H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 6.2

Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten

nestor

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

URL für Kapitel 6.2 „Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten“ (Version 2.3): urn:nbn:de:0008-2010061780
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010061780

Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.

sonstige Interessen führen teilweise zu konkurrierenden Inhalten oder unnötigem Umfang von Ansätzen. Die Abgrenzung von Inhalten und die zeitliche Synchronisation können zudem auch durch die Vielzahl der Standardisierungsorganisationen negativ beeinflusst werden. Auf jeden Fall ist das Prozedere der Standardisierung und der Aufbau der Standards sehr unterschiedlich. Die geforderte Offenheit von Standards ist nicht nur eine rein definitorische Angelegenheit, sondern kann weitgehende rechtliche und wirtschaftliche Konsequenzen haben. Versteckte Patente oder sonstige Hindernisse, die z.B. Mitbewerber bei einer Implementierung behindern, können sich nachteilig auf die Zuverlässigkeit und Wirtschaftlichkeit der Langzeitarchivierung auswirken. Vorteilhaft ist, dass sich die Standardisierungsorganisationen um mehr Transparenz und auch Einheitlichkeit bei der Behandlung und Darstellung von Rechten (Intellectual Property Rights – IPR) bemühen. Das folgende Kapitel präsentiert einige wesentliche Entwicklungen im Bereich der internationalen Standards und der Bemühungen, im Bereich der technischen Standards und der Metadatenstandards für die digitale Langzeitarchivierung zu entwickeln.

6.2 Metadata Encoding and Transmission Standard – Einführung und Nutzungsmöglichkeiten

Markus Enders

Ausgehend von den Digitalisierungsaktivitäten der Bibliotheken Mitte der 1990er Jahre entstand die Notwendigkeit, die so entstandenen Dokumente umfassend zu beschreiben. Diese Beschreibung muss im Gegensatz zu den bis dahin üblichen Verfahrensweisen nicht nur einen Datensatz für das gesamte Dokument beinhalten, sondern außerdem einzelne Dokumentbestandteile und ihre Abhängigkeiten zueinander beschreiben. So lassen sich gewohnte Nutzungsmöglichkeiten eines Buches in die digitale Welt übertragen. Inhaltsverzeichnisse, Seitennummern sowie Verweise auf einzelne Bilder müssen durch ein solches Format zusammengehalten werden.

Zu diesem Zweck wurde im Rahmen des „Making Of Amerika" Projektes Ebind entwickelt[1]. Ebind selber war jedoch ausschließlich nur für Digitalisate von Büchern sinnvoll zu verwenden.

Um weitere Medientypen sowie unterschiedliche Metadatenformate einbinden zu können, haben sich Anforderungen an ein komplexes Objektformat ergeben. Dies setzt ein abstraktes Modell voraus, mit Hilfe dessen sich Dokumente flexibel modellieren lassen und als Container Format verschiedene Standards eingebunden werden können. Ein solches abstraktes Modell bildet die Basis von METS und wird durch das METS-XML-Schema beschrieben. Daher wird METS derzeit auch fast ausschließlich als XML serialisiert und in Form von Dateien gespeichert. Als Container Format ist es in der Lage weitere XML-Schema (so genannte Extension Schemas) zu integrieren.

Das METS Abstract Model

Das METS „Abstract Model" beinhaltet alle Objekte innerhalb eines METS Dokuments und beschreibt deren Verhältnis zueinander. Zentraler Bestandteil eines METS-Dokuments ist eine Struktur. Das entsprechende Element nennt sich daher structMap und ist als einziges Element im „Abstract Model" verpflichtend. Jedes METS Dokument muss ein solches Element besitzen. Unter Struktur wird in diesem Fall eine hierarchische Struktur mit nur einem Start-

1 O.V.: An Introduction to the Electronic Binding DTD (Ebind). http://sunsite.berkeley.edu/Ebind/
Alle hier aufgeführten URLs wurden im Mai 2010 auf Erreichbarkeit geprüft .

knoten verstanden. Eine Struktur kann also als Baum interpretiert werden. Der Wurzelknoten sowie jeder Ast wird als Struktureinheit bezeichnet. Jede Struktur muss über einen Wurzelknoten verfügen. In der Praxis kann diese verpflichtende Struktur bspw. die logische Struktur – also das Inhaltsverzeichnis einer Monographie speichern. Im Minimalfall wird dieses lediglich die Struktureinheit der Monographie umfassen, da der Wurzelknoten in dem Baum verpflichtend ist. Weitere Strukturen sind optional. Eine weitere Struktur könnte bspw. die physische Struktur des Dokuments sein. Die physische Struktur beschreibt bspw. aus der Exemplarsicht (gebundene Einheit mit Seiten als unterliegende Struktureinheiten).

Verknüpfungen zwischen zwei Struktureinheiten werden in einer separaten Sektion gespeichert. Das „Abstract Model" stellt dazu die structLink Sektion zur Verfügung, die optional genutzt werden kann. Jede Verknüpfung zwischen zwei Struktureinheiten wird in einem eigenen Element definiert.

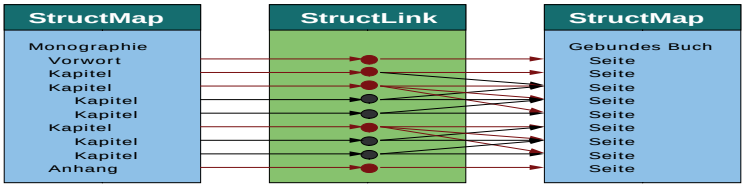

Abbildung 1: Verknüpfung von zwei Strukturen im Abstract-Model

Das „Abstract Model" macht allerdings keine Vorgaben, aus welcher strukturellen Perspektive ein Dokument beschrieben wird oder wie detailliert diese einzelnen Strukturen ausgearbeitet werden müssen.

Ferner berücksichtigt das „Abstract Model" auch Metadaten. Hierunter sind allerdings nicht nur bibliographische Metadaten zu verstehen. Vielmehr wird in deskriptive Metadaten (in der Descriptive Metadata Section) und administrative Metadaten (in der Administrative Metadata Section) unterschieden. Während die deskriptiven Metadaten bibliographische Informationen enthalten, werden Informationen zu Rechteinhabern, Nutzungsrechte, technische Informationen zu einzelnen Dateien oder Langzeitarchivierungsmetadaten in den administrativen Metadaten gespeichert. Für beide Metadatentypen können beliebige Schema, so genannte „Extension Schema" genutzt werden, die in der jeweiligen Sektion gespeichert werden. Auch die Referenzierung von Metadatensätzen ist möglich, sofern diese bspw. per URL zugänglich sind. Jede Datei sowie jeder Struktureinheit lässt sich mit entsprechenden Metadatensätzen versehen, wobei jeder Einheit mehrere Datensätze zugeordnet werden können. Als „Extensi-

on Schema“ können sowohl XML-Metadatenschema wie bspw. MARC XML, MODS, Dublin Core) sowie Binärdaten benutzt werden. Dies erlaubt auch die Integration gängiger bibliothekarischer Standards wie bspw. PICA-Datensätze.

Abbildung 2: Verweis auf Metadatensektionen im METS-Abstract-Model

Neben den Struktureinheiten und ihren zugehörigen Metadaten spielen auch Dateien bzw. Streams eine wesentliche Rolle, da letztlich in ihnen die durch das METS-Dokument beschriebenen Inhalte manifestiert/gespeichert sind. Eine Datei kann bspw. den Volltext einen Buches, die Audioaufnahme einer Rede oder eine gescannte Buchseite als Image enthalten. Entsprechende Daten können in ein METS-Dokument eingebunden werden (bspw. Base64 encoded in die METS-XML Datei eingefügt werden) oder aber mittels xlink referenziert werden. Ein METS-Dokument kann also als Container alle für ein Dokument notwendigen Dateien enthalten oder referenzieren, unabhängig davon, ob die Dateien lokal oder auf entfernten Servern vorhanden sind. Metadatensätze, die nicht in die METS Datei eingebunden sind, werden nicht als Datei betrachtet, sondern sind aus der entsprechenden Metadatensektion zu referenzieren.

Grundsätzlich müssen alle für ein METS-Dokument relevanten Dateien innerhalb der File-Sektion aufgeführt werden. Innerhalb der File-Sektion können Gruppen (File-Groups) von Dateien gebildet werden, wobei die Abgrenzungskriterien zwischen einzelnen Gruppen nicht durch das „Abstract Model“ definiert sind. Ja nach Modellierung lassen sich Dateien bspw. nach technischen Parametern (Auflösung oder Farbtiefe von Images), Anwendungszweck (Anzeige, Archivierung, Suche) oder sonstigen Eigenschaften (Durchlauf bestimmter Produktionsschritte) den einzelnen Gruppen zuordnen.

Das METS-Abstract-Model erlaubt das Speichern von administrativen Metadaten zu jeder Datei. Generelle, für jede Datei verfügbare technische Metadaten wie Dateigröße, Checksummen etc. lassen sich direkt in METS speichern. Für weiterführende Metadaten kann mit jeder Datei eine oder mehrere Administrative Metadatensektion(en) verknüpft werden, die bspw. Formatspezifische Metadaten enthalten (für Images könnten die Auflösungsinformationen, Informationen zur Farbtiefe etc. sein).

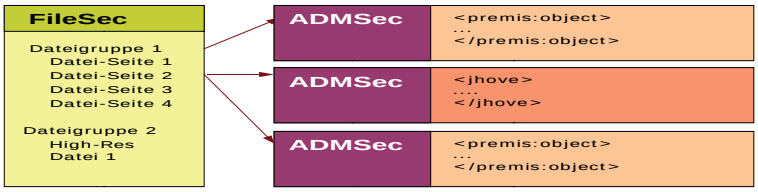

Abbildung 3: Administrative Metadata zu Dateien

Dateien sind darüber hinaus mit Struktureinheiten verknüpft. Die Struktureinheit, die eine einzelne Buchseite repräsentiert, kann somit mit einer einzelnen Datei, die ein Image dieser Seite beinhaltet, verknüpft werden. Das „METS-Abstract-Model" stellt hierzu eine N:M Verknüpfung bereit. Das bedeutet, dass eine Datei von mehreren Struktureinheiten (auch aus unterschiedlichen Strukturesektionen) aus verknüpft werden kann, genauso wie eine Struktureinheit mehrere Dateien verknüpfen kann. Im Ergebnis heißt das, dass der Struktureinheit vom Typ „Monographie" sämtliche Imagedateien eines gescannten Werkes direkt unterstellt sind.

Für die Verknüpfung von Dateien sieht das „METS-Abstract-Model" noch weitere Möglichkeiten vor. So lassen sich mehrere Verknüpfungen hinsichtlich ihrer Reihenfolge beim Abspielen bzw. Anzeigen bewerten. Dateien können entweder sequentiell angezeigt (Images eines digitalisierten Buches) oder auch parallel abgespielt (Audio- und Videodateien gleichen Inhalts) werden. Darüber hinaus kann nicht nur auf Dateien, sondern auch in Dateiobjekte hinein verlinkt werden. Diese Verlinkungen sind u.a. dann sinnvoll, wenn Einheiten beschrieben werden, die aus technischen Gründen nicht aus der Datei herausgetrennt werden können. Das können bestimmte Teile eines Images sein (bspw. einzelne Textspalten) oder aber konkrete zeitliche Abschnitte einer Audioaufnahme. In der Praxis lassen sich so einzelne Zeitabschnitte eines Streams markieren und bspw. mit inhaltlich identischen Abschnitten eines Rede-Manuskriptes taggen. Das METS-Dokument würde über die Struktureinheit eine Verbindung zwischen den unterschiedlichen Dateien herstellen.

Abbildung 4: Struktureinheit ist mit verschiedenen Dateien und Dateibereichen verknüpft

Das METS-Abstract-Model nutzt intensiv die Möglichkeit, einzelne Sektionen miteinander zu verknüpfen. Da METS überwiegend als XML realisiert ist, geschieht diese Verknüpfung über XML-Identifier. Jede Sektion verfügt über einen Identifier, der innerhalb des XML- Dokumentes eindeutig ist. Er dient als Ziel für die Verknüpfungen aus anderen Sektionen heraus. Aufgrund der XML-Serialisierung muß er den XML-ID Anforderungen genügen. Es muss bei Verwendung von weiteren Extension Schemas darauf geachtet werden, dass die Eindeutigkeit der Identifier aus dem unterschiedlichen Schema nicht gefährdet wird, da diese üblicherweise alle im gleichen Namensraum existieren.

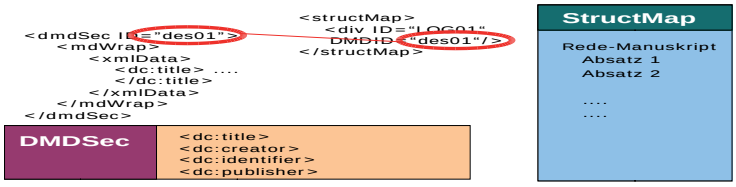

Abbildung 5: Unterschiedliche Sektionen mittels XML-IDs verknüpft

Dokumentation

Wie deutlich geworden ist, stellt das METS-Abstract-Model sowie des XML-Serialisierung als METS-XML Schema lediglich ein grobes Modell da, welches auf den jeweiligen Anwendungsfall angepasst werden muss. Die Verwendung von Extension Schema sollte genauso dokumentiert werden wie die Nutzung optionaler Elemente und Attribute in METS. Hierbei sollte vor allem auch die Transformation realer, im zu beschreibenden Dokument vorhandene Objekte in entsprechende METS-Objekte bzw. METS-Sektionen im Vordergrund stehen. Eine einzige Strukturesektion kann bspw. logische Einheiten (bspw. das Inhaltsverzeichnis eines Buches) umfassen als auch bestimmte physische Einheiten (bspw. einzelne Seiten) enthalten. Alternativ können jedoch bestimmte Einheiten in eine separate Struktursektion ausgelagert werden. Das „Abstract

Model“ erlaubt diese Flexibilität. Eine Implementierung von METS für einen bestimmten Anwendungsfall muss dieses jedoch konkret festlegen.

Um die Dokumentation zu standardisieren wurde das METS-Profile Schema entwickelt. Es gibt eine Grobstrukturierung vor, die sicher stellt, dass alle wesentlichen Bereiche eines METS-Dokuments in der Dokumentation berücksichtigt werden. Die Dokumentation selber muss derzeit noch auf XML Basis erfolgen. Die so entstandene XML-Datei lässt sich jedoch anschliessend als HTML oder PDF konvertieren.

Um ein solches Profil auf der offiziellen METS-Homepage veröffentlichen zu können, wird es durch Mitglieder des METS-Editorial-Board verifiziert. Nur verifizierte METS-Profile werden veröffentlicht und stehen auf der Homepage zur Nachnutzung bereit. Sie können von anderen Institutionen adaptiert und modifiziert werden und somit erheblich zur Reduktion der Entwicklungszeit einer eigenen METS-Implementierung beitragen.

Fazit

Aufgrund der hohen Flexibilität des METS Abstract Models wird METS in einer großen Zahl unterschiedlicher Implementierungen für sehr verschiedene Dokumententypen genutzt. Neben der ursprünglichen Anwendung, digitalisierte Büchern zu beschreiben, existieren heute sowohl METS-Profile zur Webseitenbeschreibungen (Webarchivierung) sowie Audio- und Videodaten. Während in den ersten Jahren METS überwiegend zum Beschreiben komplexer Dokumente genutzt wurde, um diese dann mittels XSLTs oder DMS-Systeme verwalten und anzeigen zu können, wird METS heute gerade auch im Bereich der Langzeitarchivierung zur Beschreibung des Archival Information Packets (AIP) genutzt. METS ist heute für viele Bereiche, in denen komplexe Dokumente beschrieben werden müssen, ein De-facto-Standard und kann sowohl im universitären als auch im kommerziellen Umfeld eine große Zahl an Implementierungen vorweisen. Ein großer Teil derer ist im METS-Implementation Registry auf der METS-Homepage (http://www.loc.gov/standards/mets/) nachgewiesen.