The International Journal of Digital Curation
Issue 2, Volume 4 | 2009

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

Abstract

The Landesarchiv (State Archive) of Baden-Württemberg has designed and implemented a metadata concept for digital content covering a heterogenous range of digital-born and digitised material. Special attention was given to matters of authenticity and to economic ingest and dissemination methods in line with the requirements of a public archive. This paper describes the outcome of discussions on metadata during the implementation period of the DIMAG repository. It addresses 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].

[1] This paper is based on the paper given by the authors at the 4th International Digital Curation Conference, December 2008; received July 2008, published October 2009.
The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Stating Requirements

The Landesarchiv (State Archive) of Baden-Württemberg is holding records from the Middle Ages to the present day at seven locations throughout the State of Baden-Württemberg. Archivists are used to working 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 (the authors of this paper) started work at the Ludwigsburg branch of the Archive. They constructed the DIMAG system, based on a LAMP (Linux, Apache, MySQL, PHP) web server architecture which provides controlled storage of digital objects and metadata.

By the end of 2007, the Landesarchiv had 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 display over the Web and, if necessary, for long-term preservation.

It is the diversity of these objects which represents the key challenge in devising a metadata concept to describe, preserve and distribute them. They all need to be located on the existing finding aid system, regardless of their media format. Logically, this system can be described as a strictly hierarchical classification tree with branches representing depositing institutions, its twigs reflecting series and sub-series, and its 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[2] and MICHAELplus[3] has already been implemented; we anticipate further participation in German and European digital library projects.

[2] Bibliotheken, Archiven, Museen (BAM) http://www.bam-portal.de
[3] Multilingual Inventory of Cultural Heritage in Europe (MICHAELplus) http://www.michael-culture.eu

Establishing Principles

The path to a solution began with a study of functional and data models[4] [5] [6] (National Library of Australia [NLA], 1999; Consultative Committee for Space Data Systems [CCSDS], 2002; National Library of New Zealand [NLNZ], 2003; Die Deutsche Bibliothek, 2005). 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 were to be destined for sharing with preservation systems outside the Landesarchiv, standard-compliant AIP (Archival Information Package) design would have to adapt to future schemas, not to the current ones. For example, it would be useless to establish a METS-compliant schema for content if these metadata were, sooner or later, to require partial re-structuring. Current international discussion (McDonough, 2008) seems to confirm this point.
- Even though the Landesarchiv was already sharing most of its finding aid metadata with other memory institutions at 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 and others.

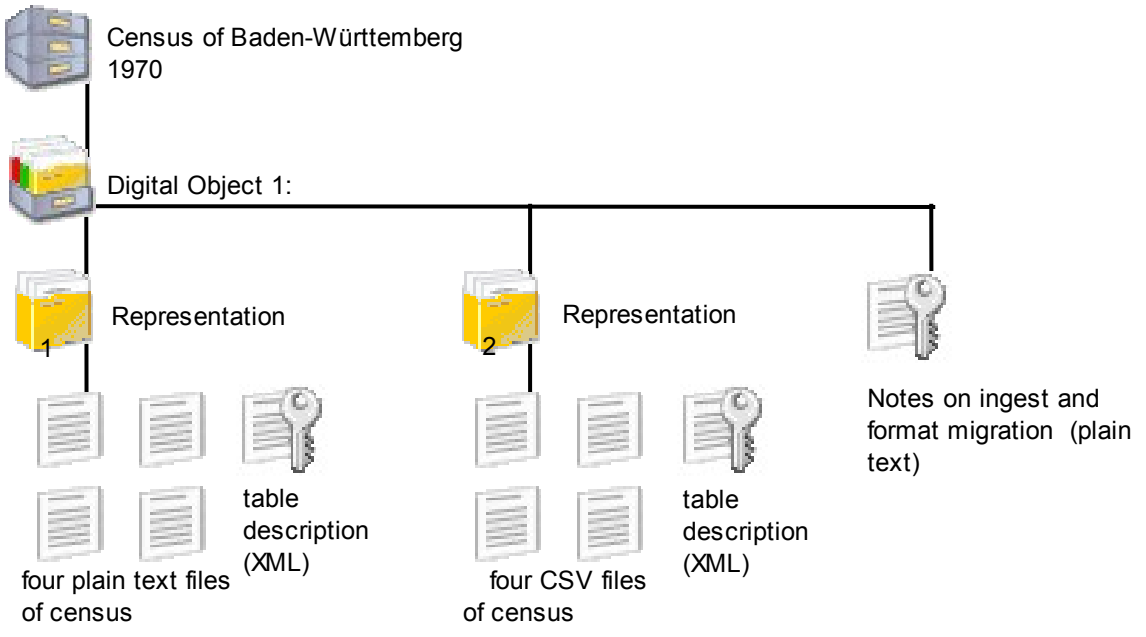

Figure 1. Logical structure of a digital object with 2 representations, 8 content files, 3 documentation files.

[4] EAD: Encoded Archival Description Version 2002 Official Site (EAD Official Site, Library of Congress) http://www.loc.gov/ead/
[5] METS: Metadata Encoding and Transmission Standard 1.6. http://www.loc.gov/standards/mets/
[6] PREMIS: Preservation Metadata Implementation Strategies Data Dictionary http://www.loc.gov/standards/premis/

The Representation Experience

Existing standards served as a source for functional requirements. The concepts of Representation and Significant Properties were derived from PREMIS. 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 to implementation than the intricate "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 was defined in Significant Properties. The totality of such 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 Landesarchiv, thus ensuring preservation of digitised and digital-born material alike.

PREMIS was also the inspiration for the elements added for the purposes of emulation (see Table 2, Representation category). At the moment, the Landesarchiv does not apply this strategy to the preservation of its objects. Nevertheless, it has decided to retain the first representation of an intellectual entity in perpetuity, in order to anticipate the possibility that emulation may prove to be effective for certain formats in the future.

The National Library of Australia (NLA) contributed another metadata element which it 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 Table 1).

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, pp. 4-17).

In order to close this gap, the team decided to store vital metadata redundantly 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 Tables 1 and 2) from XML files. This means that the functions of DM are divided: the database component only handles retrieval of metadata, while the storage of metadata is entrusted to the storage component that also holds the content.

Metadata files had to be linked with content files on storage media in an easy, robust and efficient way. Therefore, metadata, fixity metadata, and content files are all given the same base name (Figure 2 below). The decision to provide stable storage for metadata conflicted, of course, with the need to amend metadata regularly. It therefore caused difficulties with synchronisation: altering some letters in metadata stored with content on a Write Once Read Multiple (WORM) media could impose a re-write of several gigabytes.

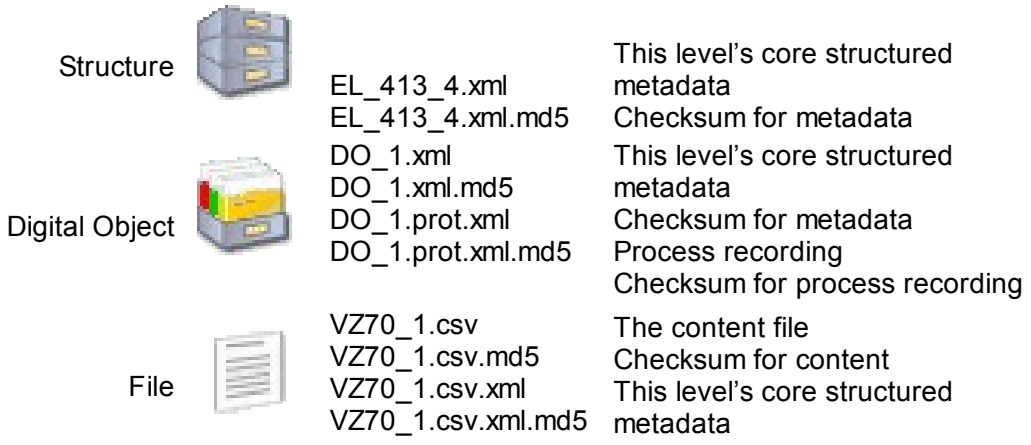

Figure 2. Structural components of DIMAG and their representing files (simplified view).

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 is enhanced with metadata and packaged for long-term storage and can be fetched for format migration or dissemination. Most of our metadata are located in files on this level, and is continuously sychronised with the database. The completely packaged representation containers are stored on the locked storage level (comprising WORM media) and make up the bulk of the content. Attached to them is a subset of technical metadata (see Table 2, Representation and Content File categories) which can only be altered through versioning or migration of the whole representation. The management database keeps track of the physical location of every file while regularly writing backups of this information to support disaster recovery.

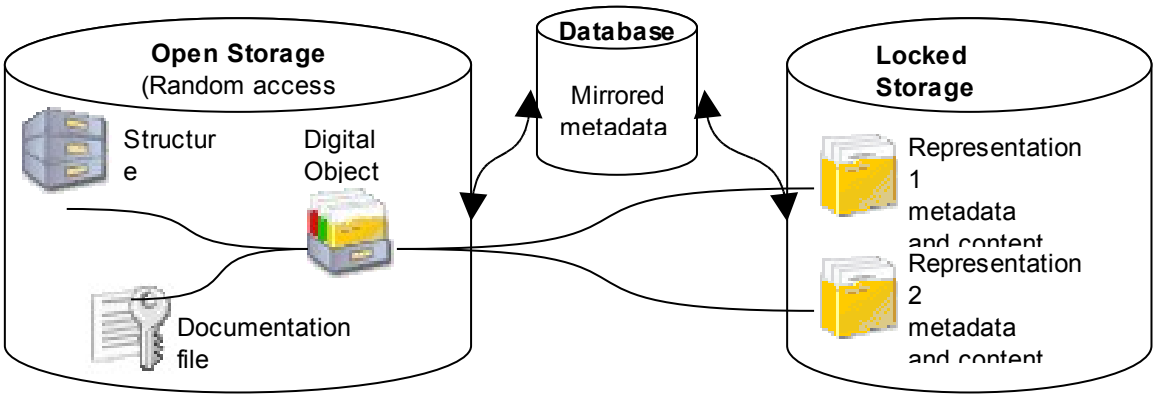

Figure 3. Storage of metadata files and packaged content in a hierarchical storage scheme.

Another challenge was to find a location for integrity metadata (checksum values) of stored metadata files. Placing the calculated value inside the XML would both change the XML and alter the checksum. DIMAG provides a simple remedy by writing the checksum values into a file assigned to the target file by duplicating the target filename and adding an extension (see Figure 2 above). 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 as well as on demand.

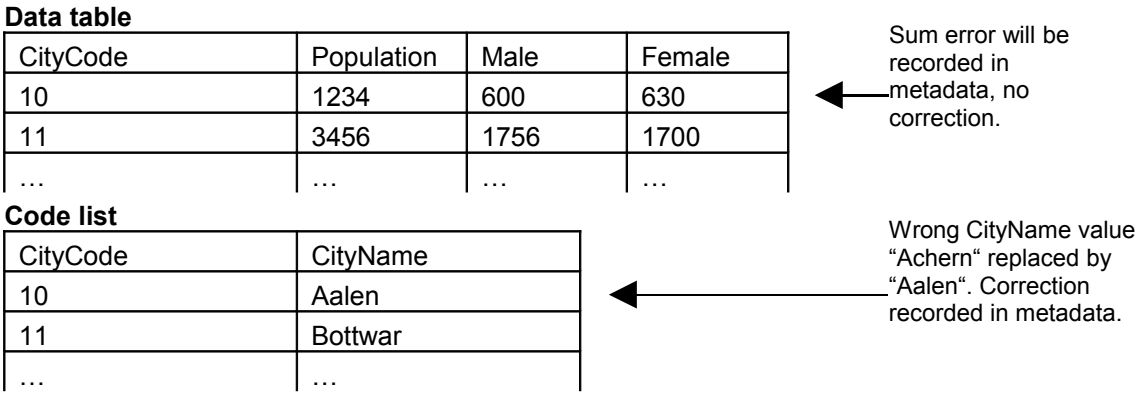

Figure 4. Example of distinction between metadata and content. These primary data consist of a data table (content) and a code list (metadata).

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 turn out to be wrong, they can be manipulated in a controlled manner, whereas content must be preserved without change. Content is sacred, metadata are free. The decision as to 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, provided corrections were properly recorded. Errors found in the microdata themselves can only be recorded, not corrected.

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 in archives often resemble complex molecular structures that lose their character when rearranged. In other words: for this type of content, preserving structure represents 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 largely unsatisfactory (Borghoff, 2005, p. 7). Larger objects with an internal sub-structure are not served by current systems (Woods & Brown, 2008, p. 68). Encoded Archival Description (EAD) is a strong tool to describe these relations, since it was conceived for archival finding aids, but it 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.

Constructing Viable Workflows

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

Vital Process Recording

Archives which seek to assert that no information has been 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 to the archive's staff through their 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. These 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 employed. Others can be recorded manually if necessary. 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.

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 will not expect an archive to retain 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* are recorded in parsed XML and simultaneously in the management database. They are 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 the 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* require 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 types.
- *Integrated Metadata* are part of ingested files. For reasons of cost, they are mostly left inside the files and only extracted if necessary. Only some values are extracted automatically via the JHOVE[7] and DROID[8] Java libraries and

[7] JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/
[8] Digital Record Object Identification http://droid.sourceforge.net/

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, authors' names in MS 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 Table 2, Documentation category). 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 cannot 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.

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. scopeArchiv will create representation folders on DIMAG on request which can be charged 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 talk to one another. Unique identifiers with a namespace prefix, assigned to metadata as well as content, will play a key role. They will also support exchange of data packages with future applications inside and outside the Landesarchiv for transfer, ingest, migration, and use.

Figure 5. A prototype all-purpose DIP format showing catalogue context ("Bestands- und Findbuchkontext") and the internal structure of the requested census primary data ("Bestellte Einheit").

However, these identifier codes will most likely 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 has yet to focus 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.

Conclusions: Points for Discussion

There are some findings of the project group that could be discussed on a broader scale.

Concepts of 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 aids and structured metadata levels, leaving additional information on a non-standardised level.

Long-term archiving is largely based on interoperability of past, present, and future systems, policies and concepts. Persistent identification is a key asset for the resulting interchange operations. However, internal identifiers and reference codes for the public should be viewed separately, though.

Repository owners often deploy XML-based standards in order to guarantee interoperability in content sharing. What actually exists, though, are local profiles or schemas based on these standards, and sharing between repositories still presents difficulties (McDonough, 2008). Paradoxically, repository developers who have to deal with heterogenous content might save 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 they on statistical primary data, office documents, or digitised journals.

Repository systems often fail to 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 relationships between content units can, in some cases, represent a matter of authenticity. Under such circumstances, a repository architecture must be able to guarantee a reliable recording of those relationships.

Landesarchiv Baden-Württemberg Metadata Elements for Mixed Digital Content

Field Name	Description
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.

Table 1. Core Metadata: General. These are 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).

Field Category / Name	Description
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 was content created?
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 to 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
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 the 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)	

Table 2. Core metadata for descriptive levels. 11 user-defined elements mandatory (U/M), 14 optional (U/O), 8 defined by system (S).

Field Category / Name	**Description**
Table	
Number of fields (columns) (U/M)	
Number of records (rows) (U/M)	Field headers do not 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

Table 3. Specific Metadata (Data Tables, Raster Graphics). 3 user-defined elements mandatory (U/M), 11 optional (U/O).

Field Category / Name	Description
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).

Table 4. Process and Fixity Metadata. One user-defined element mandatory (U/M), one optional, 4 defined by system (S).

References

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 January 16, 2008, from http://www.langzeitarchivierung.de/downloads/mat/03_summary.pdf

Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS)*. Retrieved January 15, 2008, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Fedora. (2007). *Fedora digital object relationships*. Retrieved January 15, 2008, from http://fedora.info/download/2.2.1/userdocs/digitalobjects/introRelsExt.html

Die Deutsche Bibliothek. (2005). *Long-term preservation metadata for electronic resources, v. 1.2.* Tobias Steinke (Ed.). April 7, 2005. urn:nbn:de:1111-2005051906 Retrieved September 15, 2009, from http://www.d-nb.de/standards/pdf/lmer12_e.pdf

McDonough (2008). Structural metadata and the social limitation of interoperability: A sociotechnical view of XML and digital library standards development, in *Proceedings of Balisage: The Markup Conference, 2008.* Retrieved October 14, 2008, from http://balisage.net/Proceedings/print/2008/McDonough01/Balisage2008-McDonough01.html

National Library of Australia. (1999). *Preservation metadata for digital collections*. Retrieved October 31, 2008, from the National Library of Australia Web site: http://www.nla.gov.au/preserve/pmeta.html

National Library of New Zealand. (2003). *National Library of New Zealand Metadata Standard Framework*. Retrieved Januray 15, 2008, from National Library of New Zealand Web site: http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-standards-framework/view

Severiens, T., & Hilf, E.R. (2006). *Langzeitarchivierung von Rohdaten (Long-term archiving of scientific primary data.* Rretrieved February 15, 2008, from http://nbn-resolving.de/urn:nbn:de:0008-20051114018

Woods, K., & Brown, G. (2008). Creating virtual CD-ROM collections. In *Proceedings of the Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, pp. 62-69.