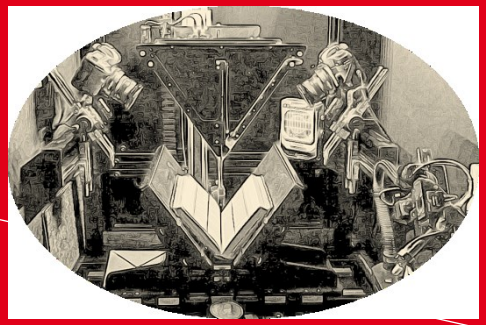
SLUB
Wir führen Wissen.

Automatische Langzeitarchivierung

für die Digitalisierung mit Kitodo



Sabine Krug, Fotos mit freundlicher Genehmigung von **Jörg Sachse**
Sächsische Landesbibliothek, Staats- und Universitätsbibliothek Dresden (SLUB)

Oktober 2016

Agenda

- Was ist ein Langzeitarchiv?
- Digitales Langzeitarchiv SLUBArchiv, Herausforderungen & Ziele
- Archivierung Retrodigitalisate
- Handreichungen & Rahmenbedingungen
- Kitodo, Architektur und Workflow
- Ingest-Preprocessing & Ingest
- Produktiver Einsatz
- Zusammenfassung und Ausblick

Digitale Langzeitarchivierung
Definition und Ziele

DIN 31644: „Organisation (bestehend aus Personen und technischen Systemen), die die Verantwortung für den Langzeiterhalt und die Langzeitverfüg-barkeit von Information in digitaler Form sowie die Bereitstellung für eine bestimmte Zielgruppe übernommen hat.“

Ziele

- Sicherung der langfristigen Verfügbarkeit und Nutzung von digitalen Objekten (50 Jahre +)
- Berücksichtigung zukünftiger Nutzungsszenarien

=> Erhalt der Korrektheit (Bitstream Preservation) und

=> Erhalt der Interpretierbarkeit und Nutzbarkeit (Content Preservation)

Digitale Langzeitarchivierung
Herausforderungen

Schneller Medien-, Format- und Systemwandel

Begrenzte Haltbarkeit der Trägermedien

Integrität der Daten nimmt durch gezielte Modifikation oder Systemfehler ab

Hardwareausfälle, Softwarefehler, Unglücksfälle

⇒ Sichern der Korrektheit (Bit stream preservation) und

⇒ Prüfen der Korrektheit bei großen Datenmengen

Veralten der Dateiformate – Software, die das Datenformat korrekt interpretieren wird nicht mehr entwickelt/gepflegt

Veralten der Daten und Metadaten

=> Erhalt der Interpretierbarkeit und Benutzbarkeit

Datendurchsatz zur Verarbeitung der täglich produzierten Daten

Digitales Langzeitarchiv SLUBArchiv
Ziele und Stand

Aufbau des Digitalen Langzeitarchivs der SLUB erfolgt im Rahmen eines Projektes (Mai 2012 bis Oktober 2014)

Ziele

- Sichern der Langzeitverfügbarkeit der Digitalen Sammlungen der SLUB (Digitalisierung mit Kitodo, Elektronische Publikationen, Digitale Sammlung der Deutschen Fotothek, Digitales Audio/Video-Material der Mediathek)
- Vorbereitung einer Dienstleistung für andere sächsische Institutionen

Stand:

- Produktiver Betrieb des Workflows für die Digitalisierung mit Kitdo.Production (Retrodigitalisate) seit 2015
- Ca. 60 000 IEs, 3 000.000 Files, 43 TB
- Zertifizierung mit dem Data Seal of Approval im Juni 2015
- Erweiterung und Anpassung für Kitodo Workflow im Rahmen des Landesdigitalisierungsprogramm Sachsen Q1 – Q4 2016

Digitales Langzeitarchiv SLUBArchiv
Grundsätze

Verwendung als **Dark Archive**, in dem die Masterdaten verwaltet und archiviert werden – die Präsentationsdaten bleiben in einem separaten Repository, können aber aus den Masterdaten erzeugt werden

Automatisierung des Ingest, d.h. der Übernahme ins Langzeitarchiv, und des Access, d.h. des Zugriffs auf die Daten aus dem Langzeitarchiv (bis auf Fehlerfälle)

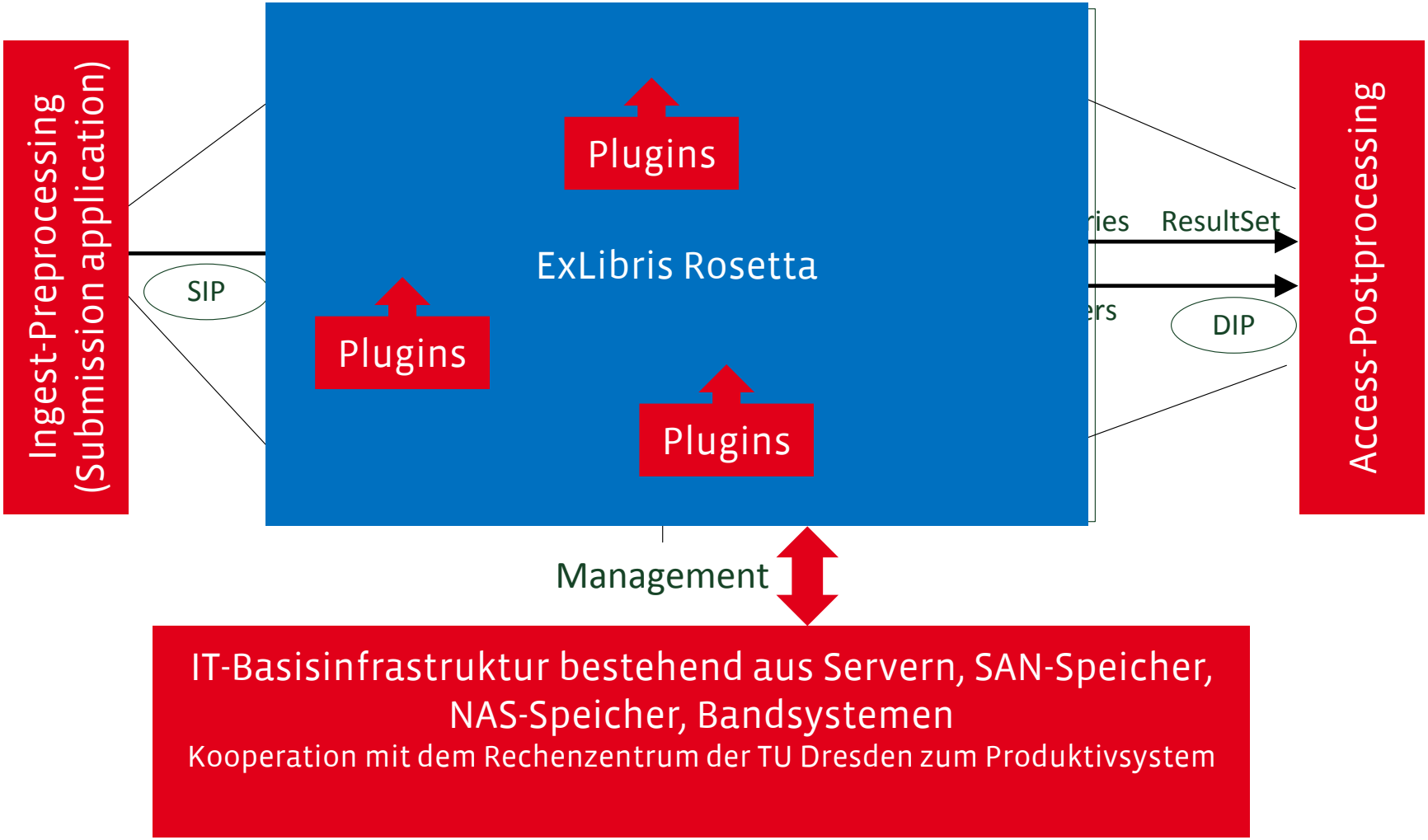
Prüfsummen werden bereits im Produktionsprozess (bei der Digitalisierung) erzeugt und bei der Übernahme ins Langzeitarchiv geprüft

Unterstützung einer **definierten Menge** von LZA-fähigen Datenformaten

Übernahme ins Langzeitarchiv nur für **erfolgreich** geprüfte Dokumente

Archivierung Retrodigitalisate

Architektur SLUBArchiv



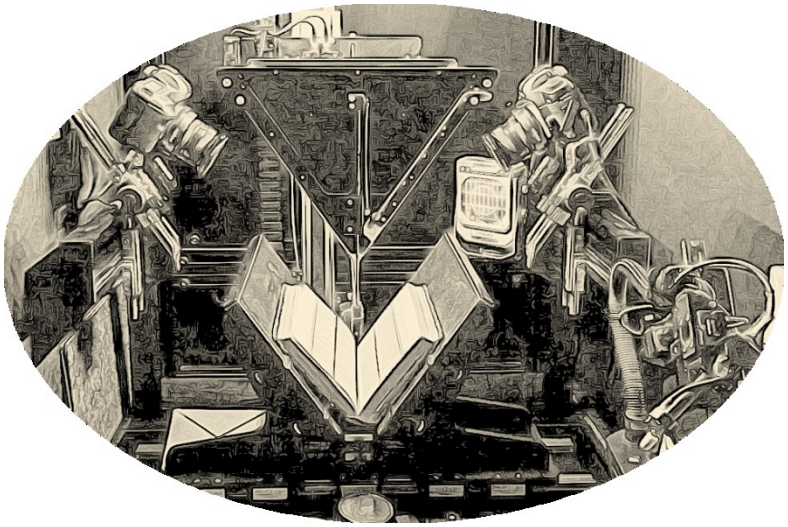
Archivierung Retrodigitalisate
Kitodo für Massendigitalisierung

Kitodo ist eine Software zur Unterstützung des Digitalisierungsworkflows

SLUB Digitalisierungszentrum erzeugt ca. 3 Mio. Scans im Jahr

Digitalisiert werden ausgewählte Printmaterialien

Digitale Dokumente sollen automatisiert nach beendeter Bearbeitung ins SLUB Langzeitarchiv übernommen werden



Archivierung Retrodigitalisate
Rahmenbedingungen für Produzent

Handreichung für archivfähige Formate (Liste archivfähiger Formate)

- Gewährleistung Formaterkennung, Formatvalidierung und Metadatenextraktion

Spezifikation SIP als Ablieferungspaket

- Verzeichnisstruktur und Bezeichnung
- Paketinhaltinhalt und Aufbau (IE)
- Festlegung Metadaten für SIP Beschreibung (SIP.xml)

Schnittstellen im Workflow und Verantwortungsbereiche

- Zuständigkeiten für Löschung im Transferverzeichnis
- Protokolle und Auswertung
- Fehlerbehebung

Archivierung Retrodigitalisate
Handreichung TIFF

Festlegung verbindliches Subset der verfügbaren Funktionen

- Digitalisate müssen dem TIFF 6-Baseline-Standard entsprechen
- Verwendung ausschließlich Datentypen aus dem TIFF6-Standard
- Nur Images in bitonal, grayscale oder RGB zugelassen (kein Palette Color, keine Transparency Mask)
- nur genau ein Image File Directory (IFD) vorhanden
- Multipage-TIFF ist nicht erlaubt
- Grundsatz, dass Metadaten möglichst mit den bereits vorhandenen Tags abgebildet werden sollten

➢ Policy über notwendige, optionale und verbotene TIFF-Tags

➢ Tool zur Überprüfung (Open Source) „checkit_tiff“

Archivierung Retrodigitalisate
Ausgangsdaten

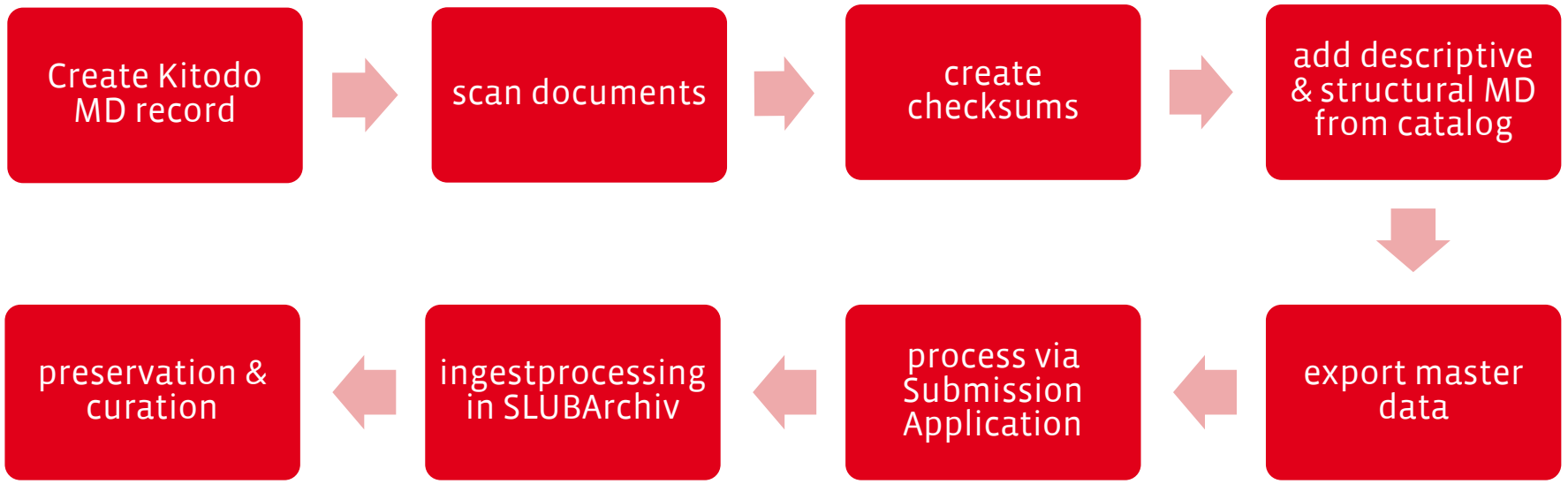
Nach der Digitalisierung und Bearbeitung (Katalogisierung, Strukturierung) wird eine digitale Einheit mit folgenden Daten an das Langzeitarchiv übergeben:



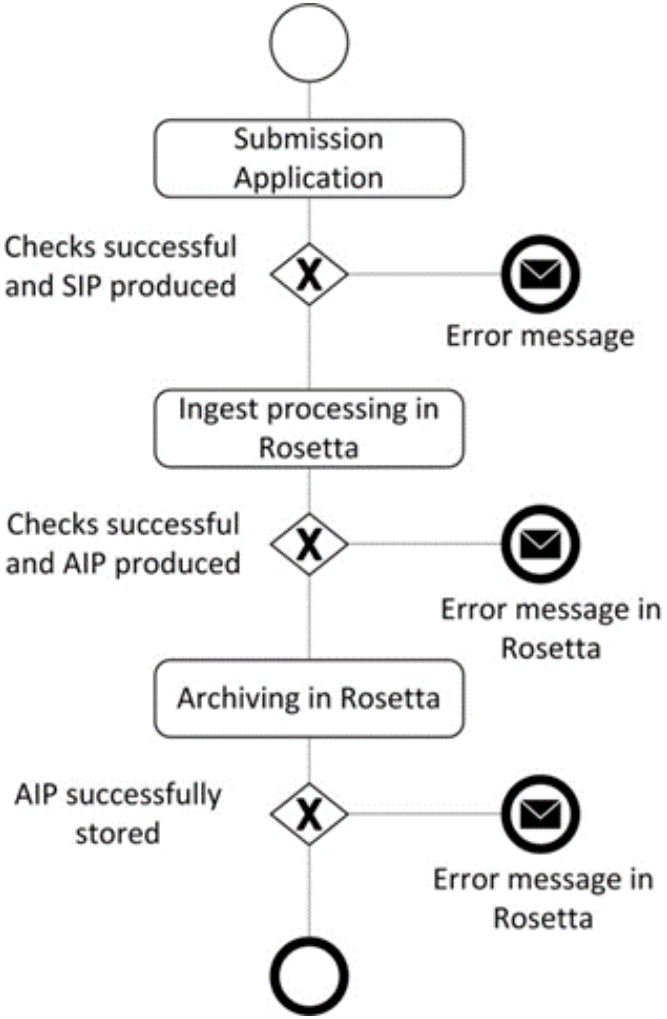
- Masterdaten in TIFF-Format
- Optional OCR-Daten im ALTO-XML-Format
- Metadaten in METS/MODS
- Prüfsummendatei

Archivierung Retrodigitalisate

Wie Kommen die Retrodigitalisate ins SLUBArchiv?



Übernahme in SLUBArchiv
Datenübernahme



Archivierung Retrodigitalisate

Submission Application (Preprocessing)

Verzeichnisstruktur Goobi (Detail eines Vorganges)

```
.
├──10008
│   ├──319037843.xml
│   ├──eingdehaz_319037843_mets.xml
│   ├──images
│   │   ├──scans_abb
│   │   ├──scans_goobi
│   │   │   ├──00000001.tif
│   │   │   └──00000002.tif
│   │   ├──scans_orig
│   │   ├──scans_tif
│   │   │   ├──00000001.tif
│   │   │   └──00000002.tif
│   │   └──tiffwriter.conf
│   ├──meta.mets.xml
│   ├──meta.mets.xml.1
│   ├──meta.mets.xml.2
│   ├──meta.xml.1
│   ├──meta.xml.2
│   ├──orig.ctimestamp
│   ├──orig.timestamp
│   ├──tif.ctimestamp
│   ├──tif.md5
│   └──tif.timestamp
:
```

Verzeichnisstruktur

```
.
├──10008
│   └──content
│       ├──eingdehaz_319037843_mets_ROSETTA_METS.xml
│       └──streams
│           ├──eingdehaz_319037843_mets.xml
│           └──LOCAL
│               ├──00000001.tif
│               └──00000002.tif
:
```

- Prüft, ob neue, beendete Kitodo-Vorgänge vorhanden im Transferverzeichnis
- Validiert mit „checkit_tiff“ mit Kitodo - Regelsatz
- Prüfung der Vollständigkeit & Integrität der Daten
- Transformation METS/MODS nach METS/DC
- Erstellt SIP
- Anstoßen der automatischen Übernahme durch Rosetta

Archivierung Retrodigitalisate
Ingest

Automatische Übernahme in Rosetta (muß für jeden Workflow konfiguriert und über Plugins/Programme angepasst werden)

- Prüfen der Vollständigkeit und Integrität der Dateien (Prüfsummen)
- Virusprüfung
- Identifikation des Datenformates (PRONOM-ID)
- Validierung = Prüfen der Korrektheit des Format mit Validierungstool (Jhove)
- Extraktion von technischen Metadaten
- Erstellen eines Archivpaketes
- Speicherung im Archivbereich

SLUB
Wir führen Wissen.

Zusammenfassung und Ausblick

Unsere Erfahrungen

- Dark Archive und lose Kopplung an das Produktionssystem hat sich bewährt
- **Automatisierung** spart Ressourcen und minimiert „menschliche Fehler“
- Handreichung zum TIFF und Bereitstellung Validierungstool minimiert TIFF Fehler
- Entwurf praxisnaher Testfälle und Dokumentation der Durchführung ist wichtig
- Größter Aufwand ist es, andere zu überzeugen in Jahrhunderten zu denken
- LZA-fähige Formate nutzen (Baseline TIFF)
- Probleme mit Ausgangsdaten vor der Aufnahme ins Langzeitarchiv lösen
- Kooperieren, Vernetzen, Austauschen

SLUB
Wir führen Wissen.

Danke!

Sabine Krug
Sächsische Landesbibliothek –
Staats- und Universitätsbibliothek Dresden (SLUB)