Enhancing our data model with Premis

Barbara Sierman, Koninklijke Bibliotheek (National Library of the Netherlands),
barbara.sierman@kb.nl

Abstract

In this paper I will give you an outline of our approach to integrate the Premis Data Dictionary in our new data model.

Barbara Sierman MA studied Dutch Literature and started her career in library automation at OCLC-PICA (then Pica) in 1979. After that she worked at several IT companies as a consultant. In 2005 she became Digital Preservation Officer at the Koninklijke Bibliotheek (National Library of the Netherlands).

Enhancing our data model with Premis

The Koninklijke Bibliotheek (KB), the National Library of the Netherlands, was founded in 1798, and is a medium sized national library, with 3 million paper volumes and 10 million electronic items. It has a staff of 275 full time equivalent people and is structurally funded by the Ministry of Education, Culture and Science. Since 1974 the KB is also a deposit library,although the Netherlands does not have a deposit law. The depot function is based on a general agreement with the umbrella publishing organisations, representing the Dutch and Flemish book trade.
When the KB in 1993 decided to accept digital publications resulting from her role as voluntary deposit of the Netherlands, an information system had to be designed to store this information. Several experiments started, for example with the (originally Dutch) publisher Elsevier. An international working group NEDLIB (Networked European Deposit Library, running from 1998-2000) started to formulate the requirements of a deposit system for digital publications in relation to long-term preservation.[1]
Almost at the same time the OAIS reference model was designed and published in 2003. Based on the NEDLIB specifications and the starting points of the OAIS model, IBM Netherlands has built the DIAS system. Elsevier and the KB signed a landmark archiving agreement in 2002, to store all the international digital publications of Elsevier in the e-Depot in the next coming years. From March 2003 the ingest of digital material has started. Currently almost 9 million articles have been stored. The KB has archiving agreements with 8 international Science, Technology and Medicine publishers.

For several years we have been using our current data model. In this period the developments on meta data went fast and the ideas about meta data in the KB developed as well.
One of the starting points, formulated in the NEDLIB study, was to avoid duplication of meta data. As a consequence of this, the decision has been made to store the descriptive meta data in the digital library environment. The original descriptive meta data from the publisher are stored within the AIP. The digital publications in the e-Depot can be accessed via the KB library catalogue.
The structural meta data of the publisher are expanded in a special procedure to give the user of the library the opportunity to see the article in the context of the related issue or the related volume of the journal. These structural meta data are not stored in the e-Depot, but only in the catalogue. The current structural meta data are not that complex yet. At this moment, the content of the e-Depot consists mainly of articles from e-journals. One article (mostly PDF and with a size of around 1 MB) equals one AIP.
When we started with the DIAS system, the Premis Data Dictionary did not exist. A KB/IBM study defined preservation meta data[2]. However, in our data model there was limited space to store preservation meta data. We do store some elements like file type, version of file type, size of the object, date of creation. Other preservation meta data are stored in a separate, stand alone application, the Preservation Manager. The concept of the Preservation Manager will be discussed later on.

So the current situation at the KB is, that we have

- a data model with limited space for meta data,
- bibliographic and structural meta data stored in the library catalogue,
- more preservation meta data stored in the Preservation Manager.

New developments, both at the KB as well as in the digital library world, require an evaluation of this situation. The main causes to reflect on our approach are:

- Use of standards: the KB wishes to use international standards for bibliographic and structural meta data. Study of the relevance of the elements of the Premis Data Dictionary.

- New materials: a wide variety of new material will be stored in the e-Depot in the coming years.
- IBM developed a new version of DIAS with new functionality.

Use of standards
At the KB, a discussion has been started about the standards in use As a result it was decided to replace the current (proprietary) meta data model with Dublin Core and MPEG21-DIDL. The Premis Data Dictionary is becoming increasingly accepted as a standard for preservation meta data. Therefor we are comparing the Premis meta data model with our Preservation Manager and the other meta data.

New materials
As a National Library, the KB undertakes a range of activities and special projects, all resulting in large quantities of digital objects.
For instance, we have mass digitization projects. All of them have the long-term preservation of the digitized objects in scope and the resulting digital objects will be stored in the e-Depot.
In 2003 a project has been started to digitize the Dutch Parliamentary Papers, published between 1814 up until 1995 (from 1995 onwards they are digitally available). This project will result in 2.3 million pages of TIFF files.
In 2006, the KB received a budget for a digitization project, the Data bank of Digital newspapers from 1618 until now. This project will result in 8 million pages of TIFF files.
Several other initiatives at the KB, like the Memory of the Netherlands, will add several million pages to the e-Depot.
A cooperation between the Dutch universities and the KB, the DARE project, resulted in the universities storing their scientific publications in the KB e-Depot for long-term preservation. In the near future a link is foreseen between these publications and the related scientific data. Although these scientific publications are mostly PDF files, several other file formats are also in use.
As part of our voluntary deposit, we are receiving e-books, audio-books, e-magazines etc. in a range of new formats.
And last but not least, a project for web archiving has started last year. The KB will harvest a selection of Dutch websites and add them to the e-Depot in WARC format for long-term preservation.
These developments will have a significant impact on the e-Depot. For in stance, the diversity of the file formats will increase, the structure of the objects will become more complex and the various kinds of meta data will also be augmented. Therefor our current meta data model is not satisfactory anymore.

New DIAS
The upgrade to a new version of DIAS will offer us some new functionality. At the same time, it is an excellent moment to evaluate our data model, and the extra meta data we intend to store in relation to our new materials.

The Data model work group

Approach
A work group has been established to define a new set of meta data, based on the new standards of the KB. In my presentation, I will focus on the Premis study. We have used Premis as a checklist and have made a comparison between the elements in Premis and the elements we are currently using. The Premis Data Dictionary describes the core set of meta data as "things that most working preservation repositories are likely to need to know in order to support digital preservation".This encouraged us to look at the organization around the e-Depot as a whole, without too much focus on the data model.
While using the Dictionary as a checklist, we made an interesting discovery. We discovered that during the life cycle of the DIAS system some implicit decisions had been taken. So this

study encouraged us to make some starting points of our e-Depot more explicit. For example, the element “preservation level” (Premis definition: A value indicating the set of preservation functions expected to be applied to the object). We thought the preservation level to be not applicable for us, because we use the same preservation level for our whole collection. But the ingest of new categories of materials requires explicit decisions to be taken about this topic.
Another example is the element “inhibitors” (Premis definition: Features of the object intended to inhibit access, use, or migration). Currently, the KB uses two accessibility levels: the open access articles are accessible via Internet e.g. Biomed, the restricted access articles are accessible within the premises of the KB (e.g. Elsevier). In the e-Depot all the articles are accessible by the general public within the premises of the KB. As yet, we have, no differentiated rights policy regarding restricted access documents from publishers. After all, we are a library for the general public and, as such, do not intend to store material that will not be accessible for everyone. However, publishers are working on a more differentiated access policy, whereby a selection of articles will start with the status “restricted access” and, after a certain period, will become “open access” articles. So we need to add “inhibitor meta data” on object level, which may change over time.

Subsequently, the members of the working group formulated the scenario’s of the e-Depot.[3] The scenario’s described the basic tasks of the e-Depot, like “storing the digital material” and ”keep the objects accessible over the years”. But the working group also distinguished some specific scenario’s, as a result of our contracts with the suppliers of the digital material. For example, the obligation to return a copy of the stored objects to the original supplier, in case he lost his original data.
Then, to every scenario, one or more functions were added, such as “search functionality” or “ownership determination”. Based on this scheme, we should be able to match the Premis meta data to one or more of these functions. Premis meta data elements which can not be matched, will not be relevant. The result of this exercise was, that we were able to link almost every element in the Premis dictionary (on the highest level) to one or more functions.

Environment
However, a special word about the Premis element “environment” (Hardware/software combinations supporting use of the object, according to the Premis definition). As said before, in our design of the DIAS system, the Preservation Manager is a stand alone application to store the information about the representation of the object, comparable with the information in the element “environment”. If a future user asks for a publication, the Preservation Manager will contain the necessary information to render the object.
As we foresee, when a future user requires an object, the rendering software will consult the Preservation Manager to collect software from the software repository, to define a suitable emulator or to perform migration on the fly. This functionality is not designed yet.

How does the Preservation Manager work?
Several levels of information are stored, where the key element is the *file format* of the digital object. On a conceptual level, in the Preservation Layer Model, it is determined, which information is needed, regarding the required software (operating system, the viewer application etc) and the hardware environment. Based on this conceptual model, for each *file format*, a so called ‘view path’ is created. A view path is an instance of the Preservation Layer Model containing all required technical information. This view path shows, which specific software and hardware is needed to render a certain file format on a target environment. Preferably, for each file format, we will collect information for several different environments. With this information we’ll create more than one view path for each file format. The idea behind this is, that having several view paths available, we will always have an alternative view path to render the digital object, in case one of the view paths become obsolete. In the mean time we can collect new information for a new view path.

As the basis for the description of the software and hardware, we have chosen a so called Reference Workstation (a Compaq PC, with an Intel Pentium processor), which is fully described separately outside the Preservation Manager. In our public reading rooms, several reference work stations will be placed for users who want to view objects from the e-Depot.
Over the years, we have become less comfortable with the current implementation of the Preservation Manager. For instance, as yet we are not able to use the information from Pronom and the future Global Digital Format Registry, as there is no link yet. We want to profit from international developments, but we also want to ensure the integrity of the information in the Preservation Manager, by testing and certifying every view path. Tests will be done in our own library, on the above mentioned Reference Work station.
We will use our Preservation Manager to store the information of the Premis element "Environment", which is, in the current version of the Data Dictionary, rather limited. In the Preservation Manager, we have the opportunity to store any additional information we consider useful. We think the concept of the Preservation Manager is satisfactory, but we will work on some serious improvements of the application later this year.

How to collect the meta data?
Our next step will be to determine how to collect the necessary meta data. Can we extract the meta data automatically from the data we will receive from the suppliers? Should we persuade the suppliers to add extra meta data? After all, the preservation meta data are required in, as the Premis Data Dictionary states it, "the preservation process".
An important activity in the preservation process is the execution of preservation strategies, like, for example, migration and emulation. A question is, are the elements of the Premis Data Dictionary sufficient to perform migration or emulation?
At the KB we have started two projects, one on migration and one on emulation. For each project, I will try to explain whether the Premis Data Dictionary gives sufficient information to support these strategies.

Emulation
Emulation is best described as imitating a certain computer platform or program on another platform or program. In this manner, it is possible to view documents or run programs on a computer not designed to do so.
The KB and the National Archive of the Netherlands (Nationaal Archief) are working together in building a modular emulator and the first results will be shown this year.
The digital object to be rendered by such an emulator, will have software and hardware specifications associated with it. The emulator needs the same information, in the building phase of the emulator itself, and later in order to make the right match between the object requirements and the parameters of the emulator. The current version of the Premis Data Dictionary defines too few elements to support the use of an emulator. Although our emulator is not ready yet, we have the impression that the information stored in the Preservation Manager will be sufficient to choose the right emulator and software programs like operating system and additional applications. Apart from the information about the Reference Work Station, we're inclined to think that the information "on the box", in which the software to create the digital object was sold, will give sufficient information. Later this year we will investigate this into greater detail.

Migration
At the KB we also have started a project on Migration. One of the main goals of the migration project is to find out, which of the file formats in the e-Depot qualify for migration. But when an object is migrated, it is required that the new version of the object meets a set of criteria. These criteria are often associated with attributes regarding context, content, behavior, appearance and structure. The values of the attributes are different for every file format. And where is this information stored? We are still discussing this issue, but we are not confident that the Preservation Manager or Premis Data Dictionary are prepared for this information. Later this year we will publish our results.

Conclusions
You can’t foresee everything, standards evolve and during the lifetime of a repository you should trim your sails to the wind. Our practical experience with the e-Depot proves to be very useful in weighing the Premis elements and making decisions about their applicability. Premis is not a stone written rule. It should be seen it in the context and environment in which it will be used. Sometimes this proves to be difficult. Sometimes we wondered by ourselves, why can’t we see the usefulness of this element, is it not suitable in our environment or did we miss the point? Why don’t we have this element already? Did we overlook this in the past?
Your e-Depot requires constant attention and it is important to record your decisions, to help next generations to understand your decisions. The Premis Data Dictionary was very useful for us to get this clear. Our next task is to design a new data model in which the elements of the Premis Data Dictionary will be incorporated. We will keep you informed via the very useful Premis Implementors' Group!

April 2007

[1] Information on digital preservation at the KB can be found at http://www.kb.nl/dnp/e-depot/e-depot-en.html
[2] *Preservation requirements in a deposit system* (IBM/KB Long-Term Preservation Study 3), at http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study-en.html

[3] A similar exercise was reported by the ASPR group, see PREMIS Requirement Statement: Project Report. National Library of Australia/Australian Partnership for Sustainable Repositories
July 2006 http://www.apsr.edu.au/publications/presta.pdf