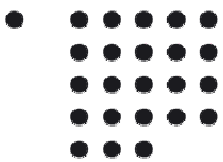
Einführung in die digitale Langzeitarchivierung



Fachhochschule Köln
Cologne University of Applied Sciences

Institut für Informationswissenschaft
Institute of Information Science

Prof. Dr. Achim Oßwald
Fachhochschule Köln
Institut für Informationswissenschaft
achim.osswald@fh-koeln.de

Problemstellungen (1)

Datensicherung im privaten Bereich

Wie lange werden wir diese Medien nutzen können?

Quelle der Grafiken: Wikipedia

Problemstellungen (2)

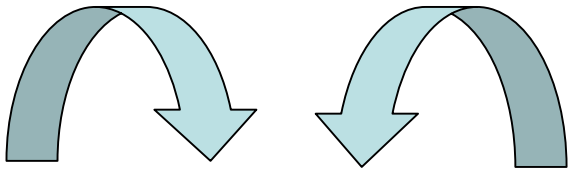
- **Ursachen für Datenverlust:**

44%	Hardware- und Systemfehler
32 %	Menschliches Versagen
14 %	Softwarefehler
7 %	Viren
3 %	Naturkatastrophen

Angaben von data recovery company Ontrack (1996), nach R. Scheffel
Quelle: Ross, Seamus, Gow, Ann: Digital Archaeology: Rescuing Neglected and Damaged Data Resources; JISC/NPO Studie 2/1999

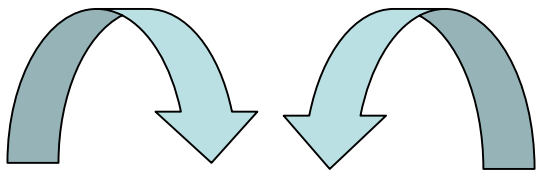
Problemstellungen (3)

Haltbarkeit der Datenträger



Verfügbarkeit der Wiedergabegeräte

Verfügbarkeit von und Nutzungskompetenz für Anwendungssoftware und ihre Dateiformate



Verfügbarkeit von und Nutzungskompetenz für Betriebssystem-software

?

Kompatibilität / Nutzbarkeit / Authentizität im Hinblick auf Langzeitverfügbarkeit

Software-Inkompatibilitäten

	Datei-format 1	Datei-format 2	Datei-format 3	Datei-format 4
Programm-version 1.0	paßt	✗	✗	✗
Programm-version 2.0	paßt noch	paßt	✗	✗
Programm-version 3.0	⚠	paßt noch	paßt	✗
Programm-version 4.0	✗	⚠	paßt noch	paßt

Quelle: Dobratz: CMS-Kolloquium, 16.5.2006

Überblick

Aktuelle Fragen der Langzeitarchivierung (LZA) und Langzeitverfügbarkeit (LZV)

- **Was ist eigentlich LZA digitaler Objekte?**
- **Was soll langfristig gesichert werden?**
- **Wer soll die Sicherung durch LZA übernehmen?**
- **Wie soll gesichert werden?**
  Oder: Welche Techniken und Strategien gibt es für die LZA digitaler Objekte?
- **Was wird / könnte das kosten?**
- **Was wird derzeit zur Langzeitarchivierung digitaler Objekte schon unternommen?**
- **Was sind derzeit die zentralen offenen Fragen?**

Was ist eigentlich Langzeitarchivierung digitaler Objekte?

- "Langzeitarchivierung digitaler Objekte umfasst alle Maßnahmen, die dazu dienen, digitale Objekte für die Nachwelt dauerhaft zu erhalten. Der Begriff ist eng verwandt mit Langzeitverfügbarkeit, die jedoch die dauerhafte Benutzbarkeit mehr in den Vordergrund stellt."
  Quelle: *nestor-Glossar; Stand 20.2.2004*
- „'Langzeit' ist die Umschreibung eines nicht näher fixierten Zeitraumes, währenddessen wesentliche nicht vorhersehbare technologische und soziokulturelle Veränderungen eintreten, die sowohl die Gestalt als auch die Nutzungssituation digitaler Ressourcen in rasanten Entwicklungszyklen vollständig umwälzen werden."
  aus: *Schwens/Liegmann in: Grundlagen der praktischen Information und Dokumentation; München 2004;*
  zitiert nach: *nestor-Glossar; Stand 20.2.2004*

Was soll gesichert werden? (1)

Objektbezogene Perspektive

- **Printprodukte**
  Pflichtexemplarbibliotheken / Bibliotheken mit speziellen Sammelschwerpunkten (DNB, LBen, Zentrale Fachbibliotheken / SSG-Bibliotheken)
- **Digitale Objekte** auf **physischen Datenträgern**
- **Online** verfügbare digitale **Objekte**
- **Archivmaterialien**
  verwaltungsbezogen zuständige Archive
- **Museumsobjekte**
  prinzipiell jedes Museum

Was soll gesichert werden? (2)

Das digitale Erbe unserer Gesellschaft …

- ... umfasst kulturelle, erzieherische, wissenschaftliche, administrative Quellen, aber auch technische, juristische, medizinische oder andere Arten von Information, die digital erzeugt oder digitalisiert wurden
  nach:
  UNESCO Charta zum Erhalt des digitalen Kulturerbes
- Digitale Materialien, d.h. Texte, Datenbasen, Foto, Film, Graphiken, Software, Webpages

=> Unter Bezugnahme auf die Sammelrichtlinien der jeweiligen Einrichtung

Was soll gesichert werden? (3)

Nationale und länderspezifische Perspektive

- **Digitale Publikationen auf physischen Datenträgern:** Sammelauftrag der DNB (bei deutscher Sprache und Bezug zu Deutschland) sowie anderer Pflichtexemplarbibliotheken
- **Netzpublikationen, soweit öffentlich zugänglich:** Sammelauftrag der DNB seit 22.6.2006 (“Medienwerke in unkörperlicher Form”; soweit Bezug zu Deutschland) sowie Zielsetzung einzelner Landesbibliotheken

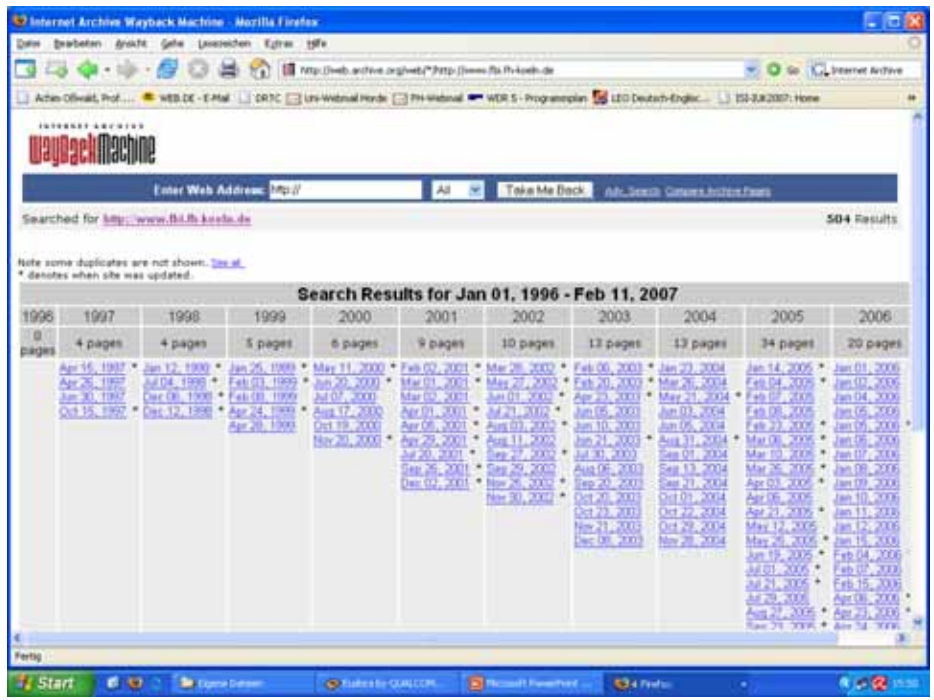
Was soll gesichert werden? (4)

- 3Sat-Film zur Aufgabenstellung der DNB

Was soll gesichert werden? (5)

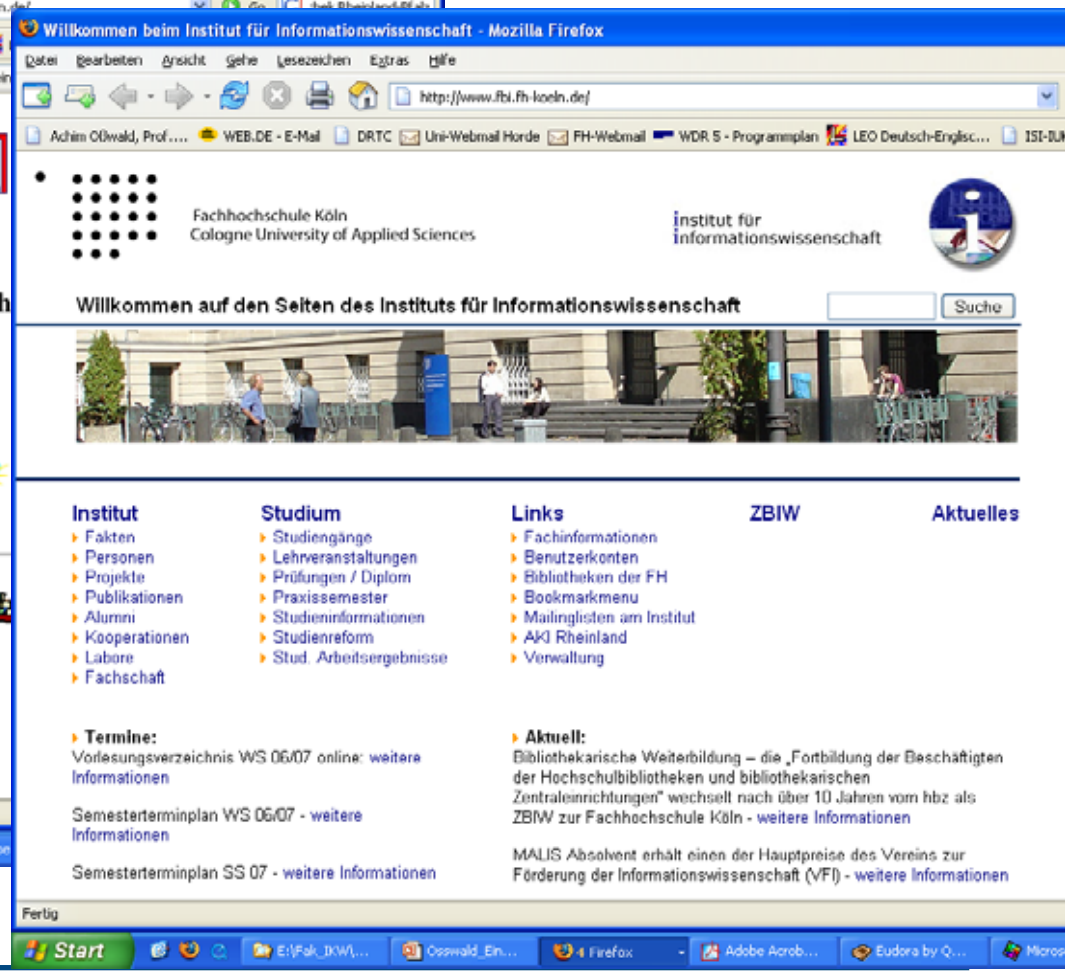
Projekt- und produzentenbezogene Perspektive

- Internet Archive (http://www.archive.org) sammelt Momentaufnahmen von Webseiten und den darüber angebotenen Materialien



- Verlage: sammeln digitale Publikationen ihres Hauses

Exkurs: Beispiele aus Internet Archive



Was soll gesichert werden? (6)

Aspekte eines digitalen Objekts

Inhalt: Information in Schrift, (bewegten) Bildern, Ton, ...

Struktur: Aufteilung und Abfolge der Daten (z.B. DTD)

Layout: Visualisierung des Inhalts + der Struktur (z.B. Stylesheets)

Metadaten: Objekteigenschaften + deren Veränderungen (z.B. DC)

Orientiert an einer Folie von S. Dobratz aus Neuroth 29.11.2005

Wer soll die LZA übernehmen? (1)

- Prinzipiell kann jede Organisation diese Aufgabe übernehmen
- **Anforderungen gemäß der nestor-AG "Vertrauenswürdige Archive"**
  http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf

1. **Organisatorischer Rahmen**
   Zielentsprechende Verfügbarkeit von Ressourcen wie Personal, Finanzen, rechtliche und methodische Absicherung der LZA

2. **Umgang mit den Objekten und ihren Metadaten**
   Integrität, Authentizität, Verfügbarkeit, Vertraulichkeit

3. **Infrastruktur und Sicherheit**

Wer soll die LZA übernehmen? (2)

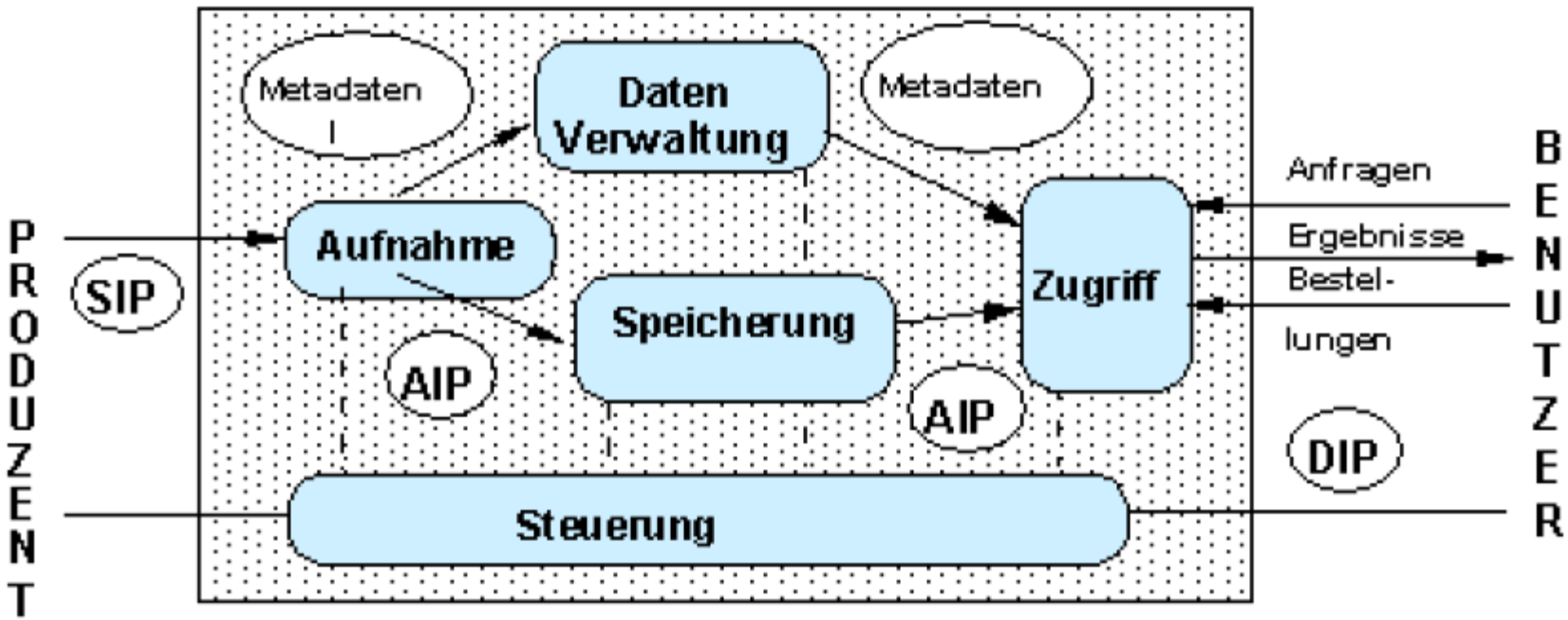
- **Staatlich finanzierte Einrichtungen:**
  - Archive (Bundes- und Landesebene)
  - Museen
  - Bibliotheken: Bundesebene = DNB; Länderebene z.B. Rheinische LB Koblenz; Württbg. LB Stuttgart; Bayr. Staatsbibl. München; Fachebene z.B. TIB, SSG-Bibliotheken
- **Privatrechtlich finanzierte Einrichtungen:**
  - Internet Archive
  - ?
- **ABER:**
  Einzelne Einrichtungen sind technisch und finanziell mit Realisierung überfordert;
  die kooperative Entwicklung und Umsetzung der Methoden zur Erschließung und Archivierung bleibt Zielsetzung

Wie soll gesichert werden?

- Eigentliches Ziel der Archivierung:
  "intellectual preservation" (Graham)
- Kopieren und manipulieren sind im digitalen Kontext extrem einfach und auf Nachfolgeversionen übertragbar
- Authentifizierung der Originalversion durch die Urheber unter Einbeziehung von Quelle, Erstellungstermin und Inhalt
- Diverse Sicherungsverfahren in Erprobung:
  - Kryptographie – problematisch wg. Zugang
  - Hashing – mathematisch Codierung
  - Hashing + Zeitstempel
  - verborgener Code z.B. Wasserzeichen

  Problem: Sicherungsverfahren kollidieren z.T. mit den technischen Archivierungsverfahren

OAIS: Modell für LZA + LZV



Open Archive Information System (OAIS)
Auch interaktiv als javascript über die Nedlib-Homepage unter
http://nedlib.kb.nl/ =>results => Functional Process Model

Techniken und Strategien der LZA (1)

- **Auffrischung der Datenträger (“digital refreshing”):**
  regelmäßiges Umkopieren der Daten eines Datenträgers auf einen Neuen gleichen Typs oder einen Datenträger mit gleichem Speicherverfahren

- **Problem:**
  LZA ist so nicht möglich; grundsätzliche Probleme der Archivierung (Verfügbarkeit von HW + SW) bleiben erhalten; eher ein Verfahren der Datenerhaltung in Vorbereitung auf LZA

Techniken und Strategien der LZA (2)

- **LOCKSS = Lots of Copies Keep Stuff Safe**
  Kopien einer Datei werden auf verschiedenen, räumlich getrennten IT-Systemen gespeichert; regelmäßiger Vergleich der Dateien, um mögliche Veränderungen entdecken zu können; Alarm beim Unterschreiten einer Mindestzahl von Kopien
- **Vorteil:**
  prinzipielle Datensicherung auch in verschiedenen Betriebssystem-Umgebungen
- **Problem:**
  Anbieter (Verleger) der digitalen Objekte müssen dem Kopierverfahren zustimmen; Funktionalitätsunterschiede bleiben u.U. unentdeckt

Techniken und Strategien der LZA (3)

- **Erhalt der technologischen Umgebung ("technisches Museum"):**
  ursprüngliche Hard- und Software-Kombination wird in ihrer Funktion erhalten

- **Vorteil:** geringe Gefährdung der Authentizität
- **Problem:**
  Funktionsfähigkeit von HW + SW ist nicht garantiert;
  Reparatur und/oder Ersatz unwahrscheinlich;
  Personen mit Bedienungskompetenz nicht fortwährend verfügbar

Techniken und Strategien der LZA (4)

- **Encapsulation**
  Ein image-file der ursprünglichen HW- und SW-Umgebung + der zu sichernden Anwendung wird angelegt sowie regelmäßig gesichert;
  Ablauf ggf. in einem virtuellen Computer (Universal Virtual Machine)

- **Vorteil:**
  geringe Gefährdung der Authentizität
- **Problem:**
  Leistungsunterschiede der HW- und Betriebssystem-Umgebungen; Kompatibilität

Techniken und Strategien der LZA (5)

- **Emulation**
  Nachahmung der ursprünglichen HW-, SW- und Betriebssystem-Umgebung;
  Metadaten der Konfiguration werden mit archiviert

- **Vorteil:**
  Kosten fallen für jeden Konfigurationsverbund nur einmalig, nämlich bei der ersten Emulation in einer neuen HW-, SW- und Betriebssystem-Umgebung an
- **Problem:**
  Leistungsunterschiede der HW- und Betriebssystem-Umgebungen, Kompatibilität

Techniken und Strategien der LZA (6)

- **Migration**
  regelmäßige Übertragung des digitalen Objekts von einer HW- und SW-Umgebung in die technologisch nächst-folgende Umgebung; Metadaten der Konfiguration und des Migrationsschritts werden mit archiviert
- **Vorteil:**
  geringere Technologiesprünge; “alte” Funktionsumgebung ist zur Rekonstruktion der Funktionalität noch verfügbar
- **Problem:**
  Kosten fallen für jeden Anwendungsfall und jeden Migrationssschritt an; ggf. Funktionsverluste in Zusammenhang mit neuer HW- und SW-Umgebung

Techniken und Strategien der LZA (7)

- **“Desiccated data” (“gedörrte Daten”) als komplementärer Teil eines Sicherungskonzepts**
  neben der Sicherung der Originaldatei sowie eines Rasterbildes der angezeigten Anwendung auch Sicherung einer ASCII-Datei mit XML-Auszeichnungen (oder “plain”)

- **Vorteil:**
  Auffangvariante zur Rekonstruktion zumindest des ursprünglichen Inhalts und/oder des Erscheinungsbildes der Anwendung
- **Problem:**
  US-westlich orientiert; ASCII statt Unicode

Techniken und Strategien der LZA (8)

- **Konversion**
  Übertragung der digitalen Daten auf analoge, vom Menschen lesbare, alterungsbeständige Speichermedien / Datenträger (Papier; Mikrofilm); im privatwirtschaftlichen Bereich sehr beliebt

- **Vorteil:**
  sehr kostengünstig
- **Problem:**
  Verlust des medienspezifischen informationellen Mehrwertes z.B. Hyperlinks; spezifische Navigations- und Suchmöglichkeiten

Was wird / könnte LZA kosten? (1)

- **Grundannahme:**
  Archivierung und Verwaltung digitaler Objekte ist aufwändiger als bei gedruckten Publikationen
- Bislang kaum langfristige Erfahrungswerte, allenfalls sind **Kostenmodelle** verfügbar
  z.B.:
  Kalkulation im Projekt
  JSTOR (The Scholarly Journal Archive):
  jährlich 25 000 US $ / Zeitschrift für Archivierung

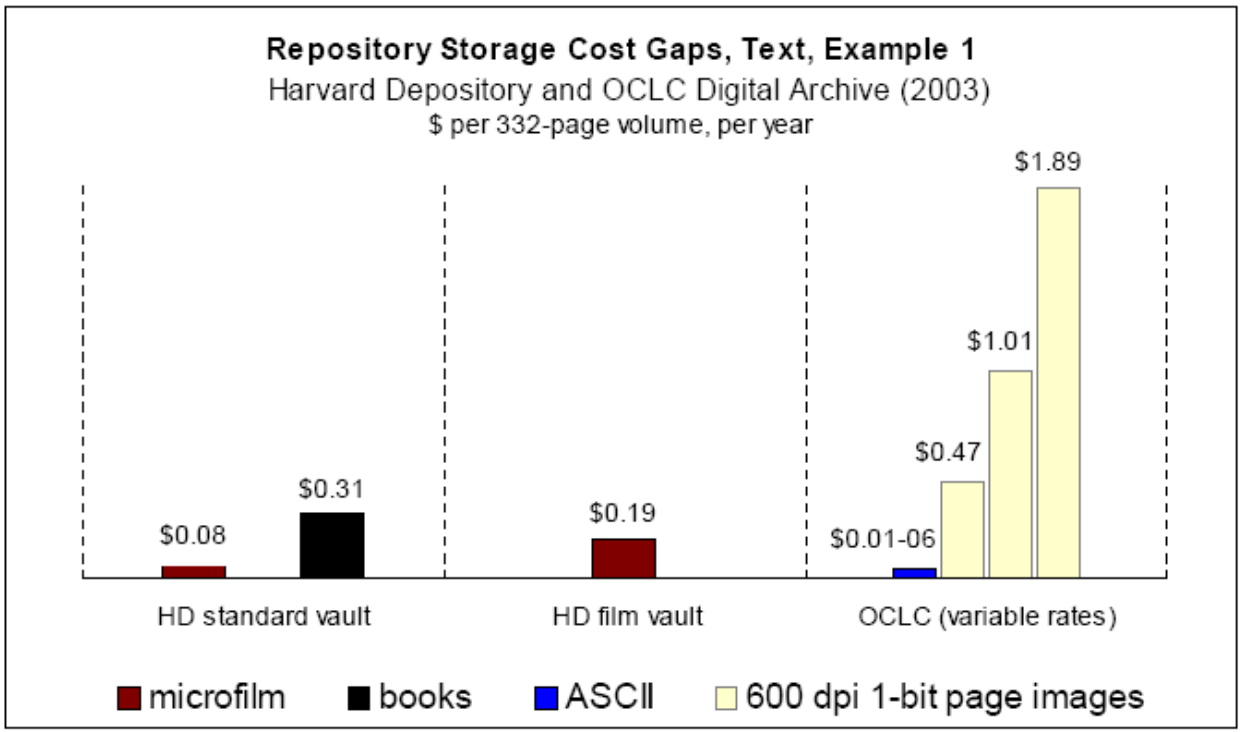
Was wird / könnte LZA kosten? (2)

Kostenfaktoren (nach Keller, Alice: Elektronische Zeitschriften, Grundlagen und Perspektiven, Wiesbaden 2005, 255f)

- ***Einmalige Kosten:***
  - "Auswahl des zu archivierenden Materials
  - Prüfung der Rechtslage
  - Festlegung der technischen Lösung …
  - Prüfung der Vollständigkeit der digitalen Objekte
  - Herstellung oder Ergänzung von Metadaten"
- ***Langfristig wiederkehrende Kosten:***
  - "Erneuerung von Hard- und Software
  - Speicherung der Dateien, regelmäßiges Überspielen oder Migration der Daten in neue Computerumgebungen
  - Verwalten des Archivs
  - Regelmäßiges Überprüfen der Sammel- und Achivierungspolitik"

Was wird / könnte LZA kosten? (3)

Chart 1. Relative Costs to Store Text "Masters":
Microfilm, Print, ASCII, and 600 dpi 1-bit Group 4 Page Image Formats



Aus: **Chapman,** Stephen: Counting the Costs of Digital Preservation: Is Repository Storage Affordable? – In: Journal of Digital information, volume 4 issue 2 Themes: Digital libraries 2003-05-07 (Peer reviewed paper), p. 7

Was wird / könnte LZA kosten? (4)

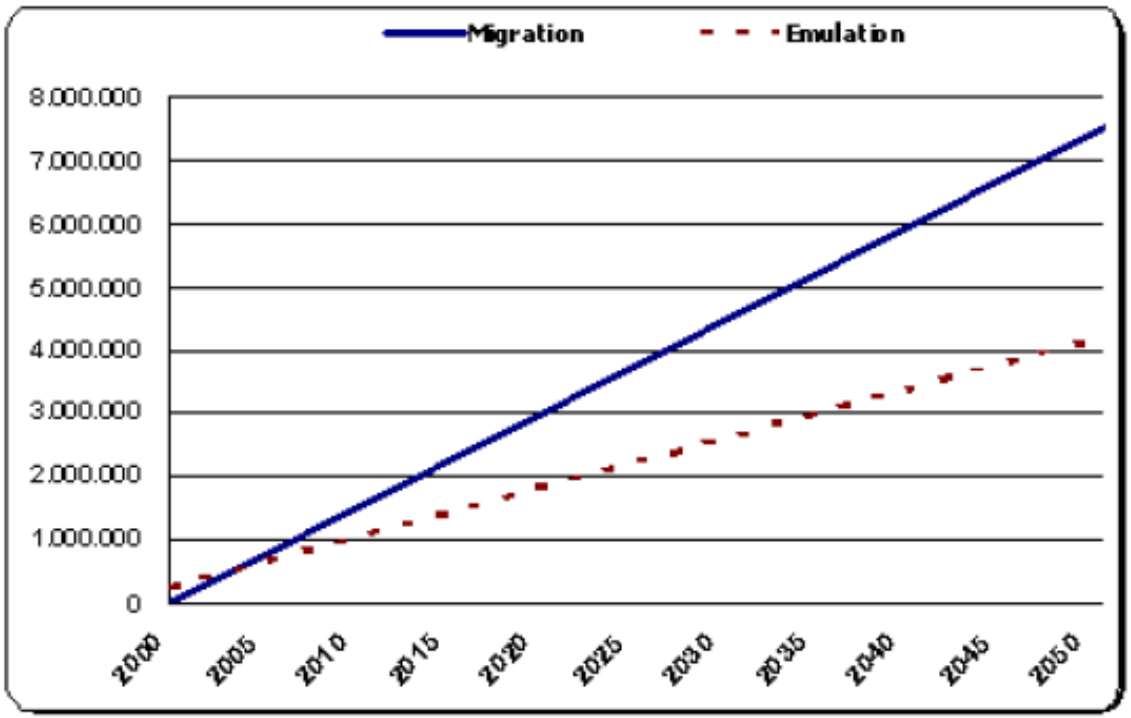


Figure 2: Costs for migration (blue line) and emulation (red dotted line) for maintaining an archive of 1,000,000 digital objects over a period of 50 years. The initial investments of setting up an emulation tool yields high costs in the first five years. But soon after that, the migration costs are higher than the emulation costs and the difference increases every year. In 50 years, the migration costs are 45% higher than the emulation costs.

Aus:
Erik **Oltmans** / Nanda **Kol**:
A Comparison Between Migration and Emulation in Terms of Costs. –
In: RLG DigiNews April 15, 2005, Volume 9, Number 2

Was läuft derzeit bzgl. LZA? (1)

… in Deutschland

- **Projekt nestor**
  Kompetenznetzwerk Langzeitarchivierung
  http://www.langzeitarchivierung.de

„Nestor, der Berater der Griechen in Troja, steht als Symbol für die beratende und unterstützende Funktion des Kompetenznetzwerkes im Bereich der Langzeitarchivierung und Langzeitverfügbarkeit. Aufgelöst ergibt das Akronym nestor "Network of Expertise in long-term STOrage and availability of digital Resources in Germany„ ...“

(www.langzeitarchivierung.de/modules.php?op=modload&name=PagEd&file=index&page_id=10)

Was läuft derzeit bzgl. LZA? (2)

- **nestor Partner**
  - Die Deutschen Bibliothek (Federführung)
  - Bayerische Staatsbibliothek München
  - Computer- und Medienservice / UB der HU zu Berlin
  - Institut für Museumsforschung, Staatliche Museen Berlin – Stiftung Preußischer Kulturbesitz
  - Niedersächsische SUB Göttingen
  - Bundesarchiv (seit 2005)
  - Fernuniversität Hagen (seit nestor II - 2006)
- Projektträger: BMBF
- Empfehlungen zum Abschluss von nestor I:
  „Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland“
  http://www.langzeitarchivierung.de/downloads/memo2006.pdf
- Nestor II seit 2006; weitere Infos über website und Newsletter

Was läuft derzeit bzgl. LZA? (3)

- **KOPAL - Kooperativer Aufbau eines Langzeitarchivs Digitaler Informationen**

http://kopal.langzeitarchivierung.de

„Ziel des Projektes kopal ist der Aufbau einer technischen und organisatorischen Lösung, um die Langzeitverfügbarkeit elektronischer Publikationen zu sichern. Dabei spielt die transparente Integration in vorhandene Bibliotheksysteme und die Nachnutzbarkeit durch Gedächtnisorganisationen eine wesentliche Rolle.“ (ebd.)

Was läuft derzeit bzgl. LZA? (4)

… in anderen Ländern

- ***PADI** Preserving Access to Digital Information*
  der National Library of Australia
  http://www.nla.gov.au/padi
- Dort gute Übersicht zu den Aktivitäten in einzelnen Ländern
- Methodisch & technologisch führend sind u.a. Australien, GB und NL

Zentrale offene Fragen (1)

- Infrastruktur des Archivsystems:
  Inwieweit ist das OAIS-Reference Model wirklich langfristig funktionsfähig?
- Standards zur Dokumenterstellung u. -kodierung:
  Ziel: Kostensenkung bei Emulation o. Migration durch Nutzung von Auszeichnungssprachen
- Metadatenstandards z.B. auf der Basis des Dublin Core Element Set und Metadatenmanagement
- Anforderungen an Dokumentformate:
  “Archivierungsfreundlichkeit”

Zentrale offene Fragen (2)

- Rechtliche Fragen des Zugriffs / DRM
- Dokumentidentifikation mittels persistenter Identifikatoren (z.B. DOI; URN)
- Transferstandards und -protokolle
- Funktionsverteilung und Organisation zwischen Produzenten und Anbietern digitaler Objekte einerseits und Organisationen der LZA andererseits
- Finanzierung der LZA
- Formen der Zusammenarbeit auf internationaler, nationaler und regionaler Ebene

Weiterführende Quellen & Hinweise

- ZfBB-Themenhefte aus den Jahren 2001 (3-4) und 2005 (Heft 3-4)
- Borghoff, Uwe M. et al.: Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente, Heidelberg 2003
- *http://deposit.ddb.de/netzpub/web_langzeiterhaltung_ep.htm*
- nestor-Projekt: *http://www.langzeitarchivierung.de*
  => *nestor – materialien*
  => *nestor - Informationsdatenbank*
- Nedlib-Projekt: *http://nedlib.kb.nl/*
  kopal-Projekt: *http://kopal.langzeitarchivierung.de*
- PADI: *http://www.nla.gov.au/padi*

Gerne beantworte ich Ihre Fragen!

achim.osswald@fh-koeln.de