

Wie haben die Zentralen Fachbibliotheken ihre digitale Langzeitarchivierung organisiert?

M. Lindlar

Technische Informationsbibliothek

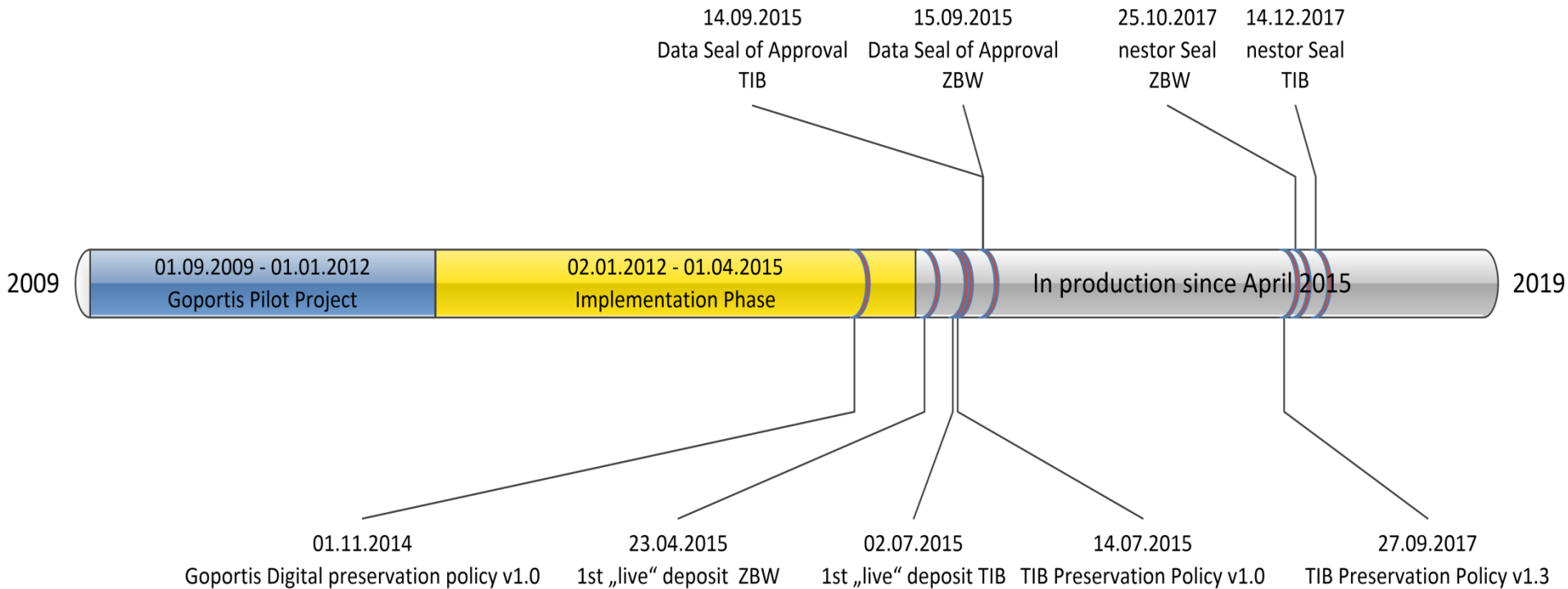
Leibniz Info-Tag zur digitalen Langzeitarchivierung, Berlin, 30.10.2018



Agenda

- 2009 – 2012: Das Goportis Pilot Projekt
- Infos zum System: Wie „machen“ wir LZA?
- Datenstand im Archiv: Was haben wir schon?
- Organisatorische Faktoren der LZA

Was war, was ist, was wird ...



Goportis Pilot Projekt 2009 - 2012

Die 3 zentralen Fachbibliotheken unterscheiden sich bzgl.:

- Fächerkanon
- verfügbaren Ressourcen, Umfang der Bestände
- Verbünde (Auswirkung z.B. auf angebundenen Katalog), technische Infrastruktur

Ziel Pilotprojekt:

- Evaluierung technologischer, organisatorischer und institutioneller Anforderungen an ein kooperativ betriebenes Langzeit-archivierungssystem
- Testimplementierung eines Systems (basierend auf Ex Libris „Rosetta“)
- Berücksichtigung Exitszenario von Anfang an

Goportis Pilot Projekt - Kriterienkatalog

- Kriterienkatalog mit 35 Punkten, u.a.
 - **Sicherheit:** Schutz der Objekte vor böswilliger oder unbeabsichtigter Beeinträchtigung.
 - **Skalierbarkeit** : Speicherung und Verwaltung von Millionen von Objekten.
 - **Offenes System** Anwendungen, Metadaten und Systemdaten müssen über Generationen von Hardware, Speichersystemen und Software-Komponenten konfigurierbar, ausweitbar und migrierbar sein.
 - **Konformität** Das System muss allgemeingültigen Standards (bspw. OAIS-Modell [Open Archival Information System] oder TRAC [Trustworthy Repositories Audit & Certification]) entsprechen.
- Vergleich der Systeme DuraSpace, IBM DIAS, Planets Framework, Roda, Portico, Rosetta, Tessella SDB gegen Kriterienkatalog

Infos zum System

3 Umgebungen – DEV, TEST, PROD

- Jede Umgebung ist redundant ausgelegt (2, 4, 4 Application Server + eigene DB + eigener AppServer)
- Solaris (virtualisierte Zonen), Oracle, Debian
- ZFS Storage via NFS angebunden gespiegelt (neues Storagesystem mit Einbindung Uni-RRZ aktuell in Planung)



Image M. Lindlar, CC BY SA

Integrationen (Auswahl):

- Katalog: hbz Aleph / gbv PICA
- Repositories: DSpace (ZBW Entwicklung), GMS Harvester (ZBMED Entwicklung), Submission Application / SIP Packer, easyDB Anbindung (TIB Entwicklung), goobi2Rosetta Plugin (Intranda / TIB Entwicklung)
- Plugins: DROID, JHOVE, itext Migrationstool (ZBW), Sophos Virensan (BSB)
- Datenlieferung: Deposit API, OAI-PMH, SFTP, Datenträger



Maßnahmen – Bitstream Preservation

- Trennung von Datenträger und digitalem Objekt durch USB-Imaging, CD-Imaging
- Mehrere unabhängige Kopien jeder Datei
- 3 verschiedene „Hashwerte“ (CRC, MD5, SHA-256) für jede Datei im Archiv die regelmäßig überprüft werden

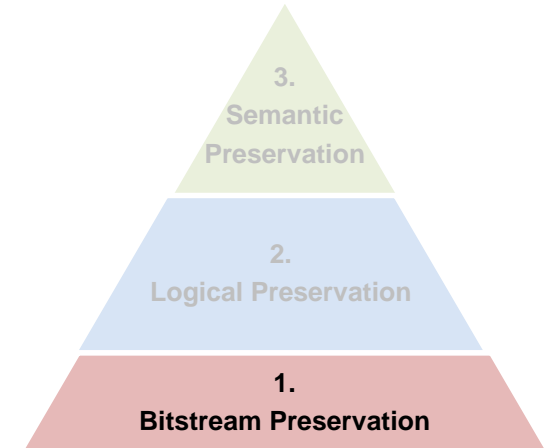


Image Pyramid: Pia Rudnik, CC BY SA

MD5 Wert:

afae2aa1c9120d7af54b6de9c5acc9f7

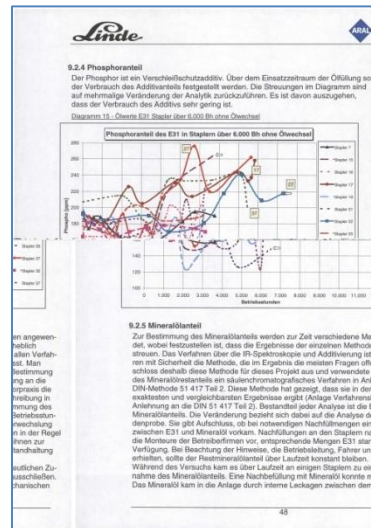


Image: Oleg Nekhayenko, CC BY SA



Maßnahmen – Logical Preservation

- Bestimmung des genauen Formats (versionsgenau)
- „Validierung“ der Datei gegen Kriterien des Standards (wo möglich)
- Extrahierung und Speicherung technischer Metadaten (z.B. Videocodec in AV Container)
- Stetige Überwachung von Technologieänderung (Formatcommunity)
- Bei Bedarf Migration in neues Format oder Emulation

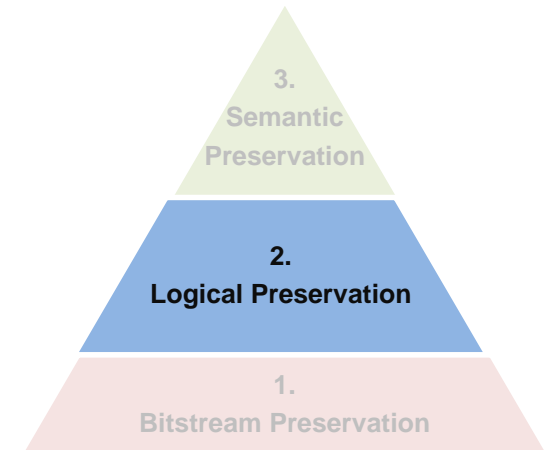


Image Pyramide: Pia Rudnik, CC BY SA

%PDF-1.4
%%EOF



Resource	Extens...	Size	Last ...	Ids	Format	Version	Mime ...	PUID	Method	Hash
C:\Files\0...	pdf	15 bytes	27.03.1...		Acrobat...	1.4	applicati...	<u>fmt/18</u>	Signature	

Erhaltungszyklus

Eingang von
Daten in das
Langzeitarchiv



(Re-)Analyse des
Contents



Festhalten der
Ergebnisse / des
Prozesses in
PREMIS
Metadaten und
DB



Preservation /
Technology /
Community /
Organization
Watch



Integration
neuer Tools /
neuer Patterns



```
</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>
```



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

Maßnahmen – Semantic Preservation

- Neben den Dateien speichern wir einen kurzen Abzug des Katalogeintrags für jedes Objekt ab
 - Erfassung in Dublin Core
 - Institution bestimmt „Mindestdatensatz“ je Workflow / Bestand
 - Validierung gegen definierte Policy im System
- Überwachung der eingesetzten Metadatenstandards auf Veränderungen, bei Bedarf Migration / Anpassung (z.B. Änderungen in PREMIS v3)

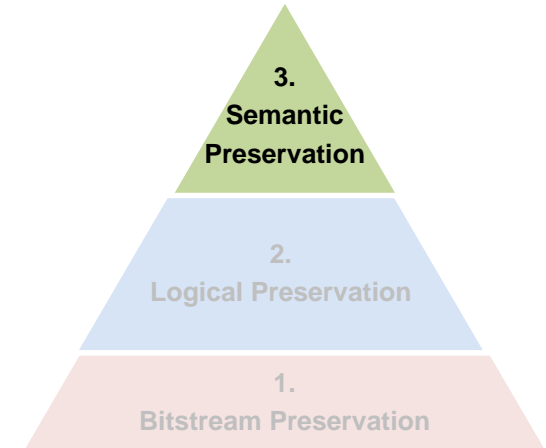
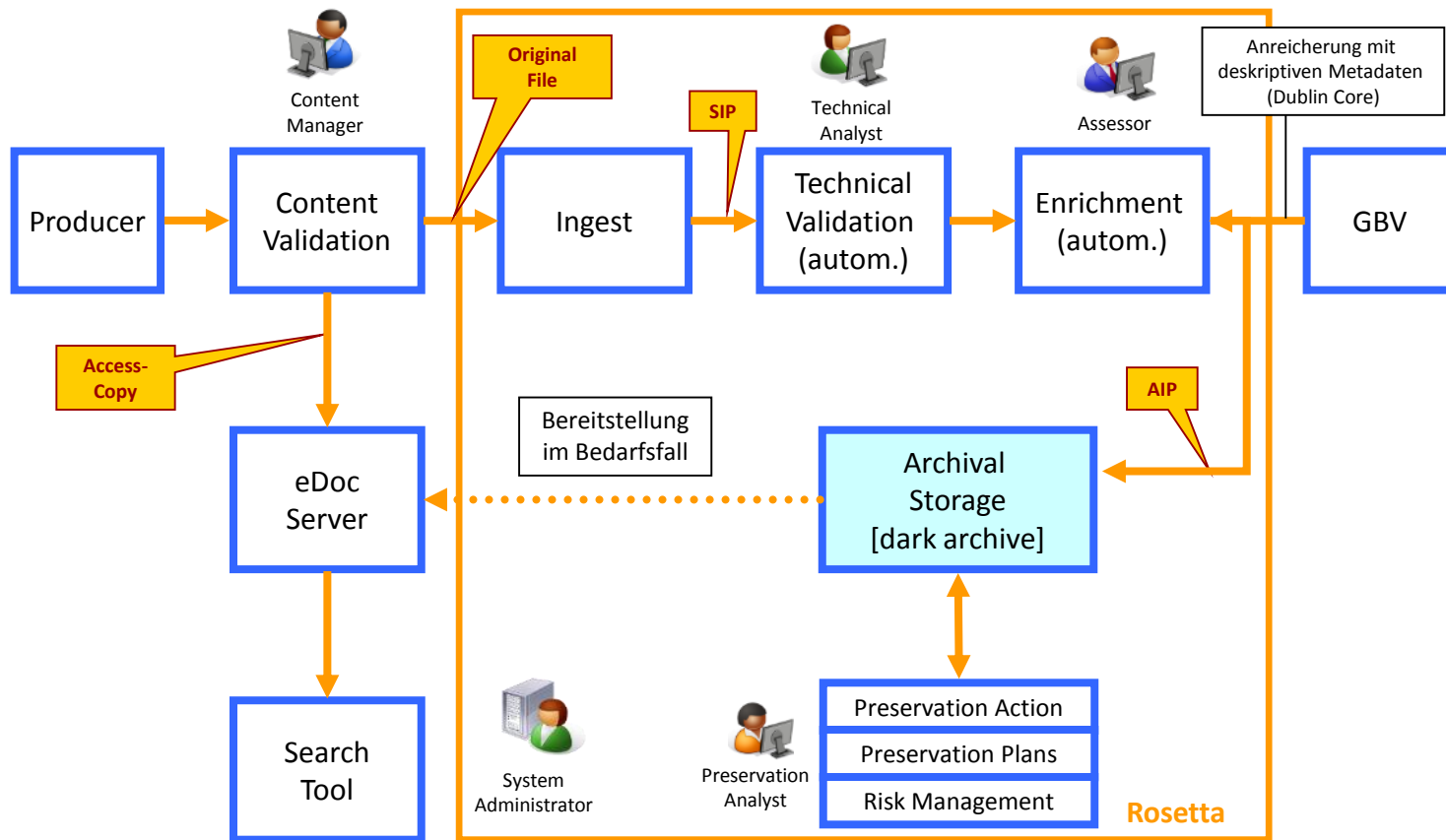


Image Pyramide: Pia Rudnik, CC BY SA

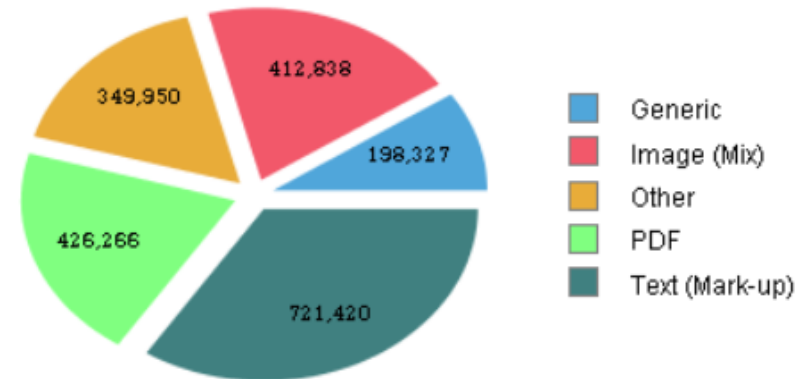
Beispielworkflow – Dark Archive



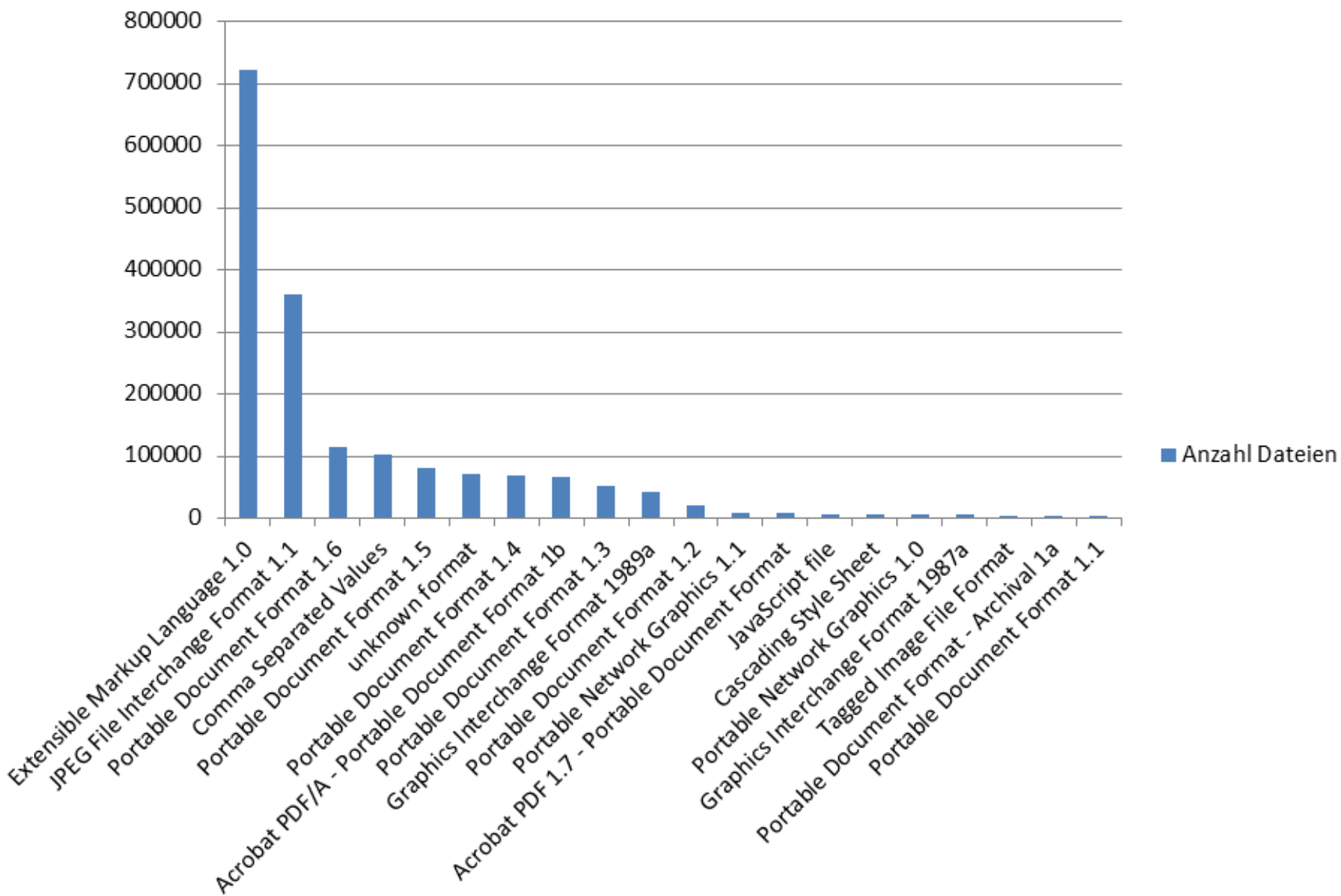
Datenstand im Digitalen Langzeitarchiv

- Dateien: 1.770.785
- Intellektuelle Einheiten: 562.993
- Anzahl Formate: 123
 - Davon ca. 50% mit weniger als 10 Instanzen im Archiv
- Volumen (nur PRESERVATION MASTER Dateien): ~ 2.5 TB
- Geplanter Zuwachs 2019:
 - > 430 TB (Digitalisate)

Files by Classifications



Top 20 Formate im Archiv



Beispiele archivierter Bestände

- TIB:
 - USB / CD Images
 - graue Literatur
- ZB MED:
 - GMS German Medical Science Proceedings & Journals
- ZBW:
 - National- und Allianzlizenzen
 - ECONSTOR Repository



Image M. Lindlar, CC BY SA

ECONSTOR
Make Your Publications Visible.



Maßnahmen – Organisatorische Faktoren

- Verständnis LZA als organisatorischer Prozess:
 - Qualitätskontrolle, auch außerhalb der LZA
 - Archivierung als Bestandteil von Lizenzen, wo möglich
 - Beratung von Teams und Datenlieferanten
- Preservation Policy - öffentliche Beschreibung unserer Prozesse:
 - Goportis Rahmenpolicy und institutionelle Policy
- Umfangreiche Dokumentation unserer Prozesse
 - <https://wiki.tib.eu/confluence/display/lza/Digitale+Langzeitarchivierung+an+der+TIB>
- Zertifizierung:
 - Data Seal of Approval (2015)
 - nestor Seal (2017)
 - Core Trust Seal (geplant 2018/2019)

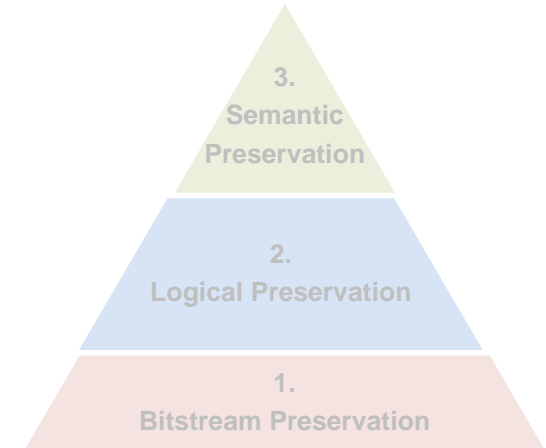


Image Pyramide: Pia Rudnik, CC BY SA



Zusammenarbeit: wie wir uns organisieren

- Wiki
 - Dokumentation von Workflows, gemeinsame Projekte
- Jira
 - Tickets von ZB MED & ZBW an TIB, Tickets der TIB an Ex Libris
- „Praktikertelko“:
 - Monatliche Telefonkonferenz
 - Feste Agenda: Tickets, Anwendungsadministration, Berichte aus den Häusern, Netzwerkarbeit
- „AG LZA Treffen“
 - 2 mal im Jahr
 - Teilnehmerkreis: Bereichsleitungen, LZA Teams, Anwendungs- & Systemadministration, Entwickler



Image: Library of Congress, CCO
https://www.flickr.com/photos/library_of_congress/45269447921/ 16

Gemeinsame Netzwerkarbeit

- **Zusammen sind TIB, ZB MED und ZBW aktiv in:**

- 9 nestor
- 6 Rosetta User Group
- 3 Open Preservation Foundation
AGs / Gremien



- **Beispiele für gemeinsame Konferenzeinreichungen:**

- iPRES 2013: „Benefits of geographical, organizational and collection factors in digital preservation cooperations: The experience of the Goportis consortium.“
- iPRES 2016: „Consortial Certification Processes – The Goportis Digital Archive. A Case Study“
- iPRES 2018: „Time-travel with PRONOM – The fourth dimension of DROID.“
- IDCC 2017: „How valid is your validation? A closer look behind the curtain of JHOVE“




Fragen ? Anmerkungen !



Kontakt:

M. Lindlar – TIB Hannover

 Michelle.lindlar@tib.eu

 0511 762 19826

 Lindlarm

 mickylindlar



Welttag Digitale Erhaltung

29. November 2018



ZBW