Bundesarchiv
Kerstin Schenke

Koblenz, d. 15.10.2010

Das Metadatenkonzept des Bundesarchivs

Vortrag bei der Sitzung des AWV Arbeitskreises 6.2 „Dokumentation und Archivierung von Webpräsenzen“ am 19.10.2010 in München

1.) Einleitung

Ihr Arbeitskreis „Dokumentation und Archivierung von Webpräsenzen“ bat um Informationen zur Genese des Metadatenkonzepts für das Digitale Archiv des Bundesarchivs.
Vielen Dank für Ihr Interesse und Ihre Einladung in Ihren Arbeitskreis.

Das Bundesarchiv arbeitet an seinem Digitalen Archiv seit vielen Jahren, der Produktivbetrieb läuft seit Oktober 2008. Wir haben uns dabei zunächst auf die Übernahme elektronischer Akten konzentriert und zur Zeit bereiten wir das Digitale Archiv gerade für die Übernahme von Datenbanken vor.
Das Feld der Webarchivierung wurde bis heute bewusst der Deutschen Nationalbibliothek überlassen. Ihr wurde diese Aufgabe per Gesetz übertragen und in ihrem Sammlungsprofil ist die Sicherung von Webseiten des Bundes festgehalten.
Das Bundesarchiv und die Deutsche Nationalbibliothek gehören beide zum Geschäftsbereich des Bundesbeauftragten für Kultur und Medien.

2.) Das Metadatenkonzept des Bundesarchivs. - Der Werdegang

2.0.) Voraussetzungen

Das Digitale Archiv des Bundesarchivs wurde konform zu OAIS (Open Archival Information System) - dem Referenzmodell der NASA und ISO-Standard - konzipiert, in dem die Ablage von digitalen Archivobjekten in AIP's (Archival Information Package) vorgegeben wird. Die Vorgaben des OAIS für AIP's sind jedoch so allgemein gehalten, dass das Metadatenkonzept für die Anforderungen des Bundesarchivs konkretisiert werden musste.

Im Jahre 2005 galt es zunächst, Inhalt und Technik für die Übernahme von elektronischen Akten zu beschreiben. Dafür haben sich die Kollegen - insbesondere Andrea Hänger, Kathrin Schroeder und Karsten Huth - in den Jahren 2005 und 2006 die bestehenden Standards angesehen:

- ELMER - von der Deutschen Nationalbibliothek
- VERS - aus Australien
- Metadaten im Kontext von Archisafe und den
- DOMEA-Aussonderungskatalog.

2.1.) XDOMEA 1.0

XDOMEA 1.0 war zunächst der Standard für alle Metadaten, mit dem elektronische Akten inhaltlich beschrieben und vom DMS/VBS in das Digitale Archiv des Bundesarchivs übernommen werden konnten. XDOMEA 1.0 enthält u.a. die Elemente

- Vorgangskennzeichen
- Vorgangsbetreff
- Laufzeiten
- Vertraulichkeitsstufen.

Vorgegeben wird die Stufung „Akte, Vorgang, Dokument", die wir heute ggf. über den Strukturierungseditor herstellen können.

XDOMEA 1.0 ließ eine Lücke von Metadaten und Informationen zur genauen Beschreibung des digitalen Archivobjekts, die für die langfristige Erhaltung digitaler

Unterlagen bedeutsam sind. Dafür benötigte man einen Standard, der diese anderen Bereiche abdeckt und sich zugleich eng an das AIP-Modell im OAIS-Standard hält.

2.2.) PREMIS

Dann haben die Kollegen PREMIS entdeckt. PREMIS entsprach den Anforderungen des Bundesarchivs von allen vorhandenen Standards am besten und ist - wie auch XDOMEA - bis heute Bestandteil von XBArch.

PREMIS steht für "**PRE**servation **M**etadata: **I**mplementation **S**trategies", der Namen einer internationalen Arbeitsgruppe, die von 2003 bis 2005 von OCLC (Online Computer Library Center) und RLG (Research Libraries Group) gefördert wurde. PREMIS hat sich seit 2005/2006 weltweit verbreitet. Eine überarbeitete zweite Version wurde im März 2008 veröffentlicht.
Ein Vorteil von PREMIS besteht darin, dass es einen internationalen Austausch darüber gibt - und auch eine deutsche Übersetzung, an der beispielsweise Karsten Huth beteiligt war.

PREMIS gibt Auskunft über die technische Herkunft und die weitere Behandlung eines digitalen Objekts.
Es werden einzelne Dateien beschrieben,

- in ihrer Größe,
- in ihrem Format und
- in der Version,
- in den Umgebungen, die für das Lesen und Aufbewahren der Dateien nötig sind, aber auch
- welcher Viewer für die Organisation der Dateien benötigt wird,
- wie sie bearbeitet wurden usw..

Das heißt, im AIP wird mittels PREMIS die ganze Historie der Datei hinterlegt.
Von Bedeutung sind dabei die „wesentlichen Eigenschaften“, der unbedingt zu erhaltenden Informationen (significant properties), um den nötigen Informationsgehalt zum AIP über die Zeit zu bringen. Beispielsweise kann es wichtig sein, dass digitale Unterlagen in Farbe vorliegen, um die farbigen Geschäftsgangsvermerke interpretierbar zu halten oder bei Bildern die Farbtiefe so genau zu beschreiben,

dass sie rekonstruierbar ist. Es ist mit PREMIS möglich, Ursprungsinformationen bis ins letzte Detail vorzuhalten, so dass beispielsweise auch eine Emulation möglich wäre.
PREMIS bietet damit die Möglichkeit, digitale Objekte unabhängig vom Inhalt und von der Objektart technisch zu beschreiben.

PREMIS unterscheidet zwischen

- einem Ereignis (event), z.B. die Datei ist konvertiert worden,
- dem Agenten, das sind die Programme oder Menschen, die etwas mit den Dateien getan haben,
- Objekten, das können Dateien, Repräsentationen, Bitstream sein,
- sowie Rechten.

Im Bundesarchiv bzw. in XBArch werden für die Beschreibung der technischen Metadaten aus PREMIS die Entitäten „Objekt“, „Ereignis“ und „Agent“ genutzt. Außen vor gelassen wurden bei uns die „Rechte“. Hier geht es eher um Copyright-Fragen, wie sie z.B. die DNB zu klären hat (die Rechte müssen für die professionelle Aufbewahrung von eBooks andere sein, als für Bibliotheksbenutzer).

2.3.) METS

Es gab auch lange Zeit Versuche, METS (**Me**tadata **E**ncoding and **T**ransmission **S**tandard) zu verwenden.
METS bietet die Möglichkeit, Daten und Dokumentationen zusammenzubringen. METS besteht aus sieben Teilen, mit denen ein digitales Buch in all seinen Teilen - Kapiteln und Seiten - beschrieben werden kann. Hier geht es darum, diese Buchteile in der richtigen Ordnung abzubilden. Im Bundesarchiv wird METS v.a. in der SAPMO genutzt, um Digitalisate zu verwalten.

Für das Digitale Archiv gab es Versuche, PREMIS und METS miteinander zu verknüpfen, wobei sich wohl alles verheddert hat. Im Hintergrund stand der Gedanke mittels METS die technischen und die inhaltlichen Metadaten von AIP’s nochmals zu umschließen. Der „Befreiungsschlag“ war schließlich die Erkenntnis, dass dieses zusätzliche Umfassen gar nicht nötig ist.

METS wird vornehmlich für den Austausch bspw. zwischen einzelnen Institutionen benötigt. Dafür gibt es im Digitalen Archiv keinen Bedarf, weil die AIPs bei uns im Haus bleiben und nicht mit anderen Institutionen ausgetauscht werden müssen. Deshalb muss auch keine Kompatibilität mit anderen Systemen geschaffen werden.

2.4.) XBArch

Um die noch immer bestehenden Lücken im Zusammenwirken aus XDOMEA 1.0 und PREMIS zu schließen, wurde eigens XBArch (**X**ML-Schema des **B**undes**arch**ivs) geschaffen.
Die Lücken bestanden insbesondere im administrativen Bereich, in dem die Angaben zur Provenienz des digitalen Objektes als auch der gesamte technische Übernahmeprozess dokumentiert werden.
XBArch ist über Jahre weiterentwickelt und angepasst worden, um eine effektive automatische Verarbeitung von XML-Strukturen, die nach XBArch gemappt werden sowie Fehlerfreiheit bei der Generierung von AIP's zu erreichen. Dabei ist ein relativ flach gehaltenes XML-Schema entstanden, das aus einer einzigen Datei besteht. Sie sehen es auf der Folie.

Ziel von XBArch ist die langfristige Speicherung von digitalem Archivgut in selbstbeschreibenden Container-Einheiten, den AIP's.

Mit XBArch können Informationen zu einzelnen elektronischen Dokumenten strukturiert abgelegt und erschlossen werden. Dafür gibt es die Elemente:

- Archivsignatur
    - eindeutige Archivsignatur für das gesamte AIP, auch über alle Migrationsstufen hinweg
    - Verknüpfung zwischen Applikation und AIP im Datenspeicher
    - künftig auch Verknüpfung zur Archivdatenbank „BASYS“
- Administrative Daten
    - enthält Angaben zum
        - Ereignistyp,
        - Hash-Werte,

        - Datum der Transaktion,
        - Transaktionsnummer,
        - Behördenangaben und
        - Kontaktinformationen
    - Informationen zur Beschreibung des organisatorischen Verfahrens, der Übernahme elektronischer Akten aus einer Bundesbehörde
    - nötig zur Plausibilitätsprüfung im Datentransfer
    - Informationen werden in der te-Datei festgehalten und nach XBArch übernommen
    - damit entsteht eine Übertragungshistorie für das gesamte Dokument (Laufzettel)
- Technische Metadaten
    - PREMIS-Metadaten zur technischen Beschreibung der Objekte sowie aller Maßnahmen für deren dauerhafte Archivierung:
        - Objekt
        - Ereignis
        - Technische Umgebung
        - Agent
- XDOMEA-XBArch
    - ist aus XDOMEA abgeleitet und zur Bildung von AIP’s angepasst worden
    - Datencontainer für die hierarchische Inhaltsbeschreibung von Akten, Vorgängen und Dokumenten
- Platzhalter für Datenstrukturen
- XBArch-SIARD.

XBArch kann flexibel für andere Objektarten erweitert werden, wie bspw. Datenbanken, wobei die Anforderungen in ihrer Art und Komplexität anders sind, als bei elektronischen Akten. Aktuell befassen wird uns im Bundesarchiv damit, das Digitale Archiv auch für Datenbanken aufnahmebereit zu machen. Hier entstehen völlig neue Anforderungen an die Metadaten und eine grundsätzliche, nachhaltige Konzeption.

2.5.) Zusammenfassung

Bei der Erstellung des Metadatenkonzepts wurde im Bundesarchiv schrittweise und sehr systematisch vorgegangen. Zunächst haben sich die Kollegen international nach Vorhandenem umgeschaut und dann geprüft, ob es für unsere Zwecke passt - oder mit welchem Aufwand es angepasst werde könnte. Dafür wurde lange und intensive Testreihen gefahren, um auch die Feinheiten zu erfassen. Letztendlich hat man sich dann für vorhandene Standards wie XDOMEA und PREMIS entschieden und diese durch die Eigenentwicklung XBArch ergänzt.
XBArch ist ein dynamisches Metadatenschema, das bei Bedarf neuen Anforderungen angepasst und ergänzt werden kann.