



**SLUB**

Wir führen Wissen.

# Workflow Spezifikation für automatischen Ingest SLUBArchiv

SLUB Dresden

Version 1.2.1, 2017-09-30



- Klarstellung Protokolldatei
- Unterverzeichnis *data* statt *content*
- Rechtschreibfehler

## 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.

## Ablieferungsprotokoll (Ingest)

### Austauschverzeichnis

Dem Dienstnehmer wird im Rahmen der Übernahmevereinbarung ein Austauschverzeichnis für den Ingest zugeteilt. Dieses Verzeichnis ist 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 oder als gepacktes ZIP-File 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.

## SIP als Verzeichnis

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

*Beispiel Austauschverzeichnis mit SIPs "122768" und "122769"*

```
/mnt/import/
├── 122768
│   ├── data
│   │   ├── DerFr_417178867-19350102_anchor.xml
│   │   ├── DerFr_417178867-19350102_mets.xml
│   │   └── DerFr_417178867-19350102_year.xml
│   └── 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
├── sip.xml
├── 122769
│   ├── data
│   │   ├── DerFr_417178867-19350103_anchor.xml
│   │   ├── DerFr_417178867-19350103_mets.xml
│   │   └── DerFr_417178867-19350103_year.xml
│   └── 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
```



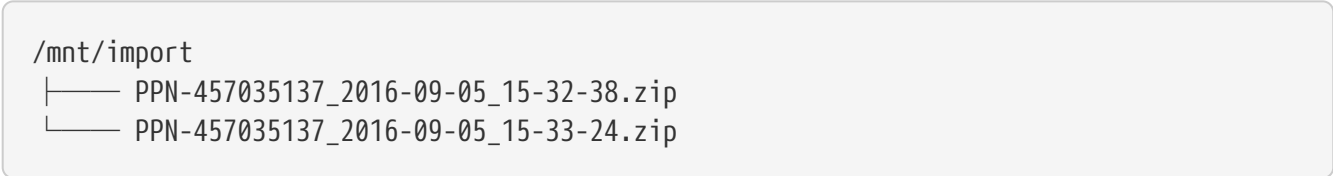
Um sicherzustellen, dass die SIPs erst prozessiert werden, wenn diese vollständig in das Austauschverzeichnis kopiert wurden, prüft die Submission Application die modification time des SIP. SIPs, deren Dateien jünger als 10 Minuten sind, werden ignoriert.

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

## SIP als ZIP

Ein gepacktes SIP wird von der SLUB automatisiert im Austauschverzeichnis ausgepackt. Die Submission Application ignoriert dabei alle gepackten SIPs, die jünger als 10 Minuten sind. Der Name der ZIP-Datei muss die Endung ".zip" tragen. Der Dateiname muss eindeutig sein. Im Folgenden ein Beispiel:

*Beispiel Austauschverzeichnis mit ZIP-komprimierten SIPs "PPN-457035137\_2016-09-05\_15-32-38" und "PPN-457035137\_2016-09-05\_15-33-24"*



In der Übernahmevereinbarung wird festgelegt, ob SIPs als Verzeichnisse oder als ZIP-Dateien übergeben werden.

Es wird empfohlen für Dateinamen folgendes Schema zu verwenden `<externalWorkflow>-<externalId>-<YYYY-MM-DD_hh-mm-ss>.zip`.

Sollten beim Entpacken Fehler festgestellt werden, wird das Fehlerprotokoll angestoßen. Dabei wird eine Protokolldatei mit dem Namen der ZIP-Datei und der zusätzlichen Endung ".ERROR" erzeugt.

Beispiel Austauschverzeichnis mit ZIP-komprimierten SIP "PPN-457035137\_2016-09-05\_15-32-38.zip", Fehlerdatei "PPN-457035137\_2016-09-05\_15-32-38.zip.ERROR" und ausgepacktem SIP "PPN-457035137\_2016-09-05\_15-33-24"

```
/mnt/import
├── PPN-457035137_2016-09-05_15-32-38.zip
├── PPN-457035137_2016-09-05_15-32-38.zip.ERROR
└── PPN-457035137_2016-09-05_15-33-24
    ├── data
    │   ├── eingdehaz_319037843_mets.xml
    │   └── images
    │       └── scans_tif
    │           ├── 00000001.tif
    │           └── 00000002.tif
    └── sip.xml
```

Das Löschen eines gepackten SIPs erfolgt nach dem erfolgreichen Entpacken sofort. Ein entpacktes SIP wird anschließend wie ein SIP-Verzeichnis behandelt.

## 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 gesicherte Kopienablage zu betrachten!

Der Dienstnehmer hat außerdem 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 Preingest 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. Dieses Verzeichnis ist per *SFTP* 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.

```
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  $\{\text{externalWorkflow}\}.\{\text{externalId}\}$ .

Für Dienstnehmer, die bei der Ablieferung komprimierte SIPs nutzen, legt das SLUBArchiv im Verzeichnis *Wiederherstellung\_YYYYMMDD* je ein komprimiertes DIP mit dem Namen  $\{\text{externalWorkflow}\}.\{\text{externalId}\}.zip$  ab. Im ZIP gibt es dann ein Verzeichnis mit dem Verzeichnisnamen  $\{\text{externalWorkflow}\}.\{\text{externalId}\}$ , welches das eigentliche DIP darstellt.

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 sowohl im Preingest als auch im Ingest Fehler auftreten (bspw. nicht übereinstimmende Checksummen oder Validierungsfehler). Diese Fehler werden per Email an die in der Übernahmevereinbarung hinterlegte Adresse übermittelt. Tritt der Fehler im Preingest auf, werden die betroffenen Vorgänge als fehlerhaft markiert. Tritt der Fehler im Ingest auf, 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.