

Archival metadata for durable data sets

E.H. Dürr, R. Dekker,
K. van der Meer

Hydrology data sets 1

((1))

DARELUX project (Delft Univ. Tech., Utrecht Univ.,
Gabriel Lippmann Inst., ASTA)

SURF project

Archiving of measurement data for longitudinal
research

- ❖ Amount of rain, snow
- ❖ Level of currents and rivers
- ❖ Speed of water discharge

Hydrology data sets 2

These hydrology data sets must be saved for the future as

1. They will be needed in the future for calculations on water management, possibly with new models (Water management is an important issue for the Low Countries)
2. If these data sets would be lost, the data sets cannot be reconstructed (they are unique)

Example

From http://devcms.library.tudelft.nl/axis/partscache/Q3_measured-200506-200506.xml:

```
<?xml version="1.0" encoding="UTF-8" ?>
- <dl:dataset xmlns:dl="http://dlnamespace/location"
  xmlns:dcmi="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/" >
- <dl:sensor name="Sensorname" sensorCode="Q3" sensorType="Sensortype"
  otherSensorAttributes="none" >
- <dl:datatype name="measured" >
  <dl:datatypecomment>none</dl:datatypecomment>
  <dl:datatypedescription date="iso-date" time="iso-timeUTC" unit1="kPa" unit2="kPa"
    unit3="C" unit4="C" variable1="Total pressure" variable2="Air pressure"
    variable3="Water Temperature" variable4="Air Temperature" />
  <dl:ms date="2005-06-01" time="00:00" values="99.452;99.092;11.4;9.1" />
  <dl:ms date="2005-06-01" time="00:10" values="99.454;99.086;11.3;9.2" />
  <dl:ms date="2005-06-01" time="00:20" values="99.463;99.104;11.4;9.3" />
  <dl:ms date="2005-06-01" time="00:30" values="99.46;99.104;11.4;9.3" />
  <dl:ms date="2005-06-01" time="00:40" values="99.451;99.095;11.3;9.2" />
  <dl:ms date="2005-06-01" time="00:50" values="99.435;99.086;11.3;9.1" />
  <dl:ms date="2005-06-01" time="01:00" values="99.435;99.081;11.1;9" />
&c., &c., &c. ...
```

1st problem: Dublin Core

((2))

SURF project, SURF money

The use of the Dublin Core Metadata Element Set (ISO standard 15836) was prescribed by SURF

Guideline to use Dublin Core was issued by SURF 

Dublin Core aims at 'resource discovery' of digital publications. But:

- ❖ DC is not tailored to data sets
- ❖ DC is not tailored to archival needs

A rather new problem, so 

Archival metadata element sets

((3.1))

- ❖ ISAD(G), General International Standard Archival Description (2nd edition)
- ❖ EAD, Encoded Archival Description
- ❖ Metadata for records - ISO standard 23081
- ❖ OAIS, Reference model for an Open Archival Information System, has a part on PDI, Preservation Description Information - ISO standard 14721

Merge DC with archival elements

((4.1))

Criteria for the archival metadata element sets

- ❖ Simplicity
- ❖ Flexibility
- ❖ Suitability
- ❖ ... ease of combination with Dublin Core (or ...)
- ❖ ... costs of assigning and processing metadata

Result for metadata element set

ISAD(G) and EAD are extended (expensive);
moreover structure is not suited for this purpose

Metadata for records has not yet a useable
metadata element set

OAIS Preservation Description Information (PDI)
(reference, provenance, context, fixity) suits best

! OAIS PDI to be added to Dublin Core

*By the way: this result (for ISAD(G), EAD and
OAIS) is comparable to that for software
components*

2nd problem: Fixity!

((4.3))

A data set and the metadata are stored in an XML container

Fixity: checksum, e.g. MD5 (RFC 1321)

Calculate the value of the container → insert the value in the fixity tag → you changed the value of the container!

According to relations: a rather new problem, so 

Fixity: solution

1. Take preservation of metadata serious (preferred)
2. Omit metadata from the authenticity demand

! DARELUX proposal: choice 1.

Do NOT write 0 in the fixity field: problem overlooked

- ! Procedure: Write XXXXXXXXXXXXX in the fixity field, calculate checksum, write checksum in fixity
- + Future data archeology: copy fixity value from fixity field, write XXXXXXXXXXXXX in fixity field, calculate checksum, compare

Other problems (3rd, 4th)

((4.5)) DC is ambiguous: DC Point and DC Box can both be applied

! For DARELUX, a choice was made

((4.6)) Retrieval of many (similar) measurement data sets all having neatly the same metadata; identical subject, only their dates and values are different. Archival 'meso' level ?!

! To be investigated

Wind up

For access to the DARELUX preservation of hydrology data sets

- ! Metadata element set: combine Dublin Core
- ! ...with (choice) OAIS PDI (four fields)
- ! Fixity – include the metadata
- ! DC Point vs. DC Box
- ! Archival meso level