

# Archival metadata for durable data sets

E.H. Dürr  
Euformatics  
Scriveriusmate 28  
8014 JW Zwolle  
The Netherlands  
E.H.Durr@phys.uu.nl

R. Dekker  
Delft University of Technol.  
Library  
P.O. Box 98  
2600 MG Delft  
The Netherlands  
R.Dekker@TUDelft.nl

K. van der Meer  
Delft University of Technol.  
Computer Science  
P.O. Box 5031  
2600 GA Delft  
The Netherlands  
K.vanderMeer@TUDelft.nl

## ABSTRACT

The DARELUX project envisages long-term storage of hydrology measurement data in a permanent archive. In order to make the DARELUX hydrology data sets accessible in an unknown future, a future-proof metadata element set is wanted. An analysis of current standards and best practices for metadata lead to the choice of the metadata of the Dublin Core Metadata Element Set, with an addition of archival metadata elements, based on the OAIS Preservation Description Information (PDI) metadata.

A second problem is related to the encapsulation of an OAIS AIP package into an XML container file. This is the best we can do to let containers with data travel through time. Metadata and content are entangled together as an indivisible unit. The problem is that the PDI fixity/checksum value stores the value of the entire XML container, inside of the preservation metadata section of the container. However, storage of the value of the container means that the value of the container changes. Our solution to this, a procedure to store a valid checksum value of the XML container in the XML container is described.

A third problem is the geographical coverage. The Dublin Core prescription for the geographical representation is ambiguous; a choice was made for an encoding scheme.

## Categories and Subject Descriptors

H.3.6 [Information storage and retrieval]: Library automation - *large text archives*

## General terms

Documentation

## Keywords

Metadata, Data sets, Longevity, OAIS, Preservation Description Information, Dublin Core, fixity, MD5

## 1. PRESERVATION OF HYDROLOGY DATA SETS

The DARELUX (Data Archiving River Environment Luxemburg) project concentrates on preservation of measurement data sets of precipitation (rain, snow), water levels in currents, water discharge etc.: hydrology data sets (1).

Hydrology data sets like these are needed for models of water management, and they will continue to be needed for future models; water management is an important issue for the Low Countries. To make certain the usability of these data in the future, long-term storage of the data is crucial.

Based upon our experience with archiving digital information objects in a durable way, we are currently building the DARELUX repository; for the present state see (2).

There is a growing experience with preservation of digital publications. Preservation of data sets is in some aspects different. Data sets are often deleted after their use, but there are data sets that have to be preserved, as they will remain valuable a long time after they have been generated, and they are unique, they cannot be reconstructed once they would be lost. Hydrology data sets are an example of this type of data sets.

Especially for their use in an unknown future, it is essential to make the data sets retrievable. Without the prospect of future retrieval and access, preservation efforts simply do not make sense. The demand for future retrieval and access leads to the demand of a future-proof metadata element set. Determination of a suitable metadata element set, then, is the first problem.

## 2. METADATA ELEMENT SETS

The DARELUX project is one of the SURF / DARE projects. DARE puts conditions to the metadata accompanying the data sets: the DARE projects have been prescribed to use the Dublin Core Metadata Element Set, "DC", ISO standard 15836:2003 (3). Even more, DARE Guidelines to the DC have been issued (4), prescribing how to use the DC MES.

The prescription of a metadata element set and the edition of Guidelines for its actual use should be applauded. The use of standardised metadata assignment is the best we can do to make digital information objects retrievable. This is even more valid for future retrieval. However, the DC is not without drawbacks.

The DC is meant to enable retrieval and access to digital publications. Although the website (3) states, that there are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned, the DC is not tailored to the description of data sets. Also, the DC offers no built-in support for, e.g., authenticity and provenance, i.e. archival

aspects needed to determine the value of the content in a distant future.

The fact that DARELUX deals with data sets could be overcome, but the archival aspect should be adapted. The two options for the DARELUX project are:

- Accept some other metadata element set, which gives rise to a difficult explanation to the project board that their conditions cannot be met
- Add metadata elements originating from records management or archival practice to the DC.

### 3. ARCHIVAL METADATA

#### 3.1 Archival metadata element sets

There are several well-known archival metadata element sets.

1. ISAD(G) (General International Standard Archival Description, second edition, (5)) is an international description standard for cataloguing archival materials. It is based on older, national standards. It is used a lot and has a good reputation for the description of these sources.

2. EAD, (Encoded Archival Description, (6)) is a standard for encoding archival finding aids using XML. EAD is used a lot and a valuable tool, too.

3. The ISO standard for metadata for records ISO 23081 is a recent development (7). ISO 23081 is connected to the ISO standard for records management, ISO 15489 (8), and ISO 15489 is a most important standard for records management (it stresses, in a way, the same need for records management as the famous or infamous US Sarbanes-Oxley law (9)). Records management is not the same as a durable archive, but as the two are closely related, ISO 23081 must be considered, too.

4. The OAIS (Reference Model for an Open Archival Information System, ISO standard 14721:2002, (10)) metadata set. This set includes Preservation Description Information, PDI metadata elements. The OAIS PDI is a limited set that describes the use of four metadata elements: Reference information, Provenance information, Context information and Fixity information. Fixity information is described as "documents authentication mechanisms used to ensure that the Content information has not been altered in an undocumented manner (e.g. checksum, digital signature)". OAIS has the advantage that the developers of the OAIS metadata element set obviously included in their requirements that data sets would have to be described.

Due to the specific material, data sets, OAIS representation information is probably not needed.

González et al. compare these metadata element sets (for another divergent data type, viz. software components) based on granularity, suitability and compatibility for the data type, and simplicity for users (11). That is a basis to judge the DARELUX case, too.

#### 3.2 Combination

If one metadata element set, with its own purpose and 'designated community', will not do, metadata element sets can be combined.

There are various crosswalks that can be used as start for a combination; see e.g. Day (12).

Additions to and combinations with DC have been proposed several times, especially before Qualified Dublin Core with its refinement had been established. For the data type of software components, González (11) added archival elements to the DC. Another example of DC extension to non-document-type information objects is the proposal of Bird for an extensible XDC scheme for a type of language vocabularies (13).

### 3.3 Experiences

Searle and Thompson describe a pragmatic approach at the National Library of New Zealand (14). Referring to experiences at the national Library of Australia, Cedars, the OCLC/RLG Working Group, it is emphasized that a balance should be found between the principles expressed in the OAIS Information Model and the practicalities of implementing a working set of preservation metadata. Unfortunately, there is no recipe for preservation metadata assignment.

The Dutch "Koninklijke Bibliotheek", the Dutch National Library, preserves digital publications in its Digital Information Archiving System (DIAS). The experiences at the DIAS are useful, despite that the DIAS is not meant for data sets and the DIAS is far larger than the DARELUX data base. DIAS is one of the largest repositories in the world with at present over 4 million publications. It is based upon Dublin Core-like XML-based metadata and compatible with the OAIS reference model. The Koninklijke Bibliotheek is considering the (rather extended) PREMIS model to preserve the publications in her DIAS. PREMIS (ref. 15, page 4-5: fixity, integrity, authenticity) states that these characteristics of a [digital] object have to be verified, but again the recipe is still under construction. The PREMIS working group does not seem to be operational yet.

In order to ensure encapsulation the use of digital containers as the basis of OAIS AIP's (Archival Information Packages) is defended. The use of XML containers stems from a previous project called EArchive, where the unity of metadata and content was developed (16). In the AIP's, the containers contain both the metadata and all representations of the digital archive information object.

Evidently, there is not much experience yet with metadata element sets for preserved data sets.

## 4. DESIGN

### 4.1 Choice of the element set

The demand for simplicity and flexibility (criteria also used by González) of DC, compared to ISAD(G) and EAD, would induce a choice for DC even if it had not been prescribed. ISAD(G) is broad and general and cannot be tailored easily to properties of types of documents. ISAD(G) does not primarily seem to think of retrieval. ISAD(G) and EAD both have more to offer for archival aspects than DC but it is a bit overdone. By the way, Boudrez (17) states that ISAD(G) is of limited use for digital archive documents, and the DARELUX data sets are surely digital.

An advantage of the DC is the mass of users in the present "designated community" for the DARELUX data sets. Also in that respect ISAD(G) and EAD are not superior to the DC.

There is overlap between the ISAD(G) and DC and the EAD and DC metadata element sets. Upon inspection it is clear that

their structure is different from the DC structure. It is not easy to make a consistent set of ISAD(G) or EAD combined with DC.

A disadvantage to the ISO standard 23081 for metadata for records is that only Guidelines to the metadata for records standard have been published, there is no accepted metadata element set that is ready to be used or can be tailored for this purpose and that is broadly accepted.

The OAIS metadata structure has the advantage, that the OAIS PDI metadata types are a suitable complement to the DC: simple, concise, no overlap, just an addition.

Next to the conceptual connection, cost aspects are a factor. For the preservation of hydrology data sets, DARELUX must aim at minimal effort to make the data sets accessible. Hydrology data sets do not rank high in the list of cultural heritage materials. So, the keeping of this type of raw material is faced with a second problem, a financial one, leading to a second (less important, but not to be neglected) selection criterion.

Probably, the extent of external financial support is restricted; it is not even sure whether financial support from a government body will be granted, although we (of course) feel that the need to preserve these hydrology data sets is beyond doubt. In this respect, the situation for hydrology data sets may be different from, e.g., archeological data sets and is surely different from normal archive documents.

ISAD(G) and EAD describe extensive tag libraries, their implementation would require a lot of effort. The costs to assign the small set of OAIS PDI metadata are relatively low.

So, the combination DC + OAIS PDI was chosen.

Finally: it has been brought to our attention that the DC has been extended with (among others) a provenance element. It is not part of the Simple DC, and different from the refinement and encoding scheme that is the base of the Qualified DC. We cannot judge the acceptance among users of this element; that is essential for the DARELUX users. So, we can not choose for DC alone. But the use of just one set would obviously be preferred over two; and if the DC would be extended with a dedicated, recognizable, concise set of archival metadata elements (and why not the current OAIS PDI), that could well be preferred.

## 4.2 Structure for long-term preservation

In the DARELUX project, metadata and content are stored together in self-descriptive containers in XML format. An XML container, an indivisible unit in which metadata and content are stored together, has a good chance to travel unchanged through time. It continues to be the base for automated processing. Relying on any linking mechanism (e.g. via a data base) between metadata and content requires that archiving organisations are obliged to maintain the technical linking infrastructure (data base software) over a very long time. A linking mechanism is a considerable risk for long-term preservation. That is the background of the original idea of Lourens et al. (16). Boudrez (17) evidently follows the same reasoning.

The purpose of OAIS is an archival information system. In the DARELUX repository, or in any repository for that sake, digital information objects have to be able to travel through time. Therefore, the main focus is the longevity of digital information objects. The solution that is being used to manage these digital objects (e.g. databases or linking mechanisms) is subordinate to accomplishing that goal. As the purpose of the DARELUX

project is different from the OAIS system, the way of working is different.

## 4.3 Fixity

The OAIS PDI metadata structure is suitable as an addition to the DC for the DARELUX repository.

As stated, OAIS PDI describes four metadata element types. Of these four, the fixity item poses an extra problem. The metadata is stored in an XML container together with the bit stream. So, the fixity information is stored in the container.

The fixity describes the value of an information object, e.g. by stating the calculated checksum.

In DARELUX containers, there are several possibilities.

One could calculate the checksum on the content part only, but then the metadata could be changed; that is not to be preferred.

One could calculate the checksum of the complete container, but then a problem arises. When the value of the container is calculated, and filled in in the fixity field (in the container), the value of the container changes.

One could calculate the value of the checksum for the content and the metadata separately. This again leads to the problem that the value, in this case of the metadata, changes upon completion of the field.

We propose the next, probably unambiguous way to deal with this problem. The checksum tag will be completed before calculation with XXXXXXXXXXXXX (12 times X). This XXXXXXXXXXXXX will most likely draw attention. It is evidentially not a hexadecimal value, so one kind of misunderstanding is omitted. And we hope that mistakes with a value '0' by a future data archeologist are avoided. Next, the checksum is calculated and the checksum value is stored in the checksum tag. Control of the checksum value in the future has to follow the same procedure: the checksum tag has to be saved, overwritten by XXXXXXXXXXXXX, and then the checksum value can be calculated (and the value has to correspond to the original value in the checksum tag).

This procedure has to be stored in the XML container, too, of course; in the fixity field.

OAIS prescribes a separation of the bit stream and the metadata. In our case, the bit stream and the metadata are not separated. This requires an explanation.

## 4.4 OAIS PDI overview

As a result, the following metadata element types are designed to be added to the DC:

- The OAIS Reference information field is omitted. Reference information is covered sufficiently in the DC metadata.

- The OAIS Provenance information contains e.g. information on the data set provenance and the equipment, and any data on refreshment, i.e. bit stream preservation and eventually on migration.

- The OAIS Context information contains any contextual data related to the measurement data

- The OAIS Fixity information contains several tags: the choice of the way of calculation, in the DARELUX case MD5; the recipe to calculate the fixity value (including XXXXXXXXXXXXX); and the MD5 value of the container. The MD5 algorithm is probably stable, it is widely used and well-known; moreover MD5 is as near a standard as can be for a subject in the field of internet information, as MD5 has been published as RFC 1321 (18). The publication of an RFC is the

best guarantee for the future use of a digital information structure.

## 4.5 DC geographic coverage

In the Dublin Core metadata set an item called <DCMI:Coverage> is included. This item is meant to include time period and spatial information on the geographic location for the document. In the DCMI part two options are offered: either a (geolocation) Point <DCMI:Point> or a geographic rectangular area <DCMI:Box>. For a point the following components are defined: east, north, elevation, units, zunits, projection and a name. For a box the components are northlimit, eastlimit, southlimit, westlimit, uplimit, downlimit, units, zunits, projection and a name.

Several encoding schemes are proposed in the documentation (19) like DCSV (semicolon separated), and XML either with sub-elements or with attributes. This document states: "Given the flexibility of XML many alternatives are possible. One possible form is: ...". Next, an example XML element definition with corresponding DTD is given.

In the published schema for the DC elements (20) currently the coverage item is defined as a generic element type. This is a string with optionally a language attribute. The consequence is that no sub-elements are allowed by this schema.

The consequence of this choice is that inside the DC metadata no uniform encoding scheme is defined (chosen) for geographical information items. All kinds of indexing schemes using DC metadata are difficult to develop as they have to deal with a wide variety of encoding schemes or deliver ambiguous results.

This issue may become very relevant in the near future where "mobile information retrieval" based upon the geographic (from GPS or Galileo systems) location of the user is needed. Examples are public transport, tourism, traffic information etc. Also in our hydrologic data sets geographic location of the sensors evidently plays a vital role. It is a major parameter in building models based upon these data sets now and in the future.

A choice is necessary. In the DARELUX project it was chosen to use an encoding with elements for the components and an attribute for the name. We had to introduce our own name space (dl) because the DC schema only allowed text strings here.

Like the example from the DCMI Box the result is:

```
<dl:spatial name="Maisbich" >
  <northlimit>49.8942</northlimit>
  <eastlimit>6.0506</eastlimit>
  <southlimit>49.8812</southlimit>
  <westlimit>6.0303</westlimit>
</dl:spatial>
```

## 4.6 The archival macro level

Finally, we find that data sets are different from more "normal" documents. A single data set consists of the measurements for one geographic location for a certain time interval (e.g. a month). The DC metadata for each of them is identical apart from the time period indication field. Indexing based on DC delivers many "hits". That may be unwanted.

The choice for metadata assignment may be analogous to the different levels of description of archives: micro (individual information items are considered), meso (folders are considered) and macro (complete collections are considered). More research

is needed to the question, whether new retrieval methods are needed in such situations, where "classic" indexing based on meta-data fields for content consisting of individual data sets are less suited, and how to use these methods for a digital repository.

## 5. CONCLUSIONS

In the DARELUX project metadata must be assigned to make data sets in the hydrology data repository retrievable and accessible, also in a distant future. The use of the Dublin Core Metadata Element Set was prescribed to make these data accessible, a Guideline was available, and Dublin Core is a valuable start, but it is not enough. In addition to the DC, metadata on archival aspects are necessary to make the DARELUX repository useable in the distant future. Moreover, due to budgetary constraints, metadata assignment must be cheap (a condition that may be applicable to other data sets as well).

Based on over a year of experiences with the DARELUX repository, our findings are:

1. The combination of DC plus OAIS Preservation Description Information metadata elements seems to be the best fit for retrieval of and access to the DARELUX hydrology data sets in the long term. This, by the way, is not strikingly different from the solution of González for software components.
2. The data set and its corresponding metadata are saved together in XML containers.
3. A conjuring trick enables to deal with the recording of the checksum value of the XML container inside of the container it describes.
4. An overview of the contents of the OAIS PDI elements has been given.
5. A choice was made as to the description of the geographical information by DC in the DCMI Point item.
6. More research is needed to the possibility for repositories to describe collections of data sets.

The current and foreseen implementation promise metadata that are sufficient for future use. It should enable a "data archeologist" in the far future to use the current DARELUX hydrology data sets.

*We acknowledge the valuable comments of an unknown referee.*

## 6. REFERENCES

All references have been checked 23<sup>rd</sup> February 2006

- [1] DARELUX Data Archiving River Environment LUXemburg <http://www.library.tudelft.nl/darelux/>
- [2] DARELUX archieftoegang (in Dutch) [http://www.library.tudelft.nl/darelux/3872/f\\_EN.html](http://www.library.tudelft.nl/darelux/3872/f_EN.html)
- [3] Dublin Core Metadata Element set. ISO standard 15836:2003. <http://dublincore.org/documents/dces/>
- [4] DARE use of Dublin Core version 2.0, December 2004 <http://www.surf.nl/download/DARE%20use%20of%20DC%20v.%202.0.pdf>
- [5] General International Standard Archival Description. Second edition, 1999 [http://www.ica.org/biblio/cds/isad\\_g\\_2e.pdf](http://www.ica.org/biblio/cds/isad_g_2e.pdf)
- [6] Encoded Archival Description, 2002 <http://www.loc.gov/ead/>

- [7] Metadata for records. ISO/Technical Standard 23081-1:2004.
- [8] Records management. ISO Standard 15489:2001.
- [9] R. Kahn & B.T. Blair: The Sarbanes-Oxley act: understanding the implications for information and records management. [http://www.bitpipe.com/detail/RES/1089741697\\_942.html](http://www.bitpipe.com/detail/RES/1089741697_942.html)
- [10] Reference model for an Open Archival Information System (OAIS). Also: ISO standard 14721:2002 <http://www.ccsds.org/documents/650x0b1.pdf>
- [11] R. González and K. van der Meer: Standard metadata applied to software retrieval. Journal of Information Science 30(4), (2004), 300-309.
- [12] M. Day: Metadata: mapping between metadata formats. <http://www.ukoln.ac.uk/metadata/interoperability/>
- [13] S. Bird: A simpler format for OLAC vocabularies and schemes. <http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0209&L=olac-implementers&D=1&F=&S=&P=192>
- [14] S. Searle and D. Thompson: Preservation Metadata. D-Lib Magazine 9(4), (April 2003) <http://www.dlib.org/dlib/april03/thompson/04thompson.html>
- [15] PREMIS <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>
- [16] W. Lourens and E. Dürr: Programs for Ever. ICEIS 2001 NDDL Workshop November 2001. <http://durr.dhs.org/=>earchive/publications>
- [17] F. Boudrez: Digitale archiefcontainers voor het digitaal archiefdepot (in Dutch) [http://www.expertisecentrumdavid.be/docs/digitale\\_containers.pdf](http://www.expertisecentrumdavid.be/docs/digitale_containers.pdf)
- [18] R. Rivest: The MD5 Message-digest algorithm <http://www.ietf.org/rfc/rfc1321.txt>
- [19] S. Cox: DCMI encoding scheme <http://dublincore.org/documents/dcmi-box/>
- [20] Schema for DC elements <http://dublincore.org/schemas/xmls/simpledc20021212.xsd>