Verarbeitung von Metadaten in DA-NRW

z1334002
[25. September 2014]

Inhalt

Allgemeine Definition

„Metadaten sind strukturierte Daten zur einheitlichen Beschreibung von Ressourcen jeglicher Art (z. B. Daten, Dokumente, Personen, Gemälde, Orte, Gebäude, Konzepte)“ .[1] Sie unterstützen die Recherche der beschriebenen Daten auf dem Dateisystem, in Datenbanken sowie im World Wide Web. Aufgrund der Heterogenität der Primärdaten an sich sowie der Vielfältigkeit der verschiedenen Kontexte, in denen diese Daten verwendet werden, existieren unterschiedliche Arten, Formate und Standards der Metadaten.

Unabhängig vom jeweiligen Metadatenformat werden inhaltlich gesehen folgende Arten von Metadaten unterschieden:

- **Deskriptive Metadaten** dienen der Beschreibung von allen für die Recherche relevanten Informationen wie beispielsweise Titel, Autor und Format einer Ressource.
- **Strukturelle Metadaten** bilden die Dokumentenstruktur ab und zeigen Beziehungen zwischen Ressourcen auf.
- **Administrative Metadaten** liefern Informationen über die Herkunft sowie die verschiedenen Stationen der Verarbeitung der Ressource wie etwa Archivierung, Konvertierung etc.
- **Technische Metadaten** enthalten Informationen über die technischen Parameter der Ressource wie etwa Dateityp, Dateigröße und Auflösung.

Akzeptierte Metadatenformate

Im Rahmen des DA-NRW werden derzeit vier Metadatenformate akzeptiert:

- EAD
- METS
- LIDO
- XMP

Die genannten Formate enthalten jeweils alle vier Arten von Metadaten.

Lebensweg der Metadaten im DNSCore

Abhängig von der Verarbeitung der Primärdaten ändern sich mit diesen auch die entsprechenden Metadaten. Dabei werden in DNS der strukturelle sowie der technische Teil der jeweiligen Metadaten aktualisiert. Der von den Archivaren, Bibliothekaren etc. erfasste deskriptive Teil der

[1] http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Handbuch/metadaten.pdf?__blob=publicationFile

Metadaten bleibt unverändert. Die Administrativen Metadaten werden für das gesamte Paket in der PREMIS.xml zusammengetragen.

Langzeitarchivierung

Für die Metadaten gilt im DNSCore dasselbe Prinzip wie für die Primärobjekte: Jede einzelne Datei wird zunächst auf Byte-Ebene gesichert. Darüber hinaus wird der strukturelle sowie der technische Teil der Metadaten ggf. an die Veränderungen der beschriebenen Primärdatei bzw. Primärdateien angepasst. Mit anderen Worten: Bei Migration der Primärdatei in ein langzeitsicheres Format wird in der entsprechenden Metadatendatei der Referenzpfand sowie die Angabe des Formats aktualisiert, sodass die Metadatendatei stets eine gültige Beschreibung der Primärdatei bleibt.

Präsentation

Für die Präsentation im Portal werden die Primärdaten aus dem langzeitsicheren Dateiformat in das dafür jeweils festgelegte Präsentationsformat konvertiert. Sowohl die Primärdateien als auch die entsprechenden Metadaten erhalten eine DA-NRW interne URL. Daher ist eine erneute Anpassung der Metadaten unerlässlich. Im strukturellen Teil der Metadatendateien wird also der relative Pfad auf dem Dateisystem durch die generierte URL ersetzt. Dabei enthält die URL selbstverständlich die aktualisierte – dem Zielformat für die Repräsentation entsprechende – Dateiendung.

Anforderungen an die Metadaten im DNSCore

Im Kontext des DA-NRW sowie der Langzeitarchivierung im Allgemeinen gibt es eine zentrale Regel, die stets eingehalten werden muss:

SIP-Pakete müssen in sich konsistent sein.

Für die Struktur der Metadaten hat diese Regel wenige einfache Konsequenzen:

1. Metadaten dürfen ausschließlich die im SIP mitgelieferten Primärdaten referenzieren.

   Diese Forderung ist in keinster Weise DNSCore-spezifisch, sondern ist Bestandteil der oben angeführten allgemeinen Definition des Begriffs *Metadaten*. Mit anderen Worten bedeutet der Terminus *Metadaten*, auch *Daten über Daten* genannt, nichts anderes, als dass die Metadaten die referenzierten Primärdaten lediglich begleiten und beschreiben. Liefert man nun nur die Beschreibung ohne das beschriebene Digitalisat, enthält das Paket Informationen, die nicht zugeordnet werden können. Damit ist ein solches Paket nicht konsistent und wird aus diesem Grund von DNSCore abgelehnt.

2. Alle in der Metadatendatei enthaltenen Referenzen auf die Primärdaten müssen relativ ab der Metadatendatei angegeben werden.

   Die Forderung der Konsistenz der Metadaten beinhaltet die genaue und vor allem eindeutige Referenzierung der Primärdaten. Dies kann leicht erreicht werden, indem der Speicherort der Primärdaten im SIP stets **relativ** von der Metadatendatei angegeben wird.

3. Die Metadatendatei muss auf der obersten Dateiebene des SIP liegen.

   Auf diese Weise kann sichergestellt werden, dass die Metadatendatei von DNS als solche erkannt wird. Darüber hinaus wird so die Angabe relativer Pfade in der Metadatendatei sehr übersichtlich.

Nur die SIP-Pakete, die alle oben genannten Forderungen erfüllen, können im DNS ordnungsgemäß verarbeitet werden.

Verbreitete Abweichungen:

Oft werden Metadaten eingeliefert, in denen die Primärdaten mittels URLs (http://...) referenziert werden. Unabhängig davon, ob die eigentlichen Primärdaten ganz ausgelassen oder doch mitgeliefert werden, sind die entsprechenden SIP-Pakete nicht konsistent und werden von DNS abgelehnt. Im ersten Fall enthält das SIP-Paket Beschreibungen von nicht vorhandenen Daten. Dies macht per Definition keinen Sinn. Im zweiten Fall werden zwar sowohl Primärdateien als auch Metadaten eingeliefert, aber nur die Primärdaten können verarbeitet werden. Die mitgelieferten Metadaten können aufgrund fehlender Referenzen auf „echte“, sich im Paket unter angegebenem Pfad befindlichen Primärdaten nicht aktualisiert werden.

Validierung der Metadaten in DNS

Jedes SIP, das in das DNS eingeliefert wird, durchläuft eine Validierung seiner Metadaten. Dabei wird in der Metadatendatei jede einzelne Referenz auf eine Primärdatei auf die tatsächliche Existenz der jeweils referenzierten Primärdatei geprüft. Sollte auch nur eine einzige Datei unter dem angegebenen Pfad nicht zu finden sein, wird das gesamte SIP als inkonsistent abgelehnt.

Detaillierte Beschreibung der akzeptierten Metadatenformate

METS / MODS

Allgemeine Informationen

Metadata Encoding & Transmission Standard (METS) ist ein XML- Dateiformat zur Beschreibung von digitalen Sammlungen von Primärobjekten. Nähere Informationen zum Dateiformat METS finden Sie unter http://www.loc.gov/standards/mets/.

Jedes METS-File hat folgende Struktur:

```
<mets>
  <metsHdr>          (Beschreibung des METS-Dokuments)
  <dmdSec>           (Deskriptive Metadaten)
  <amdSec>           (Administrative Metadaten)
  <fileSec>          (Auflistung aller referenzierten Primärobjekte)
  <structMap>        (Strukturelle Metadaten)
  <structLink>       (Verknüpfung von Elementen)
  <behaviorSec>      (Verbindung zu ausführbaren Elementen)
</mets>
```

Der METS-Standard legt nicht fest, welche Form die einzelnen Abschnitte aufweisen müssen. Es kann durchaus sein, dass die Abschnitte unterschiedliche XML-Formate haben.

Die in das DNSCore eingelieferten METS-Pakete werden auch als METS/MODS-Pakete bezeichnet, weil das Element `<dmdSec>` dem *Metadata Object Description Schema (MODS)* gehorcht. Nähere Informationen zum Dateiformat MODS finden Sie unter http://www.loc.gov/standards/mods/.

Verarbeitung in DNSCore

Bei der Verarbeitung eines METS/MODS-Pakets wird bei jeder Migration der Primärdaten die METS-Datei aktualisiert. Dabei wird lediglich der `<fileSec>`-Knoten angepasst. Dieser Knoten besitzt den Kindknoten `<fileGrp>`, der die Auflistung aller Referenzen auf Primärfiles enthält. Jedes dieser Files wird mit jeweils einem `<file>`-Knoten beschrieben. Der `<file>`-Knoten enthält eine Reihe verschiedener Informationen. Aktuell werden in DNSCore insgesamt drei Felder aktualisiert: MIMETYPE, LOCTYPE und href.

Der Mimetype gibt den Typ der referenzierten Datei an. Hier erfährt man, ob es sich um ein Bild-, Audio- oder Videoformat handelt und welches genau das ist.

Der Loctype gibt den Typ der Referenz an: der Attributwert „OTHER“ steht für eine Referenz auf dem Dateisystem, der Wert „URL“ verrät, dass es sich bei der Referenz um eine URL handelt.

Schließlich enthält das Attribut „href“ die Referenz auf die Primärdatei.

Im Folgenden wird anhand eines Beispiels gezeigt, welche Ersetzungen in METS-Files vorgenommen werden.

Beispiel:

Ursprüngliche Gestalt: Angenommen, die Beispiel-METS referenziert ein Bild im BMP-Format. Die SIP-Struktur sieht wie folgt aus:

```
data/mets.xml
data/Bilder/Bild.bmp
```

Der fileSec-Knoten der eingelieferten METS-Datei:

```
<mets:fileSec>
  <mets:fileGrp>
    <mets:file MIMETYPE="image/x-ms-bmp">
      <mets:FLocat LOCTYPE="OTHER" xlink:href="Bilder/Bild.bmp"/>
    </mets:file>
  </mets:fileGrp>
</mets:fileSec>
```

Langzeitarchivierung: Da BMP kein langzeitsicheres Dateiformat ist, wird die Primärdatei Bild.bmp nach TIFF migriert. Folglich muss die beschreibende METS-Datei angepasst werden. Das AIP sieht nun so aus:

```
data/mets.xml
data/Bilder/Bild.tif
```

Der aktualisierte fileSec-Knoten der METS-Datei:

```
<mets:fileSec>
  <mets:fileGrp>
    <mets:file MIMETYPE="image/tiff">
      <mets:FLocat LOCTYPE="OTHER" xlink:href="Bilder/Bild.tif"/>
    </mets:file>
  </mets:fileGrp>
</mets:fileSec>
```

Präsentation: Für die Präsentation wird aus dem TIFF eine JPG-Datei erzeugt. Da die beiden Dateien im WWW abrufbar sein sollen, muss die relative Referenz durch die Angabe einer absoluten URL ersetzt werden. Das PIP sieht wie folgt aus:

```
data/mets.xml
data/Bilder/Bild.jpg
```

Der aktualisierte fileSec-Knoten der METS-Datei:

```
<mets:fileSec>
  <mets:fileGrp>
   <mets:file MIMETYPE="image/jpeg">
     <mets:FLocat LOCTYPE="URL" xlink:href=" http://data.da-nrw.de/[...] /[new Filename].jpg"/>
   </mets:file>
  </mets:fileGrp>
</mets:fileSec>
```

EAD / METS

Allgemeine Informationen

Encoded Archival Description (EAD) ist ein XML-Dateiformat zur Beschreibung von Findbüchern. Nähere Informationen zum Dateiformat EAD finden Sie unter http://www.loc.gov/ead/.

Die *EAD* stellt die übergeordnete Metadatendatei dar, die weitere Metadatendateien beschreibt und referenziert. Diese liegen im oben beschriebenen Dateiformat *METS* vor, daher auch die Bezeichnung der entsprechenden Pakete als EAD/METS-Pakete. Die in der EAD referenzierten METS-Dateien beschreiben und referenzieren wiederum die eigentlichen Primärdaten.

Verarbeitung in DNSCore

Bei der Verarbeitung eines EAD/METS-Pakets werden bei jeder Migration der Primärdaten die in der EAD referenzierten METS-Dateien entsprechend der oben angeführten Beschreibung aktualisiert.

Außerdem muss für die Präsentation des Pakets auch die EAD-Datei aktualisiert werden. Sowohl im SIP als auch im AIP müssen die METS-Dateien in der übergeordneten EAD-Datei relativ referenziert werden. Für die Publikation müssen *alle* relativen Referenzen durch absolute URLs ersetzt werden. Dies gilt nicht nur für die in den METS-Dateien enthaltenen Referenzen auf die Primärdateien, sondern auch auf die in der EAD-Datei enthaltenen Referenzen auf die METS-Files.

Im Folgenden wird anhand eines Beispiels gezeigt, welche Ersetzungen in EAD-Files vorgenommen werden.

Beispiel:

Ursprüngliche Gestalt: Die SIP-Struktur sieht wie folgt aus:

```
data/ead.xml
data/mets/mets.xml
data/mets/bild.bmp
```

Die Referenz auf die METS-Dateien werden im EAD im Knoten <daogrp> angegeben. Dieser sieht in der EAD-Datei aus dem Beispiel-Paket sowohl im SIP als auch im AIP wie folgt aus:

```
<daogrp>
  <daoloc title="mets.xml" role="mets" href="mets/mets.xml">
</daogrp>
```

Für die Präsentation muss der Knoten aktualisiert werden:

```
<daogrp>
  <daoloc title="mets.xml" role="mets" href=" http://data.da-nrw.de/[...] /mets.xml">
</daogrp>
```

LIDO

Allgemeine Informationen

Lightweight Information Describing Objects (LIDO) ist ein XML-Dateiformat zur Beschreibung von digitalen Sammlungen von Primärobjekten.

Die Eignung von LIDO als Metadatenstandard für Museumsdaten ist nicht unumstritten, da LIDO im Gegensatz zu allen anderen unterstützten Metadatenstandards absolute URLs zu Inhaltsdaten mitführt. Die Verwendung von URLs zur Referenzierung von Primärdaten ist in der LIDO-Spezifikation (http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf) fest vorgegeben. Das für die Referenzierung vorgesehene Element <linkResource> wird in der Spezifikation als „A url reference in the worldwide web environment“ bezeichnet.

Aus den im vorliegenden Dokument bereits genannten Gründen kann eine solche Art der Referenzierung von Primärdaten nicht für die Langzeitarchivierung verwendet werden. Aus diesem Grund werden alle Einlieferer, die ihre Daten im DA-NRW langzeitarchivieren möchten, gebeten, alle Elemente <linkResource> jeweils mit einem relativen Pfad auf die mitgelieferte Primärdatei zu befüllen. Anderenfalls wird das gebildete SIP als inkonsistentes Paket von DNSCore abgelehnt.

Verarbeitung in DNSCore

Bei der Migration von Primärdaten für die Langzeitarchivierung und die Präsentation wird in der eingelieferten LIDO.xml pro Primärdatei ein <linkResource>-Element aktualisiert.

Im Folgenden wird anhand eines Beispiels die im LIDO vorgenommenen Ersetzungen aufgezeigt:

Beispiel:

Ursprüngliche Gestalt: Die SIP-Struktur sieht wie folgt aus:

```
data/lido.xml
data/Bilder/bild.bmp
```

Die Referenz auf die Primärdatei wird im Element < linkResource > angegeben:

```
<lido:linkResource>Bilder/bild.bmp</lido:linkResource>
```

Langzeitarchivierung: Da BMP kein langzeitsicheres Dateiformat ist, wird die Primärdatei Bild.bmp nach TIFF migriert. Folglich muss die beschreibende LIDO-Datei angepasst werden. Das AIP sieht nun so aus:

```
data/lido.xml
data/Bilder/bild.tif
```

```
<lido:linkResource>Bilder/bild.tif</lido:linkResource>
```

Präsentation: Für die Präsentation wird aus dem TIFF eine JPG-Datei erzeugt. Da die beiden Dateien im WWW abrufbar sein sollen, muss die relative Referenz durch die Angabe einer absoluten URL ersetzt werden. Das Element <linkResource> sieht in der LIDO-Datei im PIP sieht wie folgt aus:

```
<lido:linkResource> http://data.da-nrw.de/[...] /[new Filename].jpg </lido:linkResource>
```

XMP

Allgemeine Informationen

Verarbeitung in DNSCore