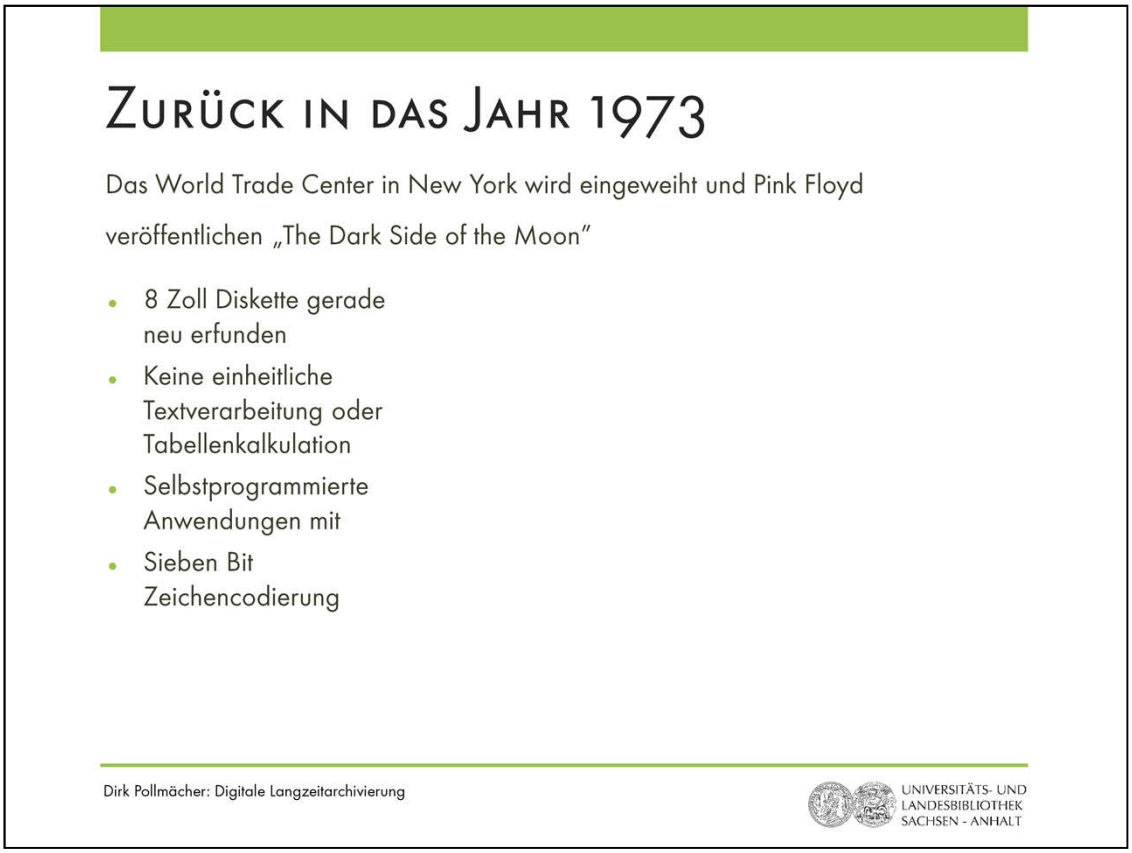
Digitale Langzeitarchivierung

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

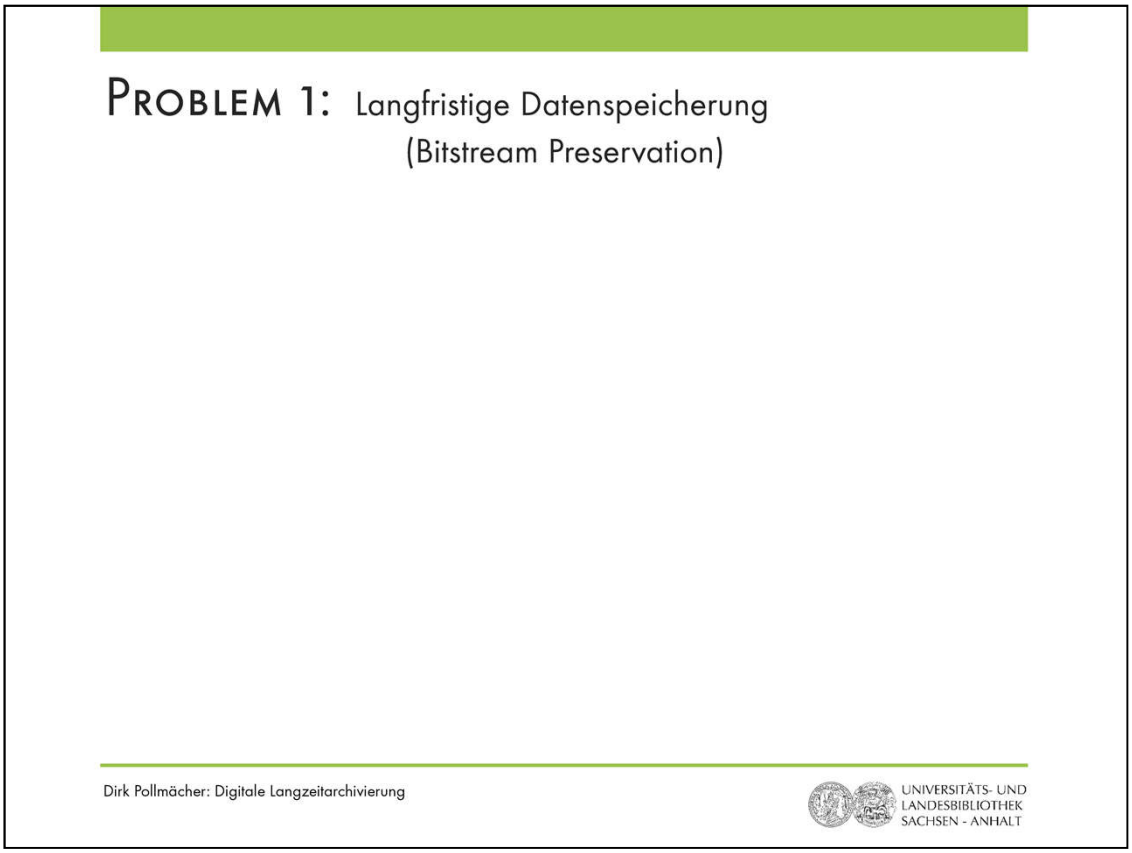


1. Wie komme ich an die physischen Daten?
    - Ferritkernspeicher
    - Lochkarten
    - Russische Rechentechnik

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

Schutz vor Datenverlust

- Hardwareausfall
- Menschen (!)
- Hacker, Viren, ...

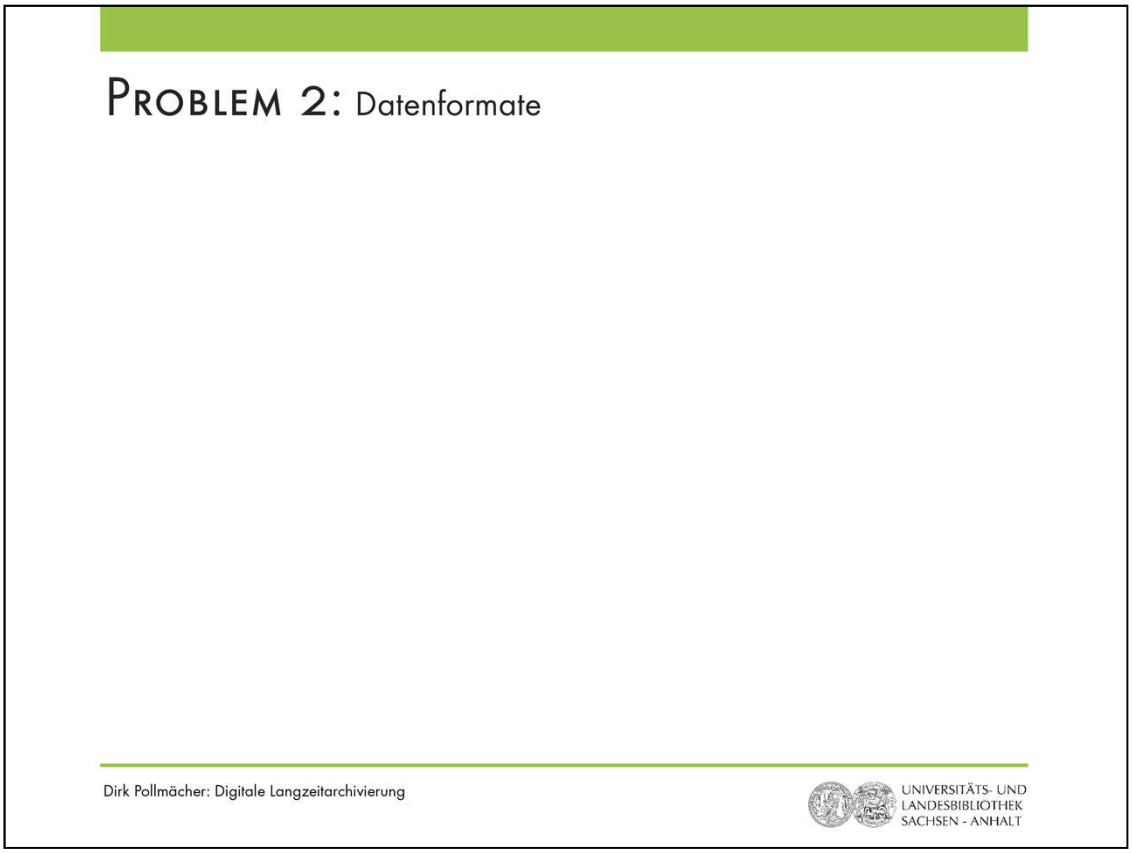
Sicherstellung der Integrität

- 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



Ohne Formatinformationen sind Daten oft nutzlos.

Auch mit Formatinformationen wird häufig eine Ausführungsumgebung benötigt.

- Interpress-Seitenbeschreibungssprache (später Postgres und Pdf)
- CODASYL-Datenbank
- Programmquellcode
  - 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

UNIVERSITÄTS- UND LANDESBIBLIOTHEK SACHSEN - ANHALT

Anfänge Ender 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

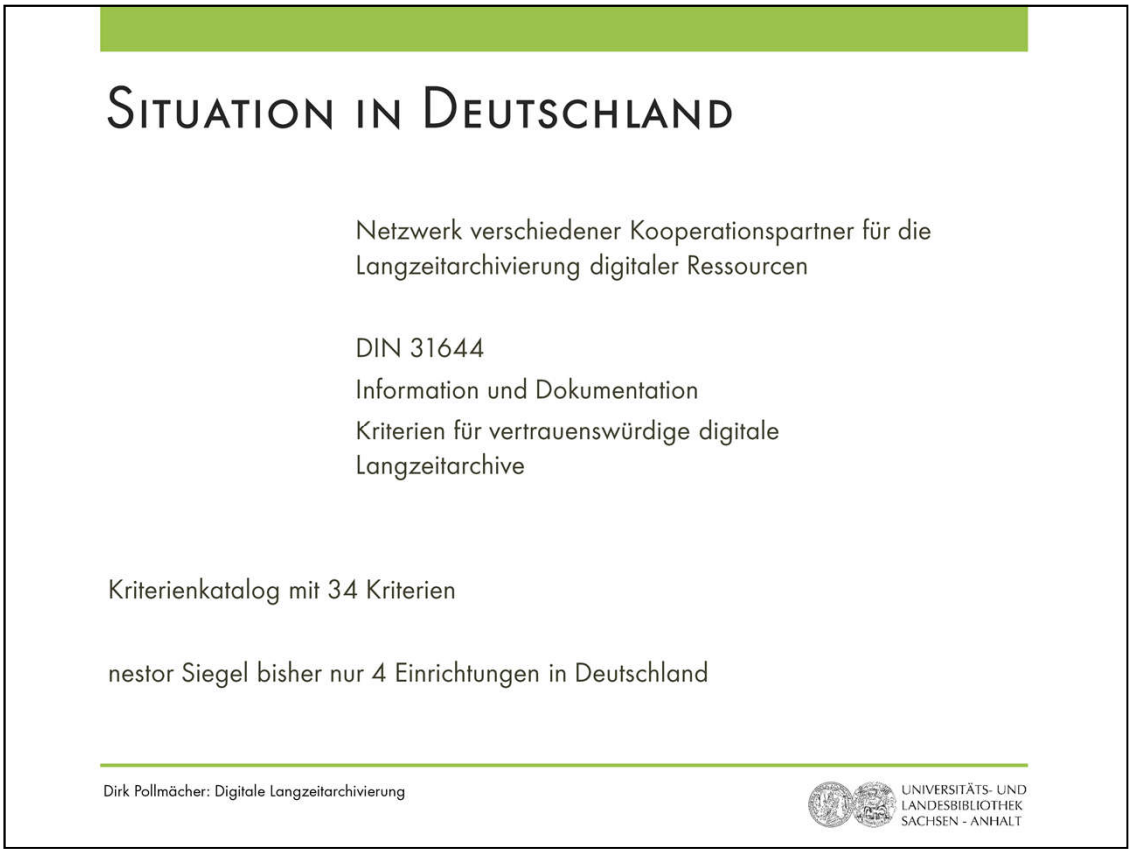
- Insbesondere OAIS ISO 14721
- DIN 31644

OAIS-Modell (Open Archival Information System)

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



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

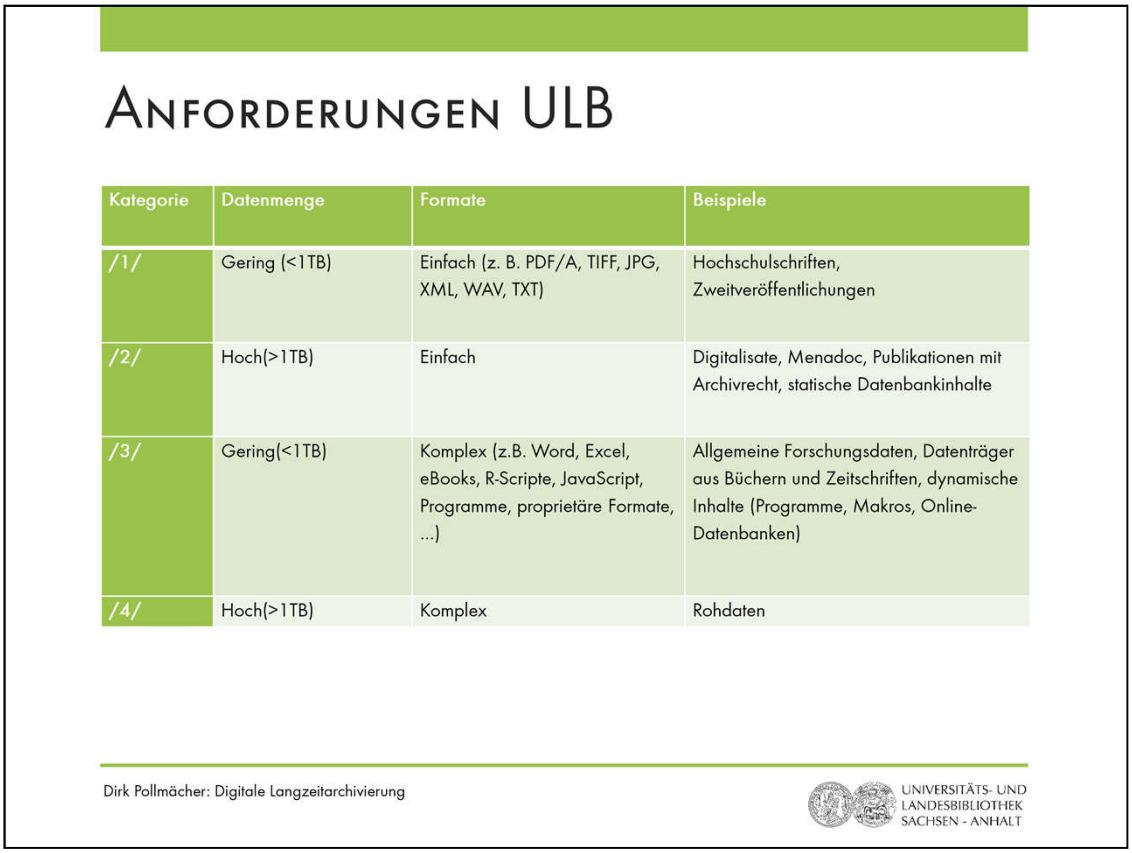
Technische Spezifikation der Pakete

Datenmanagement

Protokollierung

Metadatendefinition

IT-Infrastruktur



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

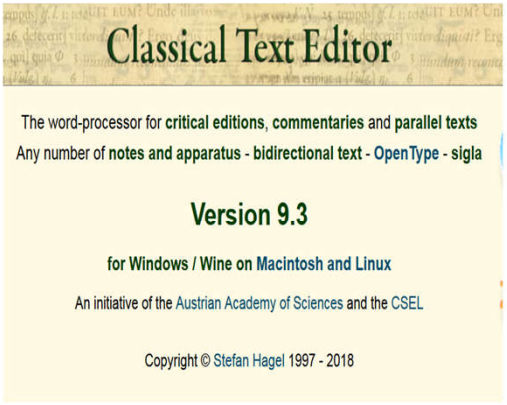
Übernahme aus DSpace oder vom Author

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



Beispiel 2 – MRT-Studie

Speicherung der Rohdaten (mehrere TByte) vor Ort

Konvertierung signifikanter Informationen in langzeit-archivierbare 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)