

DIGITALE LANGZEITARCHIVIERUNG

OPEN SCIENCE WORKSHOP
11. OKTOBER 2018

Dr. Dirk Pollmächer
Universitäts- und Landesbibliothek Sachsen-Anhalt
dirk.pollmaecher@bibliothek.uni-halle.de



UNIVERSITÄTS- UND
LANDESBIBLIOTHEK
SACHSEN - ANHALT

<http://dx.doi.org/10.25673/13398>



ZURÜCK IN DAS JAHR 1973

Das World Trade Center in New York wird eingeweiht und Pink Floyd veröffentlichen „The Dark Side of the Moon“

- 8 Zoll Diskette gerade neu erfunden
- Keine einheitliche Textverarbeitung oder Tabellenkalkulation
- Selbstprogrammierte Anwendungen mit
- Sieben Bit Zeichencodierung

Dirk Pollmächer: Digitale Langzeitarchivierung



1. Wie komme ich an die physischen Daten?

Ferritkernspeicher

Lochkarten

Russische Rechentechnik

2. Wie sind die Daten zu interpretieren?

Datenformat

7 Bit Zeichensatz in Kyrillisch

Keine Dokumentation

Ohne Formatinformationen: Endloses Meer von Nullen und Einsen

3. Hat jemand die Daten systematisch aufbewahrt?

„Kiste mit unbeschrifteten Disketten auf dem Dachboden“

Suche nach richtiger Diskette = Suche nach Nadel im Heuhaufen

Originalaufnahmen der Mondlandung bis heute verschwunden...

PROBLEM 1: Langfristige Datenspeicherung (Bitstream Preservation)

Dirk Pollmächer: Digitale Langzeitarchivierung



Problem 1: Wie bewahre ich Rohdaten für 50 oder mehr Jahre sicher auf?

Schutz vor Datenverlust

Hardwareausfall

Menschen (!)

Hacker, Viren, ...

Sicherstellung der Integrität

Checksums (evtl. kryptographisch)

Lösungsansätze:

Nutzung langfristig lesbarer Datenträger (Microfiche, Papier?)
(eher schwierig und unsicher)

Regelmäßige Medienmigration

Mindestens zwei Kopien (LOCKSS) auf verschiedenen Datenträgern
Regelmäßiges Überprüfen und Umkopieren

PROBLEM 2: Datenformate

Dirk Pollmächer: Digitale Langzeitarchivierung



Ohne Formatinformationen sind Daten oft nutzlos.

Auch mit Formatinformationen wird häufig eine Ausführungsumgebung benötigt.

Interpress-Seitenbeschreibungssprache (später Postgres und Pdf)

CODASYL-Datenbank

Programm Quellcode

Auch SPS/R-Scripte

Speicherstand eines Computerspiels

Tausende Dateiformate bekannt

Lösungen:

(Computermuseum) oder

Emulation oder

Migration

Regelmäßige Umwandlung in neuere Formate

Oft nur Erhalt der signifikanten Eigenschaften möglich

(Konzentration auf langlebige, verlustfreie Formate)

PROBLEM 3: Langfristige organisatorische Pflege

Dirk Pollmächer: Digitale Langzeitarchivierung



Anfänge Ende der 60er, Anfang 70er Jahre

NASA hatte bereits über 140.000 Magnetbänder zu verwalten
Erste digitale Archive (1973 Bundesarchiv)

Erste systematische Diskussionen Anfang der 80er Jahre

Idee: Informationen bei den Produzenten belassen

Systematische Entwicklung ab ca. 2000er Jahren

Entstehung diverser Standards sowie Langzeitarchive

Insbesondere OAIS ISO 14721

DIN 31644

OAIS-MODELL (Open Archival Information System)

Dirk Pollmächer: Digitale Langzeitarchivierung



Erste Version 2003 als ISO-Standard 14721

Überarbeitung 2012

Submission Information Package = Dissertation + CD

Archival Information Package = Metadaten (Descriptive Info) + Dissertation +
CD + andere Formate

Dissemination Information Package = Pdf + Metadaten

Ingest: Annahme der Daten, Überprüfung, Konvertierung, Metadaten

Archival Storage: Physische Speicherung der Daten

Data Management: Speicherung der Metadaten/Suche/Datenbank etc.

Access: Herausgabe der Daten, Konvertierung, Rechte, Suche usw.

Preservation Planning: Erhaltungsplanung, Migrationsprojekte etc.

Administration: Pflege des Gesamtsystems

SITUATION IN DEUTSCHLAND

Netzwerk verschiedener Kooperationspartner für die
Langzeitarchivierung digitaler Ressourcen

DIN 31644
Information und Dokumentation
Kriterien für vertrauenswürdige digitale
Langzeitarchive

Kriterienkatalog mit 34 Kriterien

nestor Siegel bisher nur 4 Einrichtungen in Deutschland

Dirk Pollmächer: Digitale Langzeitarchivierung



Nestor - Deutschlandweit rund 20 Einrichtungen beteiligt

Verschiedene Arbeitsgruppen z.B.

Bestandserhaltung, Elektronische Akte, Emulation, Formate,
Forschungsdaten, Recht, Zertifizierung uvm.

Gefördert durch BMBF 2003-2009

Nestor-Kriterienkatalog, Nestor-Siegel

Abstimmung mit internationalen Standards

DIN-Norm 2012 aus Kriterienkatalog entstanden

Beispielkriterien:

Verantwortung

Rechtliche Basis

Rechtskonformität

Finanzierung

Personal

Krisenregelung

Gewährleistung der Integrität

Gewährleistung der Authentizität

Technische Spezifikation der Pakete

Datenmanagement

Protokollierung

Metadatendefinition

IT-Infrastruktur

ANFORDERUNGEN ULB

Kategorie	Datenmenge	Formate	Beispiele
/1/	Gering (<1TB)	Einfach (z. B. PDF/A, TIFF, JPG, XML, WAV, TXT)	Hochschulschriften, Zweitveröffentlichungen
/2/	Hoch(>1TB)	Einfach	Digitalisate, Menadoc, Publikationen mit Archivrecht, statische Datenbankinhalte
/3/	Gering(<1TB)	Komplex (z.B. Word, Excel, eBooks, R-Skripte, JavaScript, Programme, proprietäre Formate, ...)	Allgemeine Forschungsdaten, Datenträger aus Büchern und Zeitschriften, dynamische Inhalte (Programme, Makros, Online-Datenbanken)
/4/	Hoch(>1TB)	Komplex	Rohdaten

Dirk Pollmächer: Digitale Langzeitarchivierung



Sehr viele verschiedene Daten mit z.T. sehr speziellen Anforderungen

Forschungsdaten – Medizin, Jura, Chemie, Physik

Formate – Texte, Bilder, Videos, Audio-Dateien, Geoinformationen, 3D-Scans, ...

Alte Datenträger – Disketten, CDs

Dynamische Inhalte – Webdatenbanken, Access-Datenbanken, ...

Einzeldateien mit wenigen KBytes (z.B. eine Veröffentlichung)

Sehr viele sehr kleine unterschiedliche Dateien (Webseite – mehrere Millionen Dateien)

Sehr große Datenmengen (MRT-Scans, Physikalische Daten, Bioinformatik)

⇒ Deutlich mehr Formate und Inhalte als Bücher

⇒ Digital Heritage Management

RAHMENBEDINGUNGEN

- Geringe personelle Ausstattung (0,75 VZÄ)
- Sehr inhomogene Anforderungen
 - Forschungsdaten
 - Digitalisate
 - Veröffentlichungen
 - Historische Datenträger
 - Webdatenbanken
- Viele Akteure, keine einheitliche Landesstrategie
 - Verschiedene Einzelbemühungen der Hochschulen
 - Unterschiedliche Reifegrade

UMSETZUNGSIDEEN

80/20-Regel: Beschränkung auf Basisdaten

- Digitalisate in Standardformaten (tif, jpeg)
- Veröffentlichungen in Standardformaten (pdf/A)
- Forschungsdaten in offenen Formaten (xml, csv, ...)
- Keine dynamischen Daten
- Alles andere: Verweis auf spezialisierte Angebote

Soweit möglich: Nutzung vorhandener Infrastruktur

- Storage: IT-Servicezentrum bzw. Vor-Ort-Partner
- Nachnutzung von DSpace (Forschungsdatenrepositorium) – wenn möglich
- Nach Möglichkeit Einführung neuer Systeme vermeiden – Pflegeaufwand!

BEISPIEL 1 – KRITISCHE EDITION

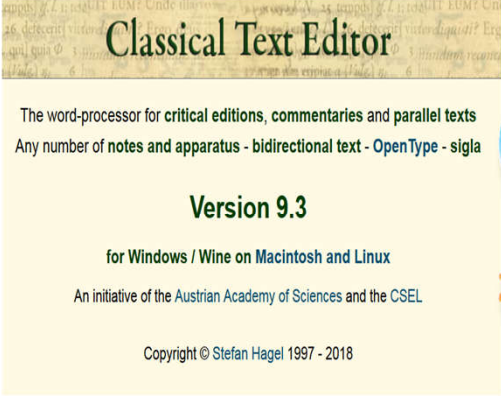
Übernahme aus DSpace oder vom
Author

Konvertierung in langzeit-
archivierbare Formate

- Standardisiertes XML
- PDF/A (ISO 19005)

Erstellung eines Archivs (AIP) mit
allen Inhalten und Metadaten

Speicherung auf Band und Platte



Classical Text Editor

The word-processor for critical editions, commentaries and parallel texts
Any number of notes and apparatus - bidirectional text - OpenType - sigla

Version 9.3

for Windows / Wine on Macintosh and Linux

An initiative of the Austrian Academy of Sciences and the CSEL

Copyright © Stefan Hagel 1997 - 2018

BEISPIEL 2 – MRT-STUDIE

Speicherung der Rohdaten (mehrere TByte) vor Ort

Konvertierung signifikanter Informationen in langzeitarchivierbare Formate, z. B.

- Unkomprimiertes TIFF
- CSV-Dateien

Erstellung eines Archivs (AIP) mit allen signifikanten Inhalten, Metadaten und Verweis auf Rohdaten

Speicherung auf Band und Platte

ALTERNATIV: NUTZUNG SPEZIALISierter DIENSTLEISTER

DISKUSSIONSPUNKTE

Wie hoch ist der Bedarf?

Wie realistisch ist die Beschränkung auf wenige wesentliche Datenformate?

Wie hoch wird der manuelle Aufwand z. B. für Konvertierungen und Datenaufbereitung sein und wer trägt ihn?

Kann DSpace (Share_it) als Grundlage für das Datenmanagement dienen oder ist Spezialsoftware notwendig (Archivematica, Ex Libris Rosetta)?

DAS MAGAZIN

- 1850 Erste Verhandlungen
 - 1871 Erster gescheiterter Umbauversuch
 - 1875 Studienreise Martin Gropius
 - 1876 Entwurf
 - 1878 Baubeginn
 - 1880 Fertigstellung
- Kosten 375.000 Mark
(ca. 2.5 - 7.5 Mio. Euro)



Dirk Pollmächer: Digitale Langzeitarchivierung

