nestor
Kompetenznetzwerk Langzeitarchivierung

nestor Handbuch:

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

1 Einführung

Herausgeber

Heike Neuroth
Hans Liegmann †
Achim Oßwald
Regine Scheffel
Mathias Jehn
Stefan Strathmann

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Im Auftrag von

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
Tel. +49 (0) 55 1 39 38 66

1 Einführung

Hans Liegmann, Heike Neuroth

1. Die digitale Welt, eine ständig wachsende Herausforderung

Die Überlieferung des kulturellen Erbes, traditionell eine der Aufgaben von Bibliotheken, Archiven und Museen, ist durch die Informationstechnologien deutlich schwieriger geworden.

In der heutigen Zeit werden zunehmend mehr Informationen digital erstellt und veröffentlicht. Diese digitalen Informationen, die Güter des Informations- und Wissenszeitalterzeitalter, sind einerseits wertvolle kulturelle und wissenschaftliche Ressourcen, andererseits sind sie sehr vergänglich. Die Datenträger sind ebenso der Alterung unterworfen, wie die Datenformate oder die zur Darstellung notwendige Hard- und Software. Um langfristig die Nutzbarkeit der digitalen Güter sicherzustellen, muss schon frühzeitig Vorsorge getroffen werden, müssen Strategien der digitalen Langzeitarchivierung entwickelt und umgesetzt werden.

Die Menge und die Heterogenität der Informationen, die originär in digitaler Form vorliegen, wächst beständig an.

In großem Umfang werden ursprünglich analog vorliegende Daten digitalisiert (z.B. Google Print Projekt[1]), um den Benutzerzugriff über Datennetze zu ver-

1 http://print.google.com

einfachen. Im Tagesgeschäft von Behörden, Institutionen und Unternehmen werden digitale Akten produziert, für die kein analoges Äquivalent mehr zur Verfügung steht.
Sowohl die wissenschaftliche Fachkommunikation wie der alltägliche Informationsaustausch sind ohne die Vermittlung von Daten in digitaler Form nicht mehr vorstellbar.
Mit der Menge der ausschließlich digital vorliegenden Information wächst unmittelbar auch ihre Relevanz als Bestandteil unserer kulturellen und wissenschaftlichen Überlieferung sowie die Bedeutung ihrer dauerhaften Verfügbarkeit für Wissenschaft und Forschung. Denn das in der „scientific community" erarbeitete Wissen muss, soll es der Forschung dienen, langfristig verfügbar gehalten werden, da der Wissenschaftsprozess immer wieder eine Neubewertung langfristig archivierter Fakten erforderlich macht. Die Langzeitarchivierung digitaler Ressourcen ist daher eine wesentliche Bedingung für die Konkurrenzfähigkeit des Bildungs- und Wissenschaftssystems und der Wirtschaft. In Deutschland existiert eine Reihe von Institutionen (Archive, Bibliotheken, Museen), die sich in einer dezentralen und arbeitsteiligen Struktur dieser Aufgabe widmen.
Im Hinblick auf die heutige Situation, in der Autoren und wissenschaftliche Institutionen (Universitäten, Forschungsinstitute, Akademien) mehr und mehr selbst die Veröffentlichung und Verbreitung von digitalen Publikationen übernehmen, erscheint auch weiterhin ein verteilter Ansatz angemessen, der jedoch um neue Verantwortliche, die an der „neuen" Publikationskette beteiligt sind, erweitert werden muss.

1.1. Langzeitarchivierung im digitalen Kontext

„Langzeitarchivierung" meint in diesem Zusammenhang mehr als die Erfüllung gesetzlicher Vorgaben über Zeitspannen, während der steuerlich relevante tabellarisch strukturierte Daten verfügbar gehalten werden müssen. „Langzeit" ist die Umschreibung eines nicht näher fixierten Zeitraumes, währenddessen wesentliche, nicht vorhersehbare technologische und soziokulturelle Veränderungen eintreten; Veränderungen, die sowohl die Gestalt als auch die Nutzungssituation digitaler Ressourcen in rasanten Entwicklungszyklen vollständig umwälzen können. Es gilt also, jeweils geeignete Strategien für bestimmte digitale Sammlungen zu entwickeln, die je nach Bedarf und zukünftigem Nutzungsszenarium die langfristige Verfügbarkeit der digitalen Objekte sicherstellen. Dabei spielen nach bisheriger Erfahrung das Nutzerinteresse der Auf- und Abwärtskompatibilität alter und neuer Systemumgebungen nur dann eine Rolle, wenn dies dem Anbieter für die Positionierung am Markt erforderlich erscheint.

„Langzeit“ bedeutet für die Bestandserhaltung digitaler Ressourcen nicht die Abgabe einer Garantieerklärung über fünf oder fünfzig Jahre, sondern die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können.
Der Bedeutungsinhalt von „Archivierung“ müsste hier nicht näher präzisiert werden, wäre er nicht im allgemeinen Sprachgebrauch mit der fortschreitenden Anwendung der Informationstechnik seines Sinnes nahezu entleert worden. „Archivieren“ bedeutet zumindest für Archive, Museen und Bibliotheken mehr als nur die dauerhafte Speicherung digitaler Informationen auf einem Datenträger. Vielmehr schließt es die Erhaltung der dauerhaften Verfügbarkeit digitaler Ressourcen mit ein.

2. Substanzerhaltung

Eines von zwei Teilzielen eines Bestandserhaltungskonzeptes für digitale Ressourcen ist die unversehrte und unverfälschte Bewahrung des digitalen Datenstroms: die Substanzerhaltung der Dateninhalte, aus denen digitale Objekte physikalisch bestehen. Erfolgreich ist dieses Teilziel dann, wenn die aus heterogenen Quellen stammenden und auf unterschiedlichsten Trägern vorliegenden Objekte möglichst früh von ihren originalen Träger getrennt und in ein homogenes Speichersystem überführt werden. Die verantwortliche archivierende Institution wird vorzugsweise ein funktional autonomes Teilsystem einrichten, dessen vorrangige Aufgabe die Substanzerhaltung digitaler Ressourcen ist. Wichtige Bestandteile dieses Systems sind automatisierte Kontrollmechanismen, die den kontinuierlichen systeminternen Datentransfer überwachen. Die kurze Halbwertszeit technischer Plattformen macht auch vor diesem System nicht halt und zwingt zum laufenden Wechsel von Datenträgergenerationen und der damit möglicherweise verbundenen Migration der Datenbestände.
Dauerhafte Substanzerhaltung ist nicht möglich, wenn die Datensubstanz untrennbar an einen Datenträger und damit an dessen Schicksal gebunden ist. Technische Maßnahmen zum Schutz der Verwertungsrechte (z.B. Kopierschutzverfahren) führen typischerweise mittelfristig solche Konfliktsituationen herbei. Ein digitales Archiv wird in Zukunft im eigenen Interesse Verantwortung nur für solche digitalen Ressourcen übernehmen, deren Datensubstanz es voraussichtlich erhalten kann. Ein objektspezifischer „Archivierungsstatus“ ist in dieser Situation zur Herstellung von Transparenz hilfreich.

3. Erhaltung der Benutzbarkeit

Substanzerhaltung ist nur eine der Voraussetzungen, um die Verfügbarkeit und Benutzbarkeit digitaler Ressourcen in Zukunft zu gewährleisten. „Erhaltung der Benutzbarkeit" digitaler Ressourcen ist eine um ein Vielfaches komplexere Aufgabenstellung als die Erhaltung der Datensubstanz. Folgen wir dem Szenario eines „Depotsystems für digitale Objekte", in dem Datenströme sicher gespeichert und über die Veränderungen der technischen Umgebung hinweg aufbewahrt werden, so steht der Benutzer/die Benutzerin der Zukunft gleichwohl vor einem Problem. Er oder sie ist ohne weitere Unterstützung nicht in der Lage den archivierten Datenstrom zu interpretieren, da die erforderlichen technischen Nutzungsumgebungen (Betriebssysteme, Anwendungsprogramme) längst nicht mehr verfügbar sind. Zur Lösung dieses Problems werden unterschiedliche Strategien diskutiert, prototypisch implementiert und erprobt.
Der Ansatz, Systemumgebungen in Hard- und Software-Museen zu konservieren und ständig verfügbar zu halten, wird nicht ernsthaft verfolgt. Dagegen ist die Anwendung von Migrationsverfahren bereits für die Substanzerhaltung digitaler Daten erprobt, wenn es um einfachere Datenstrukturen oder den Generationswechsel von Datenträgertypen geht. Komplexe digitale Objekte entziehen sich jedoch der Migrationsstrategie, da der für viele Einzelfälle zu erbringende Aufwand unkalkulierbar ist. Aus diesem Grund wird mit Verfahren experimentiert, deren Ziel es ist, Systemumgebungen lauffähig nachzubilden (Emulation). Es werden mehrere Ansätze verfolgt, unter denen die Anwendung formalisierter Beschreibungen von Objektstrukturen und Präsentationsumgebungen eine besondere Rolle einnimmt.
Die bisher genannten Ansätze spielen durchgängig erst zu einem späten Zeitpunkt eine Rolle, zu dem das digitale Objekt mit seinen für die Belange der Langzeitarchivierung günstigen oder weniger günstigen Eigenschaften bereits fertig gestellt ist. Darüber hinaus wirken einige wichtige Initiativen darauf hin, bereits im Entstehungsprozess digitaler Objekte die Verwendung langzeitstabiler Datenformate und offener Standards zu fördern. Welche der genannten Strategien auch angewandt wird, die Erhaltung der Benutzbarkeit und damit der Interpretierbarkeit wird nicht unbedingt mit der Erhaltung der ursprünglichen Ausprägung des „originalen" Objektes korrespondieren. Es wird erforderlich sein, die Bemühungen auf die Kernfunktionen (so genannte „significant properties") digitaler Objekte zu konzentrieren, vordringlich auf das, was ihren wesentlichen Informationsgehalt ausmacht.

4. Technische Metadaten

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

5. Vertrauenswürdige digitale Archive

Digitale Archive stehen erst am Beginn der Entwicklung, während Archive für traditionelles Schriftgut über Jahrhunderte hinweg Vertrauen in den Umfang und die Qualität ihrer Aufgabenwahrnehmung schaffen konnten. Es werden deshalb Anstrengungen unternommen, allgemein akzeptierte Leistungskriterien für vertrauenswürdige digitale Archive aufzustellen (vgl. Kap. 8), die bis zur Entwicklung eines Zertifizierungsverfahrens reichen. Die Konformität zum OAIS-Referenzmodell spielt dabei ebenso eine wichtige Rolle, wie die Beständigkeit der institutionellen Struktur, von der das Archiv betrieben wird. Es wird erwartet, dass Arbeitsmethoden und Leistungen der Öffentlichkeit präsentiert werden, sodass aus dem möglichen Vergleich zwischen inhaltlichem Auftrag und tatsächlicher Ausführung eine Vertrauensbasis sowohl aus Nutzersicht, wie auch im Interesse eines arbeitsteiligen kooperativen Systems, entstehen kann.
Wichtig in diesem Zusammenhang ist auch die Wahrung der Integrität und Authentizität eines digitalen Objektes. Nur wenn sichergestellt werden kann, dass

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

das digitale Objekt zum Beispiel inhaltlich nicht verändert wurde, kann man mit der Ressource vertrauensvoll arbeiten.

6. Verteilte Verantwortung bei der Langzeitarchivierung digitaler Ressourcen

6.1 National

Hinsichtlich der Überlegungen zur Langzeitarchivierung digitaler Quellen in Deutschland muss das Ziel sein, eine Kooperationsstruktur zu entwickeln, die entsprechend den Strukturen im analogen Bereich die Bewahrung und Verfügbarkeit aller digitalen Ressourcen gewährleistet. Diese Strukturen berücksichtigen alle Ressourcen, die in Deutschland, in deutscher Sprache oder über Deutschland erschienen sind, die Bewahrung und Verfügbarkeit der wichtigsten Objekte jedes Fachgebiets organisiert (unabhängig davon, ob es sich um Texte, Fakten, Bilder, Multimedia handelt) sowie die Bewahrung und Verfügbarkeit digitaler Archivalien garantiert.
Das Auffinden der Materialien soll dem interessierten Nutzer ohne besondere Detailkenntnisse möglich sein, d.h. ein weiteres Ziel einer angestrebten Kooperationsstruktur beinhaltet, die Verfügbarkeit durch Zugangsportale zu sicher zu stellen und die Nutzer dorthin zu lenken, wo die Materialien liegen. Dabei müssen selbstverständlich Zugriffsrechte, Kosten u.a. durch entsprechende Mechanismen (z.B. Bezahlsysteme) berücksichtigt werden.
Beim Aufbau einer solchen Struktur sind vor allem die Bibliotheken, Archive und Museen gefordert. In Deutschland müssen in ein entstehendes Kompetenznetzwerk Langzeitarchivierung aber auch die Produzenten digitaler Ressourcen, d. h. Verlage, Universitäten, Forschungseinrichtungen, Wissenschaftler sowie technische Dienstleister wie Rechen-, Daten- und Medienzentren und Großdatenbankbetreiber einbezogen werden.

6.2 Internationale Beispiele

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

dokumentiert und weiterverbreitet. Die Teilnahme an der DPC ist über verschiedene Formen der Mitgliedschaft möglich.
In den USA ist im Jahr 2000 ein Programm zum Aufbau einer nationalen digitalen Informationsinfrastruktur und ein Programm für die Langzeitverfügbarkeit digitaler Ressourcen in der Library of Congress (LoC) verabschiedet worden. Die Aufgaben werden in Kooperation mit Vertretern aus anderen Bibliotheken und der Forschung sowie kommerziellen Einrichtungen gelöst. Darüber hinaus hat die LoC in Folge ihrer Jubiläumskonferenz im Jahre 2000 einen Aktionsplan aufgestellt, um Strategien zum Management von Netzpublikationen durch Bibliothekskataloge und Metadatenanwendungen zu entwickeln. Der Ansatz einer koordinierten nationalen Infrastruktur, auch unter den Rahmenbedingungen einer äußerst leistungsfähigen Nationalbibliothek wie der LoC, bestätigt die allgemeine Einschätzung, dass zentralistische Lösungsansätze den künftigen Aufgaben nicht gerecht werden können.
Im Archivbereich wird die Frage der Langzeitverfügbarkeit digitaler Archivalien in internationalen Projekten angegangen. Besonders zu erwähnen ist das Projekt ERPANET, das ebenfalls den Aufbau eines Kompetenznetzwerks mittels einer Kooperationsplattform zum Ziel hat. InterPares ist ein weiteres internationales Archivprojekt, welches sich mit konkreten Strategien und Verfahren der Langzeitverfügbarkeit digitaler Archivalien befasst. Die Zielsetzung der Projekte aus dem Archivbereich verdeutlichen, wie ähnlich die Herausforderungen der digitalen Welt für alle Informationsanbieter und Bewahrer des kulturellen Erbes sind und lassen Synergieeffekte erwarten.
Ein umfassender Aufgabenbereich von Museen ist das fotografische Dokumentieren und Verfahren von Referenzbildern für Museumsobjekte. Die Sicherung der Langzeitverfügbarkeit der digitalen Bilder ist eine essentielle Aufgabe aller Museen. Im Bereich des Museumswesens muss der Aufbau von Arbeitsstrukturen, die über einzelne Häuser hinausreichen, jedoch erst noch nachhaltig aufgebaut werden.

7. Rechtsfragen

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

- Rechte zur Durchführung notwendiger Eingriffe in die Gestalt der elektronischen Ressourcen im Interesse der Langzeiterhaltung,
- Einschränkungen durch Digital Rights Management Systeme (z. B. Kopierschutz),
- Konditionen des Zugriffs auf die archivierten Ressourcen und deren

Nutzung.

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

8. Wissenschaftliche Forschungsdaten

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

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

- DFG, Empfehlung 7[3]
- OECD[4]
- Und ganz aktuell die EU[5] mit folgendem Zitat: „Die Europäische Kommission hat am 10. April 2008 die ‚Empfehlungen zum Umgang mit geistigem Eigentum bei Wissenstransfertätigkeiten und für einen Praxiskodex für Hochschulen und andere öffentliche Forschungseinrichtungen' herausgegeben. Zu diesem Thema war bereits im ersten Halbjahr 2007 unter der deutschen Ratspräsidentschaft ein Eckpunktepapier mit dem Titel ‚Initiative zu einer Charta zum Umgang mit geistigem Eigentum an öffentlichen Forschungseinrichtungen und Hochschulen' ausgearbeitet worden."

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

Interessante Fragen in diesem Zusammenhang sind zum Beispiel, wem die Forschungsdaten eigentlich gehören (dem Wissenschaftler, der Hochschule, der

3 http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

4 http://www.oecd.org/dataoecd/9/61/38500813.pdf

5 http://ec.europa.eu/invest-in-research/pdf/ip_recommendation_de.pdf

Öffentlichkeit), was Forschungsdaten eigentlich sind - hier gibt es bestimmt fachspezifische Unterschiede, welche Forschungsdaten langfristig aufbewahrt werden müssen - eine fachliche Selektion kann nur in enger Kooperation mit dem Wissenschaftler erfolgen, und wer für die Beschreibungen z.B. die Lieferung von technischen und deskriptiven Metadaten zuständig ist.