67th IFLA Council and General Conference
August 16-25, 2001

Code Number: 163-168-E
Division Number: VI
Professional Group: Preservation and Conservation
Joint Meeting with: Information Technology
Meeting Number: 168
Simultaneous Interpretation: -

Technical data and Preservation needs

Catherine Lupovici
Head, Digital Library Department
Network and Services Direction
Bibliothèque nationale de France
Paris, France
E-mail: catherine.lupovici@bnf.fr

Managing and preserving digital collections in order to guarantee long term access to researchers is, as it has always been with each new technology, both continuity in objectives and general organization of services and rupture in technical practices and staff skills requirements related to new technical needs. With the acquisition of digitally born resources which have no more analog equivalent and with the digitization of existing collections process, which is becoming a part of the preservation and conservation policies, the libraries are faced to a new preservation challenge. To ensure the long term usability of digital resources they will have to define and collect data that will be used by the preservation managers to take appropriate actions in order to maintain the bit stream which is constituting the digital object in a way that it can be rendered and interpreted whatever the technical changes in computing will be in the future. This situation is exactly the same for all the institutions that have custodial and memory missions for specific communities of users, sometimes keeping records for evidence purposes, and for which the information is already created and stored only in digital format since more then a decade. In all these communities the work for building such preservation metadata sets have been initiated most of them based on the OAIS (Open Archival Information System) standard and we are on the way of building a consensus within the Library community.

1 The new challenge of digital preservation

The preservation of a digital resource, just like the preservation of classical resources, need to preserve the technical mediation between an object and the information it contents. In the framework of what we have currently to preserve, the preservation of the physical medium, for instance a book, is the main task. Of

course we know that sometimes we need also to keep some context knowledge if we want to make the content of the object understandable. We have the example of the Rosette Stone as a symbol of what is the difference between preserving the physical object and preserving the ability to interpret the language code recorded on the medium. We have also the more recent example of the audio analog material where we need a technical mediation through a device to transform a physical vibration into sound waves and for which the characteristics of the transformation changed over the time with the disappearance of devices replaced by incompatible new ones. In that case the preservation challenge was concentrated only on the medium preservation and the device obsolescence against which the action to undertake is maintaining devices as long as possible then migrate resources to new formats corresponding to actual devices.

The digital resource is introducing a more important dissociation between the medium and the content, which is impacting all the author right legislation and libraries organization based on the physical object and its medium type, like for instance legal deposit legislation in France, or the organization of manu library conservation services.
The preservation of the medium on which the libraries concentrated until the mid 90s in conjunction with the electronic off line publishing, is now more considered as a classical computing safeguard function as we have to refresh preventively all the medium according to a strict planning.
The technical obsolescence associated with the digital resources is happening faster then the medium aging. For instance The Word® text processor new versions are issued within the average of every 3 years and are directly compatible only with the previous version. The true challenge we know we are faced to is to understand and manage the complexity of the technical obsolescence of the content information from the bit stream to the usage of it through an application. We can decide either to emulate the obsolete complex technical environment or to migrate the resource, totally or partially, depending of what is considered as the content to be preserved. In practice we will certainly have to manage both solutions and for that purpose we need to record appropriate technical data as well as archival data which constitute the **Preservation metadata.**

1.1 The components of the digital content preservation

The content of digital resources has to be processed by a computer when it is used in order to render the bits stream understandable for the user. This process is composed of a chain of technical sub-processes creating a chain of technical constraints we need to document with metadata to be able to manage the preservation. All the sub-processes will not necessarily become obsolete at the same time but one element out of use will jeopardize the access to the content. We can analyze and represent this chain in a layered model where each layer is representing a sub-process that render a service to the next upper layer on which the following sub-process in the chain is operating. From the bottom to the upper layer we can distinguish the following sub-processes with the metadata type we need to create and store along with the resource:

- **Physical layer**

The digital resource is stored on a physical or a communication medium through a physical storage format, generally standardized (for instance ISO 9660 for a CD-ROM). It can be changed if we migrate the document on another medium for instance a DVD and we need to keep the information if we will have to provide the original format.
We have also to consider at the bottom of the chain some hardware device data in case of hardware physical dependency such as the need of extra devices (for instance a multimedia application using extra devices like MIDI audio applications). For the main hardware, the information is redundant with the operating system one.

- **Binary layer**

The bit stream is organized into labeled blocks that are medium independent. The operating system (name and version) which is managing the file system is providing the service and in many cases is implicitly including the file system but not always. Example Windows NT 4.0

– **Structure layer**

The bits are aggregated into primitive data structures to be manipulated by programs. In case of preserving non compiled software we need to keep the information about the interpreter or the compiler name and version that will be required to use it.

– **Object layer**

Data are structured into objects meaningful for the application and through it to the user. The object format can be an open format or a proprietary format. The objects are rendered by the application and in the preservation environment only the object format has to be known. Example: image in JPEG, HTML pages, MPEG…

– **Application layer**

The application software manipulates the objects of the previous layer and presents them to the user. The name and version of the application can be redundant with the object format. We can have a one to one correspondence (for instance PDF format and Acrobat reader), or many to one (for instance with the JPEG format where several viewing applications can be used). In other cases only the application is known and the proprietary object format it is using remains unknown (often in early CD ROM publications)

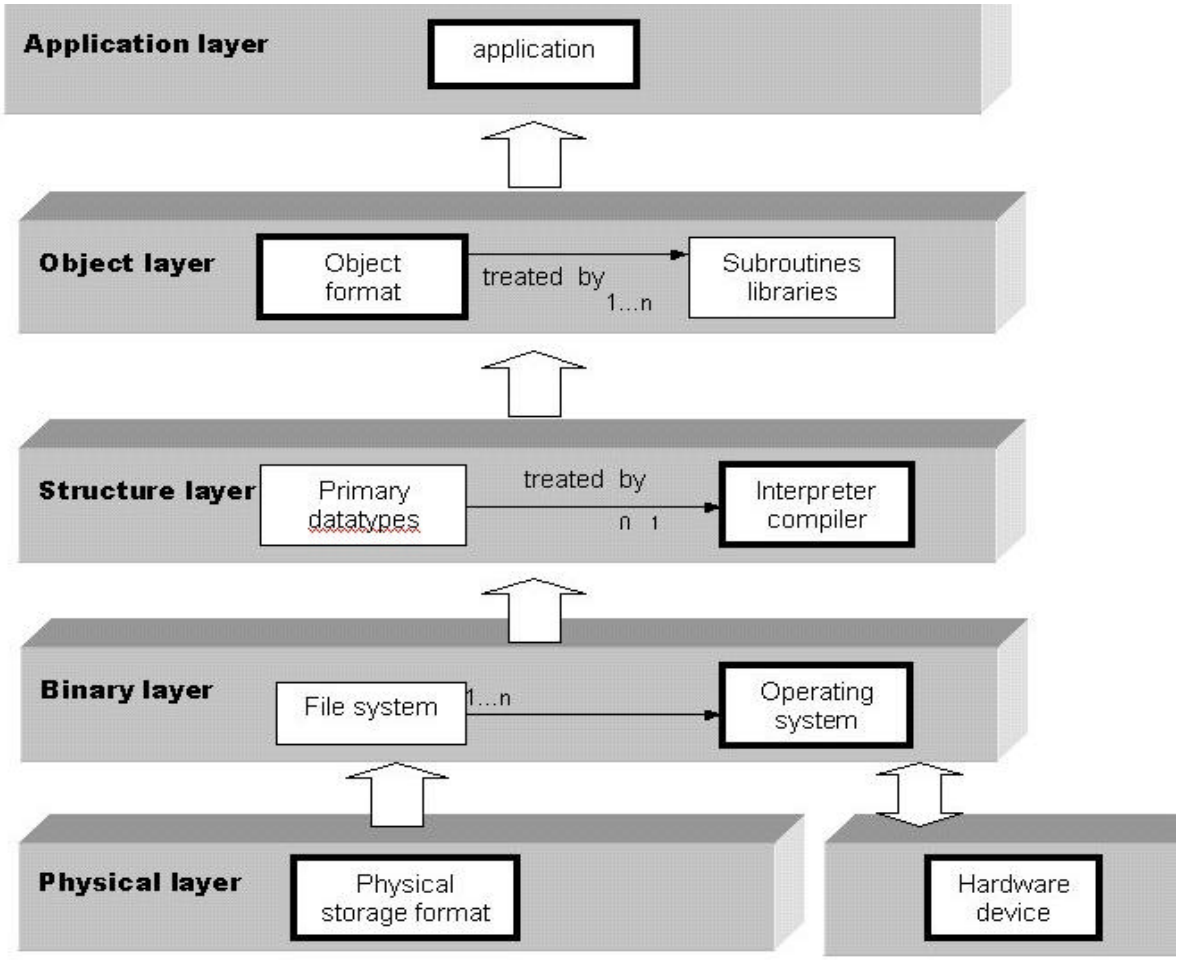

Fig. 1. The layered information model (From NEDLIB Report 2. Fig6, p. 8)

1.2 The different technical types of digital resources

This analysis leads us to distinguish between two types of electronic resources that will not require the same type of preservation actions and will not need to be archived with the same preservation metadata:

- Electronic resources that are applications dependent of specific proprietary systems for which we need metadata on the Application (application layer) and on the Operating system (binary layer). They correspond generally to off-line CD-ROM publications where a specific application manipulates unknown formats. In this case the only access to the content is the proprietary application (for instance cdu.exe in the CD ROM of the *Encyclopaedia Universalis*)

- Electronic resources with known formats that are independent of specific systems. Here we need to retain metadata on the Object format (object layer) and possibly on the application as long as the library is obliged by contract to use the original application to access the data. This type of resources corresponds generally to Web resources. Obviously igitized documents created by the library in a documented known format should be archived independently of the current digital library applications as neutral contents.

We can see that the resources most difficult to preserve are the applications dependant of specific systems, which is the case of the offline publications with, protected controlled access through proprietary applications. At the moment the Web is more open but we have to follow up the technical evolution of the commercial Web and deep Web as it can become as proprietary as some off-line publications. The Library community has to make the publishers and the authors aware that they have to deposit more open applications using known documented standard formats in order to allow the long-term preservation of their publications.

2 The Preservation Metadata

Several metadata sets are defined associated with digital resources and focussing on specific functions. Functional categories usually associated with digital repositories are already defined such as:
- Descriptive metadata facilitating resource discovery and identification. They received the most attention from the library community as they can be seen as rather closed to cataloguing
- Administrative metadata supporting the resource management within a collection of digital resources
- Structural metadata which are binding together the components of complex digital resources

But such metadata are designed for a system supporting a digital library repository that allows users to search and browse through the digital collections. They are not designed for the preservation purpose, which is more associated with the functions of an archival repository.

A first attempt to define such metadata was made by several projects in the libraries and archives communities. For instance RLG (The Research Libraries Group) released in May 1998 a set of 16 recommended metadata elements considered as essential to be captured along with digital surrogates created by the libraries in their digitization projects. This early work has led to a work preparing a NISO standard on Technical Metadata for Digital Still Images. These different approaches are reflecting specific needs in specific domains and they have to be integrated into a general framework emphasizing the archival functional aspect and covering all the range of preservation needs of the library community.

The OAIS (Open Archive Information System) reference model, developed under the auspices of NASA's Consultative Committee for Space Data Systems (CCSDS), which is becoming an ISO international standard (ISO DIS/14721), is providing such a framework. It also provides a high-level information model that can be used for the metadata framework, including the preservation metadata. Several library projects already used the OAIS information model to develop preservation metadata sets.

2.1 The OAIS information model

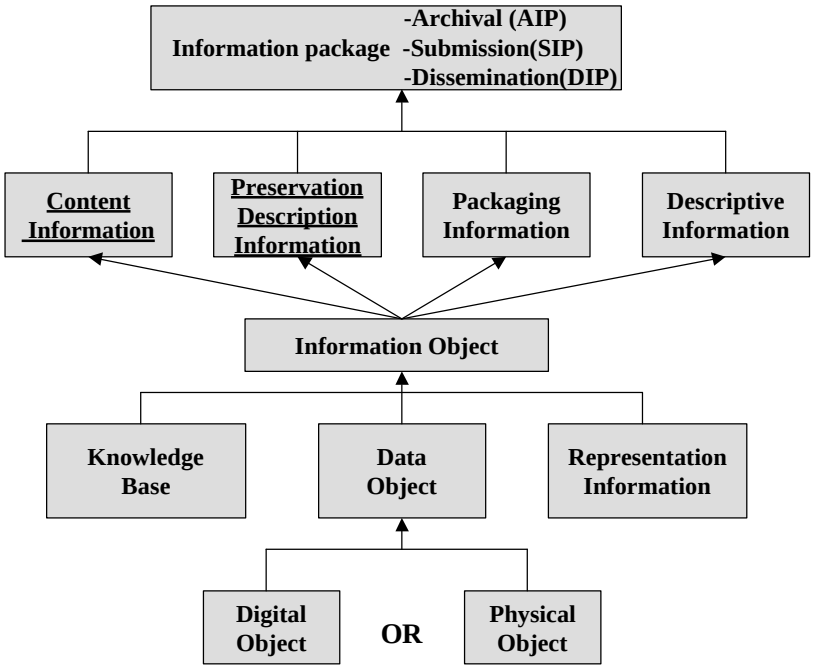

Fig.2. The OAIS Information Model (From OCLC/RLG White paper, Fig. 2, p. 12)

At the abstract level the OAIS Information Model is manipulating Information Objects. An *Information Object* is the meaningful information interpretation of a *Data Object* (a digital or a physical one) through the user's Knowledge Base and more importantly from the preservation point of view through the *Representation Information*.

Four classes of Information objects constitute the *Information Package* that the system ingests when receiving it (SIP), then stores in the archive (AIP) and delivers on demand to consumers (DIP). The information objects types are:

– The *Content Information* object, which is what has to be preserved. The definition of the Content Information is not obvious. For instance with on line journals it can be only the text and embedded

images of the articles, or it can be the content and associated presentation for instance as PDF files for the articles. But for the same journal it can also be the full application that allows to search and retrieve and to browse through the full journal collection.
- The *Preservation Description Information* (PDI), which contains information necessary to manage the preservation of the Content Information
- The *Packaging Information*, which binds the Digital Object and its associated metadata into an Information Package
- The *Descriptive Information* which facilitates the access to the Content Information via the archive's search and retrieval tools

The Content Information and the Preservation Description Information are the two information objects classes essential for long term preservation and they are providing the framework for metadata sets creation.

The Content Information corresponding to a digital object has to be rendered through the *Representation Information* and we need Representation Information metadata for instance metadata on the Operating system name and version.

The OAIS model identifies four types of PDI for which preservation metadata have to be defined:
- The *Reference Information* that enumerates and describes the identifiers assigned to the content information
- The *Provenance Information* that documents the history of the content information
- The *Context Information* that documents the relationships of the content information to its environment
- The *Fixity Information* that documents the authentication mechanisms used to ensure that the content information has not been altered in an undocumented manner

2.2 State of the art of the preservation metadata development

OCLC and RLG announced, in March 2000, their commitment to collaborate on identifying and supporting best practices for the long-term retention of digital objects. One area of collaboration is the use of metadata to support the digital preservation process and they organized an international Working group on preservation metadata. The first work of this working group was to issue by January 2001 a White paper reviewing and comparing four metadata sets in order to prepare for a consensus building. Three of them can be mapped to the OAIS information model. They are :

- The National Library of Australia metadata set issued in 1999 developed within the PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) project. The metadata are intended to support the preservation of both digitally born and digital surrogate objects.

- CEDARS (CURL Exemplar in Digital Archives) elaborated the metadata set in 2000. The CEDARS project is run by the universities of Leeds, Cambridge and Oxford in the UK. The metadata are covering the administrative, the technical and the legal information for the complete archival functions, including the preservation.

- NEDLIB (Networked European Deposit Library) published the metadata set by the end of 2000. NEDLIB project was leaded by the Royal Library of Netherlands and was gathering the National Archives of Netherlands, National Libraries of Finland, France, Germany, Italy in Firenze, Norway, Portugal, Switzerland. Publishers were also associated (Elsevier, Kluwers, Springer Verlag). The metadata were considered as the minimum mandatory for preservation management purposes in order

to handle large amounts of data items when archiving off line and Web publications through a national deposit approach.

The fourth metadata set examined in the OCLC/RLG Working Group's white paper does not follow the OAIS information model. It is developed by Harvard University's Digital Repository Services. It demonstrates how XML structures can be used to encapsulate preservation metadata into the digital object when submitting it to repository.
This white paper is a basis for building a consensus on a standard for preservation metadata that corresponds to the library community requirements.

3 Conclusion

We can see that a lot of progress has been made during the last five years on the digital preservation of all the electronic resources a Library can have in its collections. The OAIS metadata framework is offering a good basis on which the library community can build its own standard on metadata not isolated from other communities. The work already done will allow for a library standard set in a rather short time. Nevertheless the fundamental question of the nature of the work to be done in order to create the preservation metadata is not solved. Some projects point of view is that they can be created during the cataloguing descriptive process by specialized cataloguers as technical data are already input in the MARC bibliographic descriptive records. From others it is a very technical matter more related with computing expertise and it should be very important to be able to create them automatically rather then manually. We have to continue to experiment the implementation of such metadata and assess their efficiency in the preservation process during several years. There is still a long way before us to guarantee the long term access to digital resources.

References

Reference model for an open archival information system (OAIS) Red book issue 1 / Consultative committee on space data systems. 1999. 140 p. http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf (visited 20/07/2001)

Preservation metadata for digital objects : a review of the state of the art / A white paper by the OCLC/RLG Working group on preservation metadata. January 31, 2001. 50 p. http://www.oclc.org/digitalpreservation/presmeta_wp.pdf (visited 20/07/2001)