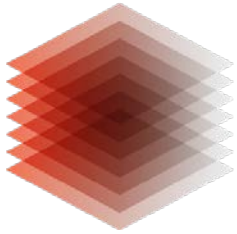
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



TIB

Digitale Langzeitarchivierung

Michelle Lindlar
Göttingen, 9. März 2017
FAG TI Sitzung

Agenda

1. DA.NRW und SIP Builder
2. Ablieferungsverfahren TIB Langzeitarchiv
3. LZA vs. Backup
4. Problemfall: Große Datenmengen
5. Zusammenfassung

DA.NRW und SIP-Builder

Der Lösungsverbund DA.NRW

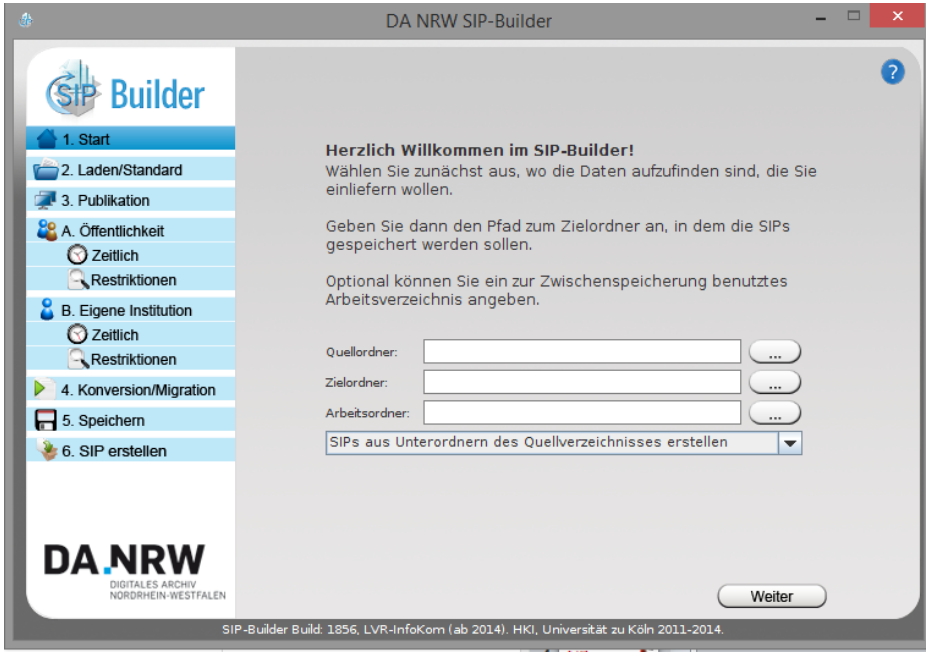
Lösungsverbund

Techn. Lösung

angelehnt an HP-SER	HKI Uni Köln / Eigenentwicklung	**Systemcore**
LWL, Stadt Köln	LVR-InfoKom	**Betreuung**
Archive	Bibliotheken	**Hauptnutzer**
„Regeln für Archivformate: ISO-normiert und langzeitstabil“ Datenbanken, Dateiablagen, AV, ….	Text: PDF, PDF/A, Doc, Docx Bild: TIFF, JPEG, JPEG 2000, BMBF, PNG, GIF Ton: PCM-Wave, MP3 Video: AVI, MXF, MPEG4	**Datenformate / Inhalte**
Nein	ja	**SIP Builder**

DNS SIP Builder

- aktuelle Version: v0.7 (last update: Oct. 2016)
- Java
- GUI und CLI
- GNU GPL v3
- als Bestandteil von DNSCore auf github: https://github.com/da-nrw/DNSCore/blob/master/SIP-Builder/



SIP Builder - Funktionsumfang

- Halbautomatisierte **Erstellung** einzelner oder mehrer SIPs

- Umfangreiche Angabemöglichkeit für **Rechte-Metadaten** und **Anforderungen an Präsentationskopien**, z.B.
  - Öffentlicher Zugriff ja / nein / Embargo
  - Präsentationskopie Bild: mit Wasserzeichen oder Fußzeile
  - Präsentationskopie Text: Einschränkung Seitenanzahl für Nutzer
  - Präsentationskopie AV: Einschränkung Playback nach Zeitwerten

- SIP-Erstellung als **Bagit v0.97** Bag mit UTF-8 encoding

```
<SIP-Builder-BAG>/
        |   bag-info.txt
        |   bagit.txt
        |   manifest-md5.txt
        |   tagmanifest-md5.txt
        \--- data/
              |   payload files
              |   premis.xml
```

SIP Builder – Nutzen und Einschränkung

- **Erstellung von SIPs** eingeschränkt anhand des AIP Aufbaus im DNS; Daten müssen auch hier für den SIP Builder von Instittutionen vorstrukturiert werden

- **Rechteinformationen** und Kriterien für **Präsentationskopien** so nur für DNS nutzbar (und auch dort in Präsentationsschicht nicht umgesetzt)

- **PREMIS** enthält nur grobe Rechteinformationen, keine Informationen z.B. zu Dateieinschränkungen wie Passwortschutz

- Bag-info auf Minimaldaten eingeschränkt (Payload oxum, Datum, Bag Größe, KEINE Informationen z.B. über Datenlieferant)

→ sinnvoll nutzbar nur im Zusammenhang mit DNS Software
→ sehr geringe Verbreitung unter Hauptnutzern (NRW ULBs); anstelle dessen Einsatz von Schnittstellen zwischen Fachanwendung und Archiv

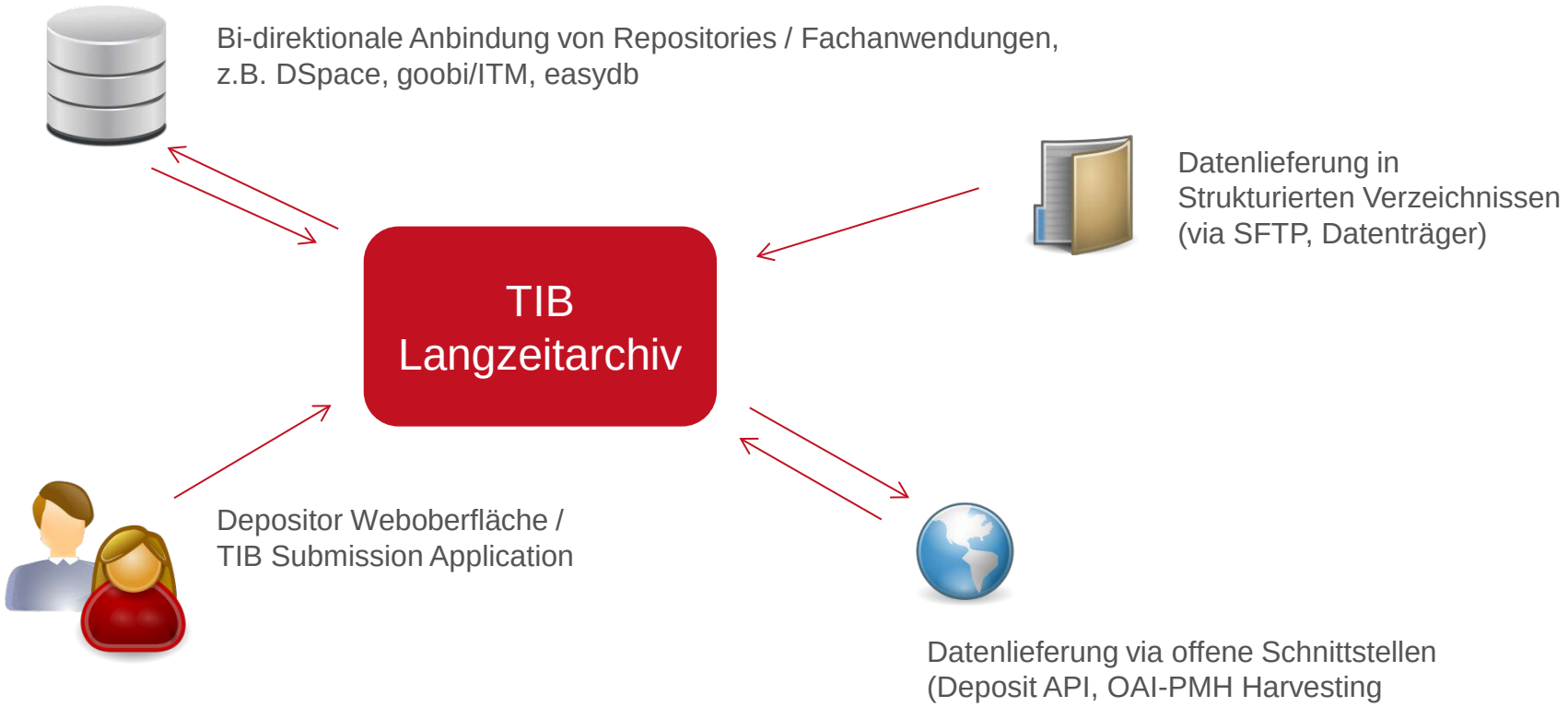
Studie IANUS Testbed DA.NRW - DNS

- DFG Projekt: Aufbau nationales Forschungsdatenzentrum für Archäologie und Altertumswissenschaften
- AP 5: Langzeitarchivierung
- Testbed Softwaresuite DNS(2013/2014)
- Fazit:
  - Geeignete Komponenten: iRODS, DA-Web
  - Teilweise geeignet: SIP-Builder, ContentBroker
  - Nicht geeignet: Präsentationsschicht mittels Fedora-Repository, Access-Management
  - DNS als Gesamtsystem nicht einsetzbar, öffentlich verfügbare Bestandteile wie iRODS können in eigener Lösung zum Einsatz kommen

Abschlussbericht: http://www.ianus-fdz.de/attachments/download/627/02_Abschlussbericht_2014-06-25_FINAL.pdf

Ablieferungsverfahren TIB Langzeitarchiv

Ablieferungsverfahren TIB Langzeitarchiv



Abgabe via strukturierte Verzeichnisse

- TIB und abliefernde Institution stimmen sich über Struktur, Inhalt, Zugriffsrecht, Lizenz, etc. ab und halten Information in Policy fest
- Institution liefert Daten in vereinbarter Struktur ab

Deposit:

- bei Übergabe von EKI oder PPN kann automatische Anreicherung mit Metadaten aus dem Verbundkatalog erfolgen
- Access Rights und license-Typ werden durch TIB in Metadaten festgehalten
- SIP wird automatisch mit allen Informationen erzeugt
- SIP wird automatisch an Langzeitarchiv übergeben

```
<EKI_oder_PPN_oder_ID>/
        |   dc.xml (optional)
        |   mets.xml (optional)
        \---PRESERVATION_MASTER/
                |    content file1
                |    content file2
                |    ...
        \---ACCESS_COPY/ (optional)
                |    access file1
                |    ...
```

Ingest Prozess

Automatische Prozess im Ingest:

- Viruscheck
- Checksummenkontrolle (CRC, MD5, SHA-256)
- Dateiformatidentifizierung
- Dateiformatvalidierung
- Extrahierung technischer Metadaten
- Validierung der deskriptiven Metadaten (basierend auf mit abliefernder Institution vereinbarter Regeln)

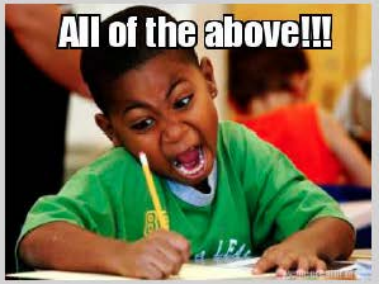
Zusätzlich möglich:

- automatisch eine DOI vergeben werden
- Accesskopien automatisch generiert werden

Alle Prozesse werden in PREMIS Daten festgehalten

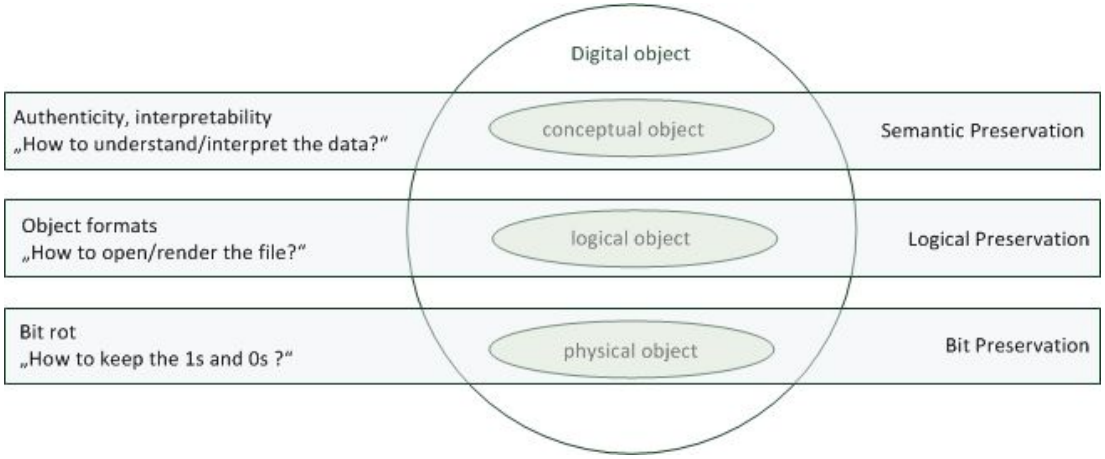
Für weitere Dokumentation siehe auch:
https://assessment.datasealofapproval.org/assessment_157/seal/html/

Ist LZA nicht nur Backup ?

		PORTICO
What small institutions think it is	*What large institutions think it is*	*What publishers think it is*
		
What IT thinks it is	*What the funding bodies think it is*	*What it really is*

Ist LZA nicht nur Backup ?

- „Gutes“ Storage ist ein Kernbestandteil der digitalen Langzeitarchivierung
- Aber digitale Langzeitarchivierung setzt sich aus Organisation + Prozess + Systemen zusammen. Sie enthält u.a.:
  - Auf der logischen / Formatebene:
    - Identifizierung, Validierung und Characterisierung des eingehenden Contents im Hinblick auf Formate
    - Risikodefinitionen / Extrahierung (z.B. pwd Schutz)
    - Überwachung der eingesetzten Verfahren / Formate („Preservation Watch“)
    - Planung und Durchführung von Migration / Emulation im Bedarfsfall („Preservation Planing & Preservation Action“)



Formatüberwachung als zyklischer Prozess

Eingang von Daten in das Langzeitarchiv

(Re-)Analyse des Contents

Festhalten der Ergebnisse / des Prozesses in PREMIS Metadaten und DB

```
</section>
- <section id="fileFormat">
  - <record>
      <key id="agent">REG_SA_DROID</key>
      <key id="formatRegistry">PRONOM</key>
      <key id="formatRegistryId">fmt/16</key>
      <key id="formatRegistryRole"/>
      <key id="formatName">fmt/16</key>
      <key id="formatVersion">1.2</key>
      <key id="formatDescription">Portable Document Format</key>
      <key id="formatNote"/>
      <key id="exactFormatIdentification">true</key>
      <key id="mimeType">application/pdf</key>
      <key id="agentVersion">6.01</key>
      <key id="agentSignatureVersion">Binary SF v.81/ Container SF v.1</key>
      <key id="formatLibraryVersion">4.1081</key>
    </record>
</section>
```

Überwachung der Community auf Tooländerungen

Integration neuer Tools / neuer Patterns

Paul Young @pmyoung84 · 14h

New PRONOM release! V89 now available. 21 new PUIDS, 35 updated entries and 19 new sigs #PRONOM #DROID nationalarchives.gov.uk/PRONOM/Default …

Problemfall: Große Datenmengen und doppelte Speicherung

Problemstellung

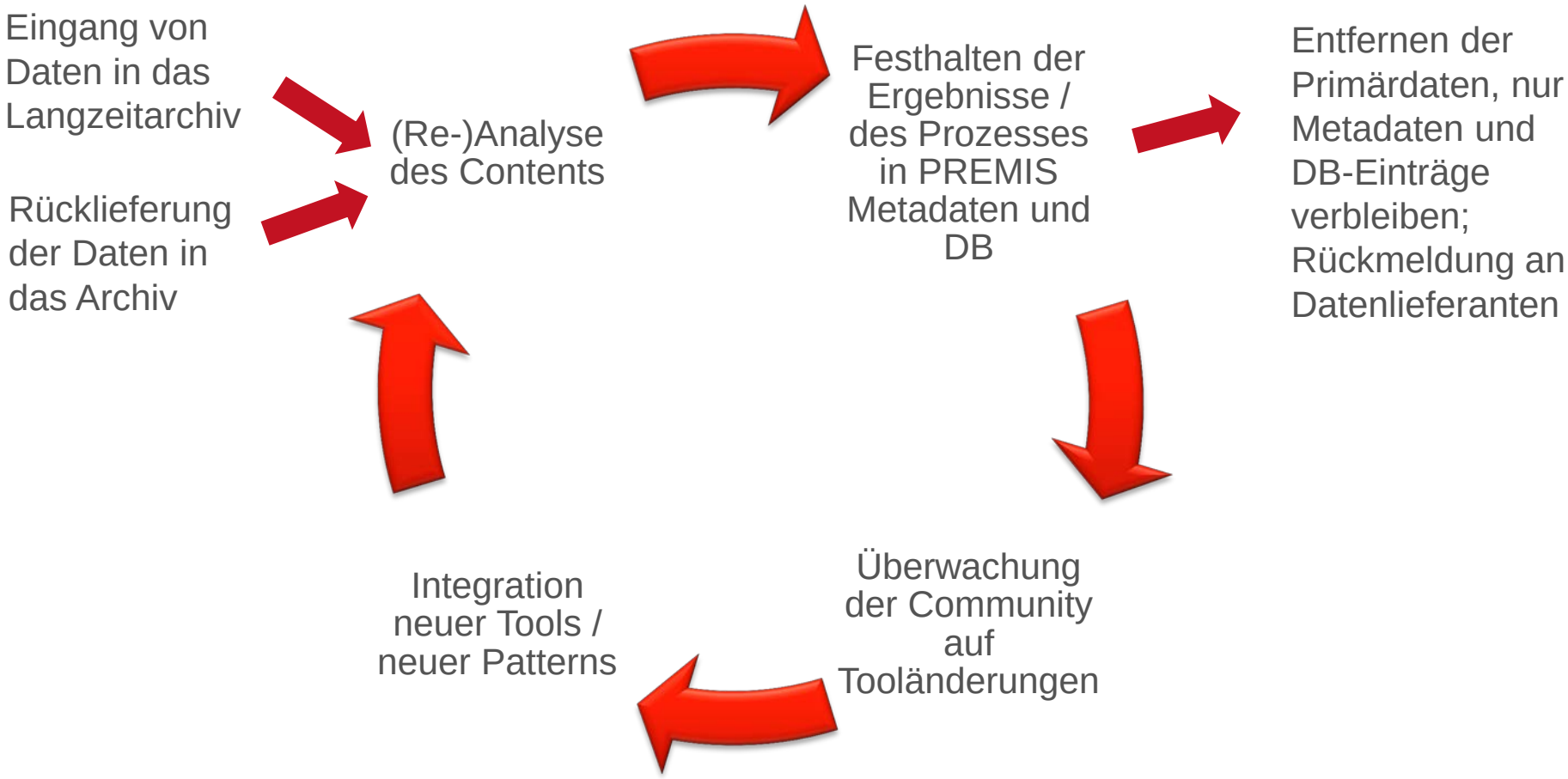
Hintergrund:

- Große Dateimengen vorhanden, z.B. aus Digitalisierungsprojekten
- Verlässliche, redundante Speicherlösung mit mehreren Kopien (Backup) schon vorhanden
- Daten müssen für andere Prozesse im Haus bleiben
- Keine doppelte Speicherung im Archiv gewünscht, da Kostenverdoppelung

Prototypische Entwicklung TIB:

- Trennung von System-Layern für Bit-Preservation und Logical Preservation
- Daten werden nur zur (Re-)Analyse in das Archiv geholt

Prototypische Entwicklung: Trennung von Storage und Data Management Ebenen



Zusammenfassung

Gegenüberstellung DNS und TIB Dienstleistung

	DA.NRW DNS	TIB Dienstleistung
Anzahl der Formate	Eingeschränkt	unbegrenzt
Aktive Formatüberwachung	Nein	Ja
Re-Validierung bei Formatänderungen	Nein	Ja
Aktive Weiterentwicklung der Software	Teilweise	Ja
Flexible Ingestverfahren	Nur mit umfangreichen Eigenentwicklungen	Ja
Erprobter Exit	Nein	Ja
Dokumentation	Sehr rudimentär	Umfangreich
Zertifizierung	Nein	Ja

Lösungsverbund != Dienstleistung

- Lösungsverbund bedeutet:
  - Eigene Vorhaltung der technischen Infrastruktur
  - Bereitstellung Ressourcen für aktive eigene Weiterentwicklung der Systeme
  - Know-how bzgl. Langzeitarchivierungsverfahren, insbesondere Formate und Werkzeuge
  - Personal für aktiven Preservation Watch, Planning und Action Prozesse
  - Gemeinsame Kosten, z.B. bei DNS durch redundante Speicherung der Daten ALLER Partner in ALLEN Systemen

Dienstleistung beinhaltet alle dieser Punkte

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Fragen? Anmerkungen?

Kontaktdaten
Michelle Lindlar
T 0511 762-19826, michelle.lindlar@tib.eu

DA NRW – Plan und Umsetzung