

4th International Digital Curation Conference

December 2008

One for Many A Metadata Concept for Mixed Digital Content at a State Archive

Kai Naumann, Christian Keitel, Rolf Lang
Landesarchiv Baden-Württemberg, Ludwigsburg, Germany

<November ><2008>

Abstract

The Landesarchiv (State Archive) of Baden-Württemberg has designed and implemented a metadata concept for digital content covering a heterogeneous range of digital-born and digitised material. Special attention was given to matters of authenticity and to economic ingest and dissemination methods under the requirements of a public archive. This paper describes the outcome of metadata discussions during the implementation period of the DIMAG repository. It treats integration of the repository's architecture with the archival classification concept, measures for long-term accessibility, the creation of adapted metadata placement, and provisions for exchange with other applications for ingest and use. The deliberately short list of metadata elements is included in this paper. Some existing standards have been evaluated under a real use environment; this paper also introduces modifications applied to them in the project context.

1. Stating Requirements

The Landesarchiv (State Archive) of Baden-Württemberg is holding records from the Middle Ages to the present on seven locations throughout the country. Archivists are used to work on files, maps, parchments, photographs, audio and video tapes. All online metadata are based on an administration system (scopeArchiv) that maintains the catalogue and keeps track of the storage locations of non-digital objects. Since 2002, acquisitions also come in digital form. In 2006, a project group (authors of this paper) started work at the Ludwigsburg branch. They constructed the DIMAG system, based on a LAMP (Linux, Apache, MySQL, PHP) web server architecture and providing controlled storage of digital objects and metadata.

By the end of 2007, the Landesarchiv has ingested 16,769 born-digital objects in 19 different series from various branches of the public sector, containing 79,950 single files and 45 million database records. Its holdings include statistical primary data, data from records management systems (RMS) and geographical information systems (GIS), office files, digitised maps and photographs, system manuals and data descriptions. The oldest dataset was created for the census of 1961. Hybrid objects occur, composed of a database with a large paper documentation. In parallel, the Landesarchiv is digitising papers and parchments for internet presentation and, if necessary, for long-term preservation.

Variety is the key challenge for a metadata concept to describe, preserve and distribute these objects. All of them need to be found and localised in the existing finding aid system, regardless of their media character. Logically, this system can be described as a strictly hierarchical classification tree with branches representing depositing institutions, the twigs reflecting series and sub-series, and the leaves describing archival units. The reference code of an object is derived from the labels of branch, twig, and leaf.

The Landesarchiv had other secondary aims:

- Fostering our reputation as a trustworthy custodian by securing integrity and authenticity of the digital records.
- Reducing cataloguing cost by using a simple encoding scheme and by ingesting metadata on transfer from public sector institutions.
- Exchanging finding aid metadata with metadata harvesters from all kinds of communities. Exchange with [BAM](#) and [MICHAELplus](#) is already implemented; further participation in German and European digital library projects is expected.

2. Setting up Principles

Finding a solution started off with the study of functional and data models ([EAD](#), [LMER](#), [METS](#), [NLA](#), [NLNZ](#), [OAIS](#), [PREMIS](#)). The most important principle, though, was to keep the system simple and open to future developments. The idea of adhering to established XML schemas and creating a defined application profile was discussed, but dismissed for three reasons:

- Data protection legislation does not allow State Archives to share the bulk of its holdings with other institutions. Thus, there was no urgent need for exchange of preservation metadata or content.
- If however, in the future, larger parts of the content should be destined for sharing with preservation systems outside the Landesarchiv, standard-compliant AIP design would have to adapt to future schemas, not to the current

ones. It would, e.g., be useless to establish a METS-compliant schema for content if these metadata would, sooner or later, need partial re-structuring. Current international discussion (McDonough, [2008](#)) seems to confirm this point.

- Even though the Landesarchiv is already sharing most of its finding aid metadata with other memory institutions on the national level, there was no recognized standard schema for finding aid metadata which could be adopted internally. Instead, an EAD export interface has been installed, providing a bridge to formats like Dublin Core or others.

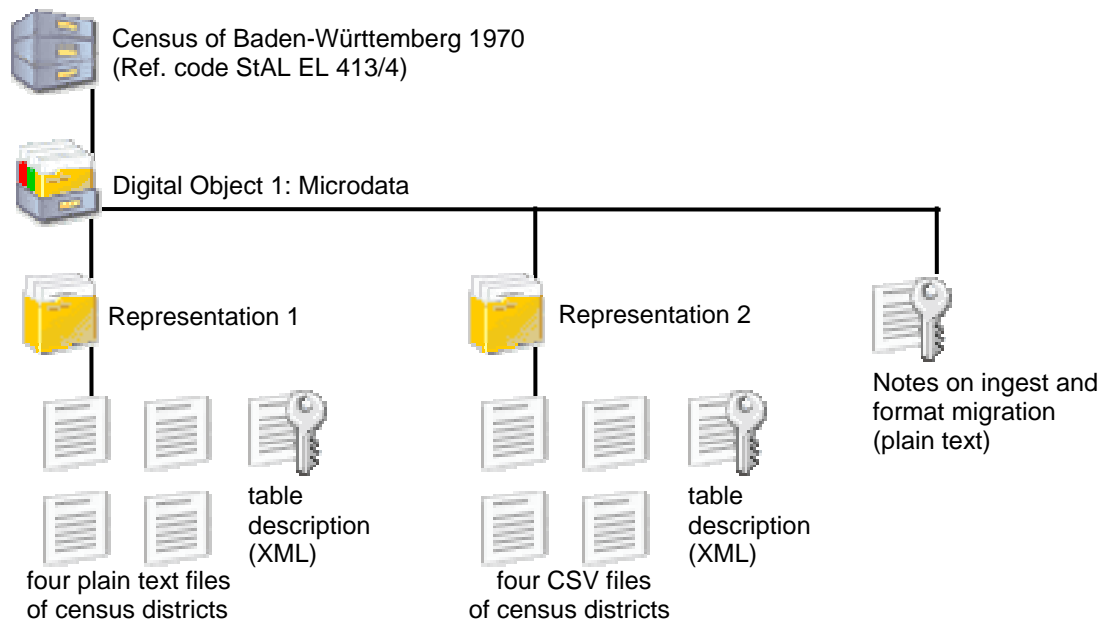


Figure 1: Logical structure of a digital object with 2 representations, 8 content files, 3 documentation files.

2.1 The representation experience

The known standards served as a source for functional requirements. From PREMIS, the concepts of Representation and Significant Properties were derived. Representation will only be used in its simple form: it is defined as an entity containing all the files necessary for the intellectual rendition of an archival object (Figure 1). This definition seemed more suited for practice than the delicate “representation network” model emanating from OAIS that PREMIS had also adopted. Over the course of 100 years, a digital object might thus accumulate several representation folders containing exactly the same information as it was defined in Significant Properties. The totality of folders will document how the information was preserved, but not all of them will necessarily need to be preserved forever. The representation concept will serve as a blueprint for future digitisation at the Landesarchivs, thus ensuring preservation of digitised and digital-born material alike.

PREMIS also inspired the elements added for emulation purposes (see Section 6.2, Representation). Right now, the Landesarchiv does not use this strategy for preservation of its objects. Nevertheless, it decided to keep the first representation of an intellectual entity forever, in order to be prepared if emulation should be working

for some formats in the future.

The National Library of Australia contributed another metadata element they described as “any characteristic that may appear as a loss in functionality or change in the look and feel of a collection, object or file“, for convenience called “Quirks“(NLA, 1999). Adapting it to the representation concept, the definition was generalised to: “Any technical or intellectual deficiency resulting from features of source data” (see section 6.1).

2.2 Enhanced OAIS

When setting up DIMAG, the team also discussed the relations between the functional OAIS entities Data Management (DM) and Archival Storage (AS). Disaster recovery for damaged content is required for AS, and DM has to maintain referential integrity of all metadata (OAIS, 2002, pp. 4-8, 4-9). OAIS does not, however, explicitly require safe recovery of all references between content and metadata. There is no direct data flow between AS and DM (OAIS, 2002, p. 4-17).

In order to close this gap, the team decided to redundantly store vital metadata in the management database and on the storage media.

Even after a total breakdown of all database functions, users will be able to use DIMAG in its emergency mode by simply viewing the file system and reading the core metadata (see Section 6.1 and 6.2) from XML files. This means that the functions of DM are split up: the database component is only covering retrieval of metadata, while storage of metadata is trusted to the storage component that also holds the content.

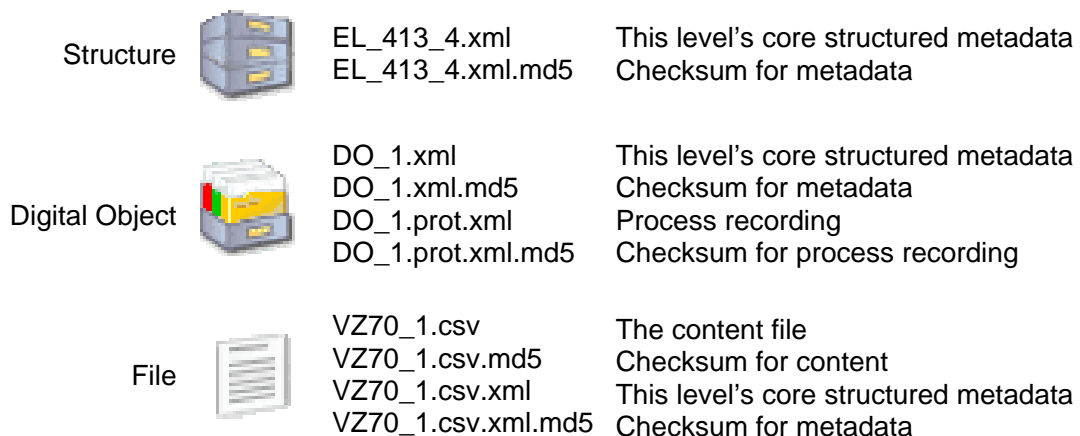


Figure 2: Structural components of DIMAG and their representing files (simplified view).

On storage media, metadata files had to be linked with content files in an easy, robust and efficient way. Therefore, metadata, fixity metadata, and content files are all given the same base name (Figure 2). The decision to provide a stable storage for metadata was, of course, conflicting with the need to change metadata regularly. It led to the challenge of synchronisation: altering some letters in metadata stored with content on a Write Once Read Multiple (WORM) media could possibly mean a re-write of several gigabytes. The problem was solved by a hierarchical storage scheme (Figure 3). The open storage level resides on online random-access media. On this level, content gets enriched with metadata and packaged for long-term storing and fetched for format migration or dissemination. Most of our metadata is located in files on this level, and is continuously synchronised with the database. On the locked storage

level, suited for WORM media, the completely packaged representation containers are stored, making the bulk of the content. Attached to them is a subset of technical metadata (see Section 6.2, Representation and Content File) which can only be altered through versioning or migration of the whole representation. The management database keeps track of every file's physical location and regularly writes backups of this information for disaster recovery.

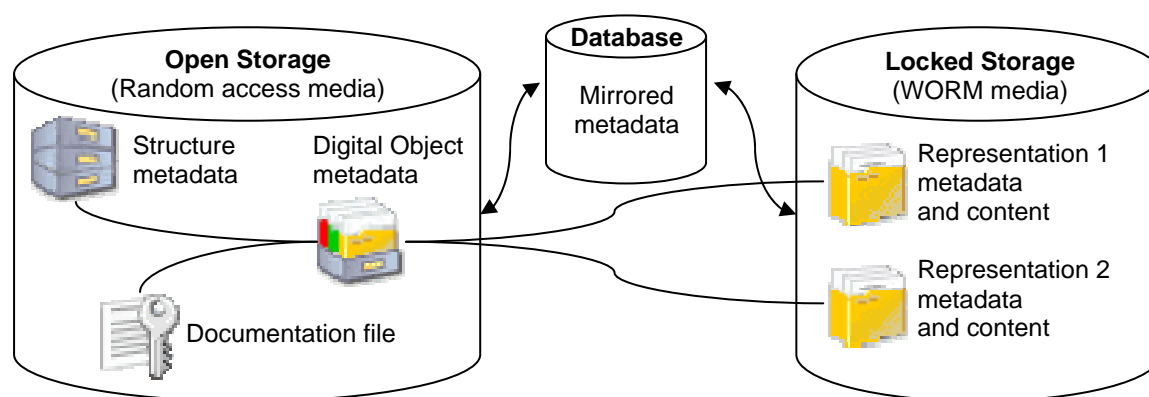


Figure 3: Storage of metadata files and packaged content in a hierarchical storage scheme.

Another challenge was to find a place for integrity metadata (checksum values) of stored metadata files. Putting the calculated value inside the XML would change the XML and alter the checksum. As a simple remedy, DIMAG writes the checksum values into a file assigned to the target file by duplicating the target filename and adding an extension (Figure 2). This operation is not only performed on metadata, but on all kinds of files on storage media. A comparison of all recorded and calculated checksums is executed before every backup and also on demand.

Data table

CityCode	Population	Male	Female
10	1234	600	630
11	3456	1756	1700
...

Sum error will be recorded in metadata, no correction.

Code list

CityCode	CityName
10	Aalen
11	Bottwar
...	...

Wrong CityName value "Achern" replaced by "Aalen". Correction recorded in metadata.

Figure 4: Example of distinction between metadata and content. These primary data consist of a data table (content) and a code list (metadata).

2.3 Sacred content, free metadata

Trustworthiness depends on the ability to preserve information without unauthorised change. Applying this basic insight to metadata, the team drew a sharp distinction between metadata and content. Content was defined as the information to

be preserved, while metadata were defined as data making this information understandable. Thus, if metadata turns out to be wrong, it can be manipulated in a controlled manner, whereas content must be preserved without change. Content is sacred, metadata are free. The decision which data are sacred has to be made individually. Primary microdata can serve as an example (Figure 4): the data records must never be altered and will, on DIMAG, be marked as content. A code list, by contrast, can be classified as metadata and marked as such. An archivist, migrating the code list to a current format, might find two wrongly assigned city names and would be authorised to correct them if corrections are recorded properly. Errors found in the microdata themselves can only be recorded, not corrected.

2.4 Preserving authenticity of structure

Archival collections strongly depend on their structure. While books or e-papers are best described as atomic units that can be re-grouped in any conceivable manner without losing meaning, electronic records at archives often resemble complex molecular structures that lose their character when being dissolved. In other words: for this type of content, preserving structure is a matter of authenticity. Existing repositories use atomic units and some can, on request, record structural dependencies through resource description frameworks stored inside the repository (e. g. the RDF used by Fedora, [2007](#)), but in general, definition of object relationships is unsatisfactory (Borghoff, [2005](#), p. 7). Larger objects with an internal sub-structure are not served by current systems (Woods/Brown, [2008](#), p. 68). EAD is a strong tool to describe these relations, since it was conceived for archival finding aids, but does not warrant authenticity. DIMAG, by contrast, requires a statement of intellectual affiliation for any metadata or content unit. A unit must have only one parent. The affiliation can be changed under certain rules, but not eluded, since it is also the source of the object's reference code. Like all other metadata, it is protected through checksums, thus securing authenticity of structure.

3. Constructing Viable Workflows

The following two subchapters deal with questions that arose when the workflow for ingest was designed. Both are expected to make ingest operations more effective.

3.1 Vital process recording

Archives who want to prove that no information has changed without permission need to record authorised manipulations of content. Will an off-the-shelf web server log do this job? The project group thought that archivists, historians and lawyers searching for evidence might find this type of transaction recording too verbose. Given that DIMAG as an application and as a machine is isolated and accessible only for the archive's staff through personal accounts, it will be sufficient to record vital processes performed by account holders.

The process list includes, amongst others, creation of objects and representations, change of reference code and metadata, deletion of metadata or content, format migrations, validation of migration results and export for use. The processes are recorded in protocol files attached to the digital objects to which they relate (prot.xml file in Figure 2). Most processes are recorded automatically every time a function is used. Others can be recorded by hand if needed. It is impossible to change or delete recorded processes, wrong protocol entries have to be cancelled by another entry. Processes that can not be attached to an object (deletion, change of affiliation) are recorded in a general transaction log.

3.2 Adapted metadata placement

Archival records are used less frequently than books or other learning objects. As a consequence, preservation cost per use case is fairly high. On the other hand, users of archival records tend to accept a modest level of availability (Severiens/Hilf [2006](#), p. 28). In the case of GIS records from 1995, people probably won't expect an archive to have all data available on a state-of-the-art geodatabase server. The project group therefore decided to create an adapted placement policy for metadata encoding. By offering more than one appropriate way of encoding metadata, the team hopes to reduce the time consumed by encoding of structure and rendition information without influencing long-term usability. There are four possible positions for metadata:

- *Core Structured Metadata* is recorded in parsed XML and simultaneously in the management database. It is only used for descriptive levels. Controlled vocabularies are only used for file format, character encoding and content type. The other elements are designed to be as open as possible. DIMAG provides, for example, a free text element called "structure" on representation level. Any specification of structure (readme texts, HTML sitemap, RDF, SQL) given by depositors or archivists can be entered to explain object characteristics.
- *Special Structured Metadata* requires parsed XML based on a schema adopted by the Landesarchiv and recorded in storage, but not in the management database. This level is currently used for data table description, but can be extended to other content type.
- *Integrated Metadata* is part of ingested files. For economic reasons, they are mostly left inside the files and only extracted if necessary. Only some values are extracted automatically via the [JHOVE](#) and [DROID](#) Java libraries and written into structured metadata. Colour depth values residing in a TIFF header will be well preserved inside the file, since obsolescence of TIFF seems to be far away. On the other hand, author's names in office file headers, if relevant to future use, will have to be extracted soon, due to rapid change in office system technology.
- *Documentation* is metadata, but can be structured in any way (for example, the code list in Figure 4). These metadata are treated very much like content files (see Section 6.2, Documentation). Even if this type of metadata is not digital, DIMAG can deal with it. Depositing institutions often submit data descriptions, manuals or other metadata on paper. If these packages prove useful and can not be easily digitised, they will be catalogued and archived on paper and mutually referenced with the digital part. Digitisation on demand for future researchers will be possible.

4. Enabling Exchange

While the first version of DIMAG was conceived as stand-alone, providing both functions of catalogue and repository, its next version will probably rely on an interface to the catalogue (scopeArchiv) and only assume repository functions for content and metadata. By request, scopeArchiv will create representation folders on DIMAG that can be changed through the DIMAG user interface. The catalogue system should also be able to synchronise its classification tree with DIMAG. These operations will require both systems to develop communicative skills. Unique identifiers with a namespace prefix, assigned to metadata as well as content, will play a key role. They will also enable exchange of data packages with future applications inside and outside the Landesarchiv for transfer, ingest, migration, and use. Most

likely, though, these identifier codes will not be used for human citation purposes. Unlike many other communities, archives have a long tradition of stable reference numbers that will continue to be the standard persistent identification for citation.

As mentioned above, the Landesarchiv is not yet focused on use scenarios. Possible solutions for Dissemination Information Packages (DIPs) might be small static websites set up with XSL-transformed XML. These packages would resemble a tiny portion of DIMAG, containing metadata and content in a portable format.



Figure 5: A prototype all-purpose DIP format showing catalogue context (section 1) and the internal structure of the requested census primary data (section 2).

5. Points for Discussion

There are some findings of the project group that could be discussed on a larger basis:

- Concepts for long term preservation metadata have to balance instant availability with easy ingest and long-term understandability. In the case of heterogeneous object types with a low expected use frequency, availability can be reduced in order to advance ingest and understandability. This leads to simple metadata sets for finding aid and structured metadata level, leaving additional information on a non-standardised level.
- Long term archiving is largely based on inter-operability of past, present, and future systems, policies and concepts. Persistent identification is a key asset for the resulting interchange operations. Internal identifiers and reference codes for the public should be viewed separately, though.
- Repository owners often believe to have warranted inter-operability for content sharing by using XML-based standards. What actually exists, though, are local profiles or schemas based on these standards, and sharing between repositories still seems to be a challenge (McDonough, 2008). Paradoxically, repository developers who have to deal with heterogenous content might avoid a waste of

time and money by neglecting standards in metadata storage. It might be wiser to foster standards only in defined metadata or content exchange projects, be it on statistical primary data, office documents, or digitised journals.

- Repository systems often don't provide maintenance of relational integrity between content and metadata. Partial mirroring of database metadata to metadata files on storage media can attenuate this problem.
- Structural relations between content units can, in some cases, be a matter of authenticity. Under these circumstances, a repository architecture needs to warrant a trustworthy recording of these relations.

6. Landesarchiv Baden-Württemberg Metadata Elements for Mixed Digital Content

6.1 Core metadata – general

Core metadata included on every descriptive level (Object, Representation, File, Documentation). 2 user-defined elements mandatory (U/M), 2 optional (U/O), 11 defined by system (S).

General

Archival ID (S)	System-generated ID of content unit
Parent Archival ID (S)	ID of parent content unit
Reference number detail (S)	Detail of reference number for actual descriptive level
Description (U/O)	
Type (S)	Descriptive level (structure, object, representation, file)
Status (S)	Under preparation; complete; withdrawn
Ingesting person (S)	
Ingest date (S)	
Manipulating person (S)	
Manipulation date (S)	
Version number (S)	Highest is most recent
XMLVersion (S)	YYYY-MM-DD
Quirks (U/O)	Any technical or intellectual deficiency resulting from features of source data.

6.2 Core metadata – descriptive levels

Core metadata for descriptive levels. 11 user-defined elements mandatory (U/M), 14 optional (U/O), 8 defined by system (S).

Structure

Examples: archives, series, subseries, finding aid

Title (U/M)

Digital Object

Intellectual entity, nested if necessary

Title (U/M)

Creation time (U/M) When did content originate?

Documented time (U/O) What time range does object cover?

Provenance (U/M) Institution at which content originated. Archival term, mapping to dc:Creator, not related with dc:Provenance.

Transferring institution (U/O) If different from provenance.

Transfer (U/O) Date of accession to archive, people involved.

Content type (U/M) Examples: photographs, GIS data, statistical primary data

Significant properties (U/O) See premis:SignificantProperties

10 One for Many – A Metadata Concept

End of closure (U/M)	Year in which record closure for the public ends
Use restrictions (U/O)	Further use restrictions
Rights (U/O)	Copyright terms
Paper parts reference (U/O)	Reference number of paper-based parts of a hybrid object.
Paper documentation reference (U/O)	Reference of paper-based metadata.
Representation	Folder containing all the files necessary for rendition of digital object.
Title (U/M)	
Structure (U/O)	May contain plain text, SQL, HTML; see premis:Relationship
Hardware environment (U/O)	See premis:Environment
Software environment (U/O)	See premis:Environment
Installation requirements (U/O)	Requirements other than hardware and software
Parent representation (U/O)	Which representation was source of this representation?
Content file	
Original file name (S)	Filename at time of ingest
Filename (S)	Filename on storage media
File format (U/M)	Multiple choice list of approved formats
File format version (U/O)	
Character encoding (U/M)	Multiple choice list of approved formats
File size (S)	File size in byte units
Documentation file	Description necessary for rendition of primary content files.
Title (U/M)	
Original file name (S)	Filename on ingest
File name (S)	Filename on storage media
File format (S)	Multiple choice list of approved formats
File format version (U/M)	Mandatory if several versions exist
Character encoding (S)	Multiple choice list of approved formats
File size (S)	
6.3 Specific metadata (data tables, raster graphics)	
3 user-defined elements mandatory (U/M), 11 optional (U/O).	
Table	
Number of fields (columns) (U/M)	
Number of records (rows) (U/M)	Field headers don't count
Field	
Name (U/M)	
Description (U/O)	
Data type (U/O)	
Length (U/O)	
Encodings (U/O)	Encoding schemes (e.g. YYMM), codelists
Relationships (U/O)	Verbal description (e.g. 1-n relation with field X in table Y)
Remarks (U/O)	
Codelist	
Name (U/O)	
Code (U/O)	
Value (U/O)	
Raster graphics	
Compression (U/O)	Compression algorithm
Digitisation date (U/O)	If applicable

6.4 Process and fixity metadata

One user-defined element mandatory (U/M), one optional, 4 defined by system (S).

Process metadata

Ending date (S)	Date and time at which process ended
Recording agent (S)	Person initiating or recording process
Processed unit (S)	
Process type (U/M)	Multiple choice list
Details of process (U/O)	Agents, causes for action, hardware, software, regulations

Fixity metadata

Checksum (S)	Checksum by md5-algorithm, saved in a separate file (foo.txt --> foo.txt.md5).
--------------	--

References

- [website]BAM (2008). *Portal zu Bibliotheken, Archiven, Museen*, retrieved 2008-01-15 from <http://www.bam-portal.de>
- [report]Borghoff et al. (2005). *Comparison of Existing Archival Systems. Edited by nestor. Network of Expertise in Long-Term Storage of Digital Resources* (English summary) retrieved 2008-01-16 from http://www.langzeitarchivierung.de/downloads/mat/03_summary.pdf
- [website]DROID. *Digital Record Object Identification*, retrieved 2008-10-14 from <http://droid.sourceforge.net/>
- [standard document]EAD (2002). *Encoded Archival Description Standard*, retrieved 2008-01-15 from <http://www.loc.gov/ead/>
- [report]Fedora (2007). *Fedora Digital Object Relationships*, retrieved 2008-01-15 from <http://fedora.info/download/2.2.1/userdocs/digitalobjects/introRelsExt.html>
- [website]JHOVE. *JSTOR/Harvard Object Validation Environment*, retrieved 2008-10-14 from <http://hul.harvard.edu/jhove/>
- [standard document]LMER. *Long-term preservation Metadata for Electronic Resources*, V. 1.2 (7. April 2005), [urn:nbn:de:1111-2005051906](http://nbn:de:1111-2005051906)
- [report]McDonough (2008). Structural Metadata and the Social Limitation of Interoperability: A Sociotechnical View of XML and Digital Library Standards Development, in: *Balisage: The Markup Conference. Proceedings 2008*, retrieved 2008-10-14 from <http://balisage.net/Proceedings/print/2008/McDonough01/Balisage2008-McDonough01.html>
- [website]MICHAELplus (2008). *Multilingual inventory of Cultural Heritage in Europe*, retrieved 2008-01-15 from <http://www.michael-culture.eu>
- [standard document]NLA (1999). *Preservation Metadata for Digital Collections*, retrieved 2008-10-31 from <http://www.nla.gov.au/preserve/pmeta.html>
- [standard document]NLNZ (2003). *National Library of New Zealand Metadata Standard Framework*, retrieved 2008-01-15 from <http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-standards-framework/view>

- [standard document]METS (2006). *Metadata Encoding and Transmission Standard 1.6*, retrieved 2008-01-15 from <http://www.loc.gov/standards/mets/>
- [standard document]OAIS (2002). *Open Archival Information System*, retrieved 2008-01-15 from <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [standard document]PREMIS (2005). *Preservation Metadata Implementation Strategies Data Dictionary*, retrieved 2008-01-15 from <http://www.loc.gov/standards/premis/>
- [report]Severiens/Hilf (2006). *Langzeitarchivierung von Rohdaten (Long-term Archiving of Scientific Primary Data)*, retrieved 2008-02-15 from <http://nbn-resolving.de/urn:nbn:de:0008-20051114018>
- [standard document]SOAP (2007). *SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)*, retrieved 2008-01-15 from <http://www.w3.org/TR/soap12-part1/>
- [report]Woods/Brown (2008). Creating Virtual CD-ROM Collections, in: *Proceedings of the Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, p. 62-69.