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

Version 2.0
Stand: 24.08.2020

Redaktion: Redaktionsteam Netzpublikationen

Deutsche Nationalbibliothek (Leipzig, Frankfurt am Main)
2020

CC-BY-SA 3.0 DE

<urn:nbn:de:101-2020080427>

Inhaltsverzeichnis

Inhaltsverzeichnis 3

0 Einleitung 4

0.1. Weitere Informationen 4

1 Grundstruktur 5

2 Übertragung 5

2.1 Protokolle 5

2.2 Ablieferung über Dienstleister 5

2.3 Hinweise zur Nutzung 6

3 Das Transferpaket 6

3.1 Ordner „content“ 7

3.2 Datei „catalogue_md.xml“ 8

4 Übermittlung von Checksummen 9

5 Zusatzmaterial 10

6 Beispiele 10

Kontakt 10

0 Einleitung

Die Deutsche Nationalbibliothek (DNB) ist die zentrale Archivbibliothek Deutschlands. Der Sammelauftrag umfasst alle Publikationen in Schrift, Bild und Ton, die seit 1913 in Deutschland, in deutscher Sprache, als Übersetzung aus der deutschen Sprache oder über Deutschland veröffentlicht wurden. Das Gesetz über die Deutsche Nationalbibliothek (DNBG) und die Pflichtablieferungsverordnung (PflAV) sind die Grundlagen für den Aufbau unserer Sammlungen und umfassen auch die Sammelpflicht für unkörperliche Medienwerke, sogenannte Netzpublikationen.

Zum Sammelgebiet der unkörperlichen Medienwerke gehören alle Darstellungen in Schrift, Bild und Ton, die im Internet zugänglich sind. Dies sind sowohl Netzpublikationen mit Entsprechung zum Print-Bereich als auch web-spezifische Medienwerke.

Konkret sammeln wir im Internet veröffentlichte Bücher (E-Books), Zeitschriften (E-Journals), Tageszeitungen (E-Paper), Hochschulschriften, Noten, Hörbücher, Digitalisate sowie ausgewählte Websites.

Für die Pflichtablieferung von Netzpublikationen bieten wir verschiedene Möglichkeiten an:

- automatisierte Verfahren für die Übertragung größerer Mengen an Publikationen sowie
- Webformulare für die Übertragung einzelner Publikationen

Für die Ablieferung größerer Mengen von Netzpublikationen haben Sie die Wahl zwischen den beiden technischen Verfahren der Nutzung eines Hotfolders und der Nutzung von OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Beide Verfahren können Sie für die Übermittlung von Monografien, Hochschulprüfungsarbeiten, Hörbüchern, Noten und Periodika nutzen.

In diesem Dokument wird die Ablieferung über den Hotfolder beschrieben.

0.1. Weitere Informationen

Weitere Informationen zur Ablieferung von Netzpublikationen finden Sie unter:

https://www.dnb.de/netzpublikationen

Hier finden Sie auch weitere für die Ablieferung relevante Dokumente.

Für allgemeine Fragen zur Ablieferung von Netzpublikationen wenden Sie sich an den Service Netzpublikationen unter np-info@dnb.de.

Für technische Fragen zur Ablieferung von Netzpublikationen wenden Sie sich an den technischen Service unter np-technik@dnb.de.

Der produktiven Ablieferung von Netzpublikationen geht eine Testphase voraus, in der alle notwendigen Absprachen zwischen Abliefernden und der DNB getroffen werden. Diese Absprachen beziehen sich auf Struktur der Transferpakete, Inhalt und technische Eigenschaften (z.B. Beschränkungsfreiheit) der elektronischen Veröffentlichungen und der Verwendung der bibliografischen Metadaten.

1 Grundstruktur

Die Ablieferung von Netzpublikationen über die Schnittstelle Hotfolder setzt die Bildung eines sogenannten Transferpaketes voraus. Bestandteile jedes Transferpaketes sind immer die digitale Publikation selbst und die bibliografischen Metadaten zur Publikation in einem von der DNB akzeptierten Metadatenformat.

Die akzeptierten Metadatenformate werden sukzessive erweitert, eine jeweils aktuelle Aufstellung findet Sie unter:

http://www.dnb.de/metadatennp

Akzeptierte Dateiformate der digitalen Publikation sind PDF, EPUB, TIFF, JPEG, PS, MP3. Die Möglichkeit, weitere Dateiformate für die Ablieferung zur übernehmen, kann während der Testphase geprüft werden.

Die Hotfolderablieferung ist unabhängig vom Publikationstyp und der Erscheinungsform der digitalen Publikation und kann damit für monografisches und für periodisches Material genutzt werden. Es ist notwendig pro digitaler Publikation ein selbständiges Transferpaket bereitzustellen. Möchten Sie beispielsweise 15 E-Books oder 15 Zeitschriftenartikel eines Zeitschriftenheftes bei uns abliefern, erhalten wir jeweils 15 Transferpakete über diese Schnittstelle.

Der Hotfolder stellt für die Ablieferenden ein Verzeichnis auf einem DNB eigenen Server zur Verfügung, auf den die Transferpakete übertragen werden können.

2 Übertragung

Die Übernahme erfolgt nach Vereinbarung in regelmäßigen Abständen, z.B. täglich, wöchentlich oder monatlich. Es werden jeweils die Transferpakete abgerufen, die seit der letzten Abfrage bereitgestellt wurden. Die gelieferten Metadaten werden direkt in den Katalog der Deutschen Nationalbibliothek importiert und dienen als bibliografischer Nachweis. Die digitalen Publikationen werden auf dem Archivserver gespeichert und sind mit den bibliografischen Metadaten verlinkt.

Für alle Ablieferenden steht nach einer Registrierung ein dezidiertes Konto mit eigenen Zugangsdaten (Benutzername und Passwort) zur Verfügung.

2.1 Protokolle

Für die Übertragung der Transferpakete werden zwei verschiedene Übertragungsprotokolle angeboten:

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

2.2 Ablieferung über Dienstleister

Wenn Sie sich für die Ablieferung über Hotfolder entscheiden, legen wir für Sie ein sogenanntes Hotfolderkonto mit entsprechenden Zugangsdaten an. Innerhalb des Hotfolderkontos steht mindestens ein Ablieferungskonto zur Verfügung. Es ist möglich mehrere Ablieferungskonten für beispielsweise unterschiedliche Publikationstypen oder Imprints anzulegen.

Wird die Ablieferung von einem Dienstleistungsunternehmen übernommen, wird das entsprechende Ablieferungskonto diesem zugeordnet. Auch hier können mehrere Ablieferungskonten unterschiedlichen Dienstleistungsunternehmen zugeordnet werden.

2.3 Hinweise zur Nutzung

Für die Übertragung der Transferpakete wird üblicherweise ein SFTP Client wie FileZilla oder WinSCP eingesetzt.

Während des Upload-Prozesses muss die Dateiendung des Transferpaketes in „tmp“ abgeändert oder um diese erweitert werden. Ist die Datei vollständig hochgeladen, muss die entsprechende Endung vom Ablieferenden wieder zurückgesetzt werden. Versteckte Dateien, die beim Ablieferungsvorgang entstehen können, müssen wieder gelöscht werden. Zudem ist es ratsam Checksummendateien zeitlich vor den dazugehörigen Transferpaketen zu übertragen.

Bitte stellen Sie sicher, dass im verwendeten Client zum Transfer der Datei die Codierung UTF-8 eingestellt ist. Soweit möglich, bitten wir darum auf Sonderzeichen oder Umlaute im Namen des Transferpakets zu verzichten. Falls es in Ausnahmefällen erforderlich sein sollte, Sonderzeichen oder Umlaute im Namen des Transferpakets zu verwenden, so muss sichergestellt sein, dass der Name in UTF-8 codiert ist. Ist das nicht der Fall, kann das Transferpaket in der DNB nicht verarbeitet werden.

3 Das Transferpaket

Ein Transferpaket setzt sich aus der digitalen Publikation und den bibliografischen Angaben aus dem dazugehörigen Metadatensatz zusammen. Die Übertragung beider Bestandteile erfolgt in einem aktuellen Containerformat. Zurzeit werden von der DNB die folgenden Containerformate akzeptiert:

- ZIP
- TAR

Der Dateiname der Transferpakete ist unter Beachtung der in 3.1 aufgeführten Regeln frei wählbar, z.B. kann die ISBN der digitalen Publikation zum Namen des Transferpaketes werden.

Beispiele Dateinamen der Transferpakete:

```
9783958750685.zip
9783958750685.tar
abfallr_2016_01-jats-001.zip
SAD497392.tar
```

Ein Transferpaket enthält auf der oberen Ebene einen Ordner mit dem Namen „content“, in dem sich die digitale Publikation befindet. Der Ordnername ist *nicht* frei wählbar.

Die obligatorischen Metadaten müssen in einer Datei mit dem Namen „catalogue_md.xml“ in XML vorliegen. Der Dateiname ist *nicht* frei wählbar.

Beispiel Struktur Transferpaket - Überblick

```
9783958750685.zip
    catalogue_md.xml
    content
```

Beispiel Struktur Transferpaket – vollständig mit digitalem Objekt

```
9783958750685.zip
    catalogue_md.xml
    content
       9783958750685.pdf
```

3.1 Ordner „content“

Alle zur digitalen Publikation gehörenden Dateien werden im content-Ordner abgelegt.

Folgende Regeln sind dabei zu beachten:

- Für die Vergabe von Dateinamen gelten folgende Beschränkungen:
  - keine Umlaute
  - keine Sonderzeichen
  - nicht mehr als 128 Zeichen
  - UTF-8-codiert
- Aus technischen Gründen dürfen die im content-Ordner enthaltenen Dateien eine Anzahl von 4999 nicht überschreiten.
- Erlaubte Dateiformate sind PDF, EPUB, TIFF, JPEG, PS, MP3.
- Im content-Ordner darf ein ZIP- oder TAR-Container enthalten sein.
- Die Verwendung von mehr als einer Container-Datei auf der obersten Ebene im content-Ordner ist ausgeschlossen. Es ist möglich weitere Dateien in einem Unterordner zu übermitteln. Die Benennung des Unterordners ist frei wählbar.

Beispiel Inhalt content-Ordner: digitale Publikation in einer Datei

```
content
   978-3-644-50121-8.epub
```

Beispiel Inhalt content-Ordner: Beispiel Buchblock und Cover in getrennten Dateien

```
content
   2711.pdf
   2711.jpeg
```

Beispiel Inhalt Auszug content-Ordner: Hörbuch mit einzelnen Trackdateien (insgesamt 64 Dateien)

```
content
   9783837143294-Track-001.mp3
   9783837143294-Track-002.mp3
   9783837143294-Track-003.mp3
   9783837143294-Track-063.mp3
   9783837143294-Track-064.mp3
```

Beispiel Inhalt content-Ordner: digitale Publikation mit Zusatzmaterial (im Containerformat)

```
content
   2711.pdf
   2711.zip
```

Bei der Ablieferung der Netzpublikation erfolgt eine automatische Prüfung der enthaltenen Dateiformate. Die DNB erwartet bezüglich der Qualität der Dateiformate einen Mindeststandard. Containerformate werden als solche geprüft, die darin enthaltenen Daten werden von unseren Prozessen nicht bearbeitet.

Die Absprachen zur Qualität der Dateiformate erfolgen im Detail in der Testphase (vgl. 0.1 Weitere Informationen). Hierzu zählt insbesondere die Möglichkeit der uneingeschränkten Nutzung von Netzpublikationen, es darf keine digitale Rechteverwaltung die Nutzung der Objekte behindern. Die Einhaltung wird bei jedem Importvorgang überprüft. Nähere Informationen dazu finden sich unter:

urn:nbn:de:101-2012102408 oder unter https://www.dnb.de/hotfoldernp

3.2 Datei „catalogue_md.xml“

Für die Ablieferung von Netzpublikationen ist es notwendig bibliografische Metadaten mitzuliefern. Die Lieferung kann in verschiedenen Metadatenstandards erfolgen, die sich an ihrer Verbreitung in den jeweiligen Communities orientieren. Die jeweils aktuell unterstützten Standards und ihre Beschreibung für die Lieferung an die DNB, abhängig vom Publikationstyp, finden sich unter:

http://www.dnb.de/metadatennp

Format und Umfang der Metadaten werden während der Testphase mit den Abliefernden vereinbart. Die DNB orientiert sich dabei an einem Kernset, in dem die Pflichtfelder formatunabhängig festgelegt sind.

4 Übermittlung von Checksummen

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. Es ist möglich das Checksummenverfahren nur für das Transferpaket an sich oder für alle enthaltenen Dateien im Transferpaket anzuwenden.

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 den Ablieferenden müssen für die zu prüfenden digitalen Objekte Checksummen aus den oben genannten Verfahren mitgeliefert werden.

Geprüft wird die auf Seiten der DNB generierte Checksumme eines Transferpaketes gegen die mitgelieferte Checksumme des abgelieferten Transferpaketes. Dabei wird das gewählte Checksummen-Verfahren angewendet, das der Abliefernde durch die entsprechende Dateiendung der Checksummendatei vorgibt. Sind beide Checksummen identisch, erfolgt die Weiterverarbeitung für die Übernahme in das Archivsystem der DNB. Dieselbe Vorgehensweise gilt für Checksummen für alle Teile des Transferpaketes, wenn auf dieser Ebene Checksummen geliefert werden. Inhalt der Checksummendatei darf nur die Checksumme selbst sein, weitere Angaben wie die Wiederholung des Dateinamens sind nicht erlaubt.

Beispiel Ablieferung von Checksummen für das md5-Verfahren.

```
9783958750685.zip
9783958750685.zip.md5
    catalogue_md.xml
    catalogue_md.xml.md5
    content
       2711.pdf
       2711.pdf.md5
       2711.jpeg
       2711.jpeg.md5
```

5 Zusatzmaterial

Zusatzmaterial zu digitalen Publikationen können Materialien sein, die zum Werk hinzukommen, z.B. Lösungsblätter oder Fragebögen, aber auch sogenannte Forschungsdaten.

Zusatzmaterial ist ablieferungspflichtig. Forschungsdaten sind dann ablieferungspflichtig, wenn sie im Zusammenhang mit einer Dissertation stehen und unterliegen einer besonderen Ablieferungspolicy.

Wenn Sie die Ablieferung von Zusatzmaterial oder Forschungsdaten planen, wenden Sie sich zur Absprache der Details bitte an np-technik@dnb.de.

6 Beispiele

Ein Testpaket zum Download finden Sie unter https://www.dnb.de/hotfoldertestdaten.

Kontakt

Für alle technische Fragen zur Ablieferung von Netzpublikationen wenden Sie sich an den technischen Service unter np-technik@dnb.de.