Abstract. Metadata – structured resource descriptions – has vital role in retrieval of scientific information. Most researchers are used to on-line public access catalogues (OPACs) maintained by libraries. The coverage of these high quality databases is being extended to the electronic materials such as Web sites, e-books and e-journals. In order to cover all Web materials, libraries and other institutions are creating Web indices and archives, which are dependent on metadata embedded in documents themselves.

While OPAC databases use MARC format adopted by large majority of large public and research libraries in the world, embedded metadata relies on new format called Dublin Core.

Long time preservation of electronic documents is an interesting technical, organisational and political issue. One technical problem is related to metadata: what do we need to know about the archived electronic resources in order to be able to keep the documents in theory for ever? If we do not know what has been stored into our archive, we have failed before the job even begins.

Librarians, archivists and other communities have in recent years been investigating the metadata elements needed specifically for archiving. A few of these efforts are described in this presentation. One conclusion that can be drawn is that the final list of preservation metadata elements is not available yet. But it has become clear that the metadata required is dependent on, among other things, the nature of the data stored and the archiving strategy (migration, emulation or something else) the archive maintenance organisation has chosen.

Introduction

The aim of this presentation is to give a short overview of long-time preservation of electronic resources and of the role structured resource descriptions – metadata - will have in guaranteeing availability of our digital heritage for future users.

Digital preservation is usually defined as managed activities to ensure continued access to electronic resources. Access is the key factor here: if a resource cannot be used anymore, it is totally pointless to preserve it.

Preserving a printed book for decades or even centuries has been relatively easy. First, paper is usually very durable material. Second, humans can extract information from a book by a simple process: reading. Third, understanding the information is possible since the written languages have not changed totally and there are human experts who can translate the documents into modern language.

Electronic resources differ in a fundamental way from printed resources. Every electronic resource has to be interpreted by an application before it can be displayed to and understood by humans. Any string of bits can be interpreted in multiple ways, depending on the resource type and the application used. And this application – for instance Word 97 for Word documents – requires operational environment – hardware, operating system running on the hardware, and diverse other applications.

If the information technology we use were stable, preservation would be easy. But our technological infrastructure is changing with ever increasing speed. Technical obsolescence
threatens our cultural heritage in many different ways.

The media electronic resources are stored on may become unreadable either because the media – diskette, tape or CD ROM disk – is physically destroyed, or because the media cannot be read any more although it still is physically in good condition. Punched paper tape is a good example of this; at least in Nordic countries it is almost impossible to find a reader for paper tapes.

File formats and compression schemes are also constantly changing. Sometimes there is a real reason for this, for instance compression techniques have improved quickly. But it is also common that changes are made in order to force customers to buy new versions of products. Reluctance to use standards – or to use them properly – can also be explained from a marketing point of view.

Advances in computer design have been spectacular, and it seems certain that the current development rate – as specified by the Moore’s law – will not slow down during the next 10–15 years. If we will be able to go on like this for the next 30 years, our children will have computers that are million times faster than the current system. It is almost certain that these machines will be able to compute at least the same things than the current systems do, but what else will they be capable of? If the future computers are speech or vision controlled, will the future users be able and willing to get accustomed to user interfaces common in 2001?

Some experts have suggested that standards will solve our problems. But some relevant technical features may never be standardized, and technical development will also change the standards we rely on. One example: there are already two, very different versions of the JPEG image compression standard even though the first JPEG version is less than 10 years old. How many JPEG versions will we have 100 years from now?

I have concentrated on the technical aspects of preserving electronic resources, and will continue to do so in the next chapters. But long-time preservation is also an organisational problem. A digital archive needs skilled and experienced staff, solid financial support and sufficient legal framework for its activities. For instance, if the national copyright legislation prevents the archive from copying or converting electronic resources, long-time preservation is doomed to fail.

The phrase “solid financial support” above is very difficult to define. Nobody knows in detail or even in rough terms how expensive it will be to preserve electronic resources. The reason for this is that we do not know how fast technology will develop, and how badly this development will hit us. It has been estimated that we must convert our documents on average every five years, but this kind of generic statement can not be proven. And even if it were true, we do not know how difficult (=expensive) it will be to handle the documents.

**Preservation Methods**

In this chapter we will describe the common preservation methods. In addition to a short overview I will also depict some weaknesses in these methods.

**Refreshing**

Refreshing strategy means periodical copying of the resources to new storage media. The resource will remain the same, not a single bit is changed.

Refreshing looks like a non-challenging approach from the technical point of view. The trick is to know when it is necessary to copy a document; there is no way to check if a tape is still readable without actually reading it. On the other hand, some resources may be protected against copying. Unless the publisher is able to deliver a non-protected version of the resource preservation efforts will definitely fail; it is almost certain that no storage media will remain readable for centuries.

While every digital archive must copy their documents regularly enough – whatever that means – to new media, refreshing fails if used as the sole preservation strategy. Without an application with which the resource can be used a copy of it is useless. Therefore we need to use other, more efficient preservation techniques.

**Migration**

I define migration as a conversion of the resource into a new software and hardware platform. This strategy is the
most popular one at the moment and routinely used in many digital archives.

Conversion of Word Perfect 9 document into Word 2000 XML document sounds like a fairly easy thing to do. But migration is not as easy as it may seem at the first sight.

We have already said that it is very difficult to forecast the cost of evaluation. We do not know how often we need to convert our documents, and we cannot estimate how difficult conversions will be from the technical point of view. While most Word Perfect documents can be converted easily to Word, some documents using special WP features may be very difficult to convert.

Generally, if conversion is made from a versatile file format to a simple one, data will be lost. For instance, if a mathematical dissertation in \( \text{LaTeX} \) format is converted into HTML, every mathematical formula is lost unless they are converted into images. This is certainly possible, and we have done it, but it may take weeks for a single work. While this can be done in some individual cases, this method does not scale up into collections consisting of thousands or millions of documents.

Migration is also unpredictable in the sense that some properties of the converted documents may be lost. Some losses may be planned, some may not be. We should have very detailed knowledge about the properties of the archived resources in order to know if the conversion tool will be able to deal with all the features the documents have. A Word document is not a simple entity consisting of text; there may be tables, images, hyperlinks, embedded metadata and other things that should also be converted into the target format.

It is also possible that a resource cannot be converted at all, either because this is totally impossible, or because the cost is prohibitive. Software available in binary mode only is a good example of a resource that must be preserved as it is. And even if the source code were available, conversion to a new software platform requires skills that are usually unavailable to digital archivists such as libraries.

Converting databases will also pose serious problems. Some experts seem to think that relational databases and SQL will simplify handling of databases. Unfortunately this is not the case. No matter what the underlying database and query language is like, a database must be extracted into a flat file that can then be loaded into a new database system. The exchange format used by the library community (ISO 2709) is an example of a tool that can be used for database preservation. But the fact that we have ISO 2709 does not mean that preserving MARC data is easy. Helsinki University Library converted the old national article index into the VTLS system. The need to retain component part structure (record representing the journal, linked to records describing the articles) made the job difficult; we needed more than twenty small home made conversion scripts in order to accomplish the task even though the source data was in ISO 2709 format. Migration of this data was vitally important, but it has also been a difficult and expensive initiative due to large amount of human effort required.

To sum up, we may say that sometimes migration will be easy, while some other times it may be very challenging. Badly written and tested converter may destroy the whole collection by inadvertently removing vitally important features from the documents. But in skilled hands, and with sufficient resources, migration may yield good result – provided that it can be applied at all.

### Emulation

Jeff Rothenberg has suggested [Rothenberg, 1995] that preservation of electronic resources should be based on emulation. This strategy is based on development of applications, which mimic old hardware and/or software in new hardware/software environments. Resources would be stored encapsulated with sufficiently detailed information about the environment in which the application was originally designed to work. Based on this information, the digital archive would be able to pick the resource itself and then the emulators and applications the resource requires.

Full potential of emulation strategy still remains to be seen. Small tests have given promising results, but no large-scale or long-range tests have been carried out. Therefore some experts remain sceptical, while some others believe it will be highly useful.

Emulation requires a very accurate description of the old environment. The aim is to emulate hardware since it is more stable than software. Moreover, detailed enough descriptions of hardware do exist and are widely available. Transmeta processor is a good proof of this; it is able to mimic Intel processor quite well. However, most emulators have been developed for operating systems. Windows emulators running on MAC have been used to test emulation, quite successfully: emulators did fail, but only when the original Windows failed as well.

A digital archive based solely on emulation would not be user friendly. In order to read a text document written by a DOS text editor, a customer would be forced to learn both DOS and the old command based text editor. This is an unrealistic requirement. Instead, special viewer applications would be developed. Documents would still be stored in the original format, but migration would be applied on the fly to present the resource to the user. This strategy is already used in viewers that are able to present data in almost any image format (and there are about 100 of them).

Using emulation for long time preservation will require seamless co-operation of a large number of emulators. Since it is not possible to emulate every old platform in the new one, it must be possible to stack the emulators on top of another. In the long run this is only possible if emulator development is a well-controlled activity. Rothenberg [2000] has given an outline on how emulator development can be formalised.

### Preservation Scenario

To give a more practical idea of how the preservation methods outlined above can actually be implemented, I will present a scenario, which specifies the actions and strategies
needed for preserving the documents. The scenarios presented by Granger [2000] and Stenvall [2001] have been used as examples.

Let us assume that an organisation has a large collection of Word 97 documents stored on 3.5 inch diskettes. The documents are used only occasionally, but in some occasions these documents are still essential and must therefore be preserved.

At phase 0, the tools needed to use the documents (Windows operating system, Word 97) are in common use. The diskettes are reliable (data has not been stored in them more than five years).

At phase 1, the employees find diskettes difficult to use, and at least some diskettes begin to reach the end of their life cycle. The organisation makes a decision to copy all documents to a new archive server. All old diskettes are thrown away and new ones are not made any more since the new texts are stored on the server.

At phase 2, the organisation has upgraded the old workstations and application programs. Word 97 is no longer in use, but the documents in this format are still readable with the new text editor. Due to staff limitations no retrospective migration is done, except for those documents, which are used. Quality of migration is checked, and in case there are serious problems the original is kept too.

At phase 3, the organisation plans an upgrade to yet another hardware and software environment. During the planning stage it is noticed that the new text editor does not support Word 97. A quick check from the archival system shows that there are still a few thousand Word 97 documents left.

There is not enough staff to migrate all these documents before the new hardware and software environment is installed. Moreover, it is known that the migration will not always give good enough results. A decision is made to acquire an emulator, which enables the continued use of the current application that does support the viewing and possibly also the editing of Word 97 files.

Although all preservation strategies have some shortcomings, they can be used to complement one another. Any institution investing on the archiving of electronic resources should test all preservation methods in order to get familiar with them.

Metadata for Preservation

Dempsey and Heery [1998] point out that one of the tasks of metadata is to tell that the resource exists. But metadata may have many other roles; it may contain subject description (classification and/or subject headings), copyright and usage information, pricing information and, in the case of electronic resources, information about the hardware and software the document requires. In short,

"...metadata is data associated with objects which relieves the potential users of having to have full advance knowledge of their existence or characteristics. It supports a variety of options."

Long time preservation of electronic resources is one of the functions metadata can support. As of this writing we do not yet have the final specification of what metadata elements are needed for long time preservation, but there are several interesting attempts, which give us a good idea of the requirements. Bullock [1999] has listed the following categories, which the preservation metadata may contain:

- Identification
- Hardware and software required for the usage of the resource
- Description of the physical properties of hand held resources such as CD ROM disks and diskettes
- File format and its version; for instance Word 97
- Information about migrations and possible information losses that occurred during them
- Information that supports authentication; for instance MD5 checksum with which we can prove that the resource has not been changed during the archival period
- Copyright information

Each of these categories may pose interesting challenges for cataloguers. We do not know yet what kind of copyright information should be provided. On the other hand, even areas we think are well known may become terra incognita in the digital world. Identification of electronic resources is currently in a turbulent state; new identification systems are emerging and existing ones are being modified in response to the new needs of electronic publishing and the Internet. In addition to choosing the correct identifier an archiving organisation should also be able to pick an appropriate resolution mechanism; a tool for linking the identifier and the resource to one another. For instance an article should be identified with ISSN-based Serial Item and Contribution Identifier (SICI), which in turn can be used either as Digital Object Identifier (DOI) or Uniform Resource Name (URN) in order to guarantee persistent linking.

In spite of the present, somewhat uncertain state of affairs most experts agree that metadata will have a vital role in long time preservation of electronic resources. This should not be too surprising: after all, metadata in general, and perhaps the national bibliography databases in particular, is very important in preserving our printed cultural heritage. In the electronic world metadata will just have a set of new tasks in addition to the old ones.

For instance, the authenticity of printed books can be taken for granted; they are not modified when stored in a deposit collection, and any changes may be easy to detect. Electronic resources may be trivially easy to modify, and therefore we must have metadata, which enables the checking of authenticity.
Preservation Metadata Element Sets

Up to January 2001 there has been at least four serious attempts to develop a metadata element set for long time preservation of electronic resources. I will describe them here in chronological order.

RLG

RLG metadata specification for preservation of electronic still images was published in 1998 [RLG Working group..., 1998]. Libraries and other organisations are creating digital copies of their printed collections; the aim of the format is to provide a tool for describing these digital images. The format contains 16 elements, of which ten are related to the technical properties of the image.

The RLG work led to the establishment of National Information Standards Organisation (NISO) Committee on technical metadata for digital still images, which will publish the first version of the American national standard in early 2001. Many experts do believe that different kinds of resources – still images, moving images, sound files, text documents – do need dedicated technical metadata element specifications. With the sole exception of still images, the specification work has not really started yet.

CEDARS

CEDARS (CURL exemplars in digital archives) was an English project, which investigated long time preservation of electronic resources via emulation, and the role of metadata in this activity.

CEDARS published the second version of their metadata element set – modestly called outline specification – in spring 2000 [CEDARS, 2000]. Contrary to the approach chosen in the RLG project, CEDARS does not have any document format specific metadata elements.

CEDARS specification is based on the Open archival information system (OAIS) model [CCSADS, 1999]. OAIS is an abstract model of an electronic archive. Many projects have used it as a reference model. Since OAIS is very generic it can be applied in all branches of knowledge. But this generality is also a weakness; in order to use it in an efficient manner a more detailed, domain specific model has to be developed.

Project NEDLIB has developed such a model for digital libraries [Werf-Davelaar, 1999]. NEDLIB Deposit System for Electronic Publications has six vital areas:

- Ingest (input of the information packages into the storage system)
- Archival storage
- Data management
- Access

- Administration
- Preservation; actions related to the long time storage of the archived resources

Resources are transferred to the archive via the ingest function. Archival and preservation actions generate metadata, which archive administrators can use via the reporting functions built into the data management module.

OCLC and RLG have in 2000 established a common working group on preservation metadata. This group will in early 2001 publish another modified version of the OAIS model for libraries.

OAIS and the metadata element sets based on it divide an archival information package into content information and preservation descriptive information. The former contains the stored object and representation information; that is, everything needed to present the resource, such as hardware and software requirements.

Preservation descriptive information is focused on describing the past and present states of the content information, ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered. PDI-information can be split into reference information (bibliographic data about the resource), provenance information (the origin of the collection or resource), context information (data supporting context linking between the different parts of the resource or collection) and fixity information, that is, data, which supports authentication of the resources.

CEDARS metadata element set contains no less than 54 elements divided into the categories listed above. Although not all elements are needed every time, it is clear that creating and maintaining preservation metadata will definitely not be a trivial task, although some elements can be produced via automated means.

NLA

National library of Australia [National library of Australia, 1999] specified in 1999 25 data elements, which a digital archive system must be able to generate in order to guarantee long time preservation of the stored resources. Some elements such as the Persistent identifier are mandatory. Some contain several sub-elements; for instance, the File description element contains sub-elements for Image, Audio, Video, Text, Database and Executables. The NLA element set is generally well adapted to different resource types.

NLA specification is interesting because it is based on the library’s long experience in creating and storing digital resources. The purpose of the element set is to support both emulation and migration strategies; its applicability will be tested both in tests and practical work in NLA.

The preservation metadata elements NLA has defined are compliant with the OAIS model, although the element set is not explicitly OAIS compliant. A useful feature in the NLA format is that it enables resource description on collection, document and file levels.
NEDLIB

EU-funded project NEDLIB (Networked European Deposit Library) has defined minimum core metadata mandatory for preservation management purposes [Lupovici & Masanès, 2000]. The aim was to specify the metadata elements, which will ensure future access to the stored resources.

The NEDLIB metadata element set is based on the OAIS model. There are eight elements for representation information, and ten for preservation and description information. The NEDLIB element set has strong technical bias, and it has been built with the needs of both migration and emulation in mind. The project was well prepared to define metadata elements for emulation due to the emulation test done within the project, and the possibility to use Jeff Rothenberg as an expert advisor.

The NEDLIB elements for representation information are [Lupovici & Masanès, 2000]:

- Specific hardware requirements – description of non-standard platform configuration or hardware requirements. Three sub-elements: microprocessor, multimedia device and peripheral device.
- Specific microprocessor requirements – description of specific microprocessor instructions set (for instance MMX instructions set for Pentium III) or co-processor
- Specific multimedia requirements – description of non-standard multimedia hardware requirements
- Specific peripheral requirements – specification of non-standard peripheral devices (for instance a ZIP storage device)
- Operating system; this element has two sub-elements, name and version
- Interpreter and compiler; three sub-elements: name, version and instruction
- Object format; sub-elements name and version
- Application; sub-elements name and version

This element list makes it obvious that the creation of preservation metadata may require both cataloguer and IT specialist skills. The products to be archived – for instance an electronic book on CD ROM disk – will provide only the basic technical information; a lot of data needs to be extracted via experimentation or by asking from the publisher’s product experts.

NEDLIB specification contains links to CEDARS and NLA formats. For instance CEDARS does not allow specification of operating system, which is surprising since CEDARS aims at preservation via emulation. On the other hand, neither CEDARS nor NLA enable encoding of specific hardware or processor requirements. It can be claimed that NEDLIB is ahead of other preservation metadata element sets in defining metadata elements needed for emulation.

Preservation Metadata, MARC and Dublin Core

From the libraries’ point of view, emergence of preservation metadata elements is of course a good thing. But it is not enough. Integrated library systems are based on MARC (MAchine Readable Cataloguing) format developed and maintained by the Library of Congress. Although cataloguing of electronic resources with MARC format is possible, many of the elements proposed in the element sets described above have not yet been included. This is quite understandable; since the formats are not stable and none of them has received international approval, it is obviously too early to pick some of the existing preservation metadata element sets and use it as the starting point for extending MARC. Integrating all preservation metadata element sets and adding all the resulting elements into MARC would not make sense either, since the preservation metadata formats are not fully compliant with one another.

Once the library community has agreed on the preservation metadata elements needed, modification of the MARC 21 format used in the United States and many other countries is a relatively simple process (although it may take some time). Many of the required elements are already present in the current MARC 21, and some will only require new sub-elements or codes to the current MARC 21 tags.

Once MARC 21 has been extended, the library system vendors will incorporate the new data elements into their applications. This will take some time, but modern library systems have been built in such a way that adding new elements is a simple process.

Compared with MARC 21, Dublin Core is a simple resource description format. Accommodating preservation metadata into Dublin Core is nevertheless easy, since the format can be extended with new elements and qualifiers. In fact, it is possible to incorporate all of MARC 21 and more into Dublin Core.

Therefore adding the preservation metadata elements into Dublin Core will be technically easy. There are already communities within the Dublin Core metadata initiative who have extended Dublin Core for their own needs. Developers of preservation metadata will be one such community.

Preservation Metadata Element Sets – a Summary

Development of preservation metadata formats started in the late 90’s. The pioneers came from the United States, Australia and United Kingdom. In continental Europe, project NEDLIB was probably the first systematic effort to analyse the challenges related to preservation of electronic resources.

The existing metadata formats have a lot of similarities, but they are not fully aligned with one another. There is a need for concerted effort; it is to be hoped that the OCLC/RLG Working group on preservation metadata will create a solid basis on which the future initiatives can build
on. Since the developers of NLA and NEDLIB preservation metadata formats have been invited to the working group, the OCLC/RLG group has good chances of success. Consolidation of metadata element sets is essential, since otherwise it will be difficult to decide which elements should be added to the MARC format traditionally used by the library community, and to the extended version of the Dublin Core metadata element set, enriched with preservation metadata. Extension of these formats is an important step in the way towards making creation of preservation metadata a routine.

Epilogue

In 1964, Nelson Mandela defended himself with an eloquent three-hour speech against accusations made by the apartheid government. Although he was jailed for life and was held in Robben Island for 27 years, the speech remains today as a testimony of one man’s passionate fight for justice for himself and his people.

But Mandela’s message to future generations was almost lost. When the tape onto which his words were stored was found in the National archive of South Africa, the tape was no longer usable. There was no hardware with which to play it.

Luckily researchers in England were capable to extract the data from the original tape and convert it into digital form. We can now hear Mandela’s message for tolerance. Will our children be able to do this as well? Will there be human and technical resources 40 years from now to guarantee that Mandela’s words will remain available also in the more distant future? If the results of scientific research are published only in electronic form, how can we make sure that the publications will be available, and that their content is not modified?

Preservation of electronic resources is a complex technical, organisational and legal problem. Just how complex – nobody knows. Libraries, and especially national libraries responsible of maintaining deposit collections, will be among the central players in this area. Metadata will be one of the most important tools with which we will fight against the time.

But alone the libraries can not preserve electronic publications; broad co-operation is needed. Publishers will be among our closest allies, since the difficulty of preserving a document depends on its technical properties. We may say that long time preservation activities start at the moment when an electronic document is designed. But we will need also important technical advances to transform preservation of electronic resources from experimental activity into a familiar routine.

References


Dr Juha Hakala, Helsinki University Library, University of Helsinki, PO Box 26, FIN-00014 Helsinki, Finland, e-mail: juha.hakala@helsinki.fi

(Received 18 February 2001)