SLUB
Wir führen Wissen.

Workflow Spezifikation für automatischen Ingest SLUBArchiv

SLUB Dresden

Version 2.0, 2020-02-17

Release Note 2.0 vom 2020-02-17

- Erläuterung führendes System
- Anpassungen für BagIt
- Wegfall komprimierter SIPs

Überblick

Dieses Dokument beschreibt die Workflowschnittstellen zwischen Dienstnehmer und dem SLUBArchiv.

Das Dokument gliedert sich in vier Teile. Im ersten Teil wird das Ablieferungsprotokoll spezifiziert. Der zweite Teil thematisiert den Zugriff auf die archivierten Daten. Teil drei beschreibt das Fehlerbehandlungsprotokoll. Teil vier beschäftigt sich mit Sondernutzungen.

Dieses Dokument ist ein Teil der Übernahmespezifikation für das SLUBArchiv. Zur Übernahmespezifikation gehören die folgenden weiteren Dokumente:

- In der **Übernahmevereinbarung** zwischen SLUB und Dienstnehmer sind die Daten, Ansprechpartner und organisatorischen Randbedingungen beschrieben.
- In der **Liste der archivfähigen und vom SLUBArchiv verarbeitbaren Formate** sind die Formate aufgeführt, die die SLUB als archivfähig bewertet und für die die Funktionalitäten der Formaterkennung, Formatvalidierung und Metadatenextraktion in einem ausreichenden Maß durch das SLUBArchiv gewährleistet werden können.
- Die **SIP-Spezifikation** beschreibt den Aufbau der Ablieferungspakete (englisch: Submission Information Package, SIP) mit denen der Dienstnehmer die zu archivierenden Dokumente für das SLUBArchiv bereitstellt.
- Das Dokument **Webservice SLUBArchiv** beschreibt Funktionen, die Dienstnehmer nutzen können, um Informationen über ihre Daten im SLUBArchiv über einen Webservice abzufragen.

Grundlagen

Vereinfachtes OAIS-Modell

Um die Arbeitsweise des SLUBArchiv zu verstehen, hilft es sich das vereinfachte [1] OAIS-Modell anzuschauen.

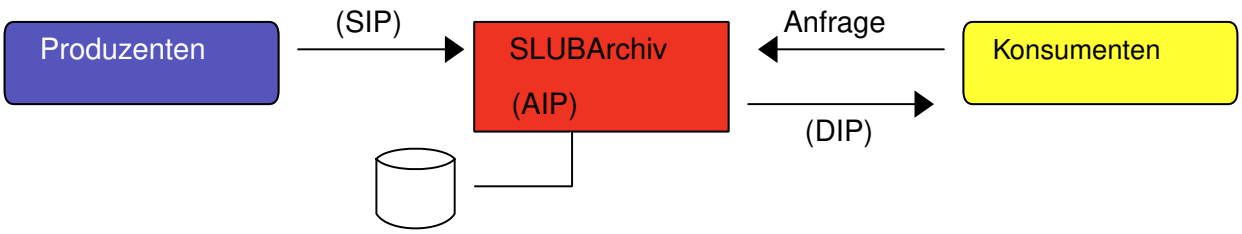

Abbildung 1. Vereinfachtes OAIS-Modell

Die Produzenten verpacken ihre zu archivierende intellektuelle Einheit in ein Submission Information Package (SIP) und senden es zum SLUBArchiv. Nach erfolgreicher Prüfung wird daraus ein Archival Information Package (AIP) gebaut, welches vom SLUBArchiv verwaltet wird. Auf Anfrage eines Konsumenten, wird ggf. ein Dissemination Information Package (DIP) gebaut und dem Konsumenten übergeben.

Produzenten und Konsumenten sind im OAIS-Modell Rollen, die es erleichtern unterschiedliche Anforderungen sauber auseinanderzuhalten. In der Praxis können Institutionen beide Rollen einnehmen.

Führendes System

Im Zuge der Formatmigration muss das SLUBArchiv Dateien von AIPs gegebenenfalls aktualisieren. Daher ist es notwendig zu verstehen, welches System wann zum führenden System wird.

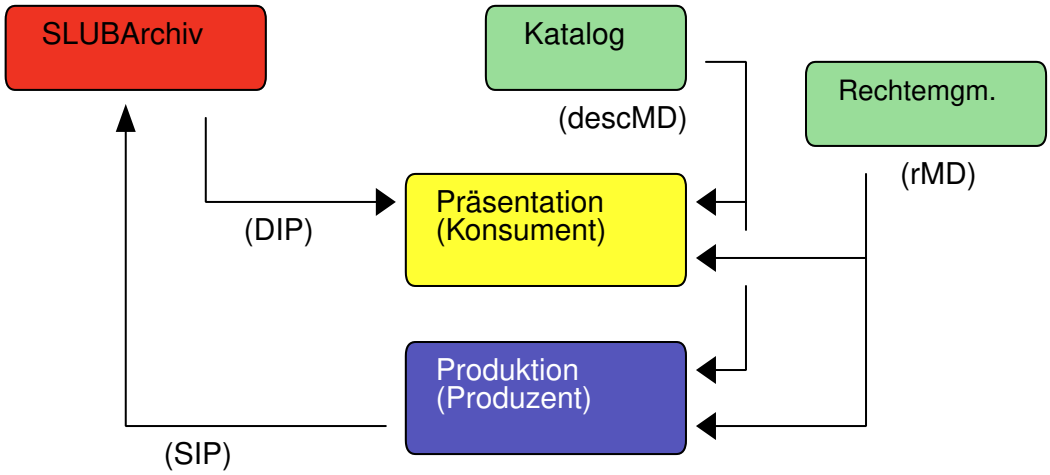

Abbildung 2. Problematik des führenden Systems

Bis zum Moment der erstmaligen Übernahme einer IE in das SLUBArchiv (Erstingest) ist das Produzentensystem das führende System für die eigentlichen Daten (zum Beispiel Digitalisate), der Katalog (oder das Produktionssystem) das führende System für die deskriptiven Metadaten (descMD), das Rechtemanagement das führende System für die Rechtemetadaten (rMD).

Mit der Archivierung wird das SLUBArchiv zum führenden System für die Daten, da diese im Zuge

von Erhaltungsmaßnahmen geändert werden müssen. Der Katalog, wie auch das Rechtemanagement ändern ihre Rolle als führendes System für die Metadaten nicht, sofern sicher gestellt ist, dass eine eindeutige ID existiert. Diese ID ist notwendig, damit eine eindeutige Zuordnung zwischen Metadaten und Daten im Archiv (via AIP) garantiert werden kann.

> Die dargestellten Systeme sind als Beispiel zu betrachten. In der Praxis ist es durchaus möglich, Rechte im Katalog abzubilden, oder mehrere Präsentationssysteme vorzuhalten.
>
> Wichtig ist nur, dass in Bezug auf das SLUBArchiv die vorhandenen Systeme je einmal in der Rolle des Produzenten bzw. des Konsumenten betrachtet werden. Dabei ist es unerheblich, wie die Systeme organisatorisch einzuordnen sind und ob es sich um technische Systeme oder Organisationseinheiten (z.B. Mitarbeiter einer Fachabteilung) handelt.

Der Workflow für eine erstmalige Einlieferung sieht so aus:

- im Produzentensystem wird eine eindeutige ID erzeugt
- die deskriptiven Metadaten werden im Katalog hinterlegt
- die Rechtemetadaten werden im Rechtemanagement hinterlegt
- die eigentlichen Daten werden im Produzentensystem hinterlegt (zum Beispiel Buchscans)
- das Produzentensystem erstellt ein SIP für einen Erstingest inklusive ID. Dies enthält die Daten, sowie einen Metadatensatz.

Während der Lebenszeit des AIPs im SLUBArchiv kann es, z.B. bedingt durch nötige Erhaltungsmaßnahmen, zu Änderungen an den Daten einer IE kommen. Dadurch entstehen Inkonsistenzen zwischen dem SLUBArchiv und dem Produzentensystem.

Um diese Inkonsistenzen aufzulösen, hat es sich bewährt die Systemgrenzen klar zu definieren und die Rollen Produzent und Konsument einzuführen.

Das SLUBArchiv vermittelt dabei als führendes System zwischen Produzenten und Konsumenten.

Möchte der Produzent Änderungen am Datenobjekt (IE) vornehmen, so fragt er in der Rolle des Konsumenten die aktuellen Daten vom SLUBArchiv ab, nimmt seine Änderungen vor (z.B. ergänzende Scans, aktualisierte Katalogeinträge oder geänderte Rechteauszeichnungen) und erzeugt als Produzent ein aktualisiertes SIP.

Im Falle eines AIP-Updates sieht der Workflow daher wie folgt aus:

- das Produzentensystem bekommt nach Anfrage via ID vom Katalog die aktuellen deskriptiven Metadaten
- das Produzentensystem bekommt nach Anfrage via ID vom Rechtemanagement die aktuellen Rechtemetadaten
- das Produzentensystem bekommt nach Anfrage via ID vom SLUBArchiv die aktuellen Daten (DIP)

- im Produzentensystem werden die Metadaten und Daten ggf. modifiziert
- das Produzentensystem erstellt ein SIP für ein AIP-Update inklusive ID. Dies enthält die modifizierten Daten, sowie einen modifizierten Metadatensatz.

Umfang der Metadaten

Im vorigen Abschnitt wurde dargelegt, welches System zum führenden System wird. Neben der verknüpfenden ID ist es trotzdem notwendig, dem SIP einen rudimentären Metadatensatz für den seltenen Fall mitzugeben, dass das Katalogsystem oder das Rechtemanagement einmal nicht zur Verfügung steht. In dem Fall ermöglichen diese Metadaten dem SLUBArchiv eine erste kontextuelle Einordnung der im SLUBArchiv gespeicherten Daten.

Die Metadaten eines SIP sollen und können keine Sicherung des Katalogs abbilden, da die deskriptiven Metadaten eines Katalogsystems zu volatil sind. Das SLUBArchiv dagegen wird als lose gekoppeltes Archivinformationsystem betrieben, welches zeitlich-stabile Datenobjekte archiviert.

Ablieferungsprotokoll (Ingest)

Austauschverzeichnis

Dem Dienstnehmer wird im Rahmen der Übernahmevereinbarung ein Austauschverzeichnis für den Ingest zugeteilt. Dieses Verzeichnis ist in der Regel per *SFTP* zugänglich.

Über dieses Verzeichnis werden die für das SLUBArchiv bestimmten Intellektuellen Einheiten (IE) als Submission Information Packages (SIP) gemäß SIP-Spezifikation der SLUB (SIP-Spezifikation) hinterlegt.

SIP

Ein SIP wird als Verzeichnisstruktur direkt im Austauschverzeichnis abgelegt. Der Kopiervorgang für mehrere SIPs sollte seitens des Dienstnehmers dabei seriell erfolgen, damit nach dem Übermitteln des ersten SIPs bereits mit dessen Verarbeitung begonnen werden kann.

Jedes SIP bildet ein Verzeichnis gemäß SIP-Spezifikation. Im Folgenden ein Beispiel:

Beispiel Austauschverzeichnis mit SIPs "122768" und "122769"

```
/mnt/import/
├── 122768
│   ├── bag-info.txt
│   ├── bagit.txt
│   ├── meta
│   │   ├── mods.xml
│   │   └── rights.xml
│   ├── data
│   │   └── images
│   │       └── scans_tif
│   │           ├── 00000001.tif
│   │           ├── 00000002.tif
│   │           ├── 00000003.tif
│   │           ├── 00000004.tif
│   │           ├── 00000005.tif
│   │           ├── 00000006.tif
│   │           ├── 00000007.tif
│   │           ├── 00000008.tif
│   │           ├── 00000009.tif
│   │           ├── 00000010.tif
│   │           ├── 00000011.tif
│   │           └── 00000012.tif
│   ├── manifest-md5.txt
│   └── tagmanifest-md5.txt
├── 122769
│   ├── bag-info.txt
│   ├── bagit.txt
│   ├── meta
│   │   ├── mods.xml
│   │   └── rights.xml
│   ├── data
│   │   └── images
│   │       └── scans_tif
│   │           ├── 00000001.tif
│   │           ├── 00000002.tif
│   │           ├── 00000003.tif
│   │           ├── 00000004.tif
│   │           ├── 00000005.tif
│   │           ├── 00000006.tif
│   │           ├── 00000007.tif
│   │           ├── 00000008.tif
│   │           ├── 00000009.tif
│   │           ├── 00000010.tif
│   │           ├── 00000011.tif
│   │           ├── 00000012.tif
│   │           ├── 00000013.tif
│   │           ├── 00000014.tif
│   │           ├── 00000015.tif
│   │           ├── 00000016.tif
│   │           ├── 00000017.tif
│   │           ├── 00000018.tif
│   │           ├── 00000019.tif
│   │           └── 00000020.tif
│   ├── manifest-md5.txt
│   └── tagmanifest-md5.txt
:
```

Um sicherzustellen, dass die SIPs erst prozessiert werden, wenn diese vollständig in das Austauschverzeichnis kopiert wurden, prüft die Submission Application die modification time des

SIP. SIPs, deren Dateien jünger als 10 Minuten sind, werden ignoriert.

Wenn die Submission Application auf ein SIP nicht zugreifen kann oder Fehler beim Zugriff festgestellt werden, wird das Fehlerprotokoll (siehe unten) angestoßen.

Löschen von Daten im Austauschverzeichnis

Das Löschen von Dateien erfolgt **ausschließlich** durch die SLUB. Das Löschen von SIPs erfolgt bei erfolgreichem Ingest sofort. Als erfolgreich gilt ein Ingest dann, wenn die IE vollständig prozessiert und im Langzeitspeicherbereich angekommen ist.

Fehlerbehaftete SIPs, die von der Submission Application entdeckt wurden, sowie Logdateien und Protokolle werden von der SLUB nach 14 Tagen aus dem Austauschverzeichnis gelöscht. Noch nicht prozessierte SIPs sind davon ausgenommen.

Treten SIP-Fehler während der Intensivprüfung in Rosetta zu Tage, behält sich die SLUB vor, diese SIPs durch einen Bearbeiter (Technical Analyst) abzuweisen. Die entsprechenden SIP-Verzeichnisse werden zeitnah aus dem Austauschverzeichnis gelöscht und der Dienstnehmer informiert.

Verantwortungsübergang

Die Verantwortung für die zu archivierenden IEs verbleibt bis zum erfolgreichen Ingest beim Dienstnehmer. Als erfolgreich gilt ein Ingest **erst** dann, wenn die IE erfolgreich vollständig prozessiert und im Langzeitspeicherbereich abgelegt ist.

Daher ist vom Dienstnehmer eine vollständige Kopie jeder zu archivierenden IE vorzuhalten, bis ihm die Rückmeldung über die erfolgreiche Archivierung vorliegt.

Das Austauschverzeichnis ist **nicht** als ausfallsicherer Speicher für die Aufbewahrung von Dateien zu betrachten! Es obliegt dem Dienstnehmer bis zum Verantwortungsübergang die Sicherheit seiner Dateien zu gewährleisten.

Der Dienstnehmer muss zudem sicherzustellen, dass die SIPs der Spezifikation (SIP-Spezifikation) entsprechen und die dort beschriebenen administrativen Metadaten für die Steuerung des Ingests korrekt angegeben sind.

Die SLUB hinterlegt im Austauschverzeichnis in der Regel täglich eine Protokolldatei der erfolgreichen Ingests (Erstingest und AIPUpdate). Diese Dateien tragen den Namen *Protokoll_SLUBArchiv_Erfolgreich-YYYYMMDD.txt*, wobei *YYYYMMDD* das Datum in ISO-Notation bezeichnet. Die Datei enthält pro Zeile einen Eintrag der folgenden Form, die Einträge sind durch Semikolon getrennt: *$externalWorkflow;$externalId;$timestamp;$sipDirectoryName* (UTF-8, Zeilenende mit Zeilenumbruch).

Beispiel Protokolldatei für erfolgreiche Ingest/AIPUpdate

```
kitodol;843722;2016-11-30T09:00:00;PPN-123456789_2016-11-28_10-00-00
kitodol;843723;2016-11-30T10:00:00;PPN-987654321_2016-11-29_17-00-00
```

Eine Abfrage der Protokolldaten kann alternativ auch über den Webservice (Beschreibung

Webservice SLUBArchiv für Produzenten) erfolgen.

Falls durch die Submission Application im Ingest Fehler festgestellt wurden, hinterlegt die SLUB zusätzlich eine Datei *Protokoll_SLUBArchiv_FEHLER-YYYYMMDD.txt*, die die in der Submission Application registrierten fehlerhaften Vorgänge aufführt.

Beispiel Protokolldatei für fehlerhafte Ingest/AIPUpdate

```
kitodol;843722;2016-11-30T09:00:00;PPN-123456789_2016-11-28_10-00-00
kitodol;843723;2016-11-30T10:00:00;PPN-987654321_2016-11-29_17-00-00
```

Der Dienstnehmer verpflichtet sich, die o.g. Protokolldateien regelmäßig zu prüfen.

Zugriffsprotokoll (Access)

Austauschverzeichnis

Dem Dienstnehmer wird ein Austauschverzeichnis für den Access zugeteilt. In der Übernahmevereinbarung ist die Zugriffsart näher spezifiziert, in der Regel ist dieses Verzeichnis per *SFTP*-Protokoll zugänglich. In diesem Verzeichnis werden auf Anforderung des Dienstnehmers vom SLUBArchiv bestimmte Intellektuelle Einheiten (IE) als Dissemination Information Packages (DIP) hinterlegt.

DIP-Anfrage

Der Dienstnehmer legt im Austauschverzeichnis als Anfragedatei eine Textdatei (UTF-8, Zeilenende mit Zeilenumbruch) mit dem Dateinamen *Wiederherstellung_YYYYMMDD.txt* an. Diese Datei enthält pro Zeile je einen mit Semikolon abgeschlossenen Eintrag für *$externalWorkflow* und *$externalId* (diese Werte entsprechen den gleichnamigen Werten in der SIP-Spezifikation, siehe da). Der Datumsanteil im Dateinamen der Anfragedatei dient dazu, mehrere Anträge zu unterscheiden. Um das Speichermedium Band zu schonen, ist es sinnvoll die Wiederherstellungen zu bündeln und maximal einen Auftrag pro Tag anzustoßen.

Beispiel Wiederherstellungsdatei für DIP-Anfrage

```
kitodol;843722;
kitodol;843723;
```

Das SLUBArchiv erstellt dann ein Verzeichnis mit dem Datum der Anfrage im Format *Wiederherstellung_YYYYMMDD* und exportiert dann darin jeweils ein DIP mit dem Verzeichnisnamen *${externalWorkflow}.${externalId}*.

Im Erfolgsfall löscht die SLUB die o.g. Anfragedatei. Im Fehlerfall wird das Verzeichnis *Wiederherstellung_YYYYMMDD* mit den teilweise exportierten DIPs gelöscht und das Fehlerprotokoll angestoßen. Dabei wird eine Protokolldatei mit dem Namen der Anfrage-Datei und der zusätzlichen Endung ".ERROR" erzeugt.

Löschen von Daten im Austauschverzeichnis

Das Löschen von Dateien erfolgt **ausschließlich** durch die SLUB. Das Löschen von Anfragedateien erfolgt bei erfolgreichem Access sofort. DIPs, fehlerhafte Anfragedateien sowie Logdateien und Protokolle werden von der SLUB nach spätestens 14 Tagen aus dem Austauschverzeichnis gelöscht.

Spezifika DIPs

Da das SLUBArchiv neben dem Erstingest auch AIPUpdates erlaubt, können bei der Archivierung von IEs mehrere Dateiversionen entstehen. Beim Access werden immer die jeweils letzten dem SLUBArchiv bekannten Stände der Dateien eines archivierten IEs zurückgegeben.

Fehlerprotokoll

Bei der Übernahme ins SLUBArchiv können im Ingest Fehler auftreten (bspw. nicht übereinstimmende Checksummen oder Validierungsfehler). Fehler werden per Email an die in der Übernahmevereinbarung hinterlegte Adresse übermittelt. Werden Fehler durch die Submission Application erkannt, werden die betroffenen Vorgänge als fehlerhaft markiert. In seltenen Fällen werden Fehler erst im Archivsystem festgestellt. In diesem Fall sind die Daten nicht mehr im Austauschverzeichnis enthalten.

Ein erneuter Ingest erfolgt sobald das SIP in einer gültigen Form wieder im Austauschverzeichnis vorliegt.

Falls im Rahmen einer DIP-Anforderung (Access) Fehler auftreten, werden diese durch die SLUB behandelt und per Email an die in der Übernahmevereinbarung hinterlegte Adresse übermittelt.

Sondernutzung

Sondernutzungen sind Nutzungsszenarien, die nicht durch die Beschreibung in den Teilen 1 bis 3 dieses Dokumentes abgedeckt sind. Aufgrund ihres projektartigen Charakters fallen zusätzliche Kosten an, die in der Regel nicht vertraglich abgesichert sind. Für die Inanspruchnahme der Sondernutzungen ist daher eine separate Vereinbarung zu schließen. In den folgenden Abschnitten werden die Sondernutzungen mit ihren Besonderheiten vorgestellt.

Wiederherstellung alter Versionen von AIPs

In bestimmten Situationen kann es notwendig werden, alte Versionsstände von IEs wiederherzustellen, die im SLUBArchiv in Form von AIPs vorliegen. Dieser Prozess kann nicht automatisiert erfolgen und erfordert seitens des Dienstnehmers eine Begründung, die in den Versionsinformationen des Archivs hinterlegt wird. Der Aufwand richtet sich nach der Zahl der betroffenen AIPs.

Umstellung externer Identifier

In einigen Fällen kann es unumgänglich sein, den *externalWorkflow* und die *externalId* (siehe SIP-Spezifikation) umzustellen. Da diese beiden Angaben vom SLUBArchiv benötigt werden, um zwischen Erstingest und AIPUpdate zu unterscheiden, muss eine solche Umstellung von der SLUB sorgfältig geplant und ausgeführt werden. Es müssen z.B. die Metadaten für alle betroffenen AIPs via AIPUpdate geändert werden. Dafür existiert zum aktuellen Zeitpunkt nur ein manueller Workflow. Der Aufwand richtet sich nach der Zahl der betroffenen AIPs.

Löschen von AIPs aus dem Archiv aus rechtlichen Gründen

In **Ausnahmefällen** kann es erforderlich sein, dass AIPs vollständig aus dem SLUBArchiv gelöscht werden müssen. Da das verwendete Archivsystem prinzipbedingt keine AIPs löschen kann, ist hierzu ein händisches Eingreifen notwendig. Um die Integrität des SLUBArchivs nicht zu gefährden, muss hierzu besonders sorgfältig und in Absprache mit dem Hersteller der Archivsoftware vorgegangen werden. Die Gründe für die Löschung werden im Archivsystem dokumentiert. Der Aufwand richtet sich nach der Zahl der betroffenen AIPs.

[1] weiterführend: nestor Handbuch, Neuroth et. al, 2009, Kapitel 4.2 Das Referenzmodell OAIS - Open Arcvhival Information System, http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:hebis:30-66418