Langzeitarchivierung
Proseminar Speicher- und Dateisysteme

Hajo Möller

WS 2010/11

Inhaltsverzeichnis

1 Einführung

1.1 Was ist Langzeitarchivierung?

Um diese Frage zu beantworten muss man wissen, dass digital vorliegende Informationen von Verlust bedroht sind. So können die Datenträger, auf denen diese Informationen gespeichert sind kaputt oder verloren gehen oder aus anderen Gründen unlesbar werden. Auch kann (mutwillige) Fehlbedienung eines Rechners zu schwerwiegendem Datenverlust führen.

Die Langzeitarchivierung beschäftigt sich mit der Erfassung und Erhaltung von wichtigen Daten und stellt sicher, dass eben diese Daten auch in Zukunft zugreifbar und lesbar sind. „Zukunft“ bedeutet hier eine nicht näher spezifizierte Zeitspanne, sondern „dauerhaft, für immer“.

Genau bedeutet dies, dass Langzeitarchive die von ihnen gesicherten Informationen auf Datenträgern einlagern, diesen Lagerbestand pflegen und regelmäßig die Datenträger erneuern und die Daten in neue, zukunftssichere Formate umwandeln müssen.

1.2 Wachstum des „Digital Universe“

Als „Digital Universe“ bezeichnet man die gesamte aktuell auf der Welt vorhandene Datenmenge. Durch die stetige technologische Weiterentwicklung wächst das „Digital Universe“ mit einer rasanten Geschwindigkeit, bisher hat sich die Menge pro Jahr veranderthalbfacht.

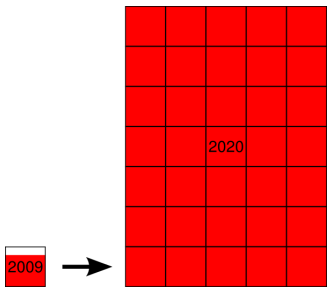

Abbildung 1: Datenwachstum bis 2020

2009 waren weltweit etwa 0,8 Millionen Petabytes an Daten vorhanden, 2010 wuchs diese Menge auf 1,2 Millionen Petabytes. Bis 2020 wird mit einem Wachstum auf etwa 35 Millionen Petabytes gerechnet. [1]

Gerade in Anbetracht der stetig anwachsenden Menge an Daten ist es wichtig, sich mit der langfristigen und zukunftssicheren Aufbewahrung und Instandhaltung von Daten, also der digitalen Langzeitarchivierung, genauer zu beschäftigen.

1.3 Haltbarkeit von Datenträgern

Eines der beiden großen Probleme der Langzeitarchivierung ist die begrenzte Lebensdauer der eingesetzten Datenträger.

Die Haltbarkeit eines Datenträgers ist von vielen Faktoren abhängig, unter anderem wird sie durch die verwendeten Materialien, den Fertigungsprozess und die Lagerung bestimmt.

Moderne Trägermedien wie CDs oder DVDs haben eine relativ hohe Datendichte, aber Lebenserwartungen, die teilweise um den Faktor 10 schwanken können. Damit sind sie zu volatil und als Datenträger für Langzeitarchive

eher nicht zu gebrauchen. Außerdem sind nur die professionell hergestellten, gepressten CDs/DVDs/Blurays zuverlässig langfristig haltbar. Die Datenschicht selbstgebrannter Medien besteht zwangsweise aus einem anderen chemischen Verbund, dessen Haltbarkeit *deutlich* unter der von gepressten Medien liegt.

Andererseits haben sich Steintafeln haben zwar als äußerst langlebig herausgestellt, die maximal mögliche Datendichte ist aber für heutige Verhältnisse nicht mehr ausreichend.

Einen kleinen Überblick über die zu erwartende Haltbarkeit verschiedener Datenträger gibt die folgende Tabelle: [2]

Medium	Erwartete Lebensdauer
Steintafeln	1.000+ Jahre
Bücher etc.	100+ Jahre
Optische Medien (professionell gepresst)	50–80 Jahre
Magnetische Medien (Festplatten, Bänder)	30+ Jahre
Optische Medien (selbstgebrannt)	5–30 Jahre
Flashspeicher (USB-Sticks, SSDs)	3–10 Jahre

Tabelle 1: Haltbarkeit verschiedener Speichermedien

1.4 Digitale Obsoleszenz

Das zweite große Problem der Langzeitarchivierung ist die digitale Obsoleszenz, sie bezeichnet den Informationsverlust durch „Aussterben“ von kompatiblen Lesegeräten und Datenformatwandel.

Auch wenn ein Datenträger sich lange halten sollte besteht die Gefahr, dass die auf ihm gespeicherten Daten verloren sind: Es könnte kein kompatibles, funktionsfähiges Lesegerät mehr existieren oder das verwendete Datenformat ist nicht mehr lesbar, es liegen keine Information über dessen Struktur vor.

Ein schönes Beispiel eines ausgestorbenen Mediums sind Lochkarten, die Daten sind weiterhin auf den bis über 100 Jahre alten Karten erhalten. Allerdings gibt es weder an heutigen Maschinen keine Lochkartenleser; häufig ist auch nichtbekannt, wie die auf der Lochkarte gespeicherten Daten zu interpretieren sind. Weitere bereits ausgestorbene oder vom Aussterben bedrohte Medien sind Disketten, Kassetten bzw. Datasetten aber auch CD-ROMs oder Flashspeicher wie Memory Sticks. [3]

1.5 Digital Rights Management

Unter Digital Rights Management, kurz DRM, versteht man Techniken, durch die die Nutzung digitaler Medien kontrolliert und eventuell eingeschränkt werden können. Urheber und Verwertungsgesellschaften können so kontrollieren, ob das von ihnen verbreitete Medium eingelesen/abgespielt bzw. ihre Software verwendet werden darf.

Effektiv bedeutet das, dass man beim Kopieren von DRM-geschützten Daten stets auf den jeweiligen Rechteinhaber angewiesen ist, was natürlich viele neue Probleme mit sich bringt. So könnte der Rechteinhaber beschließen, das DRM-System auszuschalten wodurch die vorliegenden Daten unbenutzbar würden.

2 Grundlegende Techniken

2.1 Archivieren, aber wie?

Grundsätzlich unterscheidet man vier verschiedene Prinzipien zur langfristigen Datensicherung:

- Erneuerung der Daten auf neuen Datenträgern
- Migration der Daten auf neue Systeme
- Replikation der Daten auf mehrere Systeme
- Emulation der Orginalsysteme

Jede dieser Techniken wird im Folgenden im Detail erläutert.

2.1.1 Erneuerung

In regelmäßigen Abständen, die die Haltbarkeit des eingesetzten Mediums deutlich unterschreiten sollten, werden die zu sichernden Daten *exakt* auf *gleichartige* Datenträger übertragen, um dem Verfall der Speichermedien engegenzuwirken. Die Datenträger werden also einfach *erneuert*. Diese Methode hat den Vorteil, dass die Daten in keiner Weise konvertiert werden müssen; bei der Sicherung einer Audio-CD bleibt beispielsweise sogar die Bitrate erhalten. Sofern neue Datenträger und die für den Kopiervorgang benötigte Hardware (beispielsweise Rohlinge und CD-Brenner) verfügbar sind, ist dies die einfachste Methode.

Langfristige Probleme können entstehen, wenn das Speichermedium obsolet wird und es zunehmend schwerer wird Datenträger und kompatible Hardware zu beschaffen. Ein modernes Beispiel für diese Entwicklung ist die 3,5”Diskette. Angesichts der bisherigen technischen Entwicklung ist davon auszugehen, dass dieses Verfahren nicht beliebig lange für einen Datensatz eingesetzt werden kann, ohne auf eine der drei anderen Techniken zurückzugreifen.

2.1.2 Migration

Immer wenn das Speichermedium oder das Datenformat droht obsolet zu werden wird der Datenbestand auf ein modernes Format konvertiert (*migriert*), damit die Lesbarkeit der Daten trotz der rasanten technischen Entwicklung gegeben bleibt. Ein Beispiel ist die Konvertierung von alten Textdateien im veralteten Microsoft .doc Format auf das neuere OpenDocument oder die Übertragung von Daten von Disketten auf CDs.

Neben dem großen Zeit- und Arbeitsaufwand, der dadurch entsteht, ist es problematisch, dass möglicherweise bestimmte Eigenschaften wie Textformatierung oder Funktionalitäten wie Macros in einer alten Exceltabelle verloren gehen, da das modernere Format nicht vollständig kompatibel ist. Ein ähnliches Problem tritt auf, wenn das veraltete Format nur von proprietärer, nicht mehr verfügbarer Software (beispielsweise eine Officesuite aus den 90er Jahren) vollständig gelesen werden kann. Dann können nur diese Daten migriert werden, die sich noch über einen anderen Weg lassen. Daher ist es wichtig darauf zu achten eine Migration rechtzeitig durchzuführen, um die Verfügbarkeit der Lesetechniken (Hard- und Software) sicherzustellen.

2.1.3 Replikation

Um Ausfällen und Fehlfunktionen von Hard- und Software, die zum *Totalverlust* der zu sichernden Daten führen können, vorzubeugen müssen Daten immer in mehreren Kopien vorliegen. Das Erstellen solcher Kopien nennt man *Replikation*. Durch dezentralisierung der Lagerorte (um auch gegen externe Faktoren wie Naturkatastrophen vorzubeugen) lässt sich die erwartete Haltbarkeit der Daten maßgeblich steigern.

Da es sich prinzipiell um eine mehrfache Erneuerung der Daten handelt, ist die Replikation mit den selben Problemen verbunden wie die Erneuerung. Hinzu kommt je nach Anzahl der Kopien eine Vervielfachung der Kosten für Datenträger und Lesegeräte. Sollte eine Migration nötig werden, vervielfacht sich der hierfür nötige Aufwand auch entsprechend.

2.1.4 Emulation

Emulation bezeichnet die funktionale Replikation eines anderen, meist inzwischen obsoleten Systems. Beispielsweise die Emulation einer nicht mehr eingesetzten Mikroprozessorarchitektur in Software, um die veraltete, plattformspezifische Lesesoftware ausführen zu können. Im Idealfall lässt sich der einmal erzeugte Emulator universell, das heißt für alle zu sichernden Daten, einsetzen.

Allerdings ist der initiale Aufwand - die Erstellung des Emulators - sehr hoch und eine vollständige Kompatibilität ist nahezu unerreichbar. Auch sind Emulatoren immer mit Geschwindigkeitsverlusten im Vergleich zum ursprünglichen System verbunden. Ferner lassen sich Lesegeräte in Hardware nicht emulieren, sodass das Problem der Migraton auf neuere Speichermedien erhalten bleibt.

2.2 Tatsächlich angewandte Techniken

Offenbar bietet die Anwendung einer einzelnen Methode keine Hinreichende Datensicherheit, sodass in der Realität meistens Kombinationen aus allen vieren durchgeführt werden. Je nach abgeschätzter Zukunftssicherheit der verwendeten Speicherformate kann man zwischen Erneuerung und Migration abwägen. Replikation hingegen wird im Normalfall als unerlässlich betrachtet; um eine möglichst große Anzahl und breit verteilte an Kopien zu erreichen gibt es Kooperationen der Langzeitarchive untereinander. Emulation wird als Notlösung betrachtet, wenn keine andere Möglichkeit gangbar erscheint. Sie wird hauptsächlich zur Erhaltung von proprietärer Software eingesetzt.

2.3 Integrität der Daten wahren

Durch Verfall der Datenträger, Fehler in der Software oder bösen Willen kann die Integrität der Daten beschädigt werden. Daher müssen Techniken implementiert werden, um Integritätsverlust festzustellen und nach Möglichkeit auch zu korrigieren. Zum einen können Metadaten angehängt werden, die unter anderem Zeitstempel, Migrationsgeschichte oder Dateirechte enthalten können. Problematisch ist auch hier die Auswahl eines geeigneten Dateiformats, das seinerseits wieder obsolet werden kann. Auch müssen die Metadaten selbst auf Integrität geprüft werden, sodass sich das Problem nur verschiebt.

Daher werden archivierte Dateien häufig um (mehrere) Prüfsummen (beispielsweise mittels der MD5 oder SHA-1 Algorithmen) ergänzt. So lassen sich

bei Erneuerung oder Migration die Nutzdaten gegen die Prüfsummen abgleichen und eventuell beschädigte Daten aus dem verteilten, replizierten Datenbestand wiederherstellen.

2.4 Ein (relativ sicheres) Langzeitarchiv

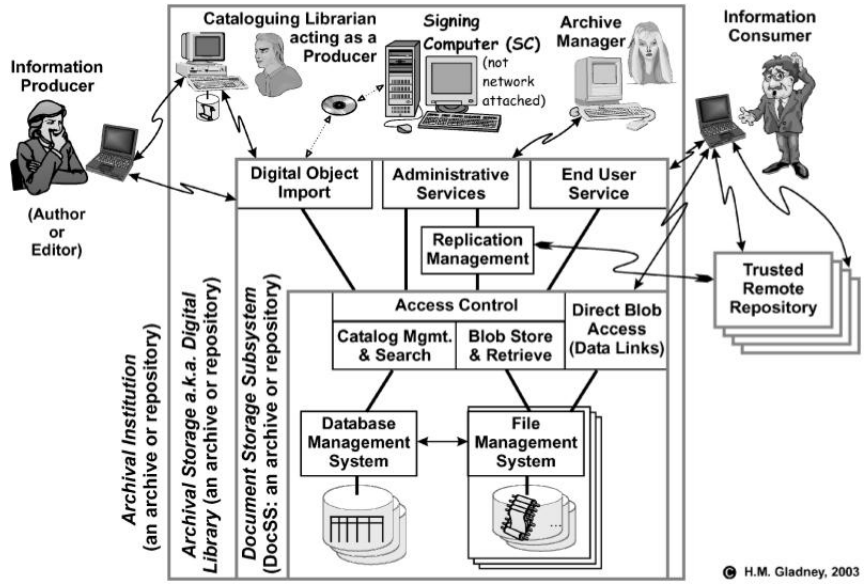

Abbildung 2: sicheres Langzeitarchiv[4]

3 Zukunftsausblick – Gelingt es, „Digital Dark Ages“ zu verhindern?

3.1 „Digital Dark Ages“

Im Hinblick auf die Zukunft ergeben sich mehrere Probleme. Seit der weit verbreiteten Verfügbarkeit von Computern und dem Durchbruch des Internets liegen große Teile des zeitgenössischen Kulturguts ausschließlich digital vor. Bei einigen dieser Dokumente lässt sich durchaus abschätzen, dass sie einmal einen gewissen historischen Wert haben werden. Daher ist es von großer Bedeutung diese Daten für die kommenden Generationen zu sichern, um nicht als *Dark Ages* ohne erhaltene Kulturzeugnisse in die Geschichte einzugehen.

Auf einem persönlicheren Level werden viele Daten von hohem ideellen Wert (z.B. Digitalfotos, elektronisch geführte Tagebücher) quasi nie wirklich sicher gespeichert. Ein normaler privater Anwender wird in den seltensten Fällen auch nur Backups seiner Daten gemacht haben.

Möglicherweise befinden wir uns schon inmitten dieser *Digital Dark Ages*, da viele Daten nur in proprietären Formaten, die drohen obsolet zu werden, vorliegen und nur sehr wenige Daten professionell langzeitarchiviert werden.

3.2 Empfehlenswerte Schritte

Um zukünftigen Datenverlust zu verhindern ist es ein wichtiger Schritt sicherzustellen, dass die Daten in Hard- und Software lesbar bleiben oder zumindest die Information offen verfügbar ist, wie ein bestimmtes Format auszulesen ist. Dazu müssen offene Standards *geschaffen* (beispielsweise durch die ISO) und auch konsequent *verwendet* werden. Denn im gegensatz zu proprietären Formaten bleibt bei offenen Standards gewährleistet, dass die Beschreibung des Formats, die benötigt wird, um die Information aus den Daten zu extrahieren, bekannt bleibt.

3.3 Geeignete Datenformate

3.3.1 Dokumente

PDF/A ist eine minimale Variante des verbreiteten PDF-Formats, die keine interaktiven oder multimedialen Inhalte zulässt, sich aber gut für Textdateien und Grafiken eignet, da sie unter anderem die benötigten Schriftarten einbettet, um den Text systemübergreifend lesbar zu halten. Außerdem bietet PDF/A direkt einen Satz standarisierter Metadatenfelder, sodass diese nicht extra angelegt werden müssen.

XML bietet sich als vielseitiges Metaformat für kompliziertere Daten an. Es ist strukturiert genug, um maschinenlesbar zu sein, aber bleibt menschenlesbar, da es sich auf den ASCII-Zeichensatz beschränkt. Eine Vielfalt von Daten, wie Grafiken auf basis von SVG, komplexe mathematische Texte auf Basis von MathML oder Hypertext auf Basis von XHTML lassen sich über XML langfristig sicher abspeichern.

3.3.2 Mediendateien

Neben SVG für Vektordateien liegt das freie PNG (Portable Network Graphics) für Rastergrafiken vor. Da es ein offener Standard ist, ist seine längerfristige Lesbarkeit gewährleistet. Außerdem arbeiten die verwendeten Komprimierungsalgorithmen verlustfrei, sodass eine eventuelle Migration auf ein anderes Format ohne Qualitätsverlust möglich sein sollte.

Ein offenes Format zur verlustfreien Kompression von Audiodateien ist das Free Lossless Audio Codec (FLAC). Leider wird es nur von wenigen Endgeräten unterstützt, da die Hersteller zumeist auf eigenentwickelte, proprietäre und DRM-belastete Formate zurückgreifen.

3.3.3 Videos

Zur verlustfreien Kompression von Videodaten liegt der offene Standard HuffYUV vor, der aber leider wenig verbreitet ist und schon seit längerem (2002) nicht mehr aktualisiert wurde. Entsprechend schlecht ist er in Hard- und Software unterstützt.

Ein anderes Format mit Potential ist der freie Videocodec VP8, der von Google gefördert wird. Obwohl er zur Zeit noch wenig Verbreitung hat ist davon auszugehen, dass durch die Rückendeckung eines einflussreichen Konzerns wie Google die Akzeptanz in Zukunft steigen wird. Leider ist VP8 verlustbehaftet, sodass die Migration auf ein anderes Format mit Datenverlust verbunden ist.

3.4 Hardware

Es wäre wünschenswert, wenn auch die Hardware offenen Standards genügen würde, um auch an dieser Stelle die langfristige Lesbarkeit (notfalls durch Nachbau des Lesegerätes) sichergestellt werden kann. Leider befinden sich die meisten freien Hardwareprojekte noch in einem Anfangsstadium. Einige erfolgeriche Beispiele sind die Projekte Arduino, Makerbot und Open Cores. Einen komplett auf offenen Standards basierenden Desktop-Computer kann man heute allerdings noch nicht herstellen.

3.5 Persönliche Empfehlung

Meiner Einschätzung nach sollte man die Daten, die langzeitarchiviert werden sollen, sorgfältig auswählen. Zu viele Daten machen das Unterfangen leicht unmöglich, sowohl von Kosten- als auch von Aufwandsseite. Wenn man aber zu viel weglässt, geht man das Risiko ein wichtige Daten zu verlieren.

Je nach Menge der Daten und gewünschter Transportabilität sollte man sich zwischen Festplatten (HDD) und Magnetbändern (LTO) entscheiden. Die Kosten belaufen sich derzeit bei HDDs auf ca. 35 Euro pro Terabyte und bei LTOs auf ca. 20 Euro für die selbe Datenmenge. Bei LTOs kommen noch rund 1000 Euro für ein Laufwerk dazu, weshalb sie eher im professionellen Umfeld anzutreffen sind. LTOs bieten zudem eine enorme Bandbreite, da sie sich schlichtweg Schubkarrenweise transportieren lassen.

Festplatten sind empfindlicher als Magnetbänder, haben allerdings als Archivmedien - nicht im laufenden Betrieb, wo sie sich schnell abnutzen - eine erwartete Lebenszeit von 10 bis 30 Jahren. Hingegen wird davon ausgegangen, dass Magnetbänder mindestens 30 Jahre sicher lesbar bleiben.

Literatur

[1] *IDC Digital Universe Study*; `http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm`; IDC und EMC

[2] *Wikipedia: Langzeitarchivierung*; `http://de.wikipedia.org/wiki/Langzeitarchivierung#Haltbarkeit_der_Tr.C3.A4germedien`

[3] *Chamber of Horrors: Obsolete and Endagered Media*; `http://www.icpsr.umich.edu/dpm/dpm-eng/oldmedia/chamber.html`; Inter-university Consortium for Political and Social Research

[4] *Trustworthy 100-Year Digital Objects: Syntax and Semantics–Tension Between Facts and Values*; Gladney, Dr. H.M.; 2003