SLUB
Wir führen Wissen.

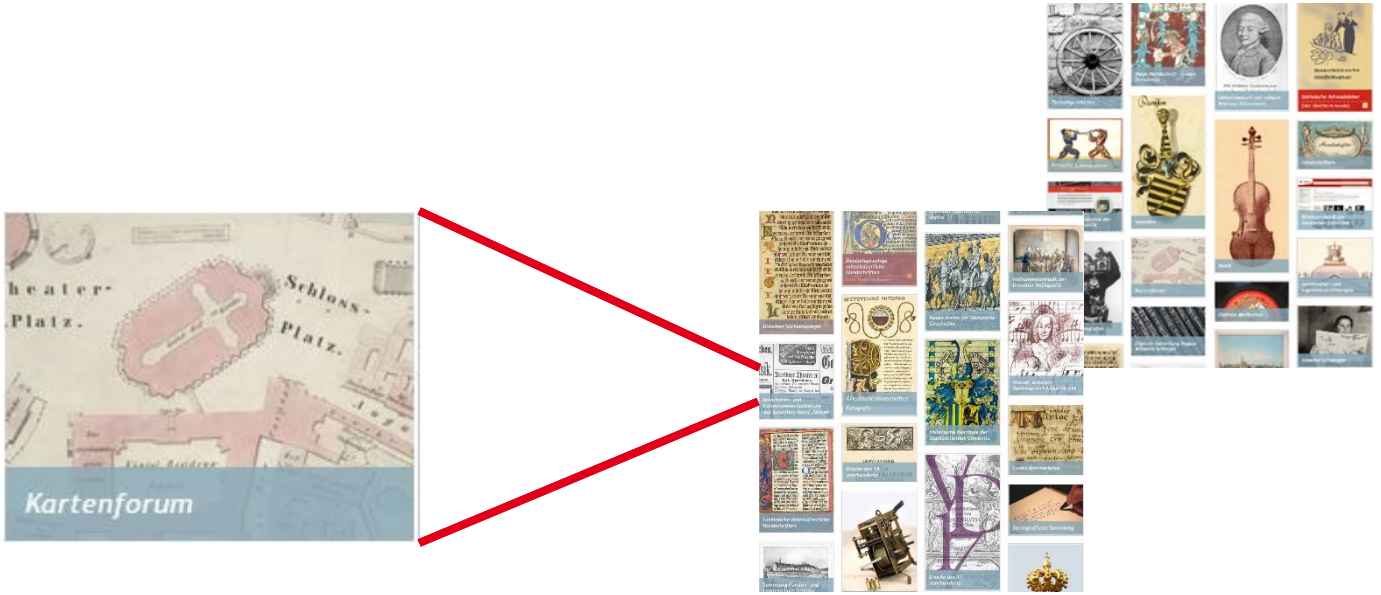
Das digitale Langzeitarchiv SLUB

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

Juni 2015

SLUBArchiv
Motivation

Digitalisierung wichtiger Sammlungen für die Literatur –und Informationsversorgung / bessere Zugänglichkeit



SLUBArchiv
Motivation

Elektronische Publikationen

Gesetz zur elektronischen Pflichtexemplarabgabe (Anpassung 2014 im Sächsischen Pressegesetz)

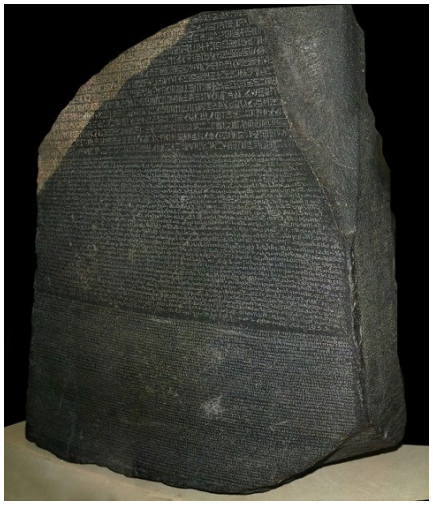
Seit dem 1. Januar 2014 sind auch Netzpublikationen aus Sachsen ablieferungspflichtig.

Vorerst sammelt die SLUB ausschließlich PDF-Dateien, deren Nutzung frei im Internet erfolgen kann. Diese Dateien werden im Sächsischen Dokumenten- und Publikationsserver Qucosa unter einer stabilen Internetadresse (URN) veröffentlicht und bereitgestellt. Dabei gelten die Leitlinien von Qucosa. Die Archivierung erfolgt über das digitale Langzeitarchiv der SLUB. Eine verlässliche Archivierung ist nur unter Einhaltung definierter technischer Standards möglich. Die SLUB übernimmt anderenfalls keine Garantie für die Langzeitverfügbarkeit.

SLUBArchiv
Motivation

Haltbarkeit: ca. 6 Jahre

Langzeitarchivierung digitaler Objekte benötigt andere Konzepte als die Langzeitarchivierung klassischer Objekte



Haltbarkeit: einige tausend Jahre

SLUB
Wir führen Wissen.

SLUBArchiv

Ziele, Grundsätze

SLUBArchiv
Ziele

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

Ziele

- Sichern der Langzeitverfügbarkeit der Digitalen Sammlungen der SLUB
- Digitalisate von Printmedien aus dem 16., 17. und 18. Jh. und weitere ausgewählte Sammlungen
- Elektronische Publikationen/Qucosa
- Digitale Sammlung der Deutschen Fotothek
- Digitales Audio/Video-Material der Mediathek

SLUBArchiv
Grundsätze

Zusätzliche **Entwicklungs- und Testumgebung** für die Entwicklung neuer Workflows, die Erweiterung vorhandener Workflows und den Test von neuen Softwareversionen

Verwendung als **reines Archiv**, 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) bzw. der Annahme (bei Elektronischen Publikationen) erzeugt und bei der Übernahme ins Langzeitarchiv geprüft

SLBArchiv
Grundsätze

Speicherung von **drei oder vier Kopien** an derzeit zwei Standorten

Unterstützung einer **definierten Menge von Datenformaten**

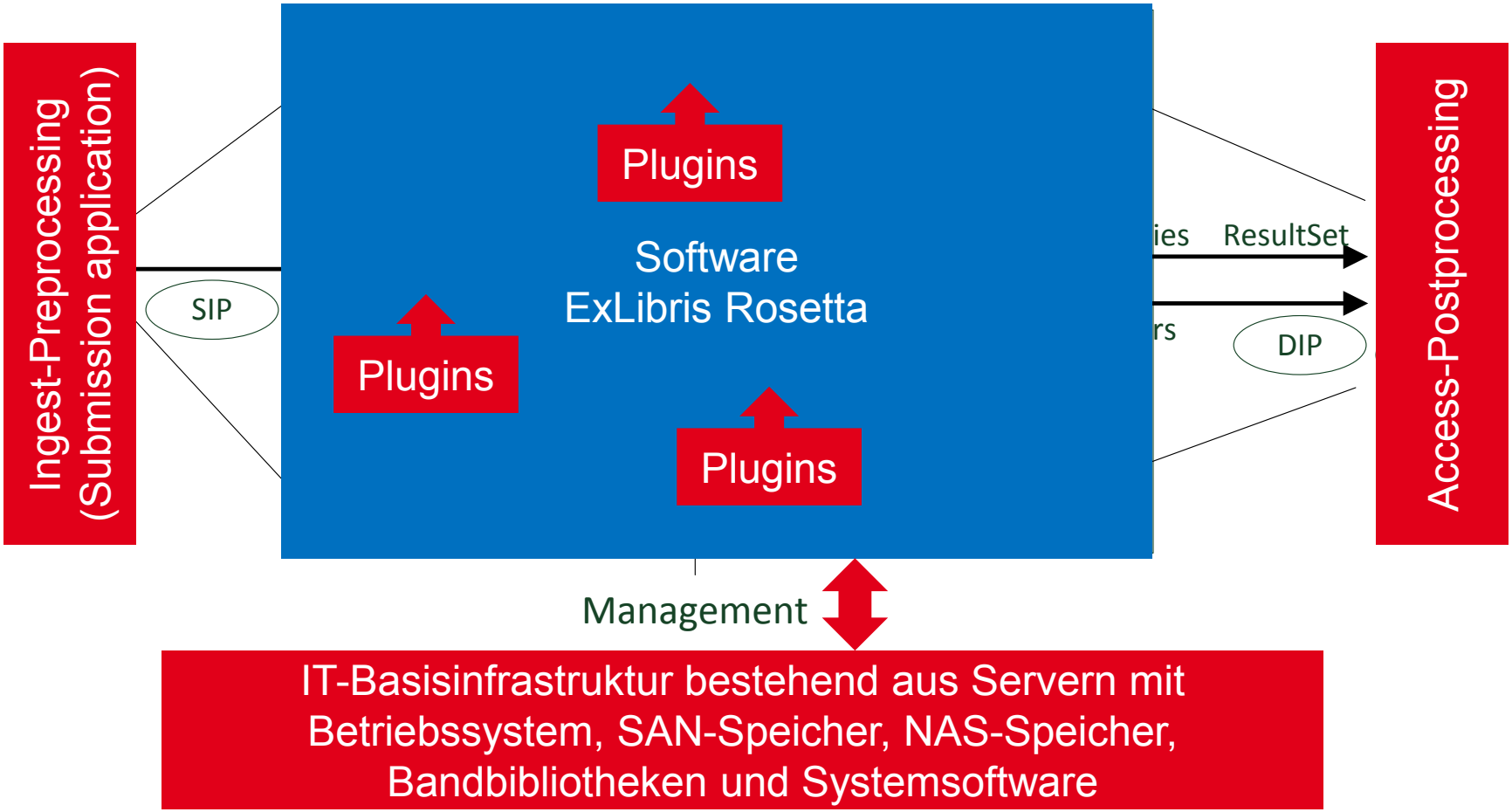
Übernahme ins Langzeitarchiv nur für **korrekte Dokumente**

SLUB
Wir führen Wissen.

SLUBArchiv

Projekt, Stand.

SLUBArchiv
Architektur



Enge und erfolgreiche Kooperation mit dem ZIH, das die IT-Basisinfrastruktur in seinem Rechenzentrum betreibt

SLUBArchiv
Vorgehen Goobi

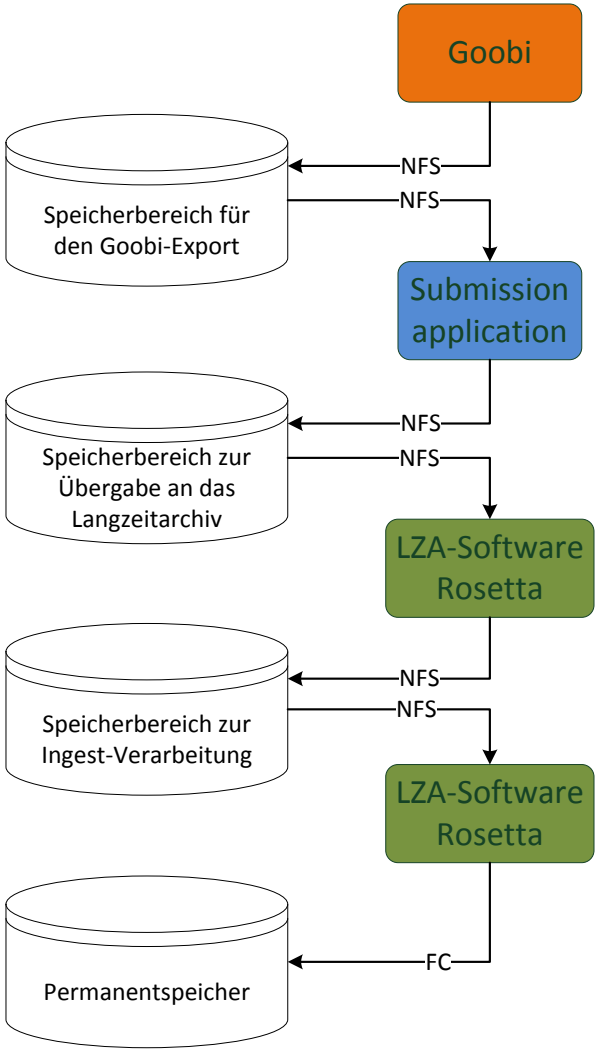
SLUBArchiv
Vorgehen Goobi: Ausgangsdaten

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

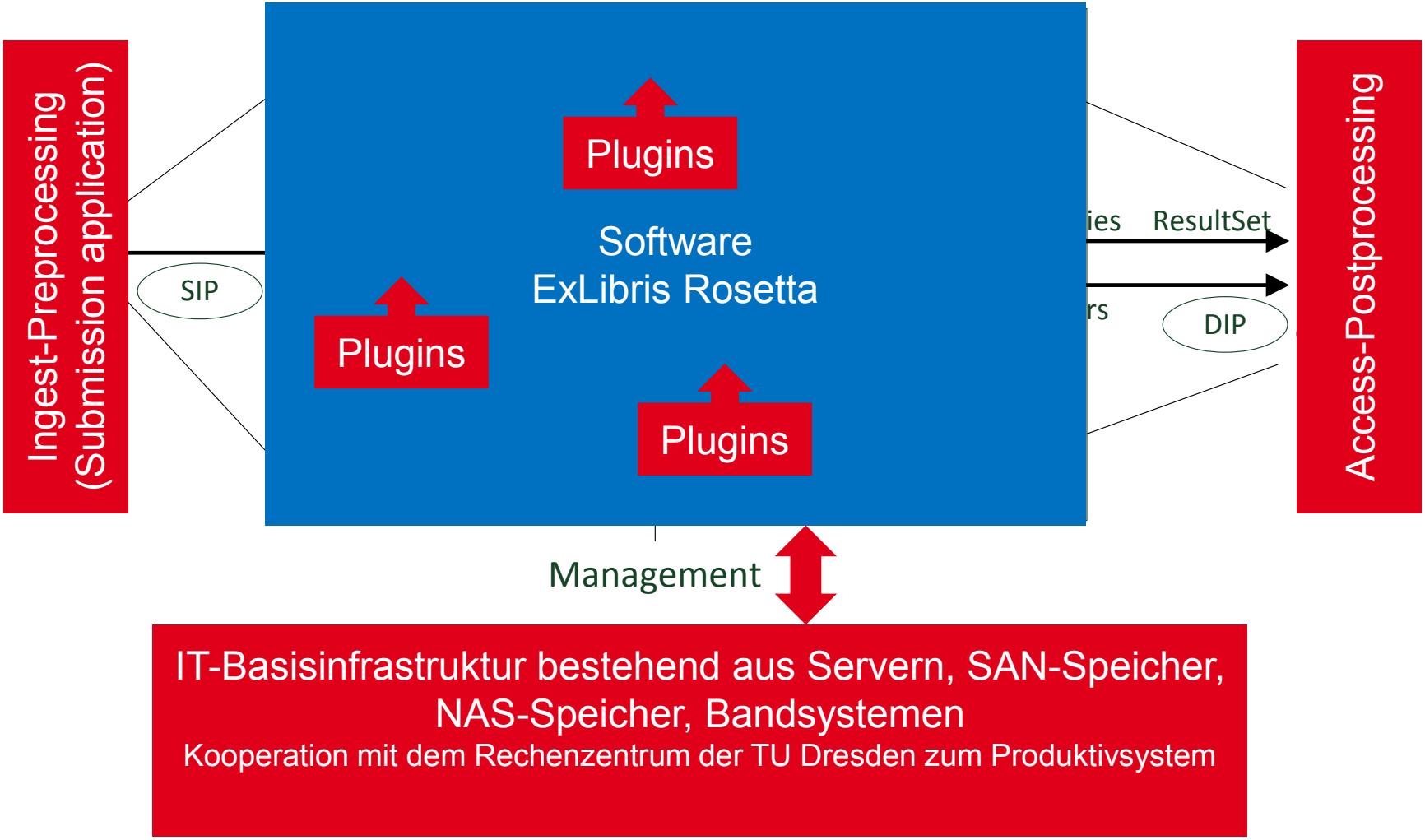


- Masterdaten in TIFF-Format (Baseline TIFF)
- Optional OCR-Daten im ALTO-XML-Format
- Metadaten in METS/MODS
- Prüfsummendatei

SLUBArchiv
Implementierung Goobi: automatisierte Datenübergabe



SLUBArchiv
Architektur



Digitales Langzeitarchiv SLUB
Implementierung Goobi: Submission Application

Automatische Vorbereitung einer digitalen Einheit für die Langzeitarchivierung durch ein Programm, die sogenannte Submission Application,

- Prüfen, ob ein neues Transferpaket vorhanden
- Prüfung der Vollständigkeit der Daten
- Prüfen der Integrität der Daten mittels Prüfsummen
- Transformation der Metadaten von Goobi METS/MODS nach Rosetta METS/DC
- Erstellen SIP (Submission Ingest Packet)
- Anstoßen der automatischen Übernahme durch die Langzeitarchivsoftware
- Protokollierung
- Fehlerbehandlung

Digitales Langzeitarchiv SLUB
Implementierung Goobi: Metadatentransformation

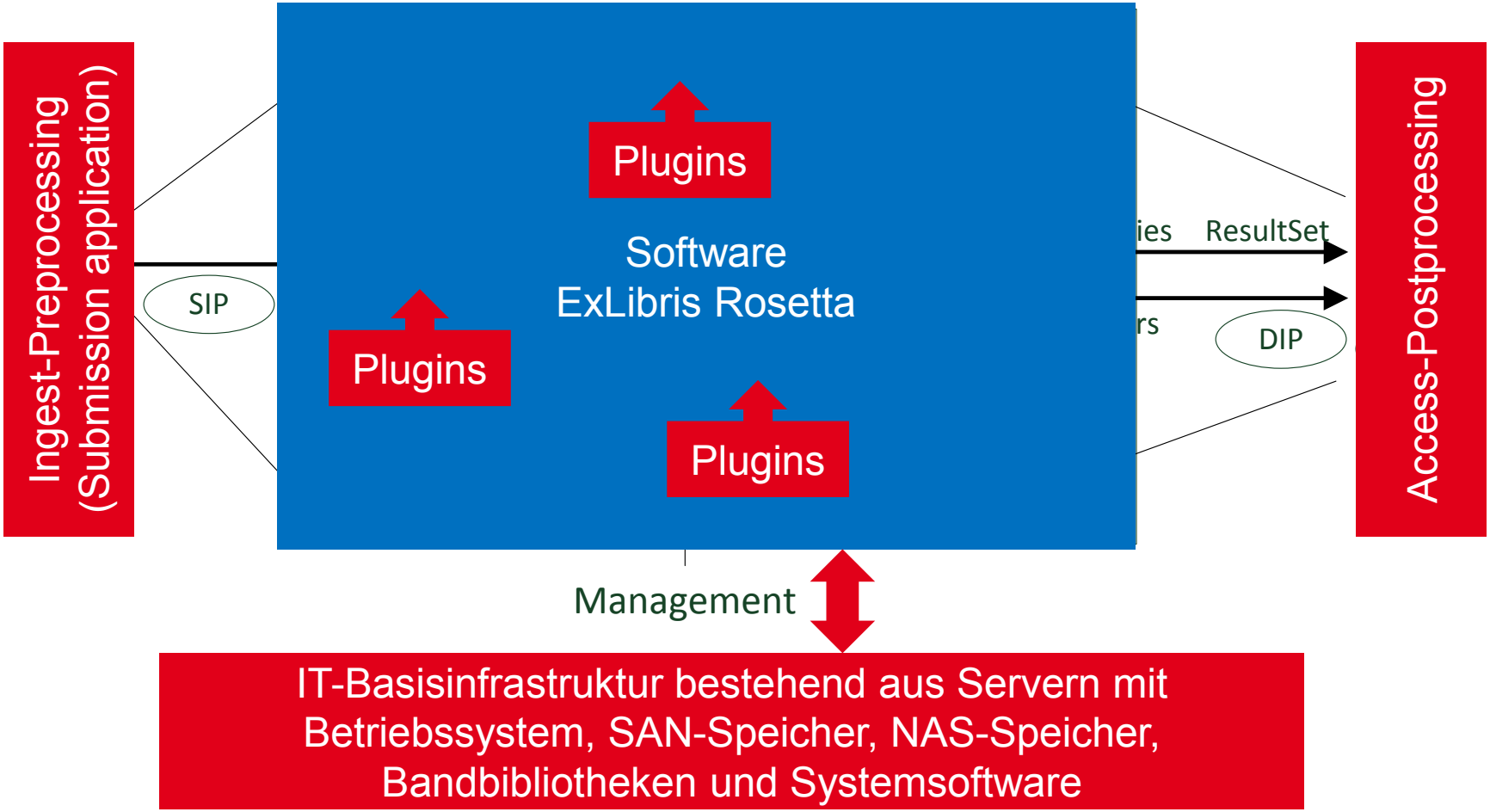
Verzeichnisstruktur Goobi (Detail eines Vorganges)

```
.
├──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.c
│   ├──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
:
```

Digitales Langzeitarchiv SLUB
Architektur



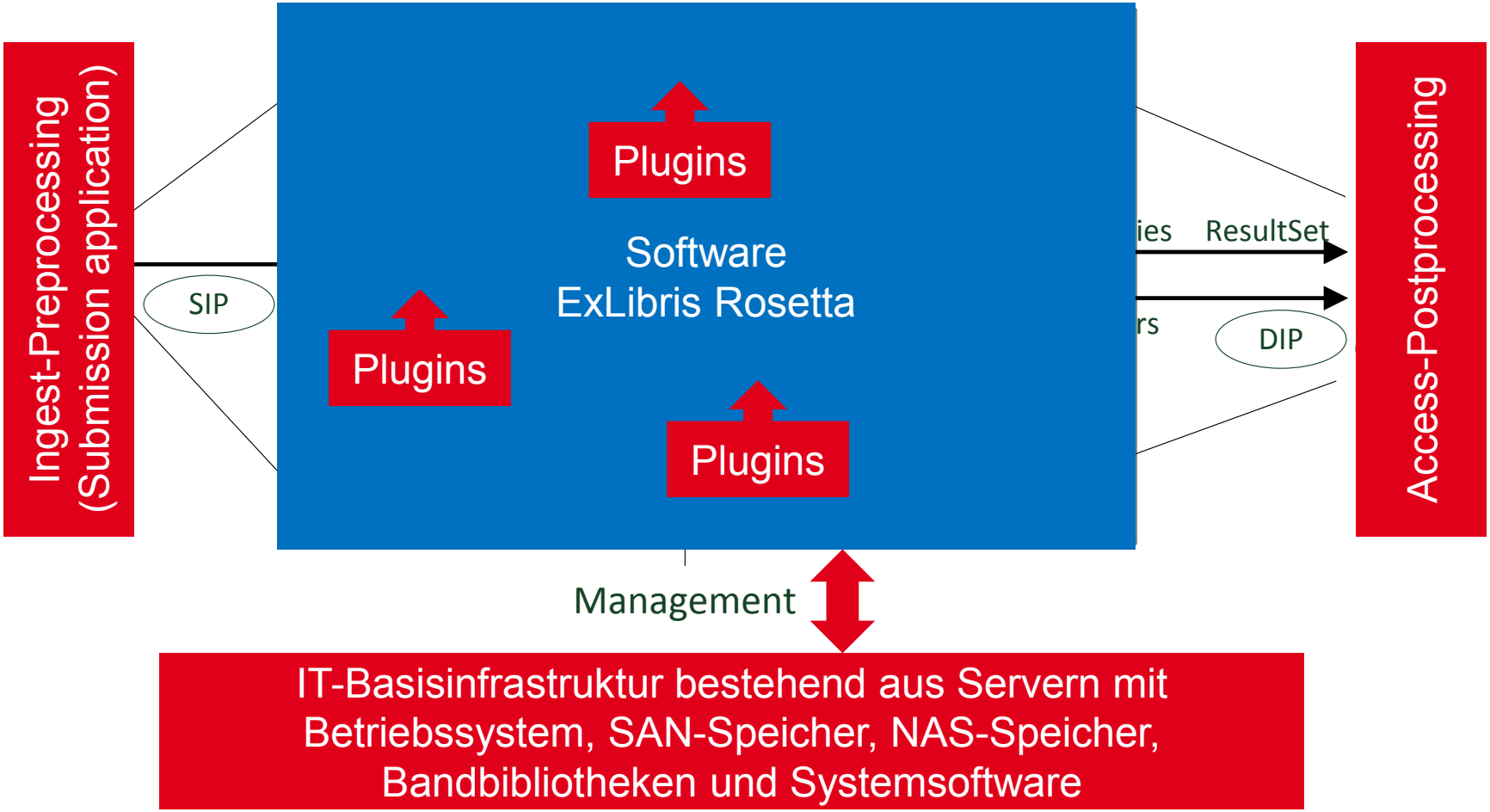
SLUBArchiv
Implementierung Ingest

muss für jeden Workflow konfiguriert und über Plugins angepasst werden

- Prüfen der Vollständigkeit und Integrität der Dateien (Prüfsummen)
- Virenprüfung
- Identifikation der Datenformate
- Prüfen der Korrektheit (Formatvalidierung)
- Extraktion von technischen Metadaten
- Erstellen eines Archivpaketes

Manueller Workflow für Bearbeitung von Fehlerfällen

SLUBArchiv
Architektur



Digitales Langzeitarchiv SLUB
Implementierung Goobi: Speicherung und Access

Speicherung

„Storage plugin“ steuert die Ablage der Daten im Archivspeicher - Lokalität der Daten steht im Vordergrund

- Alle Daten, die zu einer Einheit gehören (IE – Intellectual Entity) werden in einem Order abgelegt

Access

Erzeugen eines Nutzerpakets zum Re-Import in das Goobi-System

SLUBArchiv
Stand

Retrodigitalisate/Goobi

- Produktivbetrieb (pro Monat 3 – 4 TB)

Zertifizierung: Selbstevaluation DSA (Data Seal of Approval) eingereicht, Nestor Siegel geplant Q1/2016

Elektronische Publikationen/Qucosa

- Submission Application und Ingest-Workflows als Prototyp implementiert
- Abgaberichtlinie mit archivfähigen Datenformaten ist erstellt
- Datenanalyse für PDF Dokumente, Möglichkeiten der Validierung + Reparatur
- Anpassung des Submission Application und Aufnahme des Produktivbetriebes erfolgt nach Umstellung auf neue Repositorysoftware

Fotothek und Mediathek

- Datenanalyse erfolgt, Richtlinien für Retrodigitalisierung definiert
- Abstimmungen mit Fachabteilungen

SLUBArchiv
Erfahrungen

Stabiler Betrieb

Neu auftretende Probleme in Daten werden erkannt und behandelt, z. B.

- Falsches TIFF Datumsformat
- Multipage TIFF

Handshake-Verfahren zum Repository ist notwendig (für Fehlerbearbeitung)

Ausbau Reportmöglichkeiten (Rechenschaft, Statistik)

SLUBArchiv
Nächste Schritte

- Zertifizierung des SLUBarchiv mit dem nestor-Siegel
- Konzeption und Aufbau einer Dienstleistung
- Workflows für Geodaten , Forschungsdaten, Webseiten

SLUB
Wir führen Wissen.

SLUBArchiv

Zertifizierung

SLUBArchiv
Motivation

- Nachweis der vollständigen Implementierung
- Vollständige Dokumentation
- Aufbau von Vertrauen nach innen und außen

SLUBArchiv
Vorbereitung DSA

- Analyse der Kriterien
- Zusammenstellung fehlender Informationen aus Produktion und Nutzung
- Ansprechpartner für Rücksprache definiert
- Prioritäten gesetzt
- Erstellung fehlender Dokumentationen
- Erstellung und Einreichung des Dokuments Q2/2015

SLUBArchiv
Erfahrungen

- Langwieriger, aufwendiger Vorbereitungsprozess
- Kommunikation und Motivation im Haus

Fazit = Nutzen ist erheblich

SLUB
Wir führen Wissen.

SLUBArchiv

Zusammenfassung uns Ausblick

SLUBArchiv
Zusammenfassung und Ausblick

Digitale Langzeitarchivierung ist schon gut verstanden (OAIS, nestor-Dokumente, zahlreiche Forschungsprojekte)

Digitale Langzeitarchivierung ist auch noch Forschungsthema (Forschungsdaten, Emulation, Kosten, …)

Aufwand für den Aufbau ist erheblich, sowohl personell als auch finanziell, Dienstleistungsangebote werden entwickelt

SLUBArchiv ist produktiv im Einsatz; Ausbau ist geplant

Kooperationsvereinbarung mit der TU Dresden für den Betrieb der IT-Basisinfrastruktur für das SLUB Langzeitarchiv wird aktuell ergänzt

Lizenzerweiterung für die LZA-Software ist im Rahmen des Landesdigitalisierungs-programmes erfolgt

SLUB
Wir führen Wissen.

SLUBArchiv in Dresden

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

Juni 2015

SLUB
Wir führen Wissen.

Signifikante Eigenschaften

Aus den oben genannten Nutzungsszenarien hat die SLUB für die im Goobi-Workflow produzierten Digitalisate im TIF-Format die folgenden signifikanten Eigenschaften festgelegt:
Auflösung mind. 300 DPI; bei einer Digitalisierung aus Bestandserhaltungsgründen muss die Auflösung 600 DPI betragen.
Erhalt der Informationen der baseline TIFF-Tags und der folgenden zusätzlichen TIFF-Tags: Copyright, XMP und ICC
Erhalt der Informationen der zusätzlichen TIFF-Tags (wenn vorhanden): Exif IFD, Colormap, Extrasamples
Für die bibliographischen Metadaten, die derzeit im METS/MODS-Format vorliegen, müssen folgende Daten erhalten werden:
Bibliographischer Grundsatz an Metadaten wie Titel, Autor, Erscheinungsjahr, Verlag, Persistente ID, Prüfsummen des Original-Scans, Medienart (Buch/Karte/Zeitschrift), Serie/Reihe
Logische Struktur (Kapitel etc.)
Physische Struktur (Zurdnung einer Datei zu einer Seite des Originaldokumentes)

SLUB
Wir führen Wissen.

Nutzungsszenarien
Die Nutzungsszenarien sind in diesem Kapitel in abnehmender Priorität aufgeführt.
Lesen und Anschauen
Das wichtigste Ziel der Digitalisierung historischer Printmaterialien ist es, den Nutzern diese Dokumente unabhängig von Ort und Zeit zugänglich zu machen und damit natürlich auch die Originale zu schonen. Die optische Lesbarkeit der digitalisierten Dokumente ist somit eine wichtige Eigenschaft. Um diese Eigenschaft zu sichern, muss die Auflösung mindestens 100 DPI betragen und ein hoher Kontrast erhalten werden.
Bibliographische Einordnung
Um die Dokumente unter verschiedenen bibliographischen Aspekten zu finden, müssen die bibliographischen Metadaten erhalten werden. Dynamische Daten wie LinkedData sind davon ausgenommen. Es muss sichergestellt sein, dass man allein aus den Archivdaten notfalls einen, wenn auch rudimentären, Katalog aufbauen kann, der eine Basisrecherche über diese Dokumente nach Autor, Erscheinungsjahr, Titel, Verlag, Medienart und ggf. Serieninformationen ermöglicht. Persistente Identifikatoren sind ebenfalls zu erhalten (PPN, URN, DOI).
Maschinelle Verarbeitbarkeit
Ein weiteres wichtiges Nutzungsszenario ist die maschinelle Verarbeitbarkeit der digitalisierten Dokumente. Eine automatische OCR-Aufbereitung erlaubt beispielsweise das effiziente Auffinden von Dokumenten basierend auf deren Inhalt und die Anzeige des relevanten Textteils (unter Verwendung von Wortkoordinaten). Aber auch statistische Auswertungen durch Wissenschaftler z.B. linguistische Analysen sind so möglich. Für die maschinelle Verarbeitbarkeit muss die Auflösung mindestens 300 DPI betragen, bei Frakturschrift sind eventuell sogar 400 DPI nötig. Die Farb- bzw. Graustufeninformation muss ebenfalls erhalten bleiben, um die Fehlerrate niedriger zu halten. Um eine semantische Suche zu ermöglichen, ist die logische Struktur des digitalen Dokumentes erhalten.
Reproduktion
Ein weiteres Szenario ist die Erstellung einer möglichst originalgetreuen Kopie insbesondere zum Zweck der Bestandserhaltung, aber auch für Veröffentlichungen. Dafür ist es notwendig, die Originalgröße, die Farben und die Details zu erhalten. Die Auflösung sollte dementsprechend mindestens 600dpi betragen, der Farbraum als Profil hinterlegt sein und die Scanauflösung, Pixelzahl und ggf. bei fotografierten Digitalisaten die Objektivdaten erhalten werden. Die physische Struktur, d. h. die Zurdnung einer Scandatei zu einer Seite des Originals, muss ebenfalls erhalten bleiben, damit die korrekte Anordnung der Seiten in der Reproduktion sichergestellt werden kann.
Signifikante Eigenschaften
Aus den oben genannten Nutzungsszenarien hat die SLUB für die im Goobi-Workflow produzierten Digitalisate im TIF-Format die folgenden signifikanten Eigenschaften festgelegt:
Auflösung mind. 300 DPI; bei einer Digitalisierung aus Bestandserhaltungsgründen muss die Auflösung 600 DPI betragen.
Erhalt der Informationen der baseline TIFF-Tags und der folgenden zusätzlichen TIFF-Tags: Copyright, XMP und ICC
Erhalt der Informationen der zusätzlichen TIFF-Tags (wenn vorhanden): Exif IFD, Colormap, Extrasamples
Für die bibliographischen Metadaten, die derzeit im METS/MODS-Format vorliegen, müssen folgende Daten erhalten werden:
Bibliographischer Grundsatz an Metadaten wie Titel, Autor, Erscheinungsjahr, Verlag, Persistente ID, Prüfsummen des Original-Scans, Medienart (Buch/Karte/Zeitschrift), Serie/Reihe