Indexer

METS-Dateien

Indiziert werden die Mets-Dateien, die Metadaten über die Struktur der Werke beinhalten.

Ohne Metadaten zur Struktur sind die Seitenabbildungen oder die Textdateien, aus denen es besteht, so gut wie wertlos. Und ohne technische Metadaten über den Digitalisierungsprozess können Leser nicht sicher sein, wie genau die digitale Version die ursprüngliche Vorlage wiedergibt. Für interne Verwaltungszwecke und die Datenadministration benötigt die Bibliothek außerdem technische Metadaten, um die Digitalisate regelmäßig auffrischen oder migrieren zu können und so die wertvollen Quellen langfristig zugänglich zu erhalten.

Ein Mets-Dokument besteht aus 7 Hauptabschnitten :

”

1. **Der Kopfteil (METS Header)** - Der Kopfteil enthält Metadaten, die das jeweilige METS-Dokument selbst beschreiben, einschließlich der Angaben zum Bearbeiter oder Herausgeber des METS-Dokuments

2. **Erschließungsangaben (Descriptive Metadata)** - Der Abschnitt für die Erschließungsangaben kann sowohl Verweise auf ein externes Dokument, (etwa einen MARC Datensatz in einem OPAC oder ein EAD-Findbuch auf einem WWW-Server), wie auch in das METS-Dokument eingebettete Angaben oder beides enthalten. Es können auch mehrere externe und interne Erschließungspakete in dem Erschließungsabschnitt verwendet werden.

3. **Verwaltungsangaben (Administrative Metadata)** - Der Abschnitt für die Verwaltungsangaben liefert Informationen über die Herstellung und Speicherung von Dateien, über Urheberrechte und über die digitalisierte Vorlage. Außerdem werden hier Angaben zur Herkunft der Digitalisate erfasst (z.B. über das Verhältnis von Master und Derivaten sowie über Migrationen.) Ähnlich wie die Erschließungsangaben können diese Metadaten extern oder in das METS-Dokument integriert vorliegen.

4. **Dateienabschnitt (File Section)** - Im Dateienabschnitt werden alle Dateien mit Inhalten, aus denen das digitale Objekt besteht, aufgelistet. Einzelne zusammengehörige Dateien können dabei mit dem Element zusammengefasst werden, um etwa verschiedene Versionen auseinander halten zu können.

5. **Strukturbeschreibung (Structural Map)** - Die Strukturbeschreibung ist zentraler Bestandteil eines jeden METS-Dokuments. Sie bildet den inneren Aufbau des digitalen Objektes ab und verknüpft die Elemente der Struktur mit den Dateien, aus denen der Inhalt des digitalen Objektes besteht, sowie mit deren Metadaten.

6. **Strukturverknüpfungen (Structural Links)** - Der Abschnitt mit den Strukturverknüpfungen erlaubt es den Erstellern von METS-Dokumenten das Vorhandensein von Hyperlinks zwischen einzelnen Knoten des im Strukturabschnitt

dargestellten hierarchischen Aufbaus des digitalen Objekt zu beschreiben. Diese Funktion ist besonders für die Archivierung von Webseiten gedacht.

7. **Verhalten (Behavior)** - Ein Abschnitt über das Verhalten des digitalen Objekts kann verwendet werden, um ausführbare Anweisungen für das Verhalten mit den Inhalten in METS-Objekten zu verknüpfen. Jede Verhaltensform hat ein Schnittstellendefinitionselement, das eine abstrakte Definition eines Satzes von Verhaltensformen in dem jeweiligen Abschnitt enthält. Außerdem besitzt jede Verhaltensform ein Mechanismuselement, das ein Modul ausführbaren Codes enthält, mit dem die in der Schnittstellendefinition abstrakt formulierten Verhaltensformen ausgeführt werden können.“ [1]

- Nach dem Mets-Datei mit den Zugehörigen Images in den Hotfolder kopiert wird findet eine Indizierung statt.
- Für die Indizierung wird Apache Lucene benutzt.
- Im Index werden die Daten in Form von „key-value“ gespeichert. Für jeden Element wird es ein sog. Document erstellt.
- Dafür werden die Daten aus der Strukturbeschreibung der Mets-Datei rausgelesen und zusammen mit den Descreptiven Metadaten in den Index reingeschreiben. Was genau in den Index kommt wird durch die Konfigurationsdatei des Indexers bestimmt.

Übersicht über die wichtigsten Felder in der Konfigurationsdatei:

1. Konfiguration der Ordner

`<logFolder>C:/goobi-index/logs/</logFolder>`
Ordner für die Log Dateien

`<hotFolder>C:/goobi-index/hotfolder/</hotFolder>`
Ordner, wohin die Daten aus Goobi kopiert werden

`<successFolder>C:/goobi-index/success/</successFolder>`
Ordner für die Erfolgsmeldungen

`<tiffFolder>C:/goobi-index/tiff</tiffFolder>`
Ordner für die Tiff-Dateien, die nach der Indizierung hierher verschoben werden

`<indexedMets>C:/goobi-index/indexed_mets/</indexedMets>`
Ordner für die METS Dateien, die nach der Indizierung hierher verschoben werden.

`<errorMets>C:/goobi-index/error_mets/</errorMets>`
Ordner für die Fehlermeldungen

<updatedMets>C:/goobi-index/updated_mets/</updatedMets>
Ordner für die Mets-Dateien, die hierher nach dem Update verschoben werden

<deletedMets>C:/goobi-index/deleted_mets/</deletedMets>
Ordner für die Mets-Dateien, die aus dem Index gelöscht sind

<lucene>C:/goobi-index/deleted_mets/</lucene>
Ordner für Lucene Dateien.

2. Mapping GDZ - > ZVDD
Syntax: <TYPE in der METS-Datei>ZVDD-Typ</TYPE in der METS-Datei> d.H. im Index wird der ZVDD-Typ verwendet.

```
<docstructmapping>
                <list>
                 <_default>OtherDocStrct</_default>
                    <Abstract>Abstract</Abstract>
                    <Acknowledgment>Acknowledgment</Acknowledgment>
                                          ...
```

3. Konfiguration der nicht obligaten Felder für den Indexer
Um eigene Felder erweiterbar. Dadurch bekommt man die Möglichkeit selber zu bestimmen, welche Metadaten und wie (xpath Beschreibung) indiziert werden.

Konfiguriert werden nur Felder die per xpath innerhalb der MODS-Sections ansprechbar sind.

```
    <fields>
        <METADATA> <!-- Name des Feldes-->
            <!--
                Zum Beispiel : Die gesamte MODS Section
            //-->
           <list>
                <item>
                    <xpath>mets:xmlData/mods:mods</xpath>
                    <getchilds>all</getchilds>
                    <store>NO</store>
                    <index>TOKENIZED</index>
                    <addToDefault>false</addToDefault>
                </item>
            </list>
        </METADATA>
           ...
```

Lucene-Felder

Bei der Indexierung werden im Lucene Dokument folgende Felder erstellt:

Feldname	Beschreibung	Obligat	Sonstiges
IDDOC	Eindeutiges Identifikationsnummer für den Dokument (hier und weiter in diesem Abschnitt wird es unter Dokument ein Lucene Dokument gemeint). Ein LongInteger, was hochgezählt wird.	ja	darf nur einen geben
PARENT_PI	PI (PPN) des Parents. Wird erstellt nur bei den Werken, die aus mehreren Teilen bestehen.	nein	Eingeführt in der Version 2.0
IDPARENTDOC	IDDOC des Parents	nein	kommt zweimal vor. Bei zweitem Mal wird IDDOC und PPN gespeichert
DATEMODIFIED		ja	
DOCSTRCT	Jedes Dokument verfügt über eine Strukturbezeichung, die hier gespeichert und indexiert wird. Ggf. muss vorher ein Mapping (in der Konfiguration) durchgeführt werden (z.b. auf ZVDD oder DFG-Viewer Regelwerk).	ja	

Ausserdem werden im Indexer mehrere Felder abhängig von der Konfigurationsdatei erstellt (sehe Beschreibung der Konfigurationsdatei).

Ablauf der Indizierung

Folgende Abschnitte sind nicht für den Endkunden, sondern als Teil der internen Programmbeschreibung zu betrachten.

Bei der Indizierung der Mets-Dateien werden Strukturdaten gelesen. Jedes Objekt wird als ein Document in den Index geschrieben. Die Zusammenhänge zwischen diesen Documenten werden durch die Felder IDDOC und IDPARENTDOC beschrieben. Dadurch besteht eine Parent-Child Relation zwischen diesen Documenten. Durch die Konfigurationsdatei kann man bestimmen welche, ausser obligaten Feldern sonst Indiziert werden. Indizierung findet statt damit man eine Suche durchführen kann. Damit bestimmt die Konfigurationsdatei, wonach es gesucht werden kann.

IMPORT

- Mets-Dateien und die zugehörige Images werden in den Hotfolder von Goobi kopiert.

- Hotfolder-Mechanismus merkt das und startet die Arbeitsroutine (Überprüfung findet je 15 Sekunden statt)

- Aus dem Index wird letztes IDDOC geholt und um 1 hochgezählt

- Es wird geprüft ob es sich um ein Multivolume handelt

- Aus der Mets-Datei wird die Strukturbeschreibung geholt und geparst.

- PPN Nummer der Mets-Datei wird geholt

- Es wird überprüft ob es sich gerade um ein Update (Datei wurde schon indiziert) oder um eine Neuindizierung handelt

- Wenn es sich bei der Mets-Datei um ein Volume handelt, dann wird PPN Nummer des „Parents“ in den Feld PARENT_PI gespeichert.

- Es wird nach dem „Parent“ in dem Lucene gesucht. Wenn es gefunden ist, wird sein IDDOC in dem Feld IDPARENTDOC gespeichert.

- Es wird überprüft ob es sich bei der METS-Datei um eine sog. Anchor-Datei handelt.
- Wenn dies der Fall ist wird es im Lucene nach den Dokumenten gesucht, die einen PARENT_PI Feld haben, der gleiches PPN trägt wie die Anchor-Datei. Danach werden für jeden „Child“ Felder IDPARENTDOC geschrieben.

- Nachdem die wichtigsten Felder geparst und in Lucene geschrieben sind, wird dies für jeden Strukturelement und seine Kinder rekursiv durchgeführt.

- Mets-Datei wird in den Ordner indexed_mets verschoben. Die Images werden in den tiff-Ordner verschoben. In den Ordner success wird eine Erfolgsmeldung als eine leere Datei mit folgendem Namen geschrieben : PPN<12345>.success

UPDATE

- Update ist eine Indizierung der Mets-Dateien, die schon ein Mal indiziert wurden.
-
- Bei dem Update wird es Unterschieden, ob es sich um eine Anchor-Datei handelt oder nicht.
- Wenn es sich um eine reguläre Mets-Datei geht, werden der Dokument und alle seine „Kinder“ gelöscht. Dabei werden die Kinder mit der Hilfe von dem Feld IDPARENTDOC gesucht. Danach findet ein IMPORT statt. Bei dem Update ändert sich der IDDOC der neuindizierten Dokumente

- Wenn es sich um eine Anchor Datei handelt, wird der zugehöriger Dokument aus dem Index gelöscht, aber sein IDDOC gespeichert. Nach der Neuindizierung des Mets-Datei wird ein neues Dokument mit dem altem IDDOC erstellt.

- Nach dem Update werden die Images in den tiff-Ordner verschoben, Mets-Datei wird in den Ordner updated_mets verschoben.

DELETE

- Beim Löschen handelt es sich um Entfehrnen der Daten zu einer Mets-Datei aus dem Index. Mets-Datei wird in den Ordner deleted_mets verschoben.
- Löschen der Daten einer Anchor-Datei ist nicht möglich, wenn es noch mindestens ein „Child“ von diesem Anchor indexiert ist. Dies verhindert das man unabsichtlich mehrere Werke auf ein Mal löscht.

Anhang A: Anchor und seine Kinder

Werke die aus Goobi importiert werden, kann man als singuläre Werke oder als Teil einer Kollektion (zum Bespiel mehrere Bänder eines Werks) betrachten. Wenn es sich um eine Kollektion handelt, dann braucht man eine Mets-Datei die beschreibt, welche Werke zu dieser Kollektion gehören. Diese Mets-Datei wird als ein Anchor in dieser Documentation genannt. Die Werke, dessen Zusammenhalt diese Anchor-Datei beschreibt, werden in dieser Documentation als „Kinder“ oder „Children“ bezeichnet. Anchor Datei hat eigenen PPN Nummer und beinhaltet keine Abschnitte, die physikalische Eigenschaften und Verknüpfungen beschreiben.

Anhang B : Sonderfall - Werke mit mehreren Bändern

Manchmal beschreiben die Anchor-Dateien nicht alle Bänder, die ein Werk beinhalltet. In diesen Fällen muss der Indexer die Anchor Dateien nach dem Import/Update modifizieren, damit die Präsentationsschicht mit einer Anchor-Datei arbeitet, die auch vollständig ist. Im folgendem wird der Ablauf in solchem Fall beschrieben :

- Nach der Indizierung der Anchor Datei wird es geschaut wie viele „children“ diese hat. Wenn der Anzahl der „children“ mit dem Anzahl der Einträge nicht übereinstimmt, dann muss die Anchor Datei ergänzt werden.
- Es wird im Lucene nach „children“ gesucht und folgender Information über diese gesammlet: PPN / CURRENTNO / LABEL
- Danach wird eine neue XML-Datei mit der Endung .UPDATED in den Hotfolder geschrieben. Diese Datei ist die neue Anchor-Datei mit aktualisierten Informationenen. Diese Datei wird dann in den Ordner indexed_mets bzw updated_mets verschoben und ersetzt dabei eventuell alte Anchor Datei.
- Nach der Indexierung von je Band, der ein Teil von dem Werk ist, werden obengenannte Schritte nochmal durchgeführt. Dadurch wird die aktualität des Anchors gewährgeleistet.

Quellenangabe:

[1] Metadata Encoding & Transmission Standart, http://www.loc.gov/standards/mets/