Spezifikation von Transferpaketen und deren Übertragung an die Deutsche Nationalbibliothek mittels eines Hotfolders

Version 1.0
Stand: 14.03.2011

Redaktion: Stefan Hein, Matthias Neubauer

Deutsche Nationalbibliothek (Leipzig, Frankfurt am Main)
2011

<urn:nbn:de:101-2011031000>

Inhaltsverzeichnis:

0 Einleitung

1 Aufbau

2 Spezifikation

3 Beschreibung des Workflow-Konzeptes

I. Vorbereitung und Übertragung (Ablieferer)

II. Entgegennahme der Transferpakete (Deutsche Nationalbibliothek)

0 Einleitung

Das vorliegende Dokument definiert die Spezifikation von Transferpaketen sowie deren Übertragung an die Deutsche Nationalbibliothek (DNB) mittels eines Hotfolders.

1 Aufbau

Die Spezifikation beinhaltet zunächst die Definition der im weiteren Verlauf verwendeten Begriffe (Terminologie). Darauf aufbauend wird ein Workflow-Konzept dargestellt, das die Übertragung eines Transferpakets und deren Prüfung aufzeigt. Es soll verdeutlichen, welche Komponenten beim Transfer berücksichtig werden müssen.

2 Spezifikation

Terminologie

Ablieferer	die Institution, die zu archivierende Objekte an die Deutsche Nationalbibliothek übergibt
intellektuelle Einheit	eine Menge von Objekten, die inhaltlich zusammenhängend das zu archivierende Objekt bilden; Beispiel: ein Buch, eine Fotografie, eine Zeitschrift Eine intellektuelle Einheit kann aus mehreren weiteren intellektuellen Einheiten bestehen.
Content Information	die intellektuelle Einheit bzw. das digitale Objekt, das langzeitarchiviert werden soll
externe Metadaten	Externe Metadaten, bibliografische bzw. deskriptive Metadaten, aber auch technische Metadaten sind hier denkbar. Extern meint hier, dass diese Metadaten innerhalb einer separaten Datei vorliegen – vorzugsweise XML-kodiert. Für sammelpflichtige Materialien sind derzeit folgende Metadatenformate für die bibliografischen Angaben möglich: • ONIX for Books, Release 2.1, Revision 03 January 2006 • MARC21-XML • XMetaDissPlus Version 2.0
Transferpaket (TP)	ein Container zur Übertragung von digitalen Objekten; Die Inhalte eines Transferpakets werden im Folgenden Paket-Items genannt und lassen sich wie folgt klassifizieren: • Content Information (CI) • externe Metadaten (MD) In der Folge betrachten wir ein Transferpaket als Container einer intellektuellen Einheit, also ein Paket, das im späteren Verlauf zu einem Submission Information Package (SIP) verpackt wird und über einen eindeutigen URN adressierbar ist.

Um eine klare Trennung zwischen Content Information und externen Metadaten zu erreichen, sollten die Dateiobjekte, die zur Content Information gehören, in einem Unterverzeichnis /content/ abgelegt werden. Zur Verarbeitung (z. B. Kataloganreicherung) bestimmte externe Metadaten liegen außerhalb dieses Verzeichnisses, also auf der obersten Containerebene, können jedoch zusätzlich auch innerhalb von /content/ abgelegt werden, um sie auch als Content Information zu archivieren.

Dateiformat – TP

Als Dateiformat für das Transferpaket wird das Archivformat ZIP vorgeschlagen. Die Verwendung eines Container-Formats birgt im Vergleich zu einer einzelnen Einspielung jedes Paket Item in ein Verzeichnis folgende Vorteile:

- Komprimierung, somit kleinere TPs (Traffic!)
- kompakte TPs – Zusammengehörigkeit der Paket Items ist bereits auf Transferpaket-Ebene sichergestellt
- keine Beschränkungen durch das Dateisystem hinsichtlich der Anzahl von Verzeichnissen

Optional:

Checksummen des Transferpakets (TP-ChS)
Für die Checksumme des TP bietet sich an, diese als gleichnamige Textdatei mit der Erweiterung .md5 (für MD5) oder .sha1 (für SHA-1) separat zu übersenden.

Checksummen für Paket-Items innerhalb des Transferpakets (PI-ChS)
Pro Datei (Paket-Item) eine die Checksumme enthaltende gleichnamige Textdatei mit Endung .md5 (bei MD5) oder .sha1 (bei SHA-1)

Dateinamen-Konventionen
Es gelten die üblichen Beschränkungen:

1. keine Umlaute
2. keine Sonderzeichen
3. nicht mehr als 128 Zeichen

Dateinamen für externe Metadaten (Ablagepfad /)
Prinzipiell können mehrere Metadatenformate abgeliefert werden, verpflichtend ist nur eines der unter Punkt „externe Metadaten“ genannten Formate. Für das Metadatenset, das der DNB-Kataloganreicherung dient, gilt folgende Namenskonvention: catalogue_md.xml

Dateinamen können ansonsten beliebig vergeben werden;
Beispiel: <package-ID>.zip
<package-ID>.zip.md5

	Prüfung der Transferpakete geprüft wird die auf Seiten der DNB generierte Checksumme eines Transferpakets gegen die mitgelieferte Checksumme des Ablieferers. Dabei wird das gleiche Checksummen-verfahren angewendet, welches der Ablieferer durch die entsprechende Dateiendung der Checksummendatei vorgibt.
Verfahren bei Übertragung	Während des Upload-Prozesses sollte die Dateiendung des Transferpakets in „.tmp“ abgeändert werden, um eine vorzeitige Einreihung des Transferpakets in die Verarbeitung zu verhindern. Ist die Datei vollständig hochgeladen, muss die Endung wieder in den Ursprungszustand geändert werden. Zudem ist es ratsam ggf. mitzusendende Checksummendateien vor den jeweiligen Transferpaketen zu übertragen.
Ablieferer-ID	Jeder Ablieferer (bzw. abliefernde Verlag) erhält eine eindeutige Ablieferer-ID, welche auch als Name des HOME-Verzeichnisses innerhalb des Hotfolder-Speicherbereichs dient.
Paket-Item-Check-Summen (P-I-ChS)	Checksummen, die zur Wahrung der Integrität der einzelnen Paket-Items herangezogen werden
Transferpaket-Check-summen (TP-ChS)	Checksummen, die der Prüfung der Datenübertragung von Transferpaketen dienen
Hotfolder	ein SFTP-(SSH/FTP) oder WebDAV-Verzeichnis, das für die Einspielung von TPs für jeden Ablieferer separat zur Verfügung steht. Für die Ablieferung kann eine der beiden Varianten gewählt werden.
valides TP	ein Transferpaket, das positiv geprüft, also unter Wahrung der Integrität erfolgreich übertragen wurde
korruptes TP	negativ geprüftes, also fehlerhaftes Transferpaket (dabei wird keine Aussage darüber getroffen, ob das eigentliche Transferpaket oder die Checksummen-Datei fehlerhaft ist)
Import-Service	der Dienst, der erfolgreich übertragene Transferpakete entgegennimmt und für die Langzeitarchivierung weiterverarbeitet bzw. vorbereitet (z. B. technische Metadaten-Generierung).

3 Beschreibung des Workflows-Konzepts

I. Vorbereitung und Übertragung (Ablieferer)

1. Der Ablieferer stellt zunächst die zu archivierenden Paket-Items und ggf. deren zugehörige Checksummen zusammen.

2. Der Ablieferer packt die unter I.1. zusammengetragenen Daten in ein ZIP-Archiv. Dieses ZIP-Archiv stellt das Transferpaket (TP) dar.

3. Der Ablieferer kann ggf. aus dem unter I.2. erstellten Transferpaket eine für die Absicherung der Übertragung genutzte Checksumme (TP-ChS) erzeugen. Das verwendete Verfahren muss vorher abgesprochen werden; die Checksumme wird in eine wie unter Punkt 2 Spezifikation beschriebene gleichnamige Textdatei geschrieben. Der Dateiname muss den in „Dateiformat – TP“ definierten Konventionen entsprechen und darf nur die Checksummen-Zeichenkette selbst, also keine anderen Daten (z. B. Datum der Generierung, Tool usw.) enthalten.

4. Der Ablieferer überträgt nach ordnungsgemäßem Login (5) bei der DNB Transferpaket und ggf. Checksummen-Textdatei an den Hotfolder der DNB.

5. **Login:** Dem Ablieferer werden Zugangsdaten für einen nur ihm zugewiesenen Bereich im Remote-Verzeichnis zur Verfügung gestellt. Je nach verwendeter Übertragungsart kann der eigentliche Login-Vorgang unterschiedlich ausfallen, muss aber immer den Standard-Verfahren für die jeweilige Übertragungsart entsprechen.

II. Entgegennahme der Transferpakete (Deutsche Nationalbibliothek)

6. Die DNB prüft den Hotfolder auf neu eingetroffene TPs. Sofern neue TPs vorliegen, werden diese anhand der abgesprochenen Checksummenverfahren geprüft. Sollte keine Checksummendatei vorhanden sein, entfällt diese Prüfung.

7. Die DNB erstellt für jeden Tag einen Bericht über fehlerhafte Transferpakete (Checksumme invalide, Transferpaket nicht entpackbar), der dem Ablieferer im Übertragungsverzeichnis (Unterverzeichnis „logs“) bereitgestellt wird. Die bereitgestellten Berichte werden nach einem bestimmten Zeitraum (vorerst 7 Tage) von der DNB selbstständig gelöscht.

8. Die als fehlerhaft erkannten Transferpakete werden gelöscht. Alle anderen Transferpakete werden in ein Arbeitsverzeichnis des Import-Service verschoben.

Ansprechpartner:

Maren Brodersen (Abteilung Informationstechnik, Koordination Netzpublikationen)
m.brodersen@dnb.de

Cornelia Diebel (Abteilung Informationstechnik, Koordination Netzpublikationen)
c.diebel@dnb.de

Barbara Zeppenfeld (Abteilung Informationstechnik, Koordination Netzpublikationen)
b.zeppenfeld@dnb.de