Spezifikation der Dateiformat-Policy für die Sammlung von Netzpublikationen der Deutschen Nationalbibliothek

Erläuterungen zur Handhabung

Version 1.0
Stand: 24.10.2012

Redaktion: Stefan Hein, Matthias Neubauer, Karlheinz Schmitt

Deutsche Nationalbibliothek (Leipzig, Frankfurt am Main)
2012

<urn:nbn:de:101-2012102408>

Inhalt

1 Einleitung

Mit Inkrafttreten des Gesetzes über die Deutsche Nationalbibliothek (DNBG) vom 22. Juni 2006 wurde der Auftrag der Deutschen Nationalbibliothek (DNB) auf die Sammlung, Erschließung, Verzeichnung und Archivierung von unkörperlichen Medienwerken (Netzpublikationen) ausgeweitet. Jeder gewerbliche oder nicht gewerbliche Verleger in der Bundesrepublik Deutschland ist verpflichtet, Medienwerke in unkörperlicher Form in einfacher Ausfertigung kostenlos an die DNB abzuliefern.

Für die Ablieferung der Netzpublikationen werden von der DNB prinzipiell alle aktuellen und gültigen Dateiformate akzeptiert. Zur Sicherung der Authentizität und langfristigen Nutzbarkeit der Ablieferungen wird von der DNB während des Importes eine Qualitätsprüfung auf Grundlage der Dateiformate vorgenommen. Die Ergebnisse dieser Qualitätsprüfung fließen in einen Prozess innerhalb der DNB ein, bei dem über die Annahme bzw. die Ablehnung der Netzpublikationen entschieden wird.

Dieses Dokument stellt die technischen Entscheidungskriterien vor, die für die Annahme oder Ablehnung der Netzpublikationen heran gezogen werden. Hierfür werden im Folgenden sowohl die *Ingest-Level* vorgestellt, die für die Datei/en einer Netzpublikation[1] von der DNB vergeben werden, als auch die Regeln, die über eine Annahme und Ablehnung der Netzpublikationen entscheiden. Diese Regeln werden in Form einer Format-Policy dargestellt.

2 Ingest-Level

Im Folgenden werden technisch basierte Prüfkriterien für Dateiformate vorgestellt, die während des Importprozesses für alle Netzpublikationen gleichermaßen getestet werden. Diese Qualitätsprüfung zielt sowohl auf die Wahrung der Authentizität der entgegen genommenen Netzpublikationen als auch darauf, dass keinerlei technische Restriktionen vorhanden sind, welche die Aufgabe der DNB der langfristigen Bewahrung und Nutzung der Netzpublikationen erschweren oder gar verhindern.

2.1 Definition: Ingest-Level

Ein Ingest-Level ist das Ergebnis eines mehrstufigen aufeinander aufbauenden Prüfverfahrens, welches innerhalb der DNB realisiert ist. Repräsentiert wird eine qualitative Aussage über bestimmte markante technische Gegebenheiten der untersuchten Netzpublikation auf Grundlage des Dateiformates. Ein Ingest-Level wird jeder Datei der Netzpublikation zugewiesen.

[1] Eine Netzpublikation kann aus einer oder mehreren Dateien bestehen, z. B.: Buchblock (PDF) und Cover (JPG).

2.2 Kriterien

Zurzeit besteht das Ingest-Level-System aus fünf aufeinander aufbauenden Prüfkriterien:

1. **Dateiintegrität (DI)**
   Die vom Ablieferer übermittelten Dateien haben sich im Zuge der Datenübertragung nicht verändert.

2. **Identifikation (ID**)
   Die zu einer Netzpublikation gehörenden Dateien wurden hinsichtlich ihres Dateiformates eindeutig identifiziert.

3. **Beschränkungsfreiheit (BF)**
   Alle zur Netzpublikation gehörenden Dateien sind beschränkungsfrei, d.h. es existieren keine- von der DNB – erkennbare technische Beschränkungen, welche die Nutzung oder die Langzeitarchivierung der Netzpublikation beeinträchtigen oder unmöglich machen.

4. **Extraktion formatspezifischer technischer Metadaten (MD)**
   Formatspezifische technische Metadaten, welche für die digitale Langzeitarchivierung zwingend sind, konnten generiert werden.

5. **Format-Validität (V)**
   Das Dateiformat aller Dateien der Netzpublikation ist bzgl. seiner Formatspezifikation valide.

2.2.1 Beispiele und Erklärungen

Die oben genannten Kriterien bauen aufeinander auf. Je höher der Ingest-Level, desto mehr Kriterien wurden positiv geprüft und desto geringer ist das Risiko, dass die abgelieferte Netzpublikation langfristig nicht erhalten werden kann.

Im Folgenden wird die Vergabe der Ingest-Level erläutert:

Ingest-Level 0
Die Netzpublikation erfüllt das Kriterium: Dateiintegrität
Beispiel: Die Netzpublikation erhält Level 0, wenn die Integrität der zur Netzpublikation gehörenden Dateien nach der Übertragung an die DNB erfolgreich durch gegenseitig abgesprochene Prozesse überprüft, bestätigt und protokolliert werden konnte. Hierfür werden spezielle Verfahren (Checksummenprüfung) eingesetzt. Die gegenseitige Überprüfung der Dateiintegrität ist ein Angebot der DNB. Dieses Angebot muss nicht zwingend genutzt werden. Für den Fall, dass keine gegenseitige Überprüfung statt finden soll, wird von der DNB direkt nach Annahme der Netzpublikation eine Checksumme einseitig generiert, damit die Wahrung der Authentizität von diesem Moment an gesichert wird.

Ingest-Level 1
Die Netzpublikation erfüllt die Kriterien: Dateiintegrität und Identifikation
Beispiel: Die Netzpublikation erhält Level 1, wenn das Dateiformat erfolgreich identifiziert werden konnte. Dies kann bspw. über den MIME-Type oder den Pronom-Identifier (PUID) erfolgen. Eine einfache Überprüfung der Dateiendung wird von der DNB als nicht ausreichend eingeschätzt.

Ingest-Level 2
Die Netzpublikation erfüllt die Kriterien: Dateiintegrität, Identifikation und Beschränkungsfreiheit
Beispiel: Für die Netzpublikation konnten bereits Maßnahmen ergriffen werden, die Integrität für die Zukunft zu sichern (Level 0). Ebenso konnte bereits das Dateiformat erkannt werden (Level 1). In der weiteren Analyse der Netzpublikation konnten in diesem Schritt keinerlei beschränkende Mechanismen festgestellt werden, welche die Nutzung und Funktionalität einschränken oder verhindern. Für ein PDF-Dokument wären dies z.B. Passwort, Kopier-, oder Druckbeschränkungen, welche die Vergabe dieses Ingest-Levels verhindern würden.

Ingest-Level 3
Die Netzpublikation erfüllt die Kriterien: Dateiintegrität, Identifikation und Beschränkungsfreiheit. Zusätzlich konnten formatspezifische technische Metadaten für die Durchführung von Maßnahmen zur Langzeitarchivierung extrahiert werden.
Beispiel: Ein PDF-Dokument wurde mithilfe des abgesprochenen Checksummenverfahrens auf Dateiintegrität geprüft und ist von einem Analyse-Tool als beschränkungsfrei erkannt worden. Zudem konnten durch einen Metadatengenerator formatcharakteristische technische Metadaten erzeugt werden, so dass die Grundlagen für digitale Bestandserhaltungsmaßnahmen geschaffen werden konnte.

Ingest-Level 4
Die Netzpublikation erfüllt die Kriterien: Dateiintegrität, Identifikation, Beschränkungsfreiheit, Extraktion formatspezifischer technischer Metadaten und Format-Validität.
Beispiel: Ein PDF-Dokument wurde mithilfe des abgesprochenen Checksummenverfahrens auf Dateiintegrität geprüft und ist von einem Analyse-Tool als beschränkungsfrei erkannt worden. Zudem konnten durch einen Metadatengenerator formatcharakteristische technische Metadaten erzeugt werden. Eine Analyse der technischen Metadaten und der Netzpublikation ergab, dass bei der Erzeugung der Netzpublikation in dem abgelieferten Format die grundlegenden Spezifikationen für das Dateiformat eingehalten worden sind. In diesem Falle wird die Netzpublikation bzgl. ihres Dateiformates als valide eingestuft.

Die folgende Tabelle gibt eine Übersicht über das Zusammenspiel der einzelnen Kriterien und der daraus resultierenden Ingest-Level.

	DI	ID	BF	MD	V
Level 0	X	O	O	O	O
Level 1	X	X	O	O	O
Level 2	X	X	X	O	O
Level 3	X	X	X	X	O
Level 4	X	X	X	X	X

3 Format-Policy

Für die aktuell an die DNB abgelieferten Dateiformate wurde eine Liste bzgl. der minimal geforderten und maximal möglichen erkennbaren Ingest-Level für ein Dateiformat angefertigt. Diese Liste wurde auf Grundlage der von der DNB aktuell möglichen technischen Analysemöglichkeiten erstellt. Hieraus sind Regeln für die Annahme aber auch Ablehnung von Netzpublikationen aufgestellt.

3.1 Regel für die Ablehnung von Netzpublikationen

Generell wird die Annahme einer Netzpublikation aus technischen Gründen abgelehnt, sobald für eine Datei innerhalb der Netzpublikation ein Ingest-Level unterhalb des Ingest-Level 2 festgestellt worden ist.

Dies bedeutet, dass während des Importprozesses entweder die Identität des Dateiformates der Netzpublikation nicht eindeutig festgestellt werden konnte oder Beschränkungsmechanismen, die die langfristige Nutzung der Netzpublikation verhindern, erkannt wurden. Sofern im Vorfeld ein gegenseitiger Checksummenvergleich für die übermittelte Netzpublikation vereinbart worden ist, erfolgt ebenfalls eine Ablehnung der gesamten Netzpublikation, wenn dieser Vergleich negativ ausgefallen ist.

Die folgende Tabelle gibt die Grenzen der Annahme wieder.

	DI	ID	BF	MD	V
Level 0	X	O	O	O	O
Level 1	X	X	O	O	O
Level 2	X	X	X	O	O
Level 3	X	X	X	X	O
Level 4	X	X	X	X	X

Erreicht eine Netzpublikation lediglich Level 0 oder Level 1, erfolgt keine Aufnahme in das Archivsystem der DNB. Die Pflichtabgabe der Netzpublikation gilt somit als nicht erfüllt. Die DNB wird sich in einem solchen Fall mit der abliefernden Stelle in Verbindung setzen.

3.2 Regeln für die Annahme von Netzpublikationen

Eine Netzpublikation – bzw. alle darin enthaltenen Dateien - muss mindestens Ingest-Level 2 erreichen um in das Archivsystem der DNB aufgenommen zu werden. Werden einige Dateien der Netzpublikation lediglich mit Ingest-Level 2 oder 3 bewertet, stellt dieses keinen Hinderungsgrund für die Entgegennahme der Netzpublikation dar. Für die langfristige digitale Bestandserhaltung bleibt es die Aufgabe der DNB diese Dateien der Netzpublikation dauerhaft nutzbar zu halten und hierfür evtl. vorbereitende Arbeiten durchzuführen.

Eine Netzpublikation, die ausschließlich aus Dateien mit Ingest-Level 4 besteht, verfügt über gute Voraussetzungen langfristig nutzbar zu bleiben.

3.3 Muster-Format-Policy

Im Folgenden wird eine Muster-Format-Policy vorgestellt, die für aktuell verwendete Dateiformate von Netzpublikationen Anwendung findet. Die Liste erhebt keinen Anspruch auf Vollständigkeit, sondern stellt lediglich ein Beispiel für die gängigsten Dateiformate dar. Für jeden Ablieferer wird bei der Registrierung eine individuelle Format-Policy angelegt und abgesprochen.

Werden Dateiformate abgeliefert, die nicht in der Format-Policy verzeichnet sind, wird dies beim Import erkannt. Die Annahme dieser Dateiformate wird suspendiert, bis ein entsprechendes, gegenseitig abgesprochenes Ingest-Level in der Format-Policy nachgetragen wurde.

Die folgende Tabelle zeigt in der ersten Spalte die aktuellen Dateiformate. Die zweite Spalte gibt den minimal erforderlichen Ingest-Level wieder, der zur Aufnahme der Netzpublikation gefordert wird. Die dritte Spalte gibt den maximal möglichen Ingest-Level wieder, der auf Basis der aktuell implementierten Workflows und Analyse-Prozesse innerhalb der DNB vergeben werden kann. Da die Ingest-Level aufeinander aufbauen wird nur der höchstmögliche Ingest-Level vermerkt. In der letzten Spalte werden die mit der DNB vereinbarten Ingest-Level eingetragen, die erfüllt sein müssen um eine Aufnahme erfolgreich durchzuführen.

Dateiformat*	Min. Ingest-Level	Max. Ingest-Level	Vereinbarung
PDF	2	4	
TIFF	2	4	
JPEG	2	4	
PS	2	4	
EPUB	2	3	

*Bei einzelnen Versionen kann es zu Abweichungen kommen.

Beispiel:

Voraussetzung:
Für die Ablieferung von PDFs wird Ingest-Level 3 vereinbart. Die abgelieferte Netzpublikation enthält ausschließlich PDF-Dateien.

Annahme einer Netzpublikation:
Bei der Ablieferung dieser Netzpublikation werden alle PDF-Dateien mit Ingest-Level 4 bewertet. Da in diesem Fall alle PDF-Dateien die Mindestvorgabe von Ingest-Level 3 sogar überschreiten, wird die Netzpublikation akzeptiert und in das Archivsystem der DNB importiert.

Ablehnung einer Netzpublikation:
Bei der Ablieferung dieser Netzpublikation wird bei einer der PDF-Dateien ein Kopierschutz festgestellt, wodurch für diese Datei das Kriterium „Beschränkungsfrei“ nicht erfüllt ist. Somit wird für diese Datei lediglich Ingest-Level 1 vergeben. Alle anderen PDF-Dateien werden mit Ingest-Level 4 bewertet.
Da die Netzpublikation eine Datei enthält, deren Ingest-Level 1 unter der Format-Policy-Vorgabe von Ingest-Level 3 liegt, wird die Netzpublikation abgelehnt und ein entsprechender „Fehlerworkflow“ gestartet, bei dem sich die DNB mit der abliefernden Stelle in Verbindung setzt.

Ansprechpartner:

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

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