Pilotprojekt KOSTPROBE
Arbeitsgruppe Metadaten

Zusammenfassung der Erkenntnisse

1 Minimaler Metadaten-Katalog

In der Sitzung vom 12. Januar 2006 einigte sich die AG Metadaten auf ein umfangreiches Subset von PREMIS-Elementen, das im Projekt KOSTPROBE verwendet werden sollte. Aus der konkreten Anwendung und aus dem Informationsaustausch auf der PREMIS-Liste ergaben sich einige zusätzliche Erkenntnisse:

- Das von der AG vorgeschlagene Subset ist sehr ausführlich. Für die momentan unumgängliche manuelle Zusammenstellung der Metadaten ist es auf jeden Fall zu umfangreich.
- Die Arbeitsgruppe Erschliessung hat eine präzisere Analyse der für die digitale Archivierung notwendigen Metainformationen durchgeführt. Es ist nun klar geworden, dass die gemäss PREMIS in maschinenlesbarer Form verzeichneten Metadaten hauptsächlich diejenigen Informationen umfassen sollen, die für die automatisierte Weiterbearbeitung der Archivdaten wichtig sind, also Informationen zum Authentizitätsnachweis und für zukünftige Migrationen; anders ausgedrückt: administrativ-technische Metadaten im engeren Sinn. Alles andere wird mit Vorteil in der Dokumentation oder in den Verzeichnungssystemen in menschenlesbarer Form aufgezeichnet.
- PREMIS wird grundsätzlich als Datenaustauschstandard verstanden. Das bedeutet, dass der PREMIS-Katalog, allgemein gesprochen, diejenigen Metadaten auflistet, die ein Archiv über seine Primärdaten wissen muss. Er macht keine Aussage darüber, wo und in welcher Form diese Metadaten gespeichert werden sollen; er stellt es dem Archiv vielmehr ausdrücklich frei, gewisse dieser Metadaten in speziellen Verzeichnissen aufzuzeichnen oder gar nur implizit zu kennen[1].

Vor dem Hintergrund dieser Erkenntnisse empfiehlt die AG Metadaten einen stark reduzierten Satz aus dem PREMIS-Katalog zur Verwendung für die digitale Archivierung. Im folgenden ist dieser Satz aufgelistet. Weitere Informationen zu den einzelnen Elementen finden sich im vollständigen Katalog.

[1] PREMIS-Standard, p. ix f.

OBJECT
objectIdentifier
objectIdentifierType
objectIdentifierValue
preservationLevel[3]
objectCategory
objectCharacteristics
compositionLevel
fixity
messageDigestAlgorithm
messageDigest
format[4]
formatDesignation
formatName
formatVersion
formatRegistry
formatRegistryName
formatRegistryKey
formatRegistryRole
creatingApplication
creatingApplicationName
creatingApplicationVersion
dateCreatedByApplication
storage
storageMedium
environment
environmentNote[5]
relationship
relationshipType
relationshipSubType
relatedObjectIdentification
relatedObjectIdentifierType
relatedObjectIdentifierValue
relatedObjectSequence
relatedEventIdentification
relatedEventIdentifierType
relatedEventIdentifierValue
relatedEventSequence

EVENT
eventIdentifier
eventIdentifierType[2]
eventIdentifierValue
eventType
eventDateTime
eventDetail
linkingAgentIdentifier
linkingAgentIdentifierType
linkingAgentIdentifierValue
linkingObjectIdentifier
linkingObjectIdentifierType
linkingObjectIdentifierValue

AGENT
agentIdentifier
agentIdentifierType
agentIdentifierValue
agentName
agentType

[2] Als eventIdentifierType eignet sich der Timestamp des Events (im ISO-Format).
[3] Kann auch weggelassen werden, wenn in einem Archiv nur ein Level angewendet wird.
[4] Entweder formatDesignation oder formatRegistry müssen ausgefüllt werden.
[5] Hier ist es bedeutend einfacher und weitaus ausreichend, auf ein Informatikinventar o.ä. zu verweisen.

2 Erfahrungen aus der Anwendung in ZH und TG

Beide Archive verzichten darauf, die Tabellen, Strukturdatei und Dokumentation zu zippen. Das StAZH legt diese stattdessen in einer Ordnerstruktur ab, deren oberste Ebene die Archivsignatur als Name trägt. Unterordner umfassen die Datenobjekte, die Dokumentation und die Strukturinformationen. Entsprechend wird der Ordner der Datenobjekte als Representation verstanden, welche die einzelnen Tabellen als Files umfasst. Die Anwendung des Objektmodells von PREMIS auf die GVA-Ablieferung erwies sich als schwierig, aber machbar. Hingegen war es nicht möglich, das Datenmodell der Ablieferung in PREMIS abzubilden. Dazu sind gesonderte Strukturinformationen nötig.

3 Limiten und Probleme von PREMIS

Wie oben bereits angedeutet, hat die Arbeit der KOST und der beteiligten Archive einige Limiten von PREMIS aufgezeigt, die im Folgenden aufgelistet werden. Die Auflistung soll nicht als Fundamentalkritik missverstanden werden: Viele dieser Limiten sind in PREMIS von Beginn weg angelegt und hängen mit Struktur und Ziel von PREMIS ursächlich zusammen.

- Beschränkung auf administrativ-technische Metadaten. PREMIS ist kein Standard für sämtliche Metadaten. Er deckt explizit nur die administrativ-technischen Metadaten für die digitale Archivierung ab. Der PREMIS-Standard muss im Verbund mit anderen Standards verwendet werden. Häufig und ursprünglich vorgesehen ist z.B. seine Verwendung innerhalb des METS-Rahmens.
- Keine kategoriespezifische technische Metadaten. Für jede Art von digitalen Unterlagen (Text, Bild, Ton, Video, etc.) können zusätzliche, spezifische technische Metadaten zum Verständnis notwendig sein. In der Arbeit für die Archivierung der Gebäudeversicherungsunterlagen wurde beispielsweise festgestellt, dass die Codierung der Textfiles in den Metadaten festgehalten werden sollte. PREMIS bietet diese Möglichkeit nicht, da es ein generischer Standard ist. Kategoriespezifische Metadaten müssen gemäss zusätzlichen Standards festgehalten werden[6].
- Limiten bei der Abbildung komplexer Strukturen. Die eher bibliotheks- und forschungslastige Zusammensetzung der PREMIS working group mag für die fehlenden Möglichkeiten des Standards verantwortlich sein, komplexe Beziehungen zusammengesetzter Dateien abzubilden. Das PREMIS-Element „relationship“ und seine Unterelemente erlauben zwar die Abbildung und Charakterisierung von Beziehungen. Um Beziehungen wie diejenigen zwischen Tabellen einer relationalen Datenbank (und sei sie so übersichtlich wie das Datenmodell der archivierten GVA-Daten) abzubilden, reichen sie jedoch niemals aus.

[6] Das StAZH hält diese Information im Element formatDesignation fest; dies ist ein möglicher Workaround.

4 Weiterarbeit am Thema administrativ-technische Metadaten

Das Thema administrativ-technische Metadaten ist durch die Arbeit der AG natürlich nicht erschöpft. Die KOST soll deshalb Formen und Wege finden, sich dauerhaft weiter damit zu befassen. Die folgenden Aktivitäten sind geplant:

- Mitarbeit in der nestorII-AG Standards. Georg Büchler wurde zur Mitarbeit in dieser Arbeitsgruppe eingeladen. Er wird sich dort in erster Linie in der Diskussion rund um PREMIS (deutsche Uebersetzung, Mapping mit lmer, ev. Schritte zur offiziellen Standardisierung) engagieren. Dies soll es der KOST ermöglichen, am internationalen Erfahrungsaustausch in Sachen PREMIS dranzubleiben.
- Diskussion zusätzlicher kategoriespezifischer Metadaten. Neben den allgemeinen, durch PREMIS abgedeckten Metadaten werden für digitale Archivalien bestimmter Kategorien (Bilder, Videos, etc.) zusätzlich jeweils kategoriespezifische Metadaten benötigt. Es herrscht Konsens, dass diese Thematik nicht im voraus angegangen werden soll, sondern wiederum in konkreten Projekten, die sich mit solchem Archivgut befassen.
- Regelmässiges Update. Es muss eine Form gefunden werden, wie weitere Erfahrungen und sonstige Inputs in Sachen administrativ-technischer Metadaten gebündelt und den KOST-Mitgliedsarchiven zugänglich gemacht werden können. Da die KOST ähnliche Ueberlegungen im Rahmen ihres Dateiformatprojekts anstellen muss, wird sie aufgefordert, zu gegebener Zeit einen Vorschlag für eine Update-Struktur zu machen. Die AG-Mitglieder zeigen sich grundsätzlich interessiert, sich weiter mit dem Thema Metadaten zu befassen.