The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Dataset Preservation for the Long Term: Results of the DareLux Project

Eugène Dürr,
Utrecht University and Euformatics b.v.,
Netherlands

Kees van der Meer,
Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, Netherlands

Wim Luxemburg,
Faculty of Civil Engineering and Geosciences,
Delft University of Technology, Netherlands

Ronald Dekker,
Delft University of Technology Library, Netherlands

July 2008

Abstract

The purpose of the DareLux (Data Archiving River Environment Luxembourg) Project was the preservation of unique and irreplaceable datasets, for which we chose hydrology data that will be required to be used in future climatic models. The results are: an operational archive built with XML containers, the OAI-PMH protocol and an architecture based upon web services. Major conclusions are: quality control on ingest is important; digital rights management demands attention; and cost aspects of ingest and retrieval cannot be underestimated. We propose a new paradigm for information retrieval of this type of dataset. We recommend research into visualisation tools for the search and retrieval of this type of dataset.

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.

Rationale

After years of research into long-term (digital) archiving of books, reports and other (scientific) publications, there now exists a far greater awareness of the value of the preservation of cultural and scientific heritage. At the national as well as the European level, projects have developed which have lead to the establishment of electronic archives, repositories or e-depots across a wide variety of institutions. For example, in the Netherlands the Royal Library (KB) has started an e-depot where all digital national publications are now stored and where long-term preservation is guaranteed.

For the scientific community, academia and others, it is however becoming clear that merely archiving publications is leaving out an important part of the scientific heritage: the data and the analytical models. In a speech at the Conference on Scientific Publishing in the European Research Area, European Commissioner for Information Society and Media Viviane Reding addressed the issue of open access to publicly funded research (2007). She gave eloquent expression to this growing awareness and we highlight here some of the most telling points in her address:

- Much of the discussion here has focussed on *scientific journals*. But there are other developments worthy of consideration such as new opportunities to make better use of *research data*.
- The use of scientific data through applications such as data mining as well as the option to combine journal articles with underlying research data are growing in importance in many scientific disciplines. These kinds of applications are generally believed to represent enormous potential for the future. Indeed, we can already see a trend towards *a continuum of scientific information space, from data to publications*.
- It is also important that *the scientific community becomes fully aware* of what is at stake. Data and publications from the past are relevant in all scientific disciplines, even if on occasions they only assume their full importance in the light of new knowledge. As a very straightforward instance: studies on climate change are highly dependent on observations carried out over centuries.
- At the same time we will fund *the deployment of modern infrastructures* that will allow researchers to store and share data resulting from their observations and experiments. Through the capacities programme - part of the 7th Framework programme for R&D - the Commission will devote some 50 million Euros in the coming 2 years in order to build up a top-level data infrastructure (Reding, 2007).

When we started the DareLux (Data Archiving River Environment Luxembourg) Project in October 2004 as a SURF project in the DARE group, we stated the following three major reasons for dataset preservation in academic research institutes:

- **Validation** of results published in articles.
- **Reuse** of measurement data by other disciplines in another context: the major sources of scientific progress.
- **Valorisation:** dataset collections are among the crown jewels of any academic institution. More than anything they represent the bedrock of high-quality research and often provide an edge in the competition for new funding.

In the project we wanted to explore both the technical and organisational aspects of the archiving of datasets. We discovered that hardly anything of substance had been published on dealing with the long-time preservation of datasets at that time. Moreover our goal was to provide an open archive system that could be used as a platform for the analysis and exchange of results and data among research partners in the research. A similar goal can be identified in Cedars[1], its UK corollary. A national library with a duty to maintain a deposit of information objects tends to lay emphasis on the heritage function, as is evident with the Nedlib Project[2].

The DareLux Project uses and builds on the results of the EArchive Project, successfully completed in 2002 (Dekker, van de Meer, & Dürr, 2003).

Project Objectives

The central aim of DareLux is the long-term preservation of measurement data, collected in a river basin in Luxembourg. In this DARE services project we intend to implement a solution for the persistent storage of technically complex scientific data. We opted for hydrological data as they combine spatially distributed information and continuous data gathered over a lengthy time span as well as unique information collected on a one-time basis. Consequently we were looking at the measurement collections and model input parameters, which will be incorporated into a long-term archive together with the results of analyses as well as associated publications. Within DareLux we handle data both for spatially distributed as well as longitudinal research: in hydrology – because of periodicity in the climate and frequently changing land use – various time series only have relevance over periods greater than a decade. For predictions of the effect of changes in climate or land use, the time series need a span of half a century or more.

During and after the duration of this type of research the scientists have to be confident of the usability of the data obtained and the analysis programs: long-term preservation of these measurement data is thus of utmost importance. However, its implementation is not as yet straighforward. Within this project we will seek to achieve the following objectives:

- An operational repository, complying with OAI-PMH standard, with models for ingest, retrieval and storage of measurement data.
- Evaluation by users during an extensive test period.
- A generic design for these kinds of repositories.
- Design and implementation of the support structure around the repository.
- A cost/benefit analysis as a start for an exploitation model.
- Use of open standards and open source software.

[1] Curl exemplars in digital archives: Cedars http://www.leeds.ac.uk/cedars/
[2] Networked European Deposit Library (NedLib) http://nedlib.kb.nl/

Experience with the Ingest Process

The ingest process for the hydrology datasets was developed together with the users. The delivery of a dataset can be split into several steps. Firstly users were asked for the description metadata via a web form.

Since such metadata are invariably the same for a given location, the values supplied on the first occasion were used as default values for all subsequent occasions. The only changing items were the time period and sometimes the names of those supplying the data. The preservation data was generated automatically via a small PHP script on the server. Then the user was asked to upload to the server the data file for each sensor, using some interactive Perl scripts. The format for these monthly data files for each sensor was previously agreed. The file consisted of plain text lines, with comments in the opening lines followed by the different magnitude and the units of the measured values. All the following lines were data lines with date, time and floating point values, separated by spaces.

On receipt of these files some 10 Perl scripts were employed to perform several transformations. Initially the standard values for the location and sensor data (secured earlier via a short email exchange with the scientists) were added in XML format. Then dates and times were changed into ISO format (yyyy-mm-dd and hh:mm). As a last step all the description and values were transformed into the XML dataset format as described above. Then input files for the individual sensors were combined into one final dataset for that month.

Finally the datasets were validated against the dataset schema. As a last step a container file was constructed with the description metadata and preservation metadata, together with the data itself as content. This container was also validated against its XML schema. These container files were placed in the document directory of the archive while the archive catalogue file was updated manually to make the new sets accessible.

It became clear over time that it was difficult to enforce all the agreements on the input files all of the time. Different locations, various and changing types of equipment as well as changes in members of staff obliged us to adapt the transformation scripts regularly.

We came to the conclusion that *stringent quality control* was necessary on the ingest side to ensure that the data in the archive remains both valuable and usable over a long period of time. As it must be possible to combine data from different time periods, for time series, all monthly sets must have the same structure. Assistance from the higher management of a department is necessary to maintain the required high quality of the archive: they should encourage their staff members and stress the importance of the input in the prescribed format.

We encountered the following caveats:

- **Dates and times:** It is surprising how many different ways exist to express these items. Normalisation into ISO standard 8601 is required.
- **Frequency:** Measurement intervals of 10 minutes to capture variations in river discharge are usually sufficient in hill-slope hydrology. Seconds or

less are not necessary. Too high a frequency leads to very large files for even just one month and tend to introduce all kinds of "noise" in the values.

- **Values:** Should be in mathematical floating point format, and not strings with prefix zeros, etc.
- **Units:** We should be using ISO units globally: degrees Celsius (not deci degrees) for temperature, not grams but Newtons or milli-Newtons for forces/weights (strain gauge sensors), no liters/hour but *dm3/sec*.
- **Physics:** Remove calculation and clear measurement errors (e.g. divide by zero) leading to discharges like 10^{-7}.
- **Accuracy:** No artificial accuracy like time in milliseconds for minute intervals, or seconds by tipping buckets with intervals in the order of 10 minutes.
- **Comments:** No comments within the lines containing the values. Otherwise they make the list of values difficult for the analysis software to process and virtually no one will spot them in the middle of a large data file. Common sense dictates they should be in the file headers.

In this case quality control is done manually and is thus expensive in terms of effort. Moreover, should incorrect values (errors) creep into even some of the monthly datasets, they make the entire archive virtually useless; moreover a tool like JHOVE[3] is unlikely to be of much use.

Dataset Storage as Instance of an EArchive

In the E-Archive Project, running from November 2000 till April 2002, we developed a generic solution for the storage of Archival Information Packages for the long term. An important consideration for us was the need to store metadata and one or more representations of the content together. This would reduce the likelihood of metadata becoming separated from the content over time. In linking mechanisms and database solutions, there is always the risk of "dangling" references. Especially in a scientific context with contents consisting of large sets of numbers, losing the contextual description in the metadata renders the use of such content virtually impossiblee.

The E-Archive Container Structure

We therefore concluded that one container file per document with sections for the description metadata, preservation metadata, viewer information and one or more representation of the contents represented the solution offering the greatest chances for the long-term survival of such information. The container file itself is an XML file, and, as such, self-descriptive. For the description metadata we used the subset of Dublin Core as advised by OAI-MPH[4] . Additional items were retrieved from the DCMI terms subset like geographical box and point (Cole, 2003; Cox, 2006a, 2006b). There is not yet a wide accepted standard for preservation metadata. There is however some guidance from OAIS. We included information about the archiving history of the document, a technical section with file sizes and checksums and a provenance section. Optionally data from for example, pictures, graphs or photos can be included in representation sections 2 to 4, similar to the original input. XML documents cannot be

[3] JHOVE - JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/
[4] SURF: DARE Use of Dublin Core. Version 1.0, October 2003 http://www.surffoundation.nl/Dare

embedded in each other, therefore the content is enclosed within CDATA brackets, so that original XML files with the *<?xml version=”1.0” encoding=”UTF-8”?>* header can be included as content too. Base64encoding is used for binary files, so that only UTF-8 characters appear in the resulting content part.

This provenance section and the content sections have the form:

Provenance	Contents
<provenance> <copyrightStatement>.....</copyrightStatement> <rightsWarning>.....</rightsWarning> <permittedActors>.....</permittedActors> <permittedByStatute>.....</permittedByStatute> <contractOrRightsHolders>..... </contractOrRightsHolders> <permittedByLicence>.....</permittedByLicence> <reasonForCreation>.....</reasonForCreation> <reasonForPreservation>Digital Original </reasonForPreservation> </provenance>	<ea:dataRepresentation> <ea:originalinputfiles base64encoded="no" outputfilename="$sdi.xml"outputtype="xml" tmpfilename="$sdi.tmp" > <![CDATA[<?xml version="1.0" encoding="UTF-8"?> <dl:dataset xmlns:dl="http://dlnamespace" ..... </dl:dataset>]]> </ea:originalinputfiles>

Table 1. Provenance and content sections

We used a meaningful format for the identifier of a document in the archive. The document identifier definition for the current implementation consists of *docId = 16 letters (only capitals) + 12 digits)*:

- CC country and IIII institute e.g. BTUD UBUX BUMX
- SSSS (sub-)collection and AA first two letters of last-name of author
- TTTT title: first four letters of the title and 9999 year
- 99 month 99 day of entrance and 99 sequence number per day
- As an example: *nlbtudphysduprog200412120001.xml*

The reference software set of the EArchive consists of the Java servlets which parse the containers and deliver the metadata or content as a result stream. The stream can be saved into a file if required. Furthermore there is the front-end access software which let the user select a document and a viewer. Upon request, the data section of a container is cached and further processed by a viewer program. The processed result is delivered in a DIP cache. Viewer programs can be XSLT transformations, pdf viewers or special programs for this type of content. By using viewer programs we made the archive format-independent. Any format can be used and thus subsequently retrieved. If a viewer is present the result can be processed and the original situation is emulated. The last component is a container assembly tool which can be used to combine the individual sections into one large XML file during the ingest process.

Datasets as Content

As it is based upon the previously described generic set-up of the xmlcontainer model, the storage of datasets did not require any special modification of the EArchive reference software. We made a special interpretation of the document identifier, because author and title entries are not so useful for a set of measurements. We used these 6 positions for the geographical location in a 99.99 for North and 99.99 for the East co-ordinate. As an example: *nlbtuddtmp4988060420040701.xml*

Our experience shows that the architecture for archiving documents was very flexible and fully reusable. Most of the work involved related to the creation of user interfaces for the retrieval process, adapted to the special nature of this kind of content.

Normalisation of the Datasets

In addition to ensuring the long-term availability of the datasets as contents in the archive, a few requirements have to be fulfilled in order to guarantee usability of the datasets over a lengthy period of time. Furthermore, all future users will expect that all data files in the archive have the same structure even when they originally stem from different sources. Because standardisation of data sets will require long and complex negotiations at an international level, we concluded that the maximum achievable goal for now was to define a framework in which many different ways of storing measurement values could be united. We called this Normalisation. Such a normalised framework is typical for certain scientific disciplines, in our case Hydrology. For other disciplines it can be defined in a similar way.

General Requirements on Datasets

The additional requirements for datasets are:

- **Application and Platform Independence**: In the future other currently unknown applications and platforms may show up. The data set should thus be readable without any special programs: The XML document standard is made for this purpose: readable with all kinds of basic tools, using only ASCII or UTF-8 encoding.
- **Self-descriptiveness:** A future user must be able to interpret the measurement values in a correct way. For each number in the set the magnitude it presents and the unit it is expressed in, is needed.
- **Homogeneity:** All datasets for one location for different periods must have the same structure. Otherwise the in the following section described projection search over several periods cannot be carried out and the partial results can not be combined.
- **Discrete Units:** The continuous stream of datasets over time has to be divided in reasonably sized subsets which can be stored and retrieved individually. We found that sizes of about 0.5 Megabyte per dataset are the easiest to handle. This meant in our hydrology environment one set per month with values for each 10 to 30 minutes. Files any larger than this are generally seen by users to take too long to download.

We used the concept of a *Measurement Space* to derive our normalised dataset.

The Measurement Space

We developed the concept of a measurement space as a n-dimensional space, in which every individual measurement is a point. The multidimensional space in which we place our measurements has the following axes (dimensions).

dimension 1+2: The area (location) in a DCMI box notation.

dimension 3: The year/month in which the measurements were made.

dimension 4: The study to which the measurements belong. There can be several on-going projects in a given area during a particular month. In this way they can share their data.

dimension 5: The sensor characterised by its name. The exact location is given in the point notation.

dimension 6: The Data Type. Because a sensor can produce different kinds of measurements (such as level or conductivity) it is necessary to distinguish which.

dimension 7: The measurements themselves. Each set consists of:

- a description of the magnitude and the units.
- a list with pairs: date/time and the sequence of values, separated by colons. (CSV format)(Cox & Ianella, 2006).

By using this multi-dimensional space, each point represents one measurement. Such a point is defined by a *7-tuple* of the co-ordinates along these dimensions as axes'.

These 7-tuples form the basis of our retrieval procedures. Larger sets of combinations of measurements can be found by taking a range instead of a fixed value for one or more elements in the 7-tuple. In this way it is possible to obtain equivalents of lines, planes and other shapes in the 7-dimensional space. Because a 7-dimensional space is not easy to perceive, we illustrate this approach in two pictures. One should visualise the right-hand picture as included in the star located in the left-hand picture.

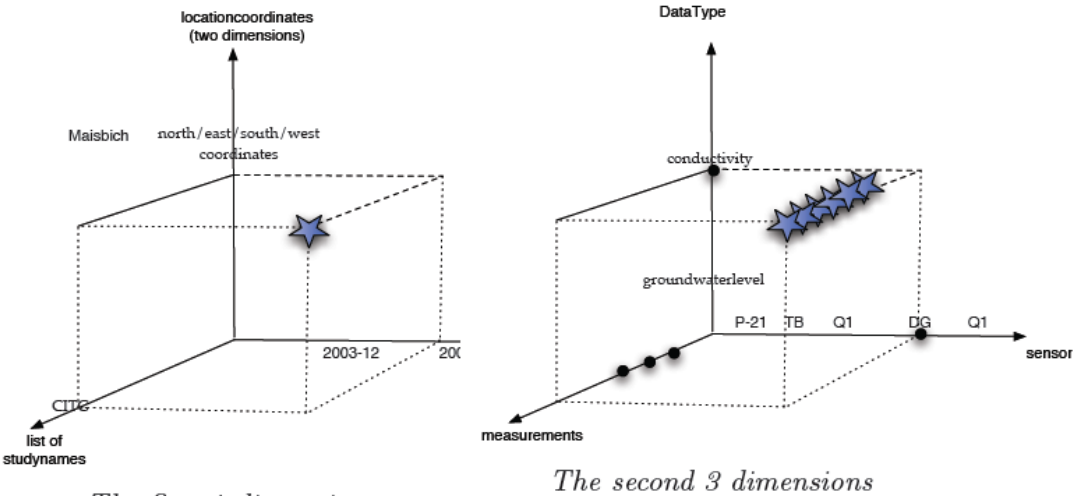

Figure 1. 7-dimensional model

Localisation Issues

For the definition of spatial items we use the Dublin Core DCMI-box and the DCMI-point definitions. They can be found at the Dublin Core Metadata Initiative (DCMI) site (Cox, 2006a, 2006b). A comment on the use of Dublin Core for durable datasets was also published (Dürr, Dekker, & van der Meer, 2006).

A *box* has the following elements (in XML notation):
< *northlimit* > . . . < */northlimit* >< *eastlimit* > . . . < */eastlimit* >
< *southlimit* > . . . < */southlimit* >< *westlimit* > . . . < */westlimit* >

The values of the elements are the geographical co-ordinates, in decimal notation. For a *point* we have :
< *north* > . . . < */north* >< *east* > . . . < */east* >

Both elements are defined in ISO standard 19115.3:2003 *Geographic Information Metadata*. This location coverage scheme is also used by the GPS and GNSS systems.

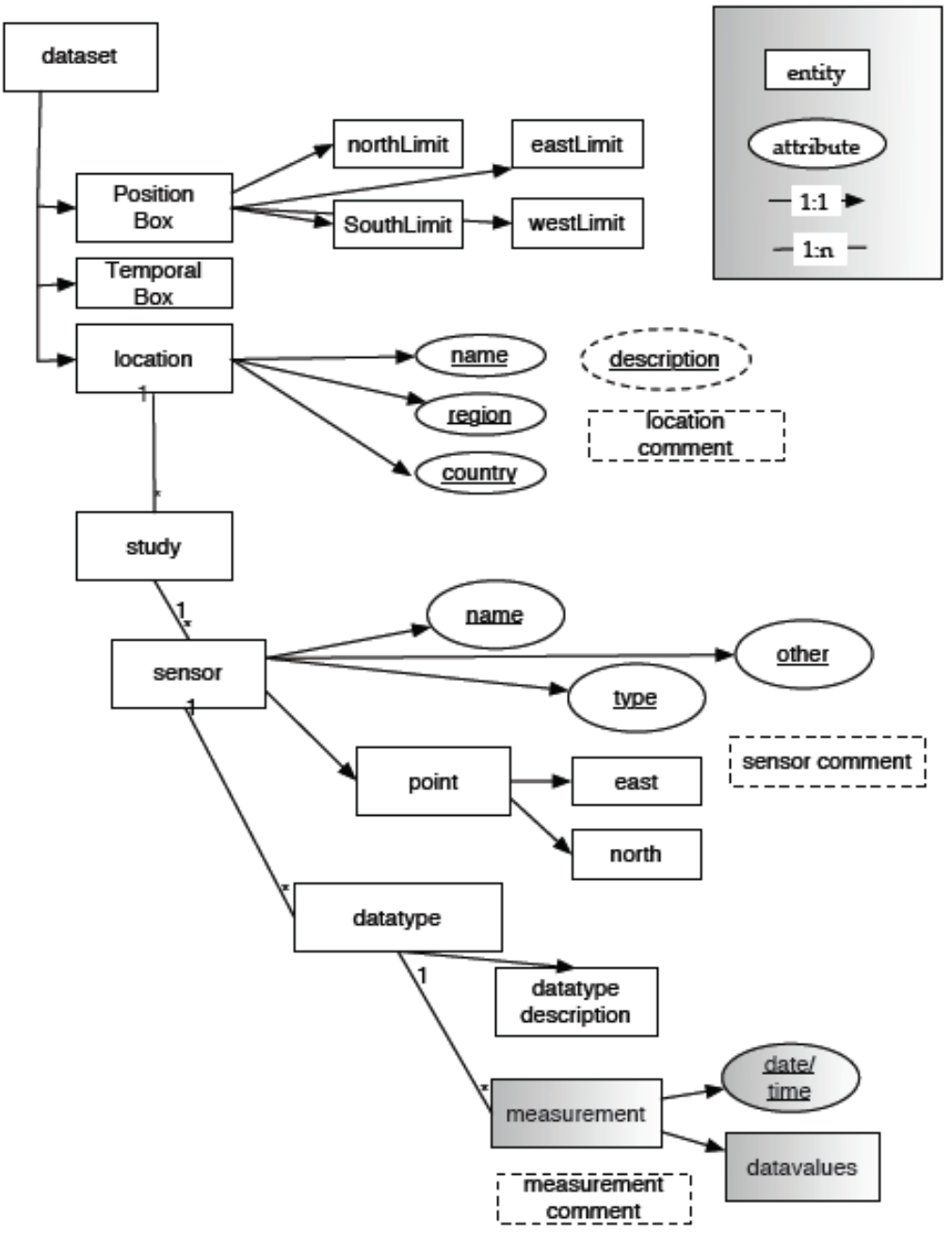

Figure 2. The entities and relations in the dataset

The Full Dataset Specification

The complete entity relationship diagram is given in Figure 2. As an example the values for a certain sensor are recorded as follows. The data type description gives magnitude and unit for all measurements:

```
<dl:sensor name="Groundwater level piezometer" sensorCode="GRW21"
        sensorType="Temperature and Pressure" otherSensorAttributes="none" >
        <dcterms:point>
                <north>49.8826</north> <east>6.0468</east>
```

```
</dcterms:point>
<dl:scomment>Type: Keller DCX-22 AA This instrument uses differential pressure
measurement of two absolute sensors. Installation: 2003-09-19</dl:scomment>
<dl:datatype name="measured">
<dl:datatypecomment>Missing values between 2005-01-11 and 2005-01-23
 Height piezometer above sensor1.25m</dl:datatypecomment>
<dl:datatypedescription date="iso-date" time="iso-timeUTC"
variable1="Total pressure" unit1= "kPa" variable2="Air pressure" unit2= "kPa"
variable3="Water Temperature" unit3= "C" variable4="Air Temperature" unit4= "C"
variable5="Water level above sensor" unit5= "mW" />
<dl:ms date="2005-01-01" time = "00:10" values="108.005;99.176;5.3;2.6;0."/>
<dl:ms date="2005-01-01" time = "00:40" values="108.011;99.191;5.3;2.7;0.89"/>
<dl:ms date="2005-01-01" time = "01:10" values="107.944;99.197;5.3;2.8;0.89"/>
<dl:ms date="2005-01-01" time = "01:40" values="107.99;99.217;5.3;2.8;0.89"/>
<dl:ms date="2005-01-01" time = "02:10" values="108.002;99.214;5.3;2.8;0.89"/>
<dl:ms date="2005-01-01" time = "02:40" values="108.078;99.245;5.4;2.9;0.90"/>
     ....900 times.....
```

Projection Retrieval: An Innovative Approach

We developed a new method for retrieving a sequence of sections as a stream result from a series of structured archival data elements within an archive. We made the assumption that an archive consists of a large collection of similarly structured archival information packages (AIP)(Consultative Committee for Space Data Systems [CCSDS], 2002). In such an archive, projection retrieval offers the option to ask for a sequence of sections where each section is part of a series of archival information items. In the response these sections are assembled into one new information package for distribution (DIP)(CCSDS, 2002).

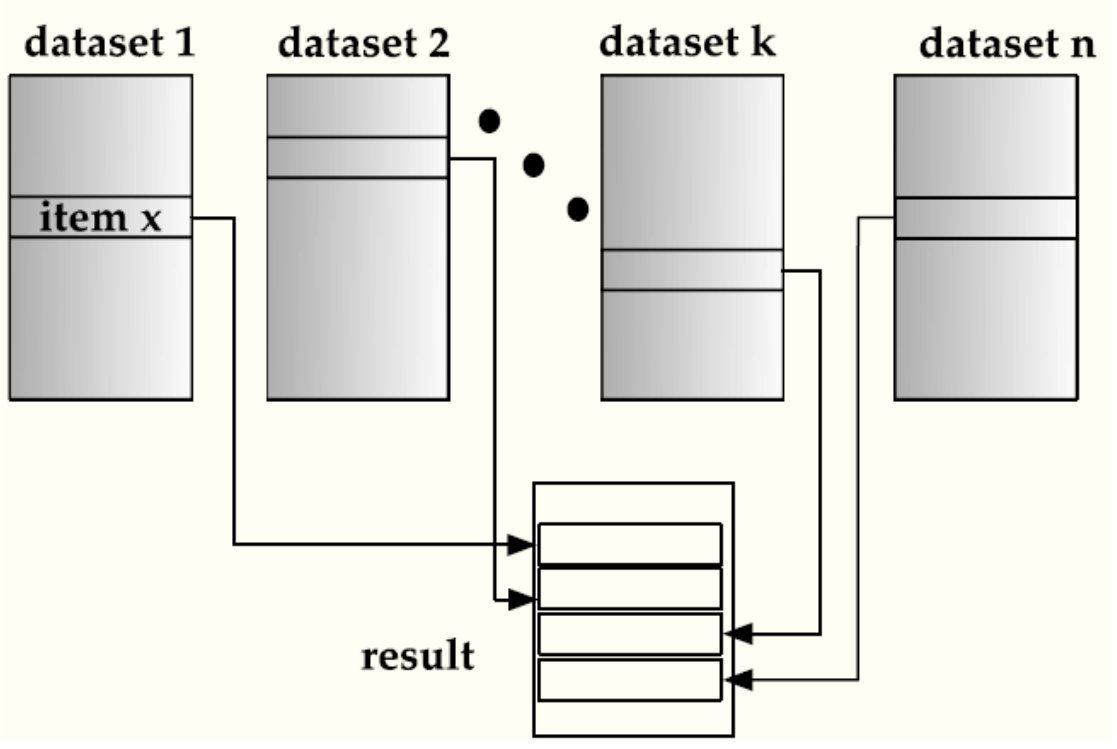

Figure 3. Projection Retrieval Visualised

If the archive information space is modelled as an n-dimensional space - one for each component-, the result of the retrieval action is the projection along one of its axes. Examples of such a request and result are:

1. In an archive with monthly hydrological measurements : give all measurements for a certain location for a certain period (expressed as start

and end date) for a given sensor and deliver it as one result file with all these measurement data.
2. In a (video) archive with daily news journals: give all news items in a given period on a certain subject and assemble these news items into one new video result, (e.g. documentary on how news on Saddam Hussein has developed over time).
3. In an archive with economic data daily snapshots, one information aspect can be retrieved and put into a time sequence, to study the long-term trends,(e.g. stock exchange data, economic reports, etc.).

Project Achievements

Most of these objectives have been reached by the closing date of the project. The next sections give the results in detail.

Contents

At the end of the project the software for the repository was fully operational. It has been decided that the archive, now called DareLux, is to be taken into "production" as a permanent service by the Library of the Technical University of Delft[5].

The archive now contains 85 fully validated hydrology data sets (approx. 2,465,000 measurements). 57 datasets are open for public access and 28 datasets have restricted access. They originate from the Hydrology Department of the Technical University of Delft, The Gabriel Lippmann Institute in Luxembourg and the Geophysics Department of Utrecht University. Apart from the measurements taken in Westerbork, The Netherlands, all measurement stem from river basins in Luxembourg. One group of datasets is linked to a publication in HESSD: measurements in Huewelerbach-2 for the periods November 2004 and June, July and August 2005. The publication is available (Gerrits, Savenije, Hoffmann, & Pfister, 2007). A detailed overview of the actual content of the archive is also possible[5].

Open Access for Harvesting

In accordance with the guidelines of the Dare Program, in which the project participated, the harvesting interface was implemented using a locally adapted version of OAICat Open Source Software (OSS)[6]. The archive is compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

User Interfaces for Retrieval

A regular evaluation of the access facilities by our hydrological partners was included in the project plan. Access to the entire archive has been provided using a web service with the SOAP protocol[7]. This design decision has opened the way for an open architecture which can be easily integrated into other services. It makes it possible subsequently to embed the web service into scientific models devised with, for example, MathLab or Mathematica. We made use of this structure to develop several different versions of the user interface.

We used a Java applet to implement access for users. The first generation used a

[5] DARELUX (Data Archiving River Environment LUXemburg) http://www.library.tudelft.nl/darelux
[6] Online Computer Library Center (OCLC): OAICat Open Source Software (OSS) http://www.oclc.org/research/software/oai/cat.htm
[7] W3C: Simple Object Access Protocol (SOAP) 1.1/2 http://www.w3.org/TR/soap/

fixed presentation with all the locations, the year/month and sensor names. Once the choice of an output format was made, the request was translated into an XML document and sent to the web service. The format choices were: the document metadata or dataset information in HTML, comma-separated values or XML for the data itself. The request was then processed by the web service and the result delivered as an XML document and the output of an XSL transformation into the requested format. These files can then be downloaded from the server to the users' own system for further processing and analysis.

The disadvantage of a fixed presentation of the locations, times and sensor names is that there is no interaction with the actual contents of the archive. Consequently, users could, unbeknown to them, formulate requests on data which were not actually in the archive. Unfortunately they would only discover the requested data were "not available" once they had completed the entire request procedure.

In the second evaluation it became clear that a more user-friendly solution was needed. We introduced an index servlet which used an XML file as a catalogue with entries for each dataset. By letting the second-generation applet interact with the index servlet, we were able to show the valid year for a given location and on selection of a particular year, the valid months. When the location-time selection was confirmed, the lists of sensor names for these datasets are displayed. On selection of the desired format, the web service is called and now will always be able to give a result.

The third user evaluation expressed the requirement for the addition of the option to place a (temporary) embargo on certain datasets. When their research was still in progress and publication was not yet feasible, scientists wanted to be able to restrict public access to certain sets of data. We satisfied this requirement by giving members of a user group a separate index file. Datasets which did not appear in the index file for that group could no longer be selected. All that was required was that users in a particular group identified themselves through a login procedure and the requisite index file for their group would be used in the interface applet.

Conclusions

The Organisational Dimension

Very strict *quality control* needs to be exercised at ingest to maintain the very *raison d'être* of long-term preservation of data. Erroneous data render the results obtained from the archive either unreliable or even useless. High-level management involvement on the scientific delivery side of operations is thus necessary to enforce this quality control upon staff members. Furthermore, changing the way datasets are used by academic staff requires a rather long time. Currently they are accustomed to *owning* the datasets themselves, instead of retrieving them from a library server. We took the initiative by including the use of the dataset archive in the training of new students in their practical assignments. This will accelerate the acceptance of using the library archive as a new working practice. Queries over the content of the archive are typically dependent on the type of scientific discipline for which the data are stored. Interfaces for the retrieval process have to be developed in co-operation with the users. In the project it was however discovered that for similar projects, i.e earth sciences-related datasets, interfaces can be reused to a certain extent.

The Business Dimension

The generic EArchive structure with the XML container structure has been shown to be a sound basis for a durable dataset archive. It can be reused for other kinds of similar services without appreciable costs. If we consider that the value of datasets increases with the length of time over which they are generated, the best approach is to treat it as a continuous data stream. There are set-up costs involved but maintenance costs are quite low: Most of the costs are on the one hand in the ingest procedure and on the other hand on the discipline-specific retrieval and presentation user interfaces. A few weeks of work is required for each new stream. The fact that there are initial costs involved can create an entry barrier for new groups of users; a proper financing scheme can alleviate this problem.

The Technical Dimension

The ingest procedure is difficult to generalise. For each discipline the dataset content will be different and has to be adapted for each new data stream.

The long-term archive aspect seems to be solved. The core of the architecture is a web service. This Service Oriented Architecture (SOA) offers much flexibility and is open to all kinds of extensions into new (commercial) services. The emphasis now shifts towards the contents, i.e. making the datasets themselves more easily accessible and providing user interfaces for retrieval, and offering options in combining and formatting or transforming. The standard document retrieval methods based upon metadata and full-text retrieval are of no use for sets of numbers. The Dublin Core metadata are almost the same for each dataset. Innovative approaches are necessary to develop the equivalent of metadata/keyword and full-text retrieval in a world where there are only numbers. Embedding dataset retrieval web service calls into the modelling and analysing software will be the next step along the way and visualisation of dataset contents may be a valuable new service for the archive.

The Operational Dimension

Maintenance of the containers is a normal system management task: storing directories with medium-sized files (0.5 to 4 Megabytes). Copies of these container files should be maintained at other locations in order to cope with disasters in the future. But this again is no different from procedures for normal documents. However, these procedures are yet to be implemented in many digital archives at time of writing; there is a decided need for *corrective maintenance*. Adaptation of user interfaces to changing requirements and usage by scientists is also necessary. A small team should be available to implement these changes and thus maintain interest and extend the usability for current and new users: in other words, *perfective and adaptive maintenance*.

Finally, the importance of quality control in the ingest process can not be emphasised enough. This has however also a down-side: it can be a bit prohibitive for new data streams and their users. However initial efforts will certainly pay off for them in the long run. It may be not so easy to convince them beforehand. We also found that much research and development needed to reach a point were working with datasets is as simple as working with plain documents. Both search as well as retrieval methods for data values do not yet exist. Catalogue development, user interfaces and remote embedded calls into other software frameworks will have to emerge in the near future as special services of archives for durable datasets.

Acknowledgements

Finally we wish to express our appreciation of all the work undertaken by our project leader, Ronald Dekker. With his sudden death at the end of 2006 we still labour under a great sense of loss.

References

Cole, T et al. (2003, April 2). *DCMI term declarations represented in XML schema language*. DCMES 1.1 XML Schema, 2003-04-02. Retrieved July 16, 2008, from http://dublincore.org/schemas/xmls/

Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS)*. CCSDS 650.0-B-1 Blue Book, January 2002. Retrieved July 9, 2008, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Cox, S. (2006a, April 10). *DCMI box encoding scheme: Specification of the spatial limits of a place, and methods for encoding this in a text string*. Contributors: Powell, A., Wilson, A., & Johnston, P. DCMI scheme. Retrieved July 16, 2008, from http://dublincore.org/documents/dcmi-box/

Cox, S. (2006b, April 10). *DCMI point encoding scheme: A point location in space, and methods for encoding this in a text string*. Contributors: Powell, A., & Wilson, A. DCMI scheme. Retrieved July 16, 2008, from http://dublincore.org/documents/dcmi-point/

Cox, S., & Ianella, R. (2006, April 10). *DCMI DCSV A syntax for writing a list of labelled values in a text string*. Version 1.0, July 2000. Retrieved July 16, 2008, from http://dublincore.org/documents/dcmi-dcsv/

Dekker, R., van de Meer, K., & Dürr, E. (2003). Digital preservation:The findings of the e-Archive Project. September 2003. Retrieved July 16, 2008, from http://durr.dhs.org/EARCHIVE/publications/e- Archivefindingsfinal13.pdf

Dürr, E.H., Dekker, R., & van der Meer, K. (2006). Archival metadata for durable data sets. In F.M.G. de Jong & W. Kraaij (Eds.), *Proceedings of the Sixth Dutch-Belgian Information Retrieval Workshop, DIR 2006*, TNO ICT, Delft, The Netherlands, ISBN-13: 978-90-75296-14-3, pp. 19-23. Retrieved July 30, 2008, from http://hmi.ewi.utwente.nl/dir%202006/dir2006.pdf

Gerrits, A. M. J., Savenije, H. H. G., Hoffmann, L. & Pfister, L. (2007). New technique to measure forest floor interception an application in a beech forest in Luxembourg. *Hydrol. Earth Syst. Sci., 11*, 695–701. Published January 17, 2007. Retrieved July 16, 2008, from http://www.hydrol-earth-systsci. net/11/695/2007/hess-11-695-2007.pdf

Reding, V. (2007). Scientific information in the digital age. *Conference on Scientific Publishing in the Research Area: Access, Dissemination and Preservation in the Digital Age.* Speech, February 16, 2007. Retrieved July 16, 2008, from http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/07/90&format=HTML&aged=0&language=EN&guiLanguage=fr