind unterschiedliche Schritte zu seiner Darstellung oder Ausführung notwendig.

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.

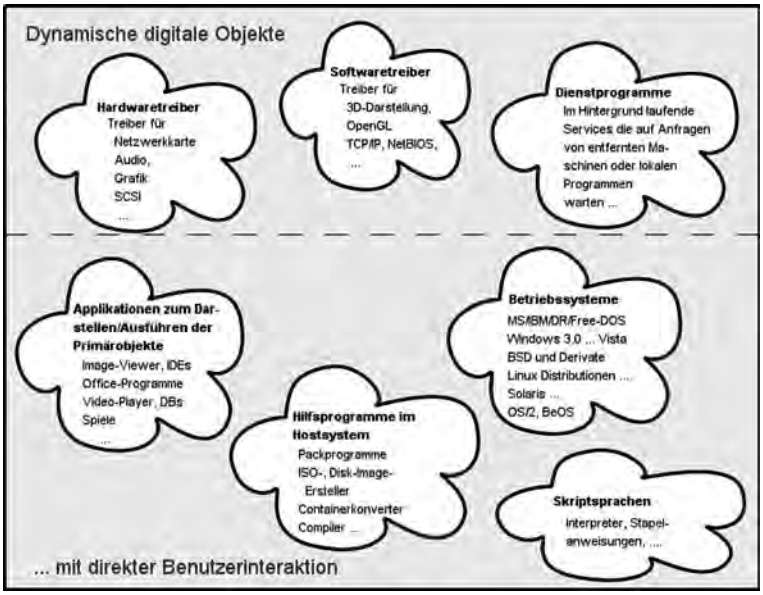


Abbildung 2: Die Klasse der dynamischen digitalen Objekte umfasst die interaktiven Objekte, die für eine direkte Benutzerinteraktion erstellt wurden. Ihre Einordnung muss nicht immer eindeutig ausfallen.

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.

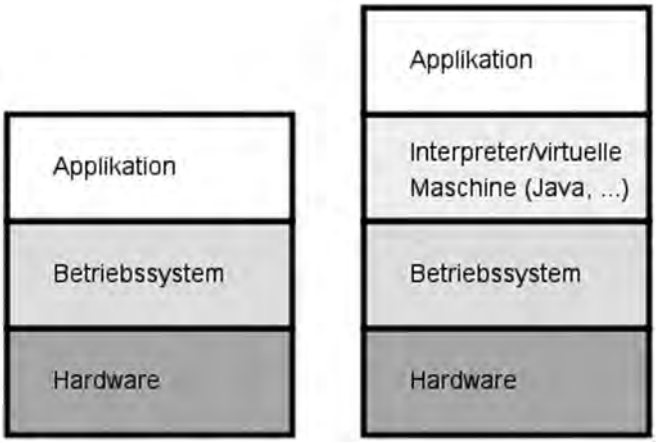


Abbildung 3: Eine typische Ablaufumgebung neuerer Rechnerarchitekturen für digitale Objekte bestehend aus Hard- und Software als Schichtenmodell. Interpreter und abstrakte Programmiersprachen wie Java können eine weitere Schicht einführen (rechts).

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.

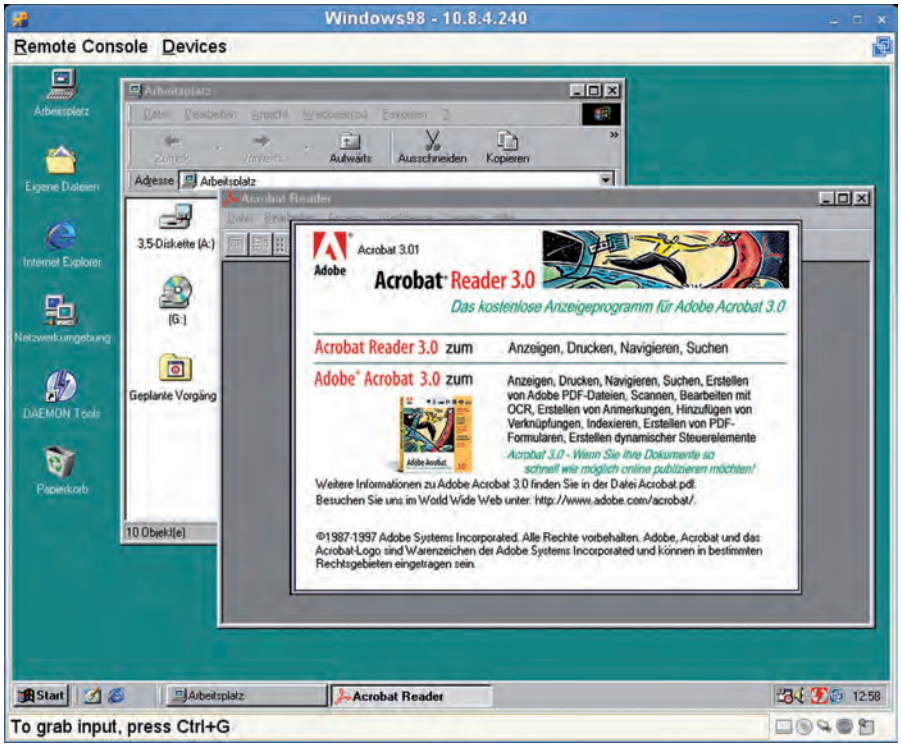


Abbildung 4: Der VMware-Server 2 ist ein X86 Virtualisierer, der einen Zugriff auf alte Nutzungsumgebungen, wie Windows98 über das Netz erlaubt.

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.

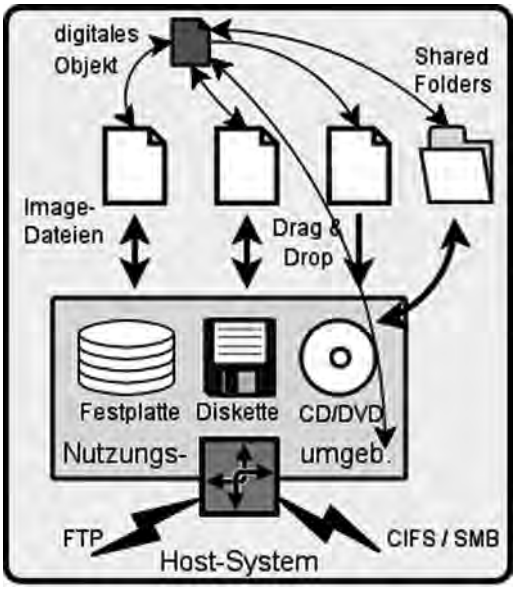


Abbildung 5: Digitale Objekte können auf verschiedene Weise zwischen der im Emulator laufenden Nutzungsumgebung und dem Host-System ausgetauscht werden.

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.

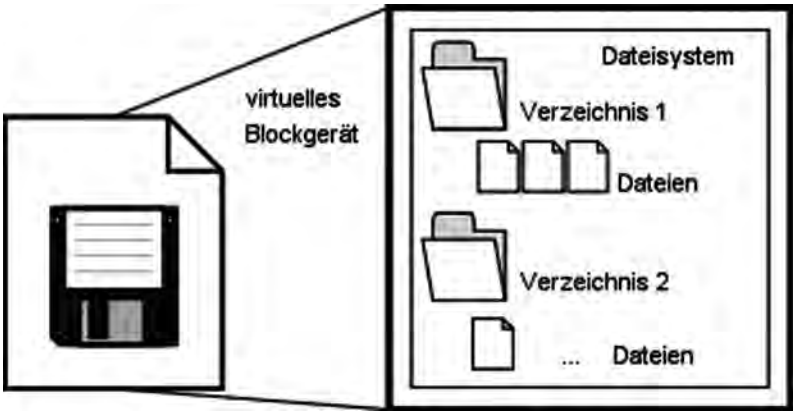


Abbildung 6: Der logische Aufbau einer Datei, die ein Disketten-Image repräsentiert. Die innere Blockstruktur kann mit einem Dateisystem versehen sein, welches Verzeichnisse und Dateien enthält.

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.

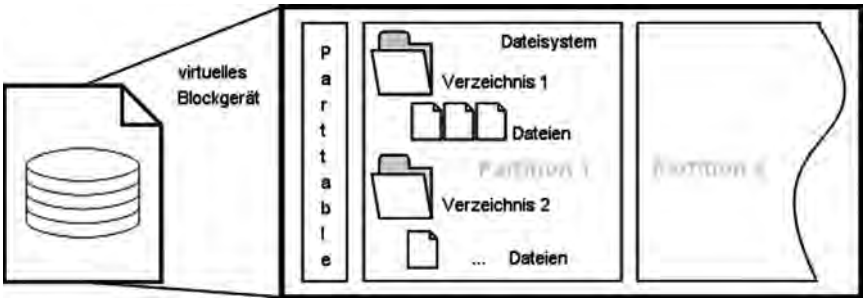


Abbildung 7: Der logische Aufbau einer Datei, die ein Festplatten-Image enthält. Dieses kann durch eine Partitionstabelle und mehrere Partitionen mit eventuell unterschiedlichen Dateisystemen strukturiert sein.

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].

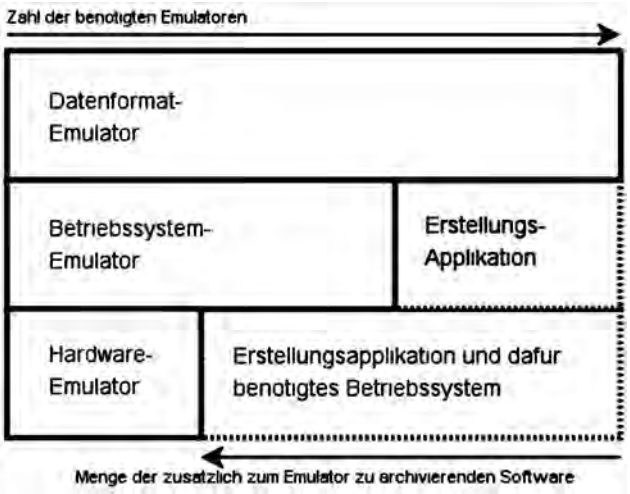


Abbildung 8: Je nach Ansatzpunkt der Emulation können eine Reihe zusätzlicher Sekundärobjekte benötigt werden.

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.

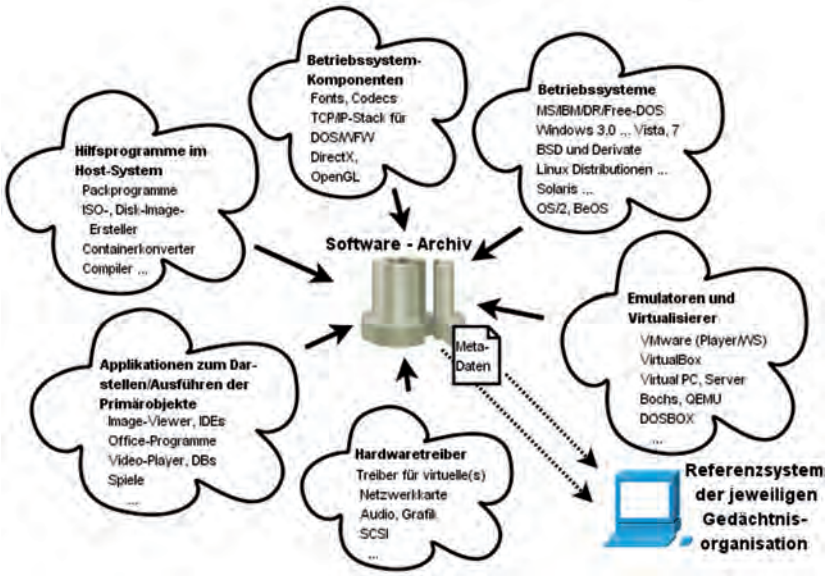


Abbildung 9: Auswahl der in einem Softwarearchiv für die X86-Architektur zu berücksichtigenden Sekundärobjekte.

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 Webarchivierung 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 18.4), 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 info-deposit.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 dieses 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"

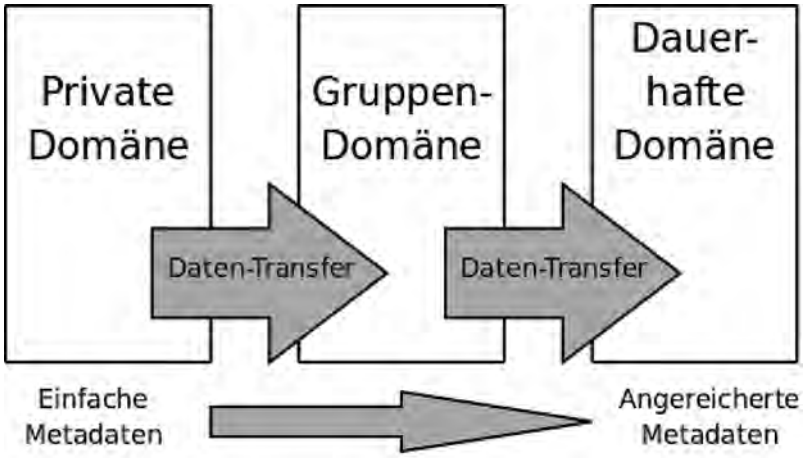
94 DFG (1998), Empfehlung 7
95 DFG (1998), Empfehlung 1
96 Kluttig (2008)
97 DFG (2009)

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://oa.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.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.

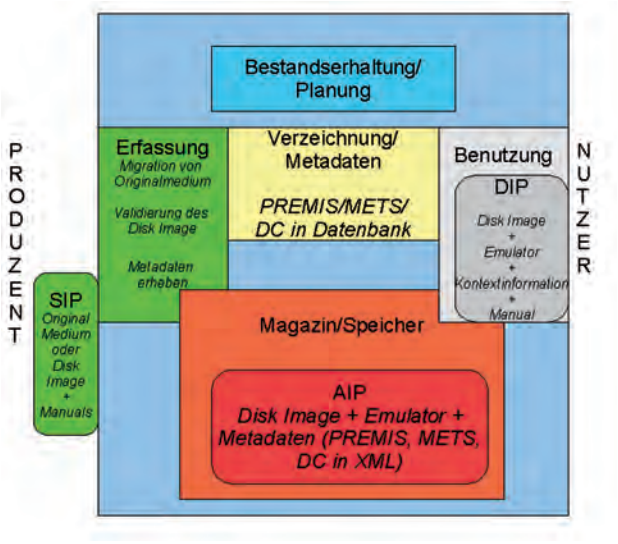


Abbildung 1: OAIS-Funktionsmodell

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.

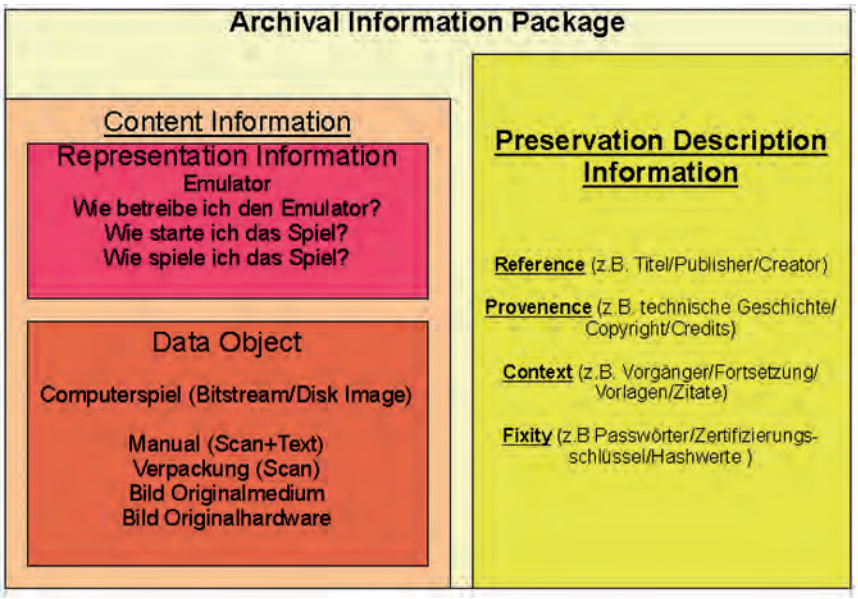


Abbildung 2: AIP für ein Computerspiel [148]

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.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. Daneben sind für Unternehmen und Verwaltungen auch andere elektronische Dokumente aufbewahrungswürdig, bspw. Bilder zu Dokumen-

159 Daher müssen die für die verschiedenen Formate verwendeten MIME-Typen beim Mailsystem von Sender und Empfänger bekannt sein.
Alle hier aufgeführten URLs wurden im Mai 2010 auf Erreichbarkeit geprüft.

160 National Archives and Records Administration (Hrsg.) (2004), S.3. Genauer heißt es "Records...which have minimal or no documentary or evidential value."

tationszwecken 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. 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]

---

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.

163 National Archives and Records Administration Regulation on E-mail (2006) §1234.24 Standards for managing electronic mail records.

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.

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.

---

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.

167 Darunter fallen beispielsweise PDF-Dokumente nicht. Eine Liste möglicher Formate ist abgedruckt in: Brand, Thorsten/ Groß, Stefan/ Zöller, Bernhard (2003), S.20.

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. 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,

---

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).

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.

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 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

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.

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.

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.

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

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.

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 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.

---

177 Boudrez, Filip (2006). S.14.

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.

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.

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

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.

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.

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]
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

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).

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

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

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.

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://www.isb.admin.ch/themen/architektur/00078/00201/00202/index.html

(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