DEUTSCHE NATIONAL BIBLIOTHEK

Spezifikation von Transferpaketen und deren Übertragung an die Deutsche Nationalbibliothek mittels eines Hotfolders im Rahmen der AREDO-Kooperation

Version 1.0
Stand: 02.04.2014

Redaktion: Karlheinz Schmitt

Deutsche Nationalbibliothek (Leipzig, Frankfurt am Main)
2014

<urn:>

Inhalt

1 Einleitung

Derzeit bietet die Deutsche Nationalbibliothek (DNB) für die Ablieferung von digitalen Objekten im Rahmen von AREDO zwei verschiedene Möglichkeiten an:

- OAI-Harvesting und
- Hotfolder-Verfahren.

Beide Verfahren bieten Partnern die Möglichkeit, einen automatisierten Geschäftsgang für eine beliebige Anzahl an digitalen Objekten zu etablieren. Beim OAI-Harvesting erfolgt die Ablieferung mittels OAI-PMH-Protokoll. Hierbei holt die DNB aktiv alle vom Partner bereitgestellten digitalen Objekte von einem Server des Partners ab und nimmt diese in das Langzeitarchiv auf. Eine detaillierte Beschreibung dieses Verfahrens ist im Dokument: Automatisiertes Abliefern über Harvesting-Verfahren zu finden und ist nicht Gegenstand dieser Spezifikation.

Das vorliegende Dokument beschreibt die Ablieferung von digitalen Objekten über ein Hotfolder-Verfahren der DNB. Im Folgenden werden die Voraussetzungen beschrieben, die für einen reibungslosen Ablauf der Ablieferung über diese Schnittstelle erfüllt sein müssen. Hierzu zählt zum einen die Festlegung des Aufbaus der übermittelten Daten, um die korrekte Verarbeitung seitens der DNB zu gewährleisten. Diese Festlegung wir im Folgenden Transferpaket genannt. In diesem Zusammenhang werden dem Partner bewusst unterschiedliche Optionen angeboten, um den beiderseitigen Aufwand so gering wie möglich zu halten. Zum anderen werden Fehlerbehandlung und Administration für das Ablieferverfahren behandelt.

2 Ablieferung von Transferpaketen

Für die ordnungsgemäße Aufnahme von digitalen Objekten in das Langzeitarchiv der DNB werden für jedes digitale Objekt vom Kooperationspartner mindestens zwei Informationen benötigt:

1. **Zu archivierende digitale Objekte:**
   Bei den digitalen Objekten kann es sich um digitale Objekte in beliebigen Dateiformaten handeln. Zum Zwecke des gemeinsamen Risikomanagements müssen jedoch die verschiedenen Dateiformate möglichst im Vorfeld abgesprochen und in eine Format-Policy eingetragen werden. (siehe Format-Policy).

2. **Prüfsumme für Transferpaket**
   Um die Dateiintegrität der übermittelten Transferpakete beginnend bei der Partnerinstitution lückenlos gewährleisten und nachweisen zu können, ist es erforderlich mittels eines Hash-Verfahrens eine Prüfsumme über das gesamte Transferpaket zu erstellen und auf dem Hotfolder gleichzeitig mit dem Transferpaket zu hinterlegen. Die genaue Spezifikation ist im Abschnitt Transferpaket-Spezifikation zu finden.

3. **Optional: Prüfsumme für jedes einzelne digitale Objekt**
   Um die Dateiintegrität eines digitalen Objektes über eine unbestimmte Zeit nachweislich sichern zu können, ist es vorteilhaft, dass die Partnerinstitution

Prüfsummen auch für jedes einzelne digitale Objekt innerhalb des Transferpaketes generiert. Anhand dieser Prüfsumme wird die Dateiintegrität zwischen der DNB und der Partnerinstitution nachgewiesen. Sofern eine Partnerinstitution Prüfsummen für digitale Objekte nicht mitliefert, generiert die DNB sofort nach Aufnahme der digitalen Objekte in die DNB Prüfsummen. Diese Prüfsumme wird im weiteren Verlauf der Langzeitarchivierung als Prüfkriterium für die Dateiintegrität der digitalen Objekte herangezogen.

4. **Optional: Metadatensatz im DC-Simple-Format:**
   Ein Metadatensatz mit beschreibenden Informationen bzgl. der im Transferpaket enthaltenen digitalen Objekte. Die Metadaten werden in das Verwaltungssystem des Langzeitarchivsystems übernommen und dienen zur Unterstützung einer erweiterten Suche innerhalb des Langzeitarchivs.

5. **Metadatensatz zur Pflichtablieferung**
   Nur im Falle der kombinierten Abgabe: Pflichtkriterium

   Sofern eine kombinierte Abgabe – Pflichtablieferung mit gleichzeitigem Wunsch zur kooperativen AREDO Langzeitarchivierung besteht, müssen die Kriterien zur Pflichtabgabe von Netzpublikationen erfüllt werden. Hierzu gehört, neben der Ablieferung der Netzpublikation, auch die Lieferung von beschreibenden Metadaten bzgl. der Netzpublikation. Die Metadaten werden direkt in den Katalog der DNB importiert und dienen als bibliografischer Nachweis. Momentan angebotene Metadatenformate werden im Abschnitt 2.5 benannt.

2.1 Übertragungsverfahren Hotfolder

Mit Betriebsetzung von AREDO bietet die DNB jedem Ablieferer die Möglichkeit, Transferpakete in ein spezielles, dem jeweiligen Partner zugewiesenes Verzeichnis innerhalb der DNB abzulegen und dadurch die Ablieferung von Transferpaketen durchzuführen. Dieser Ordner wird von der DNB überwacht, so dass einerseits aktuelle sicherheits- und rechterelevante Richtlinien jederzeit eingehalten werden und andererseits die Aufnahme von Transferpaketen automatisiert und zeitnah erfolgen kann.

Der produktiven Ablieferung von Transferpaketen geht eine Testphase voraus, in der zwischen der Partnerinstitution und der DNB zum einen bei Bedarf (kombinierte Abgabe) der verwendete Metadatenstandard abgesprochen, zum anderen die Einhaltung der im weiteren beschriebenen Transferpaket-Spezifikation getestet wird. Ebenso werden in der Testphase die vom Partner übergebenen Dateiformate identifiziert und in der Format-Policy vermerkt.

2.2 Übertragungsprotokolle

Für die Übertragung der Transferpakete auf den institutseigenen Hotfolder in der DNB werden zwei verschiedene Übertragungsprotokolle angeboten, die wegen ihrer einfachen Handhabung ausgewählt wurden:

- **S**SH **F**ile **T**ransfer **P**rotocol (SFTP)
- **Web**-based **D**istributed **A**uthoring and **Ve**rsioning (WebDav)

Für jeden Partner steht nach einer Registrierung ein separates Konto mit eigenen Zugangsdaten (Benutzername und Passwort) zur Verfügung.

2.3 Transferpaket-Spezifikation

Ein Transferpaket setzt sich grundsätzlich aus den zu archivierenden digitalen Objekten und einer Checksummendatei (außerhalb des Transferpaketes) zusammen. Hierfür ist die Übertragung des Transferpakets in einem aktuellen Container-Format üblich. Zurzeit werden von der DNB die folgenden Container-Formate akzeptiert:

- ZIP
- TAR

Der Aufbau eines Transferpaketes auf oberster Ebene ist in Abbildung 1 verdeutlicht. Auf der obersten Ebene innerhalb des Containers wird ein Ordner mit dem Namen „content“ vorausgesetzt, in dem die digitalen Objekte, die archiviert werden sollen liegen. Des Weiteren muss neben der Transferpaketdatei eine Datei mit Namen „<Transferpaket-ID>.zip.md5/sha1“ vorhanden sein, welche die Prüfsumme im ASCII-Format enthält.

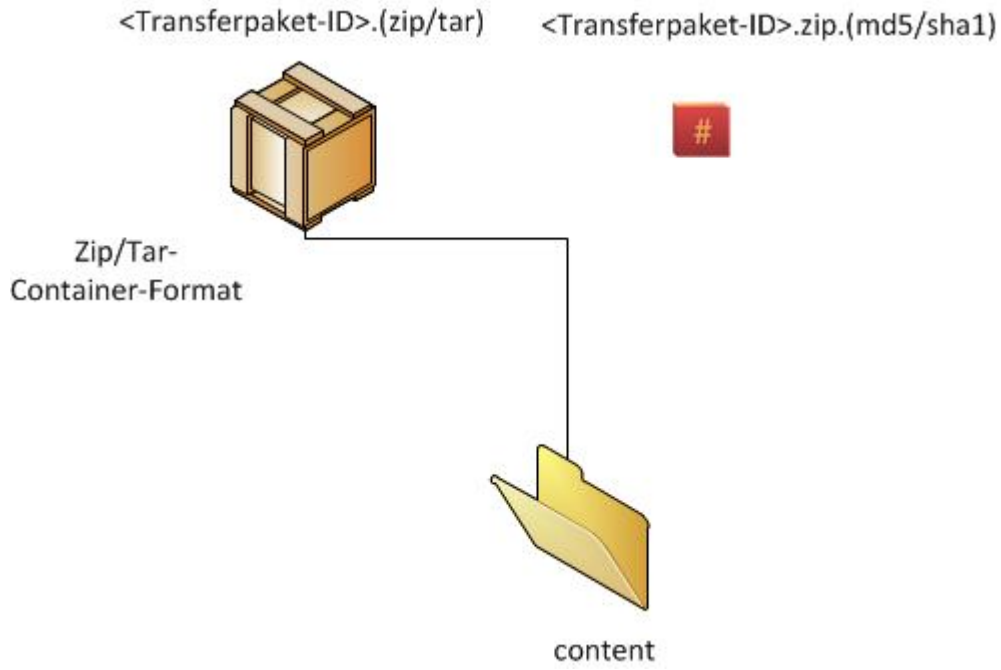

Abbildung 1: Aufbau eines Transferpaketes für AREDO auf der obersten Ebene.

Transferpakete: das Verzeichnis „content“

Alle zu einem Transferpaket gehörenden digitalen Objekte sind in dem „content“-Verzeichnis abzulegen. Der Aufbau innerhalb des „content“-Verzeichnisses kann vom Partner frei definiert werden. Im content-Verzeichnis befindliche Containerformate (z.B.: ZIP/TAR) werden als semantische Einheit betrachtet und nicht weiter entpackt. Abbildung 2 zeigt beispielhaft mögliche Optionen:

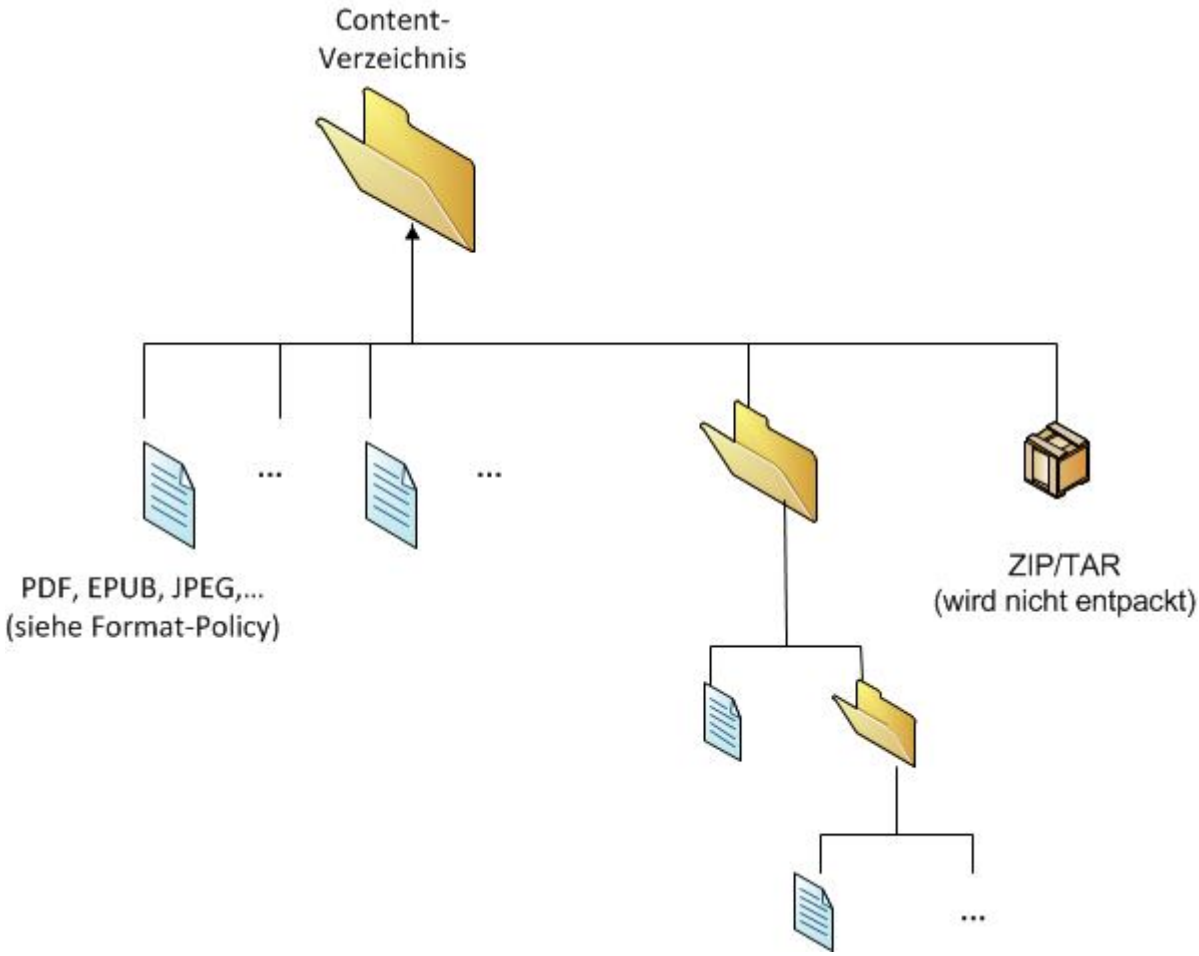

Abbildung 2: Diese Grafik verdeutlicht mögliche Strukturierungsvarianten des Inhalts eines content-Verzeichnisses.

Folgende Beschränkungen müssen jedoch eingehalten werden:

1. **Dateinamen-Konventionen**
   Es gelten die üblichen Beschränkungen:
   - keine Umlaute
   - keine Sonderzeichen
   - keine Leerzeichen
   - nicht mehr als 128 Zeichen

2. **Gesamtanzahl enthaltener Dateien im content-Verzeichnis**
   Aus technischen Gründen dürfen die im content-Verzeichnis enthaltenen Dateien eine Anzahl von 4999 nicht überschreiten.

3. **Maximale Größe eines digitalen Objektes innerhalb des Transferpaketes**
   Aus technischen und Performance-Gründen darf die maximale Größe eines digitalen Objektes innerhalb eines Transferpaketes nicht größer als 2 Gigabyte sein.

4. **Maximale Größe eines Transferpaketes**
   Aus Performance-Gründen ist die maximale Größe eines Transferpaketes auf max. 50 Gigabyte beschränkt.

5. **Erlaubte Dateiformate**
   Generell liegen keinerlei Formatbeschränkungen vor. Wie oben beschrieben können alle in der Testphase vereinbarten und vermerkten Dateiformate geliefert werden.
   Bei der Ablieferung der Transferpakete erfolgt eine automatische Prüfung, ob nicht abgesprochene Dateiformate mitgeliefert wurden. Wurde ein nicht abgesprochenes Dateiformat erkannt, erfolgt eine Rücksprache mit dem Partner, und das Dateiformat wird ggf. in die gemeinsame Format-Policy nachgetragen. Alle Dateien in den für einen Partner eingetragenen Formaten erfahren während des Ingest-Prozesses eine technische Qualitätsprüfung. Zwischen der DNB und dem Partner kann bzgl. der Qualität der Dateiformate eine Mindestqualität vereinbart werden, die erforderlich ist, um mit dem Ingest-Prozess fortzufahren. Eine genaue Absprache erfolgt in der gemeinsam zu erarbeitenden Format-Policy.

6. **Einschränkung bei Verwendung von Containerformaten im „content"-Verzeichnis**
   Containerformate, die vom Partner innerhalb des „content"-Verzeichnisses abgelegt werden, werden nicht entpackt. Die Container werden als Containerformat in das Langzeitarchiv der DNB übernommen.

Transferpakete: Die Checksummendatei

Zur Wahrung der Authentizität der abgelieferten digitalen Objekte wird von der DNB die Verwendung einer gegenseitigen Checksummenprüfung durchgeführt. Hierbei wird für jedes Transferpaket eine Prüfsumme berechnet, über die sich ungewollte Veränderungen des Transferpaketes ermitteln lassen.

Zurzeit werden von der DNB zur Checksummenprüfung die folgenden Hash-Verfahren eingesetzt:

- MD5
- SHA-1

Für den Einsatz eines gegenseitigen Checksummenvergleichs zwischen der DNB und dem Partner muss für das zu prüfende Transferpaket eine Checksumme aus den oben genannten Verfahren mitgeliefert werden. Abbildung 1 zeigt die hierfür notwendige Dateibezeichnung. Zum einen muss der Name des Transferpaketes vollständig (mit Dateierweiterung) übernommen werden und zum anderen muss, abhängig vom gewählten Hash-Verfahren, die Dateibezeichnung um den Zusatz „.md5" oder „.sha1" erweitert werden. Die Speicherung der Checksummendatei muss im ASCII-Format erfolgen.

2.4 Optionale Erweiterungen

Zusätzlich zur Checksummenprüfung für Transferpakete oder auf Wunsch alternativ kann die Checksummenprüfung für jedes digitale Objekt innerhalb eines Transferpaketes durchgeführt werden. Hierfür muss vom Partner für jedes digitale Objekt eine entsprechende Checksummendatei erzeugt und im gleichen Verzeichnis in dem das digitale Objekt liegt mitgeliefert werden. Hierdurch lassen sich jederzeit ungewollte Veränderungen am digitalen Objekt feststellen.

Für die Checksummenerzeugung gelten die gleichen Regeln wie für die Checksummenerzeugung für Transferpakete.
Abbildung 3 verdeutlicht exemplarisch die hierfür notwendige Spezifikation:

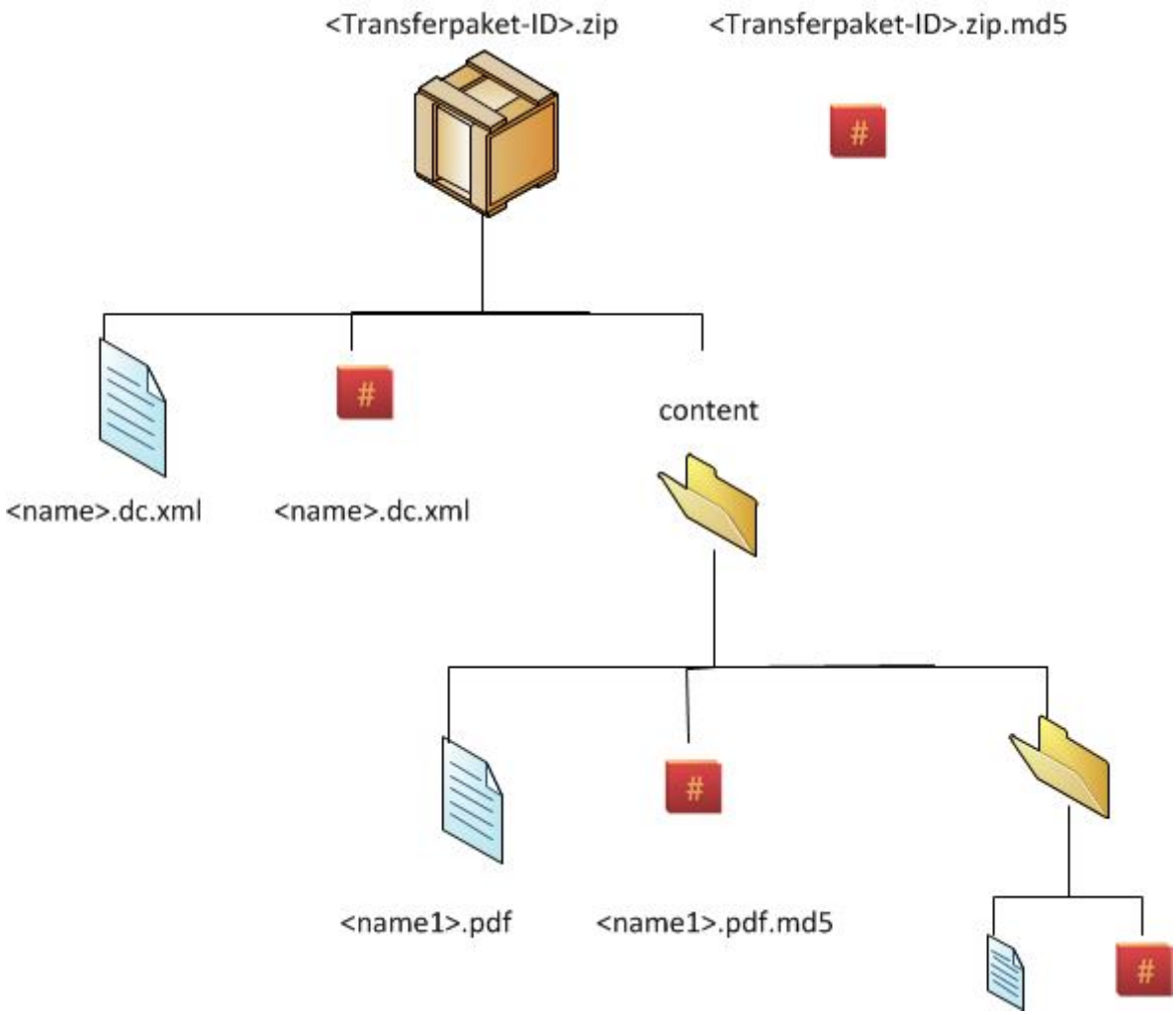

Abbildung 3: Transferpaket mit externer und interner Checksummenprüfung

Geprüft wird die auf Seiten der DNB generierte Checksumme eines Transferpakets (siehe <Transferpaket_ID>.zip) gegen die mitgelieferte Checksumme des Ablieferers (<Transferpaket_ID>.zip.md5). Dabei wird dasjenige Checksummen-Verfahren angewendet, welches der Ablieferer durch die entsprechende Dateiendung (md5/sha1)

der Checksummendatei vorgibt. Sind beide Checksummen identisch, erfolgt die Weiterverarbeitung für die Übernahme in das Archivsystem der DNB. Dieselbe Vorgehensweise erfolgt auf Checksummen für alle Teile des Transferpakets.

Transferpakete: Mitlieferung von deskriptiven Metadaten
Auf Wunsch kann von jedem Partner zum Transferpaket gehörende deskriptive Metadaten mitgeliefert werden. Die deskriptiven Metadaten werden von der DNB aufbereitet, so dass anhand dieser Metadaten das Transferpaket innerhalb des Langzeitarchivs gesucht und zurückgeliefert werden kann. Werden vom Partner keine deskriptive Metadaten mitgeliefert, erfolgt die Suche eines Transferpaketes lediglich auf Basis der zugewiesenen oder übernommenen URN des Transferpaketes.

Als einfaches Metadatenformat für die Lieferung von deskriptiven Metadaten ist das DC-Simple-Format einzuhalten.

Abbildung 3 zeigt sowohl die erforderliche Position der DC-Simple-Datei im Transferpaket als auch die Namenskonvention. Die Datei muss außerhalb des "content“-Verzeichnisses in der obersten Ebene des Transferpaketes angesiedelt sein. Der Name der DC-Simple-Datei kann unter Beachtung der in Abschnitt 2.3 angegebenen Dateinamenskonventionen beliebig gewählt werden. Die Dateierweiterung muss dagegen zwingend mit „.dc.xml“ enden.

2.5 Der Transferpaketaufbau für die kombinierten Abgabe

Im Rahmen der Kooperation kann von der Partnerinstitution auch die sogenannte kombinierte Abgabe gewählt werden. Unter diesem Aspekt wird der Besonderheit Rechnung getragen, dass Partner gleichzeitig sowohl das AREDO-Kooperationsangebot nutzen können als auch der eigenen Ablieferungspflicht nachkommen können. Mit der kombinierten Abgabe kann mit einer einzelnen Ablieferung beides erreicht werden.

Hierfür muss die Transferpaketstruktur um die Anforderungen der Pflichtabgabe von Netzpublikationen über den Hotfolder der DNB (Transferpaket-Spezifikation für Netzpublikationen) erweitert werden.

Zwingend für die kombinierte Abgabe ist die Beigabe deskriptiver Metadaten zur abzugebenden Netzpublikation, welche im Folgenden beschrieben wird.

Metadatensatz: die Datei „catalogue_md.xml“

Für sammelpflichtige Materialien sind derzeit folgende Metadatenformate für bibliografische Angaben möglich:

- ONIX for Books, Release 2.1, Revision 03 January 2006
- MARCXML
- XMetaDissPlus Version 2.2

Format und Umfang der Metadaten werden während der Testphase mit dem Ablieferer vereinbart. Die DNB hat in einem Kernset die Pflichtfelder für das Metadaten-Format

„ONIX for Books, Release 2.1“ festgelegt und stellt auf der Website ein Beispiel für ein Transferpaket mit Metadaten im Format „ONIX for Books, 2.1“ bereit.
Metadaten-Kernset:
http://nbn-resolving.de/urn:nbn:de:101-2012022219
Beispiel Transferpaket:
http://files.dnb.de/standards/beispieldatei_hotfolder.zip

Die DNB, KNV, Libri, NewBooks Services, Umbreit und das VLB haben gemeinsam ein Papier "Best practices ONIX for Books (Version 2.1) - E-Book-Standardmeldung" erarbeitet. Darin wird eine Empfehlung für ONIX Meldungen von E-Books in der ONIX Version 2.1 gegeben.
http://www.dnb.de/SharedDocs/Downloads/DE/DNB/netzpub/best_practices_onix_for_books.pdf?__blob=publicationFile

Abbildung 4 zeigt die Transferpaketstruktur für die kombinierte Abgabe. Die Datei „catalogue_md.xml“ muss, ebenso wie die optionale DC-Simple-Datei, im obersten Verzeichnis der Transferpaketstruktur abgelegt werden.

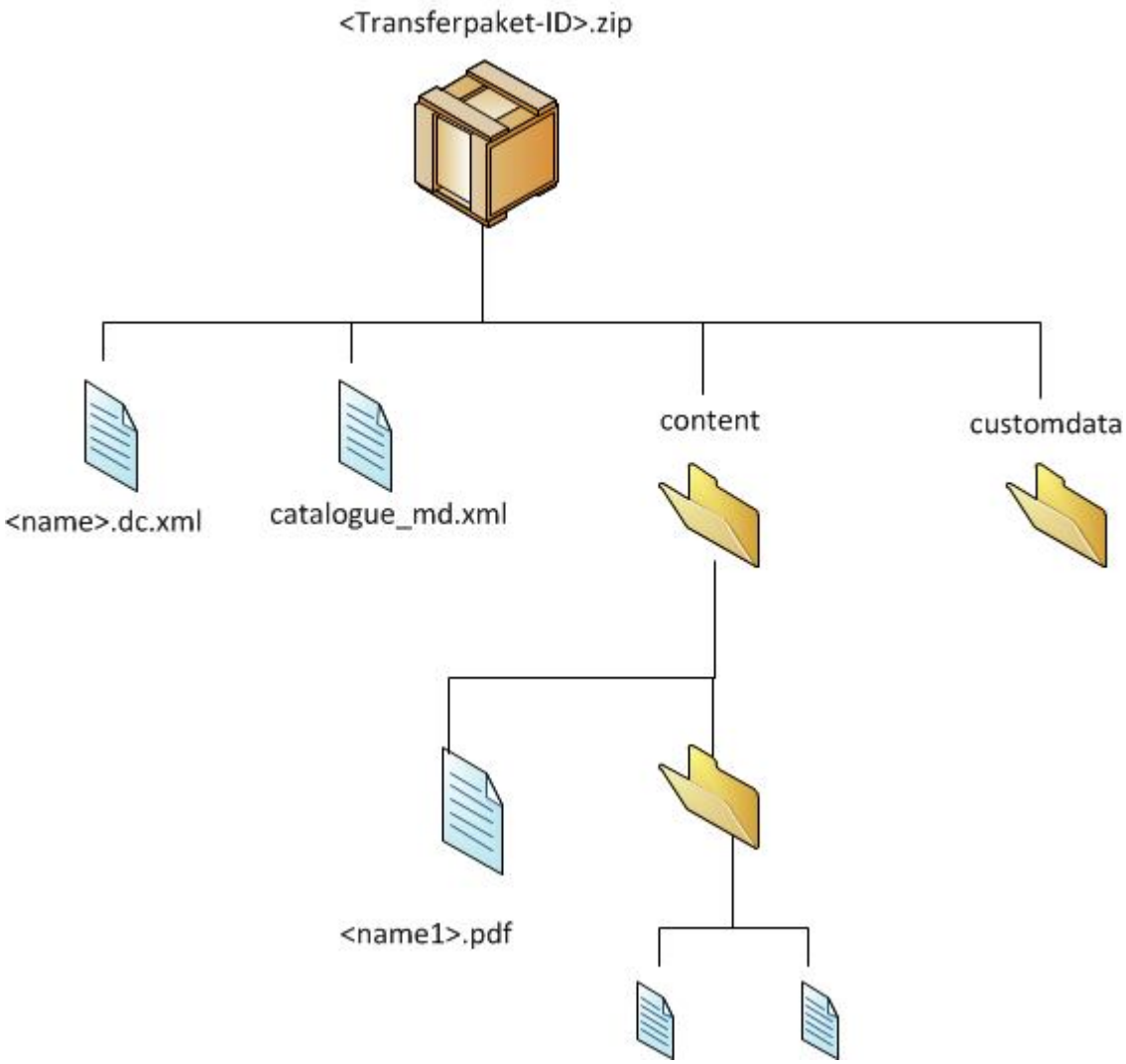

Abbildung 4: Transferpaketstruktur für die kombinierte Abgabe.

Bei der kombinierten Abgabe werden alle digitalen Objekte, die sich innerhalb des „content“-Verzeichnisses befinden als Netzpublikation und somit als zur Pflichtabgabe gehörig interpretiert. Sowohl die innerhalb der catalogue_md.xml befindlichen deskriptiven Metadaten als auch die Netzpublikation werden in die Kataloge und Repositorien der DNB im Rahmen der Sammelpflicht aufgenommen und archiviert.

Sofern es von der Partnerinstitution gewünscht ist, können weitere Informationen, die nicht zur Netzpublikation gehören, mitgeliefert werden. Hierfür kann im Transferpaket auf der obersten Ebene ein weiteres Verzeichnis „customdata“ angelegt werden.
Als Beispiel wäre hier der institutseigene Metadatensatz zur Netzpublikation zu nennen. Alle digitalen Objekte, die in diesem Verzeichnis liegen, werden von der Netzpublikation separiert und nicht in die Kataloge oder Repositorien der DNB aufgenommen. Diese Zusatzinformationen werden von der DNB für die Partner in einem gesonderten Speicherbereich langzeitarchiviert. Bei entsprechenden Vereinbarungen können bei Bedarf sowohl die Netzpublikation als auch die Zusatzinformationen an die Partnerinformation gemeinsam zurückgeliefert werden.

2.6 Hinweis für die Übertragung von Transferpaketen

Während des Upload-Prozesses muss die Dateiendung des Transferpakets in „.tmp“ abgeändert oder um diese erweitert werden. Ist die Datei vollständig hochgeladen, muss die Endung (.tmp) von dem Ablieferer wieder entfernt bzw. geändert werden. Zudem ist es ratsam Checksummendateien zeitlich vor den dazugehörigen Transferpaketen zu übertragen.

2.7 Transferpaketgenerator

Im Rahmen der AREDO-Kooperation wird von der DNB ein Transferpaketgenerator zur Verfügung gestellt, welcher lokal bei jedem Partner installiert werden kann. Dieser Transferpaketgenerator soll jedem Partner die Erstellung von validen Transferpaketen erleichtern.

Ansprechpartner:

Karlheinz Schmitt (Abteilung Informationstechnik)
k.schmitt@dnb.de

AREDO-Kontakt:
kontakt@aredo.dnb.de