

How to keep digital records understandable and usable through time?

Hans Hofman

Ministry of the Interior of The Netherlands

Paper presented at the international conference '*Long-Term Preservation of Electronic Records*'

Paris, March 8-9, 2001

1. Introduction

Francis Bacon said in the early 17th century that "The images of men's wit and knowledge remaine in bookes exempt from the wrong of time and capable of perpetuall renovation".¹ That may have been true for the invention of book-printing, but the question now is whether this also will be true for digital objects or information when looking at the IT-revolution that is taking place at this moment?

The volatility of digital information raises many concerns, especially on its sustainability. And it poses major questions about the preservation of the information we create in digital form and more in particular of our intellectual capital.

Accompanying these questions is the issue of metadata. Some people say 'without metadata content is just bits' and call this the 'biggest challenge (to) technology industry'.² Although it is not a technical question, but an intellectual, it is indeed a real challenge. If we want to preserve and maintain digital information, this information should be kept or made understandable, sustainable, usable and accessible by adding sufficient metadata of all sorts.

In this paper a bird's eye view will be given on issues with respect to both preservation and metadata in the context of the rapidly changing world of digital information. The viewpoint taken will be that of the archivist. The term archivist is used here in a broad sense, as the professional responsible for both records and archival management and as such encompassing the whole records continuum.

I will discuss briefly the characteristics of (digital) records, their preservation and the issues around the necessary related metadata.

2. The issue: the changing world

The introduction of new technologies for creating, handling, storing and making available information has made things easier in many ways, but also raised some serious problems. One area that is in particular confronted with challenges is that of memory organisations. They have to deal with the longterm preservation of information and that area is not particularly addressed by IT, on the contrary everything is focused on short term aspects. This is enhanced by the fact that IT is still in its infancy and very innovative.

In the office environment the desk top computer is equipped with all kinds of software to create documents and information. Although the desk top is still mainly a personal domain, developments with respect to e-government or e-commerce will influence radically the way business processes will be carried out. The ambitious targets set here to deliver services through the internet themselves already raise big and many challenges, but they will also have a deep impact on the creation of documents and records. It will change, and already is changing, the world not only of the business process itself, but also that of the associated

¹ Francis Bacon, *Of the advancement and Proficiency of Learning*, 1623.

² Stewart Alsop, *Without Metadata, Content is just Bits*, Fortune, Vol. 142, No.13 (2000), p.84 (<http://library.northernlight.com/MG2000112202000014.html>)

document and records management, as has been discussed in many publications during the last decade.

In the paper world documents were fixed, physical entities, but in a digital environment these documents are 'fragmented', in the sense that you need software and hardware to make to information stored in a datafile visible and readable on a computerscreen. The documents or records have become more or less intangible.

In this context we also need documentation or metadata to know how to recreate the documents. The term metadata does not really say much, except that it is data about data, and without any context it has no real meaning. With respect to records management metadata regards the preserved records and the context, i.e. the organisations and functions, in which the records are created. Apart from that it includes data about the management of the objects.

In such a volatile environment it is essential and necessary to know what should be preserved and maintained. It requires us to go back to the origin of our work. What is the objective of preserving records? Only if we have established that, it is possible to identify and define the requirements for preservation. Finally it will be necessary to identify and clarify what the consequences are for the work records managers or other preservers are doing in this area.

In short the use of IT causes the emergence of new types of records, that require new methods and procedures of handling and a complete new infrastructure that enables that. The issue is that these preconditions do not exist yet, except for parts of it, and still have to be invented. Existing methods for digital preservation, mainly in the area of relational databases, are not capable of dealing with new, emerging and more complicated forms of digital information. It also means new knowledge and skills are needed.

In the area of digital preservation, especially of long term preservation, many initiatives exist, but there are no real solutions yet, although there are some promising approaches. Our first priority is however to save what already exist as good as possible and in the meantime try to find more permanent solutions. In this respect it is necessary to deal with two issues at the same time: 1) the already created digital records that need to be preserved, and 2) the establishment of an environment that takes care of these records right from the beginning. To achieve any longterm solution requires understanding of what is happening and possible in this new world. In order to be able to adapt to the new requirements and to act (or in some cases to react) this understanding is crucial. It requires imagination and thinking 'out of the square'. Thinking along traditional lines will not be sufficient, because of the innovative character of IT.

When things become difficult and confusing it is always the best approach to go back to the roots: what do we want when recording information, or creating records? Why is that necessary and what are the (archival and other) requirements? If we understand the functions and objectives of the area we are working in, we are also able to define the requirements and subsequently the necessary metadata.

The target of our efforts is the longterm preservation of information and/or records in digital form in an authentic, usable and understandable way through time. The latter three adjectives are also the main requirements we have to achieve.

3. Characteristics of digital documents

In order to be able to preserve it is necessary to examine what the characteristic differences of digital records are compared to paper records. As mentioned earlier digital records or documents are no longer fixed entities, because they have to be reproduced and rendered on the computer screen every time again. What we need in doing so, is at least hard- and

software and a datafile. The big issue is, how do we know every time again it is the same (intellectual) object we are looking at at the screen? What did we once, when the object was captured into a system, want to preserve? It was not the file itself, because that will not tell us anything. We do not understand the zeros and ones of which it consists. What we are interested in, is the document or information that was once created, presented on the screen and used in a business activity. This distinction between what I like to call the technical and intellectual aspects of a document or record is important, if we want to achieve the above mentioned target.

At the one hand there are the records as rendered on the computer screen and at the other the digital components as stored within the computer, needed in order to be able to reproduce them. It means we do not preserve digital records or publications as such, but only the ability to reproduce them.

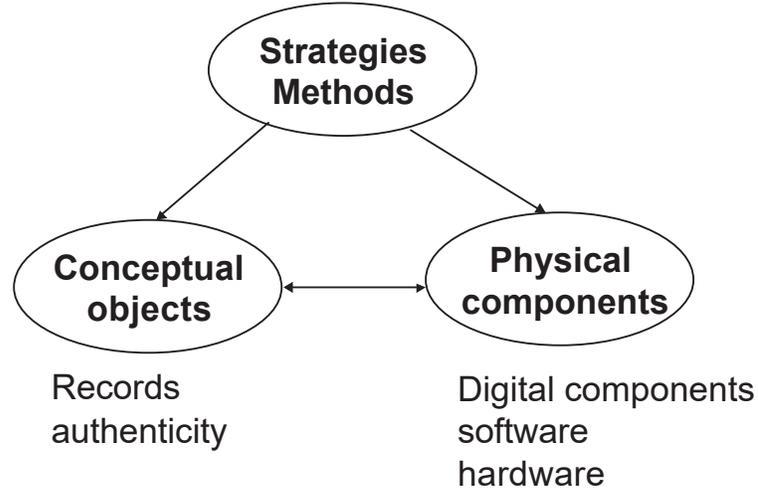
Every time one wants to read and use an electronic record it is necessary to re-assemble it. The computer with the operating system has to be started, the software started, and one or more datafiles opened before something will be shown on the screen. In reality one clicks (once or twice) on the icon of a document on the desktop and the document will show up. Nobody realises any more what is happening behind the screen. The hardware enables the software to run, the software contains the functionality to open and reproduce the datafile in the proper way. It may be that the data is stored in different files, such as may be the case for instance with a document that contains an image or a spreadsheet. These files may have also different storage formats. An example may be a Word-file which can consist of text only. As such it will be one digital component in one file. If the text document however, contains also an image for instance, that image might be embedded in the Word-file. As such the file will contain two (digital) components. If separate, then there are two digital components in two files.

This problem of different formats and objects might be reduced considerably by using (mostly de facto) standards, such as PDF, but even then it is not always possible to have only one digital component. An example of this is the use of XML. In that case there is at least a file with the tagged data and a DTD (document type definition).

Apart from the datafile and the hard- and software information about how the reproduction should take place (the method) and a medium on which the datafile is stored are required for and inextricably bound to the reproduction process.

This major change has also consequences for the management and control of the records we want to maintain. It will be more complicated. The different objects, the conceptual or intellectual objects (records or publications etc.) and the digital or technical components, and the interrelationships among and between them, need appropriate management. Especially the maintenance of all the relationships between the different digital components is crucial, in order to be able to reproduce and render the record(s) correctly on the screen. It will be the digital components, containing the data necessary to reproduce the records, that will change over time, because information technology will continuously change.

Graphically management can be shown as follows:



Figuur 1 Management of digital and conceptual objects

In short, the consequence of this 'fragmentation' is that preservation in a digital world mainly means apart from storage management, maintaining the ability to reproduce the (intellectual) object, either a record or a publication or another type.

4. Metadata

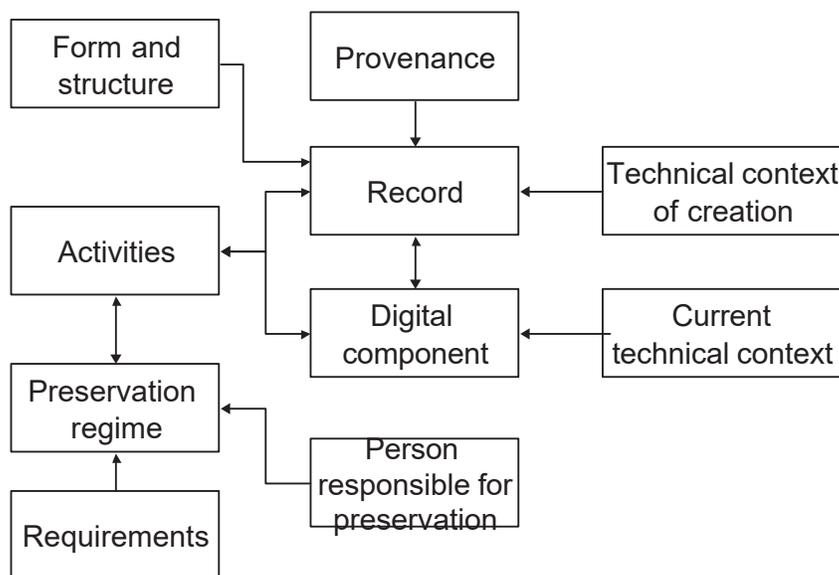
In order to manage and maintain digital records metadata play a major role. They are essential to establish what is reproduced on the screen is understandable and authentic. Volatile digital documents or objects are vulnerable, mutable, untangible and therefore require a stabilising 'mechanism'. That can to a large extent be achieved by description or metadata. Metadata regard information about the preserved records, either digital or paper, about the digital components and about both their management. The categories and elements of information needed depend on the requirements for preservation of the intellectual objects. As already mentioned at least three main requirements can be distinguished: authenticity, usability and understandability. Authenticity encompasses the ability to establish whether the record is what it purports to be, to identify its originator and the time when it was originated. Usability regards the ability of retrieving and reading the records, and understandability means the ability to understand the content in relation to the purpose for which it was created (the context). To comply to these requirements information about the provenance (who created them and why?), an identification and a description of the record(s) or object(s), and about the interrelationship between them (documentary context) is necessary. Apart from the content and context of the record as it was originally created and used, the form and structure and in some cases also the behaviour have to be described as well.

Apart from a description of the object, the management of it, since it was captured and stored, has to be properly documented. That is for instance stated in the proposed new ISO

records management standard and also in the Open Archival Information System reference model (OAIS).³

There should be a track record or audit trail of what happened since that moment. That will serve as evidence of the activities carried out with respect to preservation and maintenance of the records and the subsequent results.

In the following diagram presents based on these considerations a simplified entity-model with the main categories of metadata, that are necessary for preservation of records. It focuses on record and digital component and the management of them ('activities' and 'preservation regime').



Figuur 2 Simplified entitymodel for preservation

In practice most researchers or users will rely for the authenticity and integrity of the reproduced records or objects on trustworthiness of the organisation responsible for preserving them. They trust this organisation. Only when there is serious doubt about the way preservation is carried out, they will require explicit information and confirmation of the authenticity and integrity of the records.

The issue of metadata necessary for preservation is being discussed frequently the last couple of years. Projects such as CEDARS and NEDLIB have tried to identify the different categories and recently the OCLC/RLG working group on Preservation Metadata has conducted a study to come up with a framework. In this study three proposed sets of metadata are compared with the OAIS reference model. It concerns CEDARS, NEDLIB, and the approach taken by the National Library of Australia (NLA). The result shows that there is still some divergence of what is necessary, but on the other hand there is also a 're-assuring degree of convergence' as it is called.⁴

Several areas can be distinguished regarding metadata and they are related to the different areas or processes for which they are required. Apart from the preservation metadata, as

³ ISO Records Management standard (ISO/DIS 15489), which will be launched in October 2001. See for the OAIS Reference model (version 1.2) http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html.

⁴ '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. See www.rlg.org/

described above, areas as recordkeeping and discovery of information resources on the web have their own approach and sets of metadata. I will not go into details, but will try to give a brief overview. With respect to recordkeeping the Australian SPIRT-project is developing a set of metadata, which is partly being used by the National Archives of Australia as a basis for a standard.⁵ In the same area the Pittsburgh project, the UBC project and quite recently the Model Requirements (MoReq) for records management applications proposed sets of metadata necessary for managing and preserving records.

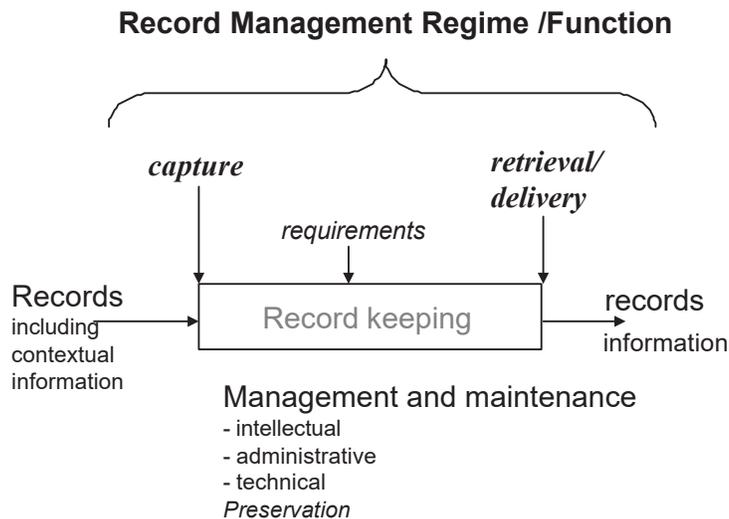
The *world wide web* poses other challenges, as for instance how to find information resources on the web. In order to enable this the Dublin Core standard has been proposed and is now widely used. Building upon this standard another initiative, called the Harmony project tries to establish a common exchange platform (?) which should enable the translation of different sets of metadata and as such improve communication between them.

Especially internet requires interoperability of metadata sets coming from different communities.

[---]

5. Strategies and methods

In order to preserve digital objects different activities have to be carried out. These activities will be governed by the preservation strategy adopted for maintenance. Roughly four main functions can be distinguished: capture, maintenance, access, and managing the preservation function. It also results in different types of control: intellectual, technical, administrative and records management control. In each case different metadata are acquired or captured. It results in different types of control: intellectual, technical, administrative and records management control. The following diagram shows these different perspectives. It is a high level model, shown as a black-box.



Figuur 3 High level model for a recordkeeping or preservation system

The Open Archival Information System (OAIS) offers a good reference model and is now widely adopted by different communities and goes into much further detail. In the area of longterm preservation many initiatives are trying to find strategies and methods that are capable of maintaining digital information so objects, such as document or publications, can be reproduced in their original, authentic form. At this moment two strategies seem to be predominant and sometimes promising, migration and emulation. Migration is now commonly used by most memory organisations for preservation. By migrating digital resources they can be kept accessible. The main characteristic of this strategy is that information is moved from an old, obsolete to a new platform. It includes mostly also a conversion to a new format. Certainly memory organisations tend to convert to standardised formats, such as XML, PDF or TIFF. It makes the management of the objects more cost-effective. The National Archives in the Netherlands requires that all digital records being transferred to be in a XML- and/or PDF-format. One of the advantages of XML is that it is rather software independent. Furthermore it enables to describe the structure and form of the objects, respectively by a Document Type Description (DTD) and a style sheet.

The other main strategy is emulation. This strategy aims at emulating old computers on new generations of hardware in using emulator software.⁶ It includes apart from the digital objects, the preservation of the original software. As such it enables to reproduce not only the original object itself, but also its behaviour. That might be important in case of multimedia or compound objects. There is a lot of discussion whether this is a viable strategy, because of the costs and the complexity. Apart from computer games there is not much practical experience.⁷

Some conclusions

The area of preservation of digital objects has caused a wide variety of activities that in the end should lead to strategies and methods that can preserve and maintain these objects. It is hard to keep up with everything that is going on, but it also leads to fruitful collaboration and information exchange at conferences and seminars. The common effort of all the different communities will therefore sooner or later come up with proper solutions and in fact the outline of it can be seen already.

Archives have to be pro-active when being involved in electronic records. It is already concluded in the ICA-guide.... Certainly regarding archival records.

The point is how to achieve that. There might be different approaches depending also on legal context. In some countries the archives have regulating power and as such can prescribe standards or other common regulations or procedures. In most cases however they have to seek more influencing ways of pro-activity.

One way is to be clear about the requirements for archival records.

In the Inter Pares project this model is analysed and decomposed further in order to understand better how to implement authenticity requirements into a system and give guidance to organisations how can be dealt with this issue. This project is building also on the OAIS reference model.

⁶ See for instance Jeff Rothenberg and Tora Bikson, *Carrying Authentic, Understandable and Usable Digital Records through time*, The Hague 1999. For some objections against emulation see David Bearman, *Reality and Chimeras in the Preservation of Electronic Records*, in: D-Lib Magazine, Vol 5 No 4, April 1999.

⁷ One of the projects that is focusing on emulation is the CAMiLEON project, a collaborative of the universities of Leeds (UK) and Michigan (USA), www.si.umich.edu/CAMiLEON.

5. Some conclusions

Maintaining and preserving digital objects requires to be pro-active, in order to be cost-effective. The point is how to achieve that? There might be different approaches depending also on legal context. Helpful might be the different existing reference models. They allow memory institutions to identify the different phases and the best places to be involved. Good examples of such approaches are the OAIS and Inter Pares models, although the latter is not yet published.⁸ These models also are the basis for applying and subsequently implementing different preservation strategies.

The question is furthermore will there be a possibility for establishment of an international metadata standard on metadata? The ISO RMS committee does see that as one of the follow-up actions for the records management standard.

Collaboration and information exchange between different institutions would be useful, both within and outside the community of memory organisations.

Helpful might be the different existing reference models. They allow archival institutions to identify the different phases and the best places to be involved. Good examples of such approaches are the OAIS and Inter Pares models, although the latter are not yet published.

The question will there be a possibility for establishment of an international metadata standard on metadata? The ISO RMS committee does see that as one of the follow-up actions for the records management standard.

It should also be clear that because of the legal situation in most countries nothing can be achieved without collaboration with the record creating agencies.

Apart from that collaboration and information exchange between different institutions would be useful as well, both within and outside the archival community.

Some references and further reading:

OAIS reference model (version 1.2, June 2001): http://ccsds.gsfc.nasa.gov/nost/isoas/ref_model.html

Inter Pares project: www.interpares.org

SPIRT project: www.sims.monash.edu.au/rcrg

CAMiLEON project: www.si.umich.edu/CAMiLEON

Mary Freeny (ed.), *Digital Culture: maximising the nation's investment. A Synthesis of JISC/NPO studies on the preservation of electronic materials*, London 1999

Seamus Ross, *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*, (NPO Preservation Guidance Occasional Papers) Glasgow 2000.

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 (www.rlg.org/)

⁸ OAIS: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html;

Inter Pares: www.interpares.org