Probleme und Strategien der Langzeitarchivierung multimedialer Objekte

Dipl.-Inf. Jens-Martin Loebel

Institut für Informatik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
jens-martin.loebel@cms.hu-berlin.de
http://waste.informatik.hu-berlin.de/jml/

Abstract: Der Beitrag behandelt Probleme und Strategien der Langzeitarchivierung digitaler Daten, gibt einen konzentrierten Überblick über die generellen Probleme und behandelt dann die speziellen Probleme der Archivierung multimedialer Objekte. Es werden technische Lösungsansätze im Zusammenspiel mit den jeweiligen Akteuren vorgestellt.

1 Einleitung

Bei der Langzeitarchivierung digital vorliegender Informationen stellen sich viele neue Probleme. Auf der einen Seite ist die Haltbarkeit und Lesbarkeit von Datenträgern im Vergleich zu analogen Medien sehr begrenzt. Sie verlieren ihre Informationen durch Umwelteinflüsse, sind sehr anfällig für chemische oder physikalische Einwirkungen und oft schon innerhalb weniger Jahre nicht mehr auslesbar. Auf der anderen Seite sind digitale Datei- und Speicherformate einem ständigen Wandel unterworfen. So sind digitale Daten ebenfalls verloren, wenn Lesegeräte und Abspielprogramme für die entsprechenden Datenträger und -formate nicht mehr vorhanden oder benutzbar sind. Diese Gefahr besteht besonders bei proprietären und unzureichend dokumentierten Formaten. Urheberrechtliche Beschränkungen und *Digital Rights Management* (DRM) bilden zusätzliche Probleme, z. B. durch die rechtliche Einschränkung bzw. technische Verhinderung, eine Kopie anzufertigen.

1.1 Definition Langzeitarchivierung

Unter dem Begriff *Langzeitarchivierung* versteht man die Aufbewahrung und die Erhaltung der Zugänglichkeit von Informationen auf unbestimmte Zeit.[1] Langzeitarchivierung bezeichnet in diesem Sinne Prozesse bzw. Vorgehensweisen, die aufgrund von technologischen und gesellschaftlichen Änderungen ständiger Anpassung bedürfen. Im Fall von digitalen Daten sind dies sämtliche Maßnahmen zur dauerhaften Speicherung, aber auch Zugänglichmachung (Katalogisierung) sowie Interpretation der Daten mit resultierender möglichst originalgetreuer Darstellung des Inhaltes in für Menschen wahrnehmbarer Form.

1.2 Probleme der Archivierung digitaler Daten

Hier offenbart sich der wesentliche Unterschied zu analogen Medien wie Büchern oder Steintafeln. Digitale Informationen sind für Menschen immer nur mittelbar, d. h. mittels eines Abspielgerätes und anschließender Interpretation durch Software wahrnehmbar [Bo05]. Sie sind dabei zwangsläufig auf einem physischen Trägermedium gespeichert. Die sich aus der Definition ergebenden Faktoren (Speicherung, Auffinden, Auslesen, Interpretation) für die Archivierung digitaler Informationen eröffnen Problemfelder, die im Folgenden kurz erläutert werden:

Jedes Speichermedium hat nur eine begrenzte Haltbarkeit. Die Haltbarkeit heute gängiger digitaler Speichermedien ist jedoch drastisch kürzer als die analoger Medien. So können beispielsweise Texte auf Steintafeln mehrere tausend, auf Mikrofilm bis zu fünfhundert[2] und auf säurefreiem Papier mehrere hundert Jahre bestehen. Derzeit gängige digitale Speichermedien lassen sich nach der Form ihrer Aufzeichnung in Kategorien einteilen. Bei magnetischen Speichern, wie z. B. Festplatten, Magnetbändern oder Disketten findet man Angaben von 5-10, bei Festplatten bzw. 1-5 Jahren, bei Magnetbändern [Ha05, S. 39; Ro95]). Optische Speicher wie *Compact Disk* (CD) oder *Digital Versatile Disk* (DVD) haben eine erwartete Lebensdauer von 30 Jahren [Ha05], wobei zwischen im Werk gepressten und durch Erhitzung der Trägerschicht „*gebrannten"* Versionen CD-R(W) bzw. DVD+/-R(W) unterschieden werden muss. Die Haltbarkeit der Letzteren ist mit ca. 5 Jahren deutlich kürzer.[3] Als letztes zu erwähnen sind Halbleiterspeicher (*Flash-Speicher*) wie *Compact-Flash* oder *USB-Sticks*. Ihre Haltbarkeit liegt bei etwa zehn Jahren [Co06 S. 33]. Man kann von einer Verdopplung der Kapazität etwa alle 12 bis 18 Monate ausgehen.

---

[1] Vergleiche NESTOR – Kompetenznetzwerk Langzeitarchivierung: Glossar; Begriff Langzeitarchivierung. URL: http://www.langzeitarchivierung.de/index.php?module=Encyclopedia&func=displayterm&id=38&vid=1 ; Stand 20.02.2004, Abrufdatum 22.06.2007.

[2] Herstellerangabe bei korrekter Lagerung bei 21 °C und 50% relativer Luftfeuchte.

[3] Artikel *Langzeitarchivierung.* In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 24. Mai 2007, 09:55 UTC. URL: http://de.wikipedia.org/w/index.php?title=Langzeitarchivierung&oldid=32247987 (Abrufdatum: 25.06.2007.

Zusätzlich sind die Schreib-/Lesegeräte, sowie die Datenträger selbst – bedingt durch schnell fortschreitende technische Entwicklungen – einem ständigen Wandel unterworfen. So liegt die erwartete Lebensdauer einer Generation von Lesegeräten und -medien bei fünf bis zehn Jahren (optische Medien) [Ha05]. Ein ständiges Umkopieren der Daten auf neue Datenträger sowie ein Wechsel auf die jeweils neueste Generation von Speichermedien als ständiger Prozess sind unumgänglich.

Weitaus größere Probleme bereitet jedoch die Fülle an Daten- und Dateiformaten bzw. -systemen. Der auf dem Datenträger gespeicherte Bitstrom muss zur Darstellung mittels eines Softwareprogramms erst logisch interpretiert werden. So können z. B. Speicherformate von Texten neben den eigentlichen Textinformationen Formatierungsanweisungen enthalten. Die verschiedenen Formate reichen dabei von standardisiert und/oder offen bzw. offen gelegt (*PDF*, *XML*) bis zu proprietären Formaten wie Microsoft Word (*Doc*). Generell ist zur Reduzierung des Verwaltungsaufwandes und zur einfacheren Wartbarkeit die Festlegung auf einige wenige standardisierte und offen gelegte Formate zu bevorzugen. Offene und standardisierte bieten gegenüber proprietären Formaten den Vorteil hoher Kompatibilität sowie in der Regel breitere Unterstützung in mehreren Betriebssystemen und Programmen. Dies erhöht die Chance der Lesbarkeit des Dokumentes, selbst wenn das ursprüngliche Erstellungsprogramm nicht mehr verfügbar ist. Die Dokumente müssen dazu notfalls in das verwendete Format umkodiert werden. Eine Möglichkeit liegt in der Verwendung von *PDF/A*. Das ursprünglich von *Adobe Systems* entwickelte *Portable Document Format* (*PDF*) ist in der Version PDF/A ein speziell zur Langzeitarchivierung von Dokumenten entwickelte ISO-Norm (ISO 19005-1:2005). PDF/A regelt die visuelle Reproduzierbarkeit digitaler Dokumente. Es werden sämtliche zur Anzeige benötigten Komponenten wie Schriftarten, eventuell verwendete Bilder sowie Metadaten in das Dokument eingebettet [Is05].

1.3 Zusätzliche Problemfelder multimedialer Objekte

Noch höhere Anforderungen an die Archivierung stellen multimediale Objekte. Unter „multimedial“ sind hierbei alle „nicht text-basierten medialen Objekte“ [Co06] wie Bilder, Audio, Video und Animationen, Datenbanken und Forschungsrohdaten sowie komplexe interaktive Software wie z. B. Computerspiele und Anwendungsprogramme zu verstehen. Da der Bereich der digitalen Medien einem besonders schnellen Innovationszyklus unterliegt und es eine Vielzahl von Formaten sowie kaum (internationale) Normierungen gibt [Co06], lässt sich die oben erwähnte Strategie der Beschränkung auf einige wenige Formate nicht sinnvoll anwenden. Zudem finden gerade im Bereich Video und Animation häufig verlustbehaftete Kompressionsalgorithmen Anwendung. Diese lassen sich, selbst wenn das Format offen gelegt und bekannt ist, nicht verlustfrei in ein neueres (ebenfalls verlustbehaftetes) Format umkodieren. Eine Speicherung in unkomprimierter Form ist aufgrund der immensen Datenmenge[4] nicht realistisch.

[4] So würde beispielsweise eine Minute Video 768x576 Pixel, 24 Bit bei 25 Bildern pro Sekunde ca. 1,85 Gigabyte Speicherplatz benötigen.

Des Weiteren können multimediale Daten wie kommerzielle Musikstücke oder Filme mittels eines *Digital Rights Management* (DRM) Systeme geschützt sein, welche die Anfertigung einer Kopie oder die Umkodierung verhindern oder sogar das Abspielen des Inhaltes auf bestimmte Rechner beschränken können.

Eine mögliche Alternative zur Umkodierung ist die Verwendung von Emulationstechniken. Dabei kann das Originalformat beibehalten werden. Mittels eines Emulators wird versucht, die ursprüngliche Hard- und Softwareumgebung möglichst originalgetreu nachzuahmen, wodurch das ursprüngliche Leseprogramm funktionsfähig bleibt [Ro95]. Dazu ist es zwingend erforderlich, dass möglichst die gesamte Softwareumgebung inklusive Betriebssystem, mindestens jedoch das Leseprogramm mit archiviert werden. Der Nachteil liegt im „hohen Aufwand bei der Erstellung" [Bo05]. Dieser könnte mit einer vereinheitlichten virtuellen Computerarchitektur, dem „*Universal Virtual Computer*" [Bo05] – ähnlich dem Ansatz der *Java Virtual Maschine* – wahrscheinlich reduziert werden. Konkrete Praxistests liegen hierzu aber noch nicht vor. Emulatoren werden jedoch bereits heute im Bereich älterer Videospiele und Konsolen eingesetzt[5].

2. Akteure und Strategien in Deutschland

Um zu verfolgen, wie die oben erwähnten Lösungsansätze und -strategien technisch umgesetzt und entwickelt bzw. gefördert werden, ist es hilfreich, sich mit den Akteuren, die ein aktives Interesse an der Langzeitarchivierung haben, auseinanderzusetzen. Traditionell fällt die Archivierung in den Zuständigkeitsbereich von Bibliotheken und privaten Archiven. In Deutschland ist dabei besonders die *Deutsche Nationalbibliothek* hervorzuheben.

2.1 Deutsche Nationalbibliothek Frankfurt

In Deutschland hat die *Deutsche Nationalbibliothek* als Anstalt des öffentlichen Rechts seit 1969 einen gesetzlich festgeschriebenen Sammelauftrag. Dieser begründet sich aus dem „*Gesetz über die Deutsche Nationalbibliothek*" (*DNBG*).[6] Die Bibliothek ist verpflichtet „*die ab 1913 in Deutschland veröffentlichten Medienwerke und [...] im Ausland veröffentlichten deutschsprachigen Medienwerke [...] im Original zu sammeln, zu inventarisieren, zu erschließen und bibliographisch zu verzeichnen.*"[7] und der Allgemeinheit zu Verfügung zu stellen (§ 4 DNBG).

[5] Bekanntestes Beispiel ist der „*Multi Arcade Maschine Emulator*" (*MAME*), offizielle Website: http://www.mamedev.org/ ; Abrufdatum 25.06.2007.
[6] Neufassung des Gesetzes vom 22. Juni 2006, ersetzt das „*Gesetz über die Deutsche Bibliothek*" (DBiblG) vom 31. März 1969. Der Gesetzestext ist durch das Bundesministerium der Justiz online verfügbar unter <http://www.gesetze-im-internet.de/dnbg/index.html>. Abrufdatum: 24. Juni 2007.
[7] Siehe § 2 Abs. 1 Satz 1 und 2 DNBG.

Des Weiteren besteht für alle in Deutschland öffentlich publizierten Medienwerke[8] eine Ablieferungspflicht des so genannten *Pflichtexemplars* an die Nationalbibliothek (§ 14 DNBG). Gesammelt werden neben gedruckten Publikationen auch Mikroformen und Tonträger, nicht aber Filme, Rundfunksendungen oder Videos[9]. In der Neufassung des Gesetzes wird (im Gegensatz zum DBiblG verwendeten Begriff *Schriftwerke)* der allgemeinere Begriff *Medienwerke* verwendet und es werden erstmals explizit „*Medienwerke in unkörperlicher Form"*,[10] d. h. elektronische Publikationen und digitale Medien, in die Ablieferungspflicht und den Sammelauftrag eingeschlossen. Innerhalb der *Deutschen Nationalbibliothek* befasst sich die Arbeitsgruppe *Langzeitarchivierung* – u. a. im Vorfeld auf die Gesetzesänderung – bereits seit 2002 mit dem Thema der Archivierung elektronischer Publikationen. Initialzündung war das Projekt „*Aufbau eines Kompetenznetzwerkes Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen*"[11] innerhalb des *Digital Library Forums*[12].

2.2 NESTOR

Wichtige Ergebnisse der Arbeitsgruppe flossen in das Nachfolgeprojekt „*Kompetenznetzwerk Langzeitarchivierung*" *NESTOR* ein [DT02]. Ziel des vom *Bundesministerium für Bildung und Forschung* (*BMBF*) bis 2009 geförderten Projektes ist der „Aufbau eines Kompetenznetzwerks zur Langzeitarchivierung digitaler Quellen in Deutschland in einer dauerhaften Organisationsform sowie die Abstimmung über die Übernahme von Daueraufgaben".[13] Dabei sind die Hauptaufgaben die Festlegung von Kriterien für vertrauenswürdige digitale Archive, von Grundsätzen für die Langzeitarchivierung, und die Einbindung von Museen und Archiven. Im Rahmen von NESTOR entstanden bisher eine Reihe von Materialien[14], welche den Verbundpartnern und der Allgemeinheit Richtlinien und Arbeitswerkzeuge im Hinblick auf die Projektziele zur Verfügung stellen. Eine aktive Unterstützung erfahren dabei *Repository-Systeme* nach dem *Open Archival Information System* (*OAIS*)[15] [DT02]. Dabei definiert das *OAIS-Modell* „die zentralen Funktionen und Abläufe eines Archivsystems und es bietet eine Terminologie und ein Strukturkonzept" [DT02].

---

[8] Wörtlich in § 15 DNBG: „ Ablieferungspflichtig ist, wer berechtigt ist, das Medienwerk zu verbreiten oder öffentlich zugänglich zu machen und den Sitz, eine Betriebsstätte oder den Hauptwohnsitz in Deutschland hat."
[9] § 3 Abs. 4 DNBG.
[10] § 14 Abs. 3 DNBG.
[11] Projektlaufzeit 1.4.2002 – 30.11. 2002. Homepage URL: http://www.dl-forum.de/deutsch/projekte/projekte_327_DEU_HTML.htm ; Abrufdatum 21.06.2007.
[12] Das Digital Library Forum der Deutschen Bibliothek bündelt Informationen zum Thema Digitale Bibliothek. URL: http://www.dl-forum.de/ ; Abrufdatum: 21.06.2007.
[13] Vergleiche NESTOR- Mission Statement, URL: http://www.dl-forum.de/deutsch/projekte/projekte_327_DEU_HTML.htm, NESTOR-Homepage: www.langzeitarchivierung.de; Abrufdatum: 15.06.2007
[14] Die NESTOR-Materialien 1-8 sind online verfügbar unter http://www.langzeitarchivierung.de/modules.php?op=modload&name=PagEd&file=index&page_id=2 ; Abrufdatum: 20.06.2007
[15] CCSDS: Reference Model for an Open Archival Information System, Blue Book. Verfügbar online http://public.ccsds.org/publications/archive/650x0b1.pdf ; Abrufdatum: 25.06.2007.

Die zu archivierenden Objekte werden als Original Bitstrom gespeichert und zusätzlich mit Metadaten versehen. Dadurch kann es unabhängig vom jeweils verwendeten Datenformat zur Speicherung eingesetzt werden. Durch die Abstraktion der einzelnen Vorgänge ist das System mit den unterschiedlichen Archivierungstechniken wie Migration der Daten durch Umkodierung oder Emulation kompatibel. Darüber hinaus sind die Daten und Archivierungsinformationen in separaten *Containern* gespeichert, was eine dezentrale Implementierung des Systems erlaubt. Zu den bekanntesten OAIS-Referenzimplementierungen zählen *DSpace* und *Fedora*[16], die beide quelloffen und kostenlos verfügbar sind. Ein detaillierter Vergleich beider Systeme im Hinblick auf ihre Funktionalität findet sich in den Nestor-Materialien Nr. 3.

3. Fazit und Ausblick

Die Langzeitarchivierung digitaler Daten eröffnet neue Problemfelder. Ständiges Umkopieren auf aktuelle Datenträgermedien und -formate ist vonnöten. Des Weiteren ist die Migration durch Umkodierung bestehender Daten auf neue Datenformate bis auf Textdokumente nur eingeschränkt möglich. Technische Ansätze zur Problemlösung bilden Repositories, welche die Speicherung, Katalogisierung und Verwaltung der Daten übernehmen. Bei der Langzeitarchivierung multimedialer Objekte bietet die Emulation Hoffnung, jedoch liegen diesbezüglich noch keine weitergehenden Erkenntnisse vor, so dass weiterhin konkrete Strategien größtenteils ungeklärt bleiben.

Literaturverzeichnis

[Bo05] Borghoff, U. M. et.al.: Langzeitarchivierung. In (Bode, A. Hrsg.): Informatik-Spektrum - Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen, Band 28, Heft 6. Springer Verlag, Heidelberg, 2005; S. 489-492

[Ro95] Rothenberg, J.: Ensuring the Longevity of Digital Documents. In: Scientific American, Jg. 1995, Heft 1. Scientific American Inc, New York, 1995; S. 42-47

[Ha05] Harvey, R.: Preserving Digital Materials. K. G. Saur Verlag, München, 2005

[Co06] Coy, W.: Perspektiven der Langzeitarchivierung multimedialer Objekte. In (Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland Hrsg.): Nestor-Materialien, Nr. 5, 2006. Abrufbar im Internet: http://nbn-resolving.de/urn:nbn:de:0008-20051214015

[DT02] Dobratz, S.; Tappenbeck, I.: Thesen zur Zukunft der digitalen Langzeitarchivierung in Deutschland. In (Kaegbein, P. et.al. Hrsg.) Bibliothek – Forschung und Praxis, Jg. 26, Nr. 3. K. G. Saur Verlag, München, 2002; S. 257-261. Abrufbar im Internet: http://www.bibliothek-saur.de/2002_3/257-261.pdf

[Is05] International Organization for Standardization: Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1). ISO, 2005

[16] DSpace Homepage: http://www.dspace.org/ ; Fedora: http://www.fedora.info/ ; Abfragedatum 01.06.2007.