

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

Version 1.0

Stand: 18.10.2012

Redaktion: Stefan Hein, Matthias Neubauer, Karlheinz Schmitt

Deutsche Nationalbibliothek (Leipzig, Frankfurt am Main)
2012

<urn:nbn:de:101-2012101804>

Inhalt

1	Einleitung.....	4
2	Ablieferung von Netzpublikationen	4
2.1	Übertragungsverfahren Hotfolder.....	5
2.2	Übertragungsprotokolle	5
2.3	Transferpaket-Spezifikation	5
2.4	Optionale Erweiterungen	8
2.5	Hinweis für die Übertragung von Transferpaketen	9

1 Einleitung¹

Derzeit bietet die Deutsche Nationalbibliothek (DNB) für die Ablieferung von Netzpublikationen drei verschiedene Möglichkeiten an:

- Webformular,
- OAI-Harvesting und
- Hotfolder-Verfahren.

Neben der Ablieferung über ein Webformular, welches vorwiegend für Einzelablieferungen Verwendung findet, bieten die zwei anderen Verfahren sowohl für die Ablieferer² als auch für die DNB die Möglichkeit, einen automatisierten Geschäftsgang für eine beliebige Anzahl an Netzpublikationen zu etablieren.

Das vorliegende Dokument beschreibt die Ablieferung von Netzpublikationen 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. In diesem Zusammenhang werden dem Ablieferer 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 Netzpublikationen

Für die ordnungsgemäße Aufnahme von Netzpublikationen in den Katalog und die Archive der DNB werden für jede digitale Publikation³ vom Ablieferer zwei Informationen benötigt:

1. Netzpublikation:

Bei den Netzpublikationen muss es sich um Publikationen handeln, die in der DNB als eigenständige Werke betrachtet und verzeichnet werden. Evtl. vorhandenes „supplemental material“ muss im Vorfeld abgesprochen werden (siehe Transferpaket-Spezifikation).

2. Metadatensatz:

Ein Metadatensatz mit beschreibenden Informationen bzgl. der Netzpublikation. Die Metadaten werden direkt in den Katalog der DNB importiert und dienen als bibliografischer Nachweis. Momentan angebotene Metadatenformate werden im folgenden Abschnitt benannt.

¹ Die vorliegende Spezifikation ersetzt das Dokument „Spezifikation von Transferpaketen und deren Übertragung an die Deutsche Nationalbibliothek mittels eines Hotfolders“ unter <http://nbn-resolving.de/urn:nbn:de:101-2011031000>

² Dabei kann es sich u.a. um einen Verlag, eine Institution oder eine Organisation handeln.

³ Wenn im Folgenden von einer digitalen Publikation oder Netzpublikation gesprochen wird, ist damit die Gesamtheit aller zur Publikation gehörenden Dateien zu verstehen. Ist von digitalen Objekten die Rede, dann gehört auch die Metadaten-Datei dazu.

2.1 Übertragungsverfahren Hotfolder

Seit dem Jahr 2012 bietet die DNB jedem Ablieferer die Möglichkeit, Netzpublikationen in ein spezielles, dem jeweiligen Ablieferer zugewiesenes Verzeichnis innerhalb der DNB abzulegen und dadurch die Ablieferung von Netzpublikationen durchzuführen. Dieser Ordner wird von der DNB überwacht, so dass einerseits aktuelle sicherheits- und rechterelevante Richtlinien jederzeit eingehalten werden und andererseits die Aufnahme der Netzpublikationen automatisiert und zeitnah erfolgen kann.

Der produktiven Ablieferung von Netzpublikationen geht eine Testphase voraus, in der zwischen dem Ablieferer und der DNB zum einen der verwendete Metadatenstandard abgesprochen, zum anderen die Einhaltung der im weiteren beschriebenen Transferpaket-Spezifikation getestet wird. Ebenso werden in der Testphase die vom Ablieferer übergebenen Dateiformate identifiziert und vermerkt.

2.2 Übertragungsprotokolle

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

- **SSH File Transfer Protocol (SFTP)**
- **Web-based Distributed Authoring and Versioning (WebDav)**

Für jeden Ablieferer steht nach einer Registrierung ein separates Konto mit eigenen Zugangsdaten (Benutzername und Passwort) zur Verfügung. Liefert ein Dienstleister für mehrere Verlage ab, wird für jeden einzelnen Verlag ein Hotfolder mit entsprechenden Zugangsdaten bereit gestellt.

2.3 Transferpaket-Spezifikation

Ein Transferpaket setzt sich aus einer Netzpublikation und dem dazugehörigen Metadatensatz zusammen. Hierfür ist die Übertragung beider Informationen 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 Netzpublikation liegt. Des Weiteren muss eine Datei mit dem Namen „catalogue_md.xml“ vorhanden sein, die den Metadatensatz in einem standardisierten XML-Format enthält.

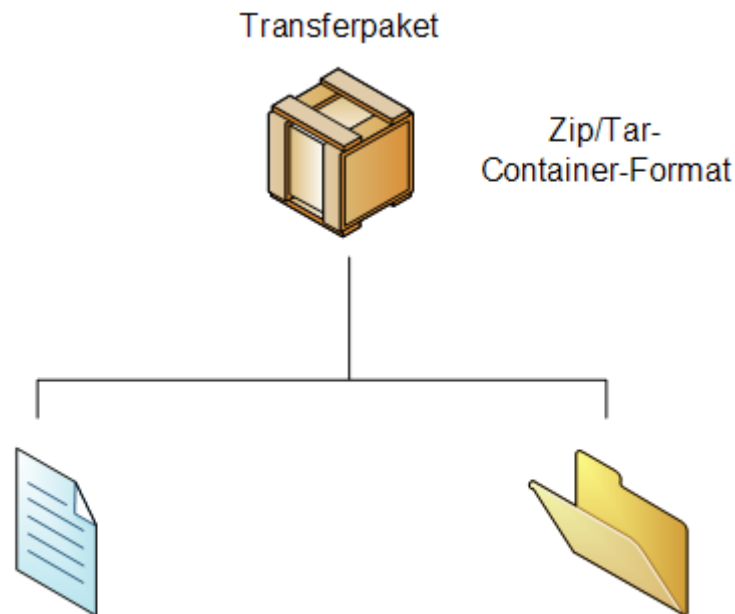


Abbildung 1: Aufbau eines Transferpaketes auf der obersten Ebene.

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

Netzpublikation: das Verzeichnis „content“

Alle zur Netzpublikation gehörenden Dateien sind in dem „content“-Verzeichnis abzulegen. Der Aufbau der Publikation innerhalb des „content“-Verzeichnisses kann vom Ablieferer frei definiert werden. Das „supplemental material“ unterliegt jedoch wie erwähnt individuellen Absprachen, die hier nicht näher beschrieben werden. Abbildung 2 zeigt beispielhaft mögliche Optionen:

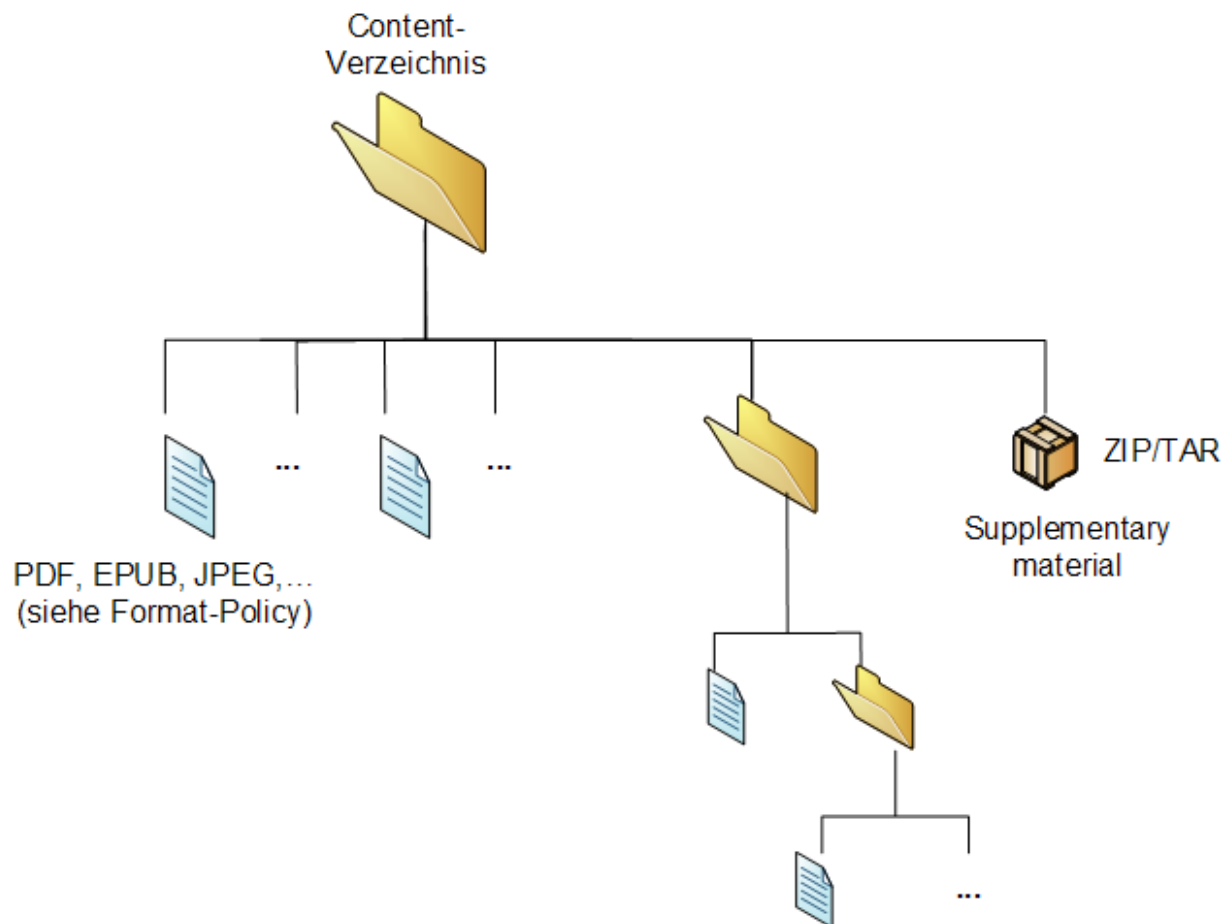


Abbildung 2: Diese Grafik verdeutlicht mögliche Strukturierungsvarianten des Inhalts eines content-Verzeichnisses.

Folgende Beschränkungen müssen jedoch eingehalten werden:

1. Dateinamen-Konventionen

Es gelten die üblichen Beschränkungen:

- keine Umlaute
- keine Sonderzeichen
- 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. 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 Netzpublikation erfolgt eine automatische Prüfung, ob nicht abgesprochene Dateiformate mitgeliefert wurden. Wurde ein nicht abgesprochenes Dateiformat erkannt, erfolgt eine Rücksprache mit dem Ablieferer, und das Dateiformat wird ggf. nachgetragen. Alle Dateien in den für einen Ablieferer eingetragenen Formaten erfahren während des

Ablieferungsprozesses eine Qualitätsprüfung. Von der DNB wird bzgl. der Qualität der Dateiformate ein Mindeststandard erwartet. Hierzu zählt insbesondere die Einschränkung der Benutzung von Netzpublikationen (z. B. Kennwortschutz, DRM).

Nähere Informationen zur Qualitätsprüfung finden sich in den folgenden Unterlagen:

<http://nbn-resolving.de/urn:nbn:de:101-2012102408>

4. Einschränkung bei Verwendung von Containerformaten im „content“-Verzeichnis

Die Verwendung von Containerformaten im „content“-Verzeichnis unterliegt den folgenden Beschränkungen:

- Die Netzpublikation besteht ausschließlich aus einem einzelnen ZIP/TAR-Container, der die zu archivierenden Dateien enthält.
- Es existieren beliebig viele zu archivierende, ungepackte Dateien und zusätzlich ein oder mehrere ZIP/TAR-Container, welche Zusätze zu archivierende Dateien (häufig als „supplementary material“ bezeichnet) enthalten.

Hinweis:

Die Verwendung von mehr als einer Container-Datei auf der obersten Ebene im „content“-Verzeichnis ist damit ausgeschlossen.

2.4 Optionale Erweiterungen

Zur Wahrung der Authentizität der abgelieferten Netzpublikationen wird von der DNB die Verwendung einer gegenseitigen Checksummenprüfung angeboten. Hierbei wird für jedes Transferpaket und ggf. für jede darin enthaltene Datei eine Prüfsumme berechnet, über die sich ungewollte Veränderungen 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 Ablieferer müssen für die zu prüfenden digitalen Objekte (Metadaten und alle Teile der Netzpublikation) Checksummen aus den oben genannten Verfahren mitgeliefert werden. Abbildung 3 verdeutlicht exemplarisch die hierfür notwendige Spezifikation:

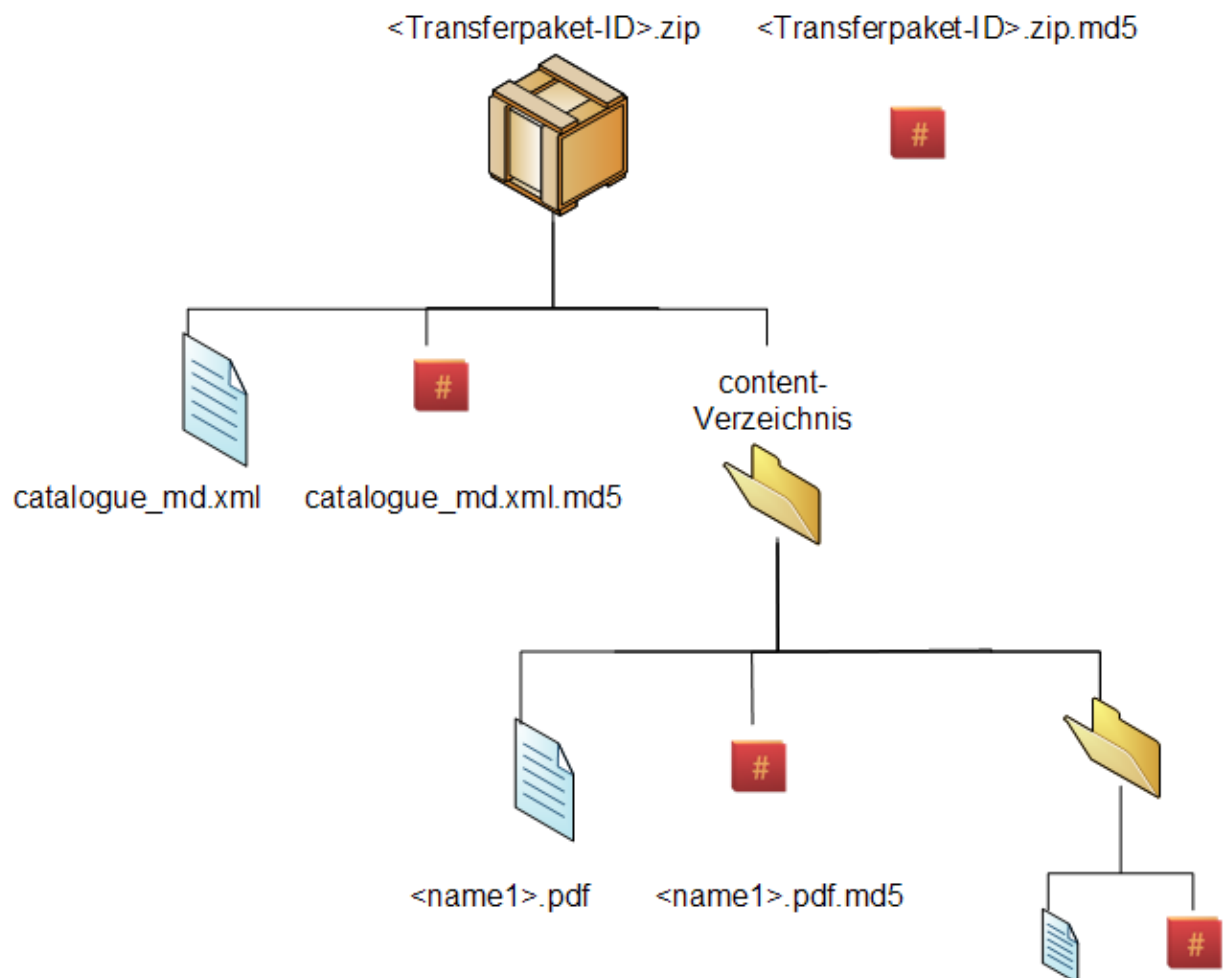


Abbildung 3: Transferpaket mit externer und interner Checksummenprüfung

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.

2.5 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.

Ansprechpartner:

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

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