Langzeitarchivierung

Hajo Möller

Proseminar Speicher- und Dateisysteme

Gliederung

Einführung

- Was ist Langzeitarchivierung?

Grundlegende Techniken

- Archivieren, aber wie?

Zukunftsausblick

- Gelingt es, „Digital Dark Ages“ verhindern?

Was ist Langzeitarchivierung?

Digitale Informationen gehen verloren

- Datenträgerverfall
- Infrastrukturänderungen

Langzeitarchivierung beugt dem vor

- Erfassung
- Erhaltung
- Sicherung der dauerhaften Verfügbarkeit von Daten

Wachstum des „Digital Universe“

2009:

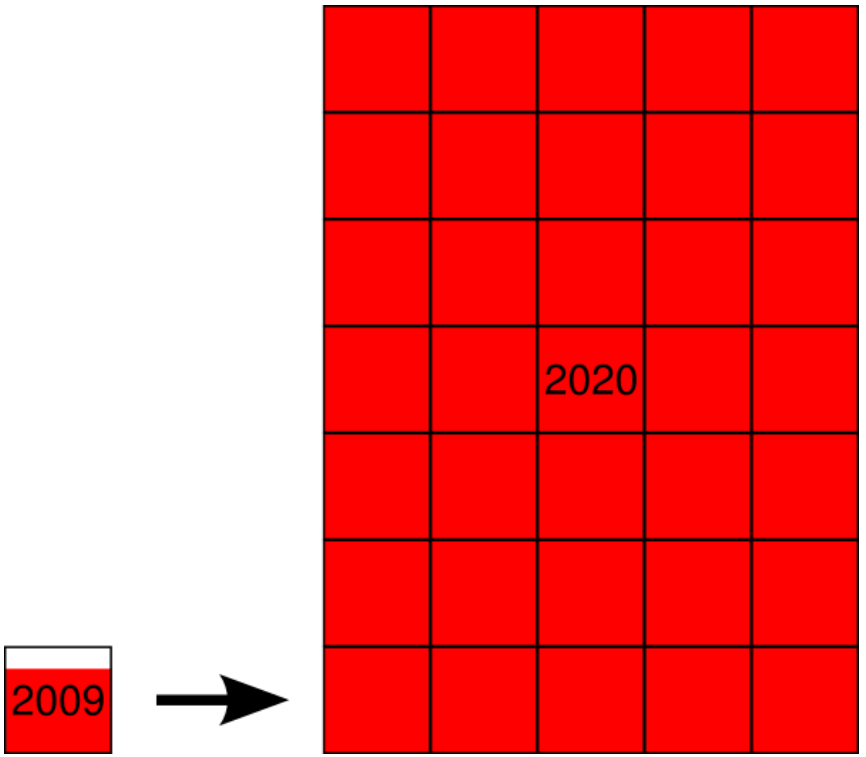
0,8 Millionen Petabytes (= Zettabytes)

2010:

1,2 ZB

2020:

~35 ZB



Haltbarkeit von Datenträgern

- Datenhaltbarkeit sinkt seit den Steintafeln
- Teilweise extreme produktionsbedingte Fluktuation der zu erwartenden Lebensdauer

Medium	(geschätzte) Haltbarkeit in Jahren
Steintafeln	> 1.000
Optische Medien (gepresst)	~ 100
Magnetbänder	~ 30
Optische Medien (gebrannt)	~ 15

Digitale Obsoleszenz

Kompatible Lesegeräte sterben aus

- Disketten, Kassetten, Lochbänder, …

Datenformate werden unlesbar

- ASCII $\leftrightarrow$ EBCDIC, MS Word, AppleWorks, …

Digital Rights Management erschwert Zugriff

- E-Books, Filme, Musik, Spiele, …

Archivieren, aber wie?

Vier unterschiedliche Techniken zur Sicherung

- Erneuerung der Daten auf frischen Datenträgern
- Migration auf neue Systeme
- Replikation auf mehrere (gleichartige) Systeme
- Emulation der Originalsysteme

Integrität der Daten wahren

- Metadaten und Prüfsummen

Erneuerung der Daten

Vorteile:

- Originaldaten bleiben erhalten, keine Konversion irgendeiner Art
- Dadurch relativ einfacher Vorgang, sofern alle nötigen Mittel vorhanden sind

Auftretende Probleme:

- Datenträger noch verfügbar?
- Lese- und Schreibgeräte funktionstüchtig?
- Gilt beides auch noch für die Zukunft?

Migration auf neue Systeme

Vorteile:

- Es stehen zukunftssichere Formate zur Verfügung, eventuell ist inzwischen ein Umstieg auf offene Formate möglich

Auftretende Probleme:

- Großer Zeit- und Arbeitsaufwand
- Möglicher Verlust von Originalfunktionalität
  - Neues Format unterstützt bestimmte Funktion nicht
  - Ursprüngliches, proprietäres Format kann nicht mehr zuverlässig gelesen werden

Replikation auf mehrere Systeme

Vorteile:

- Dezentralisierung der Daten
- Wahrscheinlichkeit des vollständigen Datenverlusts sinkt

Auftretende Probleme:

- Erhebliche Erschwerung der Erneuerung, Migration und Zugriffskontrolle
- Bei Replikation auf gleichartige Systeme treten die selben Probleme wie bei der Erneuerung auf
- Vervielfachung der Kosten

Emulation der Originalsysteme

Vorteile:

- Konvertierung selten nötig
- Im Idealfall ist der Emulator universell einsetzbar

Auftretende Probleme:

- Teilweise extreme Mehrarbeit im Vergleich zu Erneuerung und Migration
- Geschwindigkeitsverlust
- Vollständige Kompatibilität praktisch nicht erreichbar

Tatsächlich angewandte Techniken

Viele Langzeitarchive setzen auf eine Mischung aus allen verfügbaren Techniken

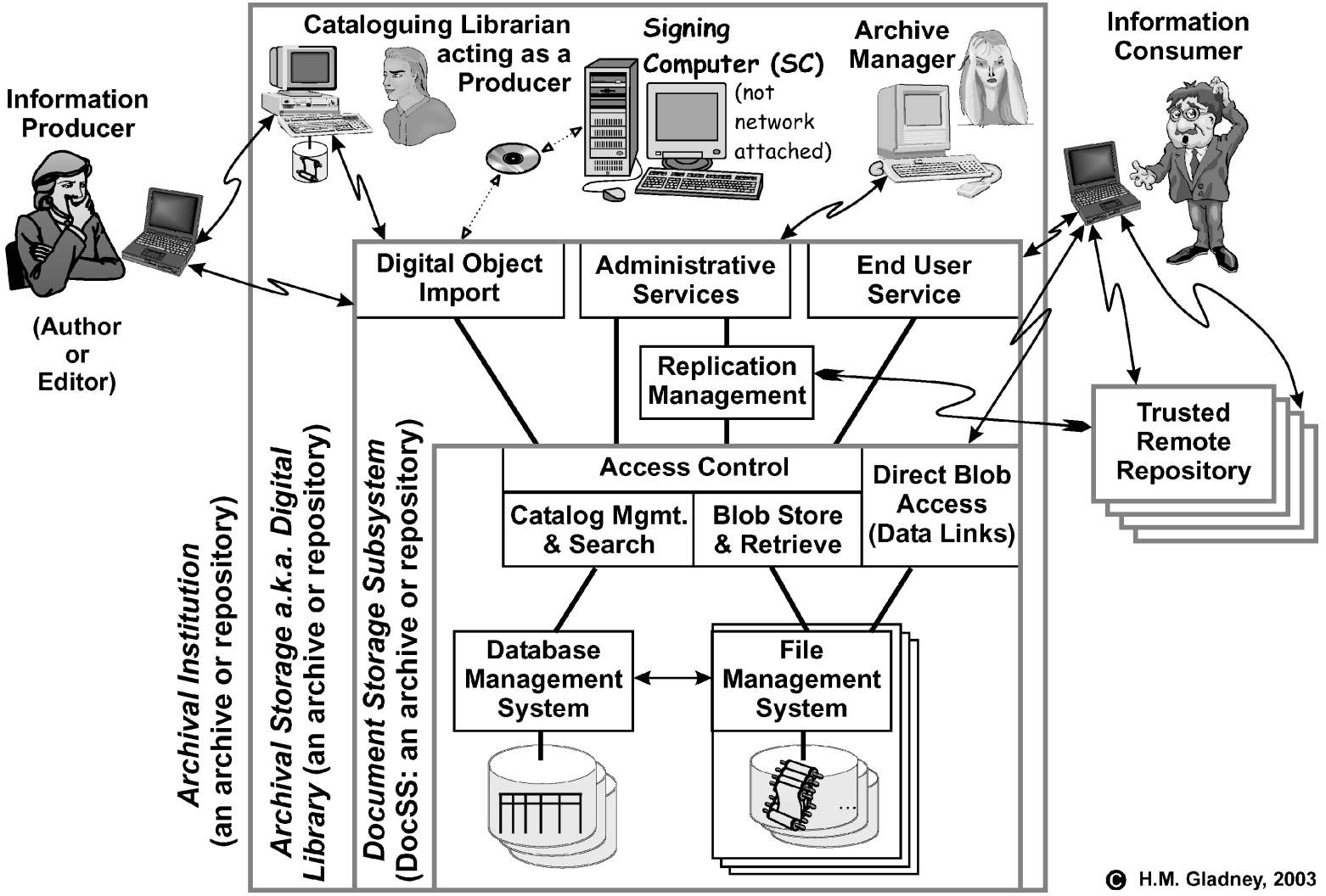
- Je nach geschätzter Zukunftssicherheit der verwendeten Formate Erneuerung oder Migration
- Replikation, eventuell Zusammenarbeit mit anderen LZA
- Emulation, wenn keine andere Technik möglich ist
  - Hauptsächlich verwendet bei proprietärer Software

Integrität der Daten wahren

Daten müssen zuverlässig vor Veränderung (mutwillig oder durch Verfall) geschützt werden

- Anhängen von Metadaten an die zu sichernden Nutzdaten
  - Problem: Welches Datenformat eignet sich? Wie schützt man die Metadaten vor Änderungen?
- Bilden von Prüfsummen über die Originaldaten
  - Mehrere, sich ergänzende Prüfsummen einsetzen, wie z.B. MD5, SHA-1, CRC, …
  - Durch redundant gehaltene Daten ist zuverlässigere Wiederherstellung möglich
  - Daten kryptographisch signieren, um Vertrauen zu schaffen

Ein (relativ sicheres) Langzeitarchiv



„Digital Dark Ages“

- Große Teile des zeitgenössischen Kulturguts liegen (ausschließlich) digital vor
- Heute wichtige „neue“ Dokumente, in Zukunft von historischem Wert, können verloren gehen
- Daten von hohem ideellem Wert, wie z.B. Digitalfotos, werden selten wirklich zukunftssicher gespeichert
- These: Wir sind schon mitten drin

Empfehlenswerte Schritte

Zukünftigen Datenverlust vermeiden:

- Offene Standards schaffen
  - Internationale Organisationen wie ISO, IEC, …
  - Peer-Review-fördernde Veröffentlichungen wie RFCs
- Freie Formate und Systeme verwenden
  - Durch freie Formate ist gewährleistet, dass Daten interpretierbar bleiben, so lange die Beschreibung des Formats bekannt ist
  - „Embrace, Extend, Extinguish“ kann relativ einfach verhindert werden

Geeignete Datenformate I

Dokumente:

- PDF/A
  - Keine interaktiven oder multimedialen Inhalte
  - Nötige Schriftarten eingebettet
  - Obligatorische, standardisierte Metadaten-Felder
- XML
  - Menschen- und maschinenlesbar, da ASCII-kodiert
  - Metaformat, also „Grundgerüst“ für spezialisierte Formate wie SVG, MathML, XHTML, …

Geeignete Datenformate II

Mediendateien (verlustfrei):

- Portable Network Graphics (PNG):
  - \+ Offener Standard, geeignet für alle Rastergrafiken
  - – Ineffiziente Kompression bei Fotos

- Free Lossless Audio Codec (FLAC):
  - \+ Freie Software, relativ hohe Kompression (auf ~50%)
  - – Von wenigen Herstellern akzeptiert, da DRM-frei

Geeignete Datenformate III

Videos:

- HuffYUV
  - + Verlustfrei, offener Standard
  - – Relativ unbekannt, kein Release seit 2002
- VP8
  - + (Jetzt) offen, da Google-gefördert wohl in Zukunft stark verbreitet
  - – Noch nicht beim Consumer angekommen
  - – verlustbehaftet

… und die Hardware?

Freie Hardware mit offen vorliegenden Bauplänen

- Existiert bereits, teilweise sehr erfolgreich
  - Arduino
  - RepRap, Makerbot
  - Open Cores
  - …
- Komplett freie Desktop-Computer momentan noch nicht möglich

Persönliche Empfehlung

- Zu archivierende Daten gut auswählen
- Je nach gewünschter Tragbarkeit und Datenmenge entweder HDDs oder LTOs
  - HDD: ~35€/TB
    - HDDs benötigen kein weiteres Laufwerk, sind aber empfindlicher als Tapes
  - LTO-4: ~20€/TB + ~1.000€ Laufwerk
    - LTOs lassen sich mit Schubkarren transportieren, immense Bandbreite
- Daten in freien Formaten auf min. 2 Datenträgern ablegen und die Datenträger getrennt aufbewahren
  - Daten können selbstverständlich verschlüsselt gespeichert werden

Quellen

- http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm
- http://web.archive.org/web/20031222194846/http://www.oit.umass.edu/publications/at_oit/Archive/fall98/media.html
- Kuny, Terry - A Digital Dark Ages? Challenges in the Preservation of Electronic Information
- Gladney, Henry M. - Trustworthy 100-Year Digital Objects: Evidence After Every Witness Is Dead
- Pinheiro, E. und Weber, W. und Barroso, L. A. - Failure Trends in a Large Disk Drive Population
- nestor-Kriterien: Kriterienkatalog vertrauenswürdige digitale Langzeitarchive
- http://de.wikipedia.org/wiki/Langzeitarchivierung