# Permanent Access Technology for the virtual heritage

May 2004

The endeavour for continuous rendering.

Research Task IN4140
J.R. van der Hoeven

Assistance
TU Delft: Dr. K. van der Meer
IBM: Dr. R.J. van Diessen

TU Delft

IBM

# I. Preface

This document is the result of a research study on accessibility of digitally preserved information on the long-term. This research forms a prelude for my graduation task and is executed under the supervision of Dr. K. van der Meer from the faculty Electrical Engineering, Mathematics and Computer Science of the TU Delft. The research is done under the group of Information Systems Design of the TU Delft and within the scope of the Long-Term Preservation (LTP) project at IBM Netherlands N.V.

With the results of this research I hope to have created more awareness on the fragileness of digital documents, which to my opinion is very important to prevent loss of information now and in the future. Furthermore I hope to have given a good reflection of the current status on permanent access technology in the world. With this reflection, it is intended to help organizations making the right decision about which digital preservation strategy to follow. Finally I hope that this information leads to new initiatives and gives birth to new ideas and practical solutions on digital preservation.

Writing a report is a solely effort, but doing research requires more. During my research I gained a lot of support and appreciation from other people, for which I am grateful. Therefore I would like to thank all people that have helped me with my research. In special I want to thank the following people. First of all Dr. K. van der Meer from the TU Delft for his assistance and supervision. Also, I want to thank Dr. R.J. van Diessen for offering the opportunity to carry out this research at IBM Netherlands N.V. and reviewing the report. And of course Dr. H. van Wijngaarden (Royal Library of the Netherlands) and Drs. T.J. van Tuijn (IBM) for offering valuable information which has been very useful. Finally I want to thank Ms. N.J.C. Kol for investing her time to review this report and motivating me which led to this result.

Jeffrey van der Hoeven
May 2004

# II. Executive summary

Much of our time we are surrounded by a virtual environment. Surfing the Internet, typing a letter, playing computer games or watching the news, all of these happen in an environment that does not physically exist, but is represented via a computer or television screen. It has become a part of our culture and just like any other physical artifact that is specific for our culture, a part of it should be saved for future generations.
This report addresses the difficulties of preservation of digital documents and how to access and understand them in the future. Therefore the following research question is chosen: "Which of the current strategies regarding permanent access technology taken worldwide ensure accessibility over the long-term?"

Digital documents are different from traditional documents. They consist of streams of bits stored on a hardware device. The meaning of such a document is kept by the logical form in which the bits are stored. This can only be interpreted using appropriate hardware devices and software applications. Besides that, metadata (data about data) is often supplied together with the original document.
Preserving digital documents is a difficult task. The explosive growth of digitally stored information makes it almost impossible to efficiently organize this information. Therefore selections have to be made about which information should be preserved and which not.
As a second, IT developments on both hardware and software make preservation not easier. Preserving a document today could be inaccessible over two or three years from now.
Finally, authenticity of a document is at stake. A digital document can only remain authentic if its integrity is safeguarded and if it can be verified as 'the real one'. This is hard to satisfy because a digital document highly depends on its environment (hard- and software). Therefore this environment should be exactly recreated or the digital document should be transformed to newer formats without loss of its intrinsic value.

Considering these aspects, how can we ever preserve digital documents so that access and understanding of them is safeguarded? This question is addressed by permanent access. Different preservation strategies are considered worldwide based on permanent access technology to guarantee permanent access. In short these are:

- **Technology preservation:** build computer museums with all hard- / software created.
- **Saving the hard copy:** print everything on paper or microfilm.
- **Encapsulation:** supply every document with a self-explanatory description of the file.
- **Migration:** convert each document to a newer logical form.
- **migration on request:** do the same as migration, only at retrieval time.
- **Emulation:** virtually recreate the original environment of the digital document.
- **XML:** store the document separated from structure, content and layout.
- **Digital Rosetta Stone:** build a knowledge archive with specifications of hard-/ software
- **Universal Virtual Computer:** view documents on a platform independent manner.

All of these approaches have their own advantages and disadvantages, which makes that there is no on-size-fits-all solution. The first three approaches (technology preservation, saving the hard copy, and encapsulation) are less suitable than the others, because it is practically impossible or loss of information is inevitable.
Considering the six other strategies, migration seems to be suitable for common document formats which are widely supported while authenticity has not the highest priority. Emulation can be seen as a last resort for uncommon file formats, whereby authenticity of a document is important and initial costs are not an issue. XML is different in its kind because it tends to a uniform standard used worldwide. Despite the history of standardization (standards come and go), XML can become the standard of standards if it stays in business. It seems to be very suitable for preservation of e-mail, spreadsheets and text documents, although less for other

document formats, e.g. image files. The Digital Rosetta Stone (DRS) is a good theory, but a complete implementation of the model seems far away. Instead, migration on request is very practical and already tested with success for image formats. A disadvantage is that it is platform dependent. This does not hold for the Universal Virtual Computer (UVC) based approach. This strategy is the only one that is platform independent, applicable for all digital documents while offering maximum authenticity. But to make the UVC approach successful, it requires decoders and Logical Data Schema's to be developed at preservation time. This demands a lot of effort. Therefore more experience with the UVC should be gained to convince others of its potency and gain more support in development.

In general it can be stated that there is a growing attention on permanent access. Many organizations are developing (or planning) digital preservation repositories and are becoming aware of the difficulties of preservation of digital documents. Frontrunners are exploring the possibilities of different preservation strategies and many libraries and archives are watching the outcomes closely. In this report the most important preservation strategies have been discussed to find an answer on the main question of this research.
Based on these outcomes it seems clear that no one-size-fits-all solution is possible. Digital documents differ from each other in too many ways and are used for many different purposes by many different users. Organizations that are waiting for "the" solution will not be successful in preservation of digital documents. Risk management should be applied to find out which strategy is most appropriate for each type of document. Thereby considering how important the authenticity of a document is.

Although we are heading the right way, more work has still to be done in the field of digital preservation. First of all, more understanding is needed on preservation strategies. Besides this, the core of the problem should not be forgotten. The creation of so many file formats depending on all kinds of hard- and software over the last decades leaves us with the preservation struggles today. We are now in the position to solve this problem. No matter which actions are required, most important is that valuable information will remain valuable, accessible and understandable for future generations, helping our civilization forward.

# III. Table of contents

# 1 Introduction

If there ever would be a golden century of information technology, it would be now. Information can be of great value. Knowledge of history, expectations for the future, time schedules and statistics are all examples of what could be valuable information. Information forms our intellectual knowledge which could be shared amongst others. It is also endurable for an infinite length of time, but the way in which it is preserved is not. It will lose its value if it is not well conserved. When we speak of longevity of information we mean that it is not straight forward that this information will remain accessible and understandable forever.

During the twentieth century the way in which information is preserved and communicated has changed enormously. Information without communication is useless. The ability to share information forms the key factor to its value and is a great motivation behind the rapid developments of information technology. Digitization of information has proved to be an effective solution for both storage and communication. Storing information in digital form implies many advantages. In comparison with books and other paperwork which are transient, digital documents are not. They can be multiplied endlessly without loss of quality. Also, new technologies make it possible to communicate digital objects with anybody at any place. Digitally stored information is used everyday and everywhere in the world and seems to be the greatest evolution of the twentieth century.

But digitally stored information has its side effects which have become obvious the last decades. Because information in digital form is growing explosively in number and size it must be kept organized and stored in a well considered way [I-1]. This is a new and difficult task, especially for libraries and archives which are supposed to take care of preservation and retrieval of information. Preserving printed publications and hard copy artifacts have successfully taken care of, but preserving digital records forms a new challenge to libraries and archives. New guidelines and conditions are required which have to be specified by a wide community not only consisting of libraries and archives, but also software houses and government.

Secondly, information technology is still developing in a rapid speed. Each year an enormous amount of new releases of hardware devices and software applications appear on the market [I-2]. Hardware producers and software vendors are persuading their customers to use their new products, meanwhile accepting the redundancy of previous versions. Older media become obsolete as well as different older software application formats. This fast development is going on for years and does not seem to stabilize in the near future, which makes preservation of digital information even harder. The world is faced with a new problem of how to preserve digital information and moreover how to keep it accessible over time.

## 1.1 Research question

To offer more insight in the problem of digital preservation and seek for a solution to maintain accessibility, an answer is given to the following research question:

**“Which of the current strategies regarding permanent access technology taken worldwide ensure accessibility over the long-term?”**

To stop deterioration of digitally stored information, it must be carefully preserved which requires well defined guidelines and conditions. Today many libraries, institutions and other organizations are taking action to prevent loss of digitally stored information. Different agreements on standards and guidelines have been established and are taken into practice [I-3]. Some of the parties involved are already implementing a repository [I-4]. Although much work still has to be done, these first developments show us that bridges have been built between different parties of knowledge and solutions are on their way to be implemented.

But preservation is only the first step of ensuring that digitally stored information will remain valuable in the future. It also must be kept accessible over time. Therefore another step has to be taken, which ensures accessibility. Most organizations have become aware of this twofold problem and are discussing what should be done to prevent inaccessibility of their information [I-3]. Several approaches are openly discussed in theory and some practical steps are taken, but so far it is unclear which approach will be most appropriate (if any) [A-1]. Because digital preservation is such a broad topic which is engaged with everyone who is working with information stored in digital form, it requires a lot of co-operation between involved parties to come to a well-thought solution. To gain more insight in which approaches are taken concerning accessibility to digitally stored information on the long-term, worldwide actions will be investigated to reflect the current status and discover where this road is heading to.

## 1.2 Scope of the research

Digital preservation of information is a broad topic which holds many aspects to be considered. Because of that and the limited time available for this research task some limitations have been drawn.
Firstly, this research is mainly focused on a subset of all digital content, named static digital objects. The meaning of this classification will be outlined later in this report. In general terms software applications and Internet content is left out of the sub set and will not be taken into account.
As a second limitation this research will be narrowed to a merely technical level instead of focusing on all levels like administrative, organizational, juridical, procedural and policy issues.

By not considering these aspects a deeper insight can be reached into the technical issues specific for a selected group of information in digital form, within the available time.

## 1.3 Research subparts

To get more insight in the aspects concerning digital preservation and long-term accessibility, the research task is decomposed into a number of questions, which are answered in the different sections of this research report:

- How to preserve the virtual heritage? Related sub questions are:
  - What do we mean with the virtual heritage?
  - What are static digital objects?
  - Which problems hamper preservation of static digital objects?
  - Which design criteria must be taken into account for developing a digital preservation repository?
  - Are there any digital preservation repositories?
- What different technical solutions exist to keep static digital objects accessible?
  - What are the pros and cons of these?
- Which preservation strategies have already been taken into practice worldwide?
  - Are there any projects active in this field?

    - What are the benefits and disadvantages of each strategy?
- What does the future look like concerning accessibility of static digital objects on the long-term?

## 1.4 Research approach

This research assignment is performed as a prelude to the final assignment: the graduation project. In the first place, the research forms an introduction to the field of preservation and access of digital documents. Secondly, it entails the basis on which the graduation task can start, building on the outcomes of this report.

To come to an answer to the research question, the research is done by executing the following global approach:

1. Orienting in the field of digital preservation and access
2. Defining a problem area and scope
3. Splitting up the research question in different sub questions
4. Finding answers: search resources to amplify or attenuate the sub questions
5. Draw conclusions

These steps are done within a time constrain of three months. During these steps this report has been written.

## 1.5 Document outline

This report is divided into four chapters. The first part, chapter 2, describes what digital preservation means and why it is important to preserve digital content. It also indicates the difficulties of preserving digital documents and which design criteria are important to consider when creating a safe place for digitally stored information.
The second part, chapter 3, is an introductory to the different technical approaches for retaining access to digital documents on the long-term.
In chapter 4 preservation strategies taken worldwide are considered, naming current initiatives and discussing the pros and cons of each different strategy.
Finally, chapter 5 covers the conclusions and recommendations on the topic of this research.
At the end of this report a glossary can be found.

# 2 The virtual heritage

Much of our time we are surrounded by a virtual environment. Surfing the Internet, typing a letter, playing computer games or watching the news, all of these happen in an environment that does not physically exist, but is represented via a computer or television screen. It has become a part of our culture and just like any other physical artifact that is specific for our culture, a part of it should be saved for future generations.
This chapter is intended to give more insight in what digital preservation of the virtual heritage means, what the characteristics are of digitally stored information, in which way it can be preserved and which design criteria are important to successfully develop a safe place. Finally a short overview of past and current initiatives and standards is given, regarding the scope of this research.

## 2.1 What does digital preservation mean?

As their goal to preserve important aspects of our culture, libraries and archives are concerned with the task to preserve not only physical aspects of a culture, but also digital objects like electronic documents and computer applications from the virtual environment. The InterPARES project [A-8], an active player in the field of digital preservation, defined digital preservation as "the processes and activities which stabilize and protect reformatted and 'born digital' authentic electronic materials in forms which are retrievable, readable, and usable over time."
Because libraries and archives have little experience with digitally stored information, it makes preservation a true challenge. To outline why preservation of digitally stored information is a difficult process, a better understanding is needed about what a digital document is and which struggles exist on the way of preserving digital objects.

### 2.1.1 Aspects of digital documents

Roughly two categories of digital objects can be distinguished: static and dynamic. Static objects are stable and do not change over time, like a text document or image. In the rest of this report, this group of objects is denoted as digital documents.
Instead, dynamic objects do change by means of executing instructions. Therefore they contain (possibly machine-specific) instructions and could have an interactive user interface [A-2]. Programs as well as documents containing scripts or macros are all dynamic digital objects and require different approaches to be preserved then static digital objects. Animations can be interpreted as dynamic objects too because they change over time, but does not serve the same meaning as dynamic means in this document. If an animation consists only of a sequence of still images, it could be seen as a static digital object. On the other hand, Macromedia Flash animations allow user interaction and are therefore dynamic.

In principle, a digital or electronic document, or electronic record, is nothing more than a stream of bits in a particular order. The meaning of such a document, hidden inside the bit stream, is written in a logical form and can only be viewed making use of specific hard- and software. For preservation matters a digital document can thus be seen as a composition of a bit stream and a logical format [A-3]. These two components have to be well preserved otherwise loss of information will be inevitable.
Accessibility to these documents reveals a third component: the functionality to interpret the representation of the digital document [A-4]. By interpretation of a digital document normally is meant the rendering process which translates the bit stream from 1's and 0's into a more understandable representation for the human senses. Hardware like motherboard, controller cards and devices as computer screens and keyboard are examples of necessary components to

run software and offer interaction with a user. Software, consisting of an operating system and one or more applications can then interpret the bit stream and reconstruct the actual meaning of the document. Understanding the details of a format is important, because misinterpreting it makes the document useless.

Besides the digital document itself, other document specific aspects have to be taken into account. Metadata are a common part of records kept by libraries and archives. Metadata can be seen as specific information about a record. It could be information about the content of the record itself, but also administrative, technical or preservation information, like author, date of creation, version history or appropriate software used to view it. Metadata can be useful when searching for particular information or helpful to index records by category. Therefore metadata is of important value for a document and needs to be preserved as well.

### 2.1.2 Preservation struggles

Developing a repository for static digital objects implies many difficulties. Preserving printed work like books and magazines is successfully taken care of by libraries and archives, but preserving digital objects forms a totally new challenge.

#### Scale: how much information?

It is said that in 1981 Bill Gates of Microsoft predicted that 640 kilobytes of computer memory would be enough for everybody [I-13]. In 2002 an approximated 5 exabytes, a number with eighteen zero's, of new information has been stored on print, film, magnetic and optical storage media, of which 92% on magnetic disks [I-1]. Although Gates claims he did not speak that prediction out loud, back in the eighties he surely would not have believed the enormous amount of digital storage (although he would have liked it without doubt if he was told in addition what his share in profits is today).
Preserving such an enormous amount of digital content requires a well thought approach. First question to be answered is of course what is valuable information? Not all information has to be preserved, assuming that equivalents and redundant data can be omitted. Selections must be made and categories created, which in itself is a hugh task for libraries and archives.

#### IT developments: how fast?

That preservation of digital documents is a difficult process will directly become obvious when looking at the developments of the last decades. During the twentieth century, the Information Technology sector (IT) has grown enormously. From the second half of that century until today it is almost impossible to name all IT developments. Hardware devices such as tape recorders and 5 ¼ inch magnetic floppy disks were introduced but have already been abandoned by most computer users. Now information stored on these media is hard to access, if not lost because of mechanical, physical (magnetic) or chemical failure [A-10].
In advance, software programs like operating systems as MS-DOS or OS2 were widely used only twenty years ago. Today they are replaced by much more advanced platforms like for example Microsoft Windows, Mac OS or Linux. The same holds for application software such as text processors, spreadsheets, databases and many others. Unfortunately backward compatibility has shown not to be a primary requirement for software developers, causing programs written for older platforms having troubles when trying to run them on newer platforms. As a result documents created with these programs can not be viewed properly anymore and future accessibility to them is jeopardized.

A simple example shows how serious this problem is [R-1]. Using WordPerfect 5.1 to view an electronic document created with the same program in 1993 will not simply work on Microsoft Windows XP. The program could be installed, but severe changes to your configuration are needed [I-14]. Another possibility is to open the document with a different text processor like

Microsoft Word 2000. This approach works well when the WordPerfect to Word converter is installed as add-on, but the result differs from its original. Although the document is only eleven years old and still recognized by MS Word 2000 its layout is severely damaged (figure 2-1). The formatted text seems to be the same, but images inside the document have been altered or completely disappeared.

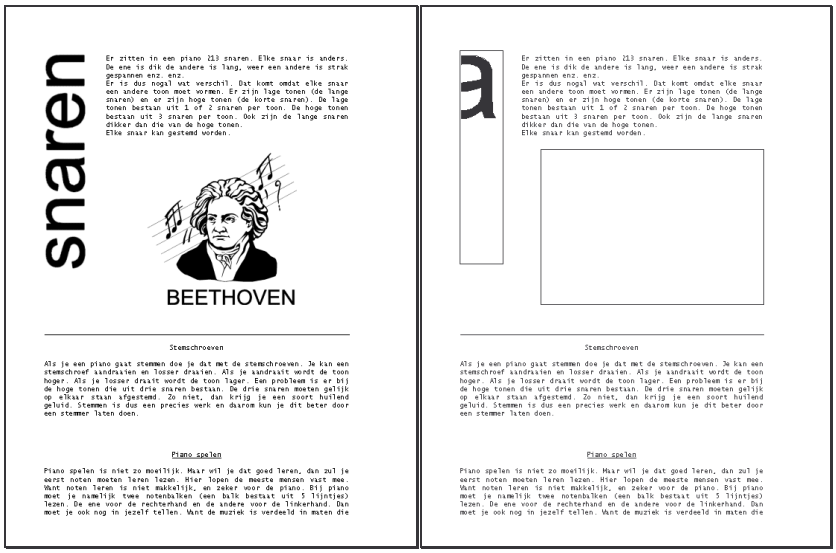



**Figure 2-1: Electronic document viewed in 1993 (left) and in 2004 (right)**

## Authenticity: what is real?

Authenticity forms a central issue in preservation. It is a broad topic with many debates ongoing about questions like: what is authenticity exactly? And how can we guarantee that documents remain authentic after they have been preserved for a long period of time?

It goes beyond the topic of this research to discuss authenticity in full glory and find answers to these questions. But as it forms an important aspect of the understanding of digital documents, a short vision will be given.

Authenticity can be explained by the fact that a document is what it is supposed to be. By definition of the Digital Preservation Testbed, authenticity is “the representation of a document conform the original determined version and opposed function of that document” [A-22].

Thereby authenticity is often linked to a couple of important issues. First, integrity should be safeguarded. Integrity stands for the fact that a document is intact and holds the original meaning and functionality. This does not free that a document could not be changed, however it should not affect the meaning or function of the document itself.

A second issue is verification, which gives us the possibility to determine if the document is what we think it is. A related aspect in this is identification. To verify if a document is indeed authentic, we should be able to identify that document as ‘the real one’. This requires that not only the document itself should be preserved, but also the context in which it was originally intended. Without information about the document itself, verification can not take place [A-2].

Another issue is reliability which refers to a document’s authority and trustworthiness, i.e. the ability to stand for the fact it is about [A-23]. Authority and trustworthiness are sometimes considered to be a combination of integrity and identification, because if one of these fails, the document is not trustworthy. On the other hand authority and trustworthiness are more subjective than integrity and trustworthiness because they are influenced by the human perception [A-2].

A final aspect of authenticity forms the intrinsic value of the document. This reflects the purpose of the document in its original way, but could be lost if the authentic document is changed. Satisfying all these issues is a difficult task, maybe even impossible because no document can be authentic for each and every unforeseeable future purpose [A-2].

## 2.2 How to design a preservation repository?

In the first place a digital preservation repository or archive should be developed to prevent loss of digitally stored information and with it a part of culture. Moreover, creating a preservation repository has the advantages of concentrating information in an organized manner and enables the possibility to interact with other repositories much easier. Well defined standardized repositories could be interconnected with each other creating an almost unlimited source of knowledge. Geographical locations are of no matter anymore and knowledge can be shared much easier and faster then ever before.

But how can we design such a preservation repository? Developing a digital preservation repository requires consideration of previous stated aspects of digital objects and its struggles to preserve them. Ideally, future users should be able to access, decipher, view, interpret and understand digital objects from the past. This can only be achieved when repositories and preservation process is independent of computing platform, media technology and format paradigms [A-5]. This implies that standards have to be developed, widely used, maintained and general concepts for information value including selection and authenticity need to be defined. These design criteria are hard to meet [A-2].

### 2.2.1 Standards are everywhere

Standardization is in itself a good basis for improving interoperability, interdisciplinary consensus on concepts, techniques and procedures and platform independence [A-11]. Without agreeing on some sort of standard it would be impossible to interpret and communicate information. The power of a standard lies in its wide spread use, but forms the weak spot of it at the same time. Practice has shown that establishing and using a standard with a wide surface is a difficult task. Standards require a lot of cross-domain effort and cost money. Besides that, different versions of one standard are often created, each interpreted differently and extended with proprietary value, resulting in a standard that is not so standard at all.

As discussed by Lorist and Van der Meer [A-11] four types of standards concerning digital longevity can be distinguished:

- standards for concepts, procedures and architecture
- standards for preservation of the digital document
- standards for preservation of access
- standards for interoperability

Not surprisingly, each of these types of standards are recognized and implemented differently by stakeholders involved. It goes beyond the topic of this report to mention all standards (if possible), but a short impression of the first three types of standards can be useful to create a better understanding on standardization of digital preservation.

#### Standards for concepts, procedures and architecture

**ISO DIS 15489** Enables organizations to standardize the terms and definitions used in records management, regulation, policies, responsibilities, requirements, design and implementation [A-11]. The forerunner of this standard was AS 4390, mainly used in Australia and North America. The ISO DIS 15489 is based on this former standard. It is consistent with AS 4390 which has now been superseded [I-20].

**DoD 5015.2-STD** Serves the same meaning as the above standard, but has been developed by the Department of Defence of the USA [A-20].

**OAIS** Stands for Open Archival Information System [I-15] and forms a reference model which describes the information flow as well as abstract components needed to develop a digital preservation repository. More details can be found in the next section of this chapter.

### Standards for preservation of content

**TIFF** Tagged Image File Format is an image format to represent raster images. TIFF has become a formal standard defined by Aldus and Microsoft, but today copyrights are in hands of Adobe [I-21]. Support is guaranteed and they offer great potential to be stable and accessible over the long-term, but are not suitable for all purposes.

**JPEG / JPEG 2000** Joined Photographic Experts Group [I-38] serves the same purpose as TIFF, but uses enhanced techniques for compression. Has become extremely popular due to the wide spread use on the Internet. Recently a new version has been introduced: Jpeg 2000, which satisfies features as lossless compression, thumbnail and metadata.

**PDF** Portable Document Format (PDF) is a proprietary standard of Adobe [I-39]. Through years it has become a 'de facto' standard as many libraries and archives use PDF as their official submission format. Although Adobe proclaims that PDF will be supported for a long time, no hard guarantee exist that PDF viewers will be available for all different platforms and backward compatible for different versions in the future. On the other hand, the PDF specifications are publicly available and many companies make PDF creation, viewing and manipulation tools.

**PDF/A** PDF/A is a standard being established to set guidelines for archiving and preserving digital documents in PDF format [R-4] [R-5]. PDF/A forms a subset of PDF, restricted in a several ways (like no audio, video, encryption allowed), but instead requires that metadata must be set. This metadata is defined using XML and enables the possibility to search and retrieve PDF/A documents much easier. PDF/A is not proprietary to Adobe.

**MS Word** Although never claimed to be a standard because the format is created solely by Microsoft, MS Word has a large share in the world's market on text processing formats [I-40]. Development and support is still controlled by software vendor Microsoft, which hold the same drawbacks as PDF, although a limited form of backward compatibility is supported. Besides MS Word other MS Office formats are widely used like Excel and PowerPoint.

**XML** A widely used, but still evolving standard is the eXtensible Markup Language (XML) [I-32]. It is a text-based markup language with the promising characteristic to separate structure, content, layout and context of an information object. Later on in this report more will be said about XML and its use in long-term preservation.

### Standards for preservation of access

**MARC** The Machine-Readable Cataloguing standard [I-41] defines the representation and communication of bibliographic and related

information in machine-readable form. It can be used for storage and exchange of bibliographic records. MARC is an industry-wide standard used by the British Library, Library of Congress and the National Library of Canada.

**Dublin Core** The Dublin Core (DC) [I-43] defines a set of 15 metadata elements for resource description and discovery, designed for cross-disciplinary networked discovery. As the name reveals it is just a core of elements. Often the DC set is extended to offer all necessary information that is needed to guarantee retrieval.

**EAD** The Encoded Archival Description (EAD) [I-34] was first developed in 1996 as nonproprietary encoding standard for machine-readable finding aids. In 2002 a new version was released based on XML, which specifies an extensive set of elements to structure data and could be used as supplement on the Dublin Core metadata set.

**METS** Metadata Encoding & Transmission Standard (METS) [I-44] is developed by the Digital Library Federation and attempts to build upon the work of 'Making of America II', a project which addressed issues concerning digital preservation. METS uses an XML based format to describe all necessary elements of a digital document. It is structured by 7 major sections in such a way that it is self-explanatory. An important aspect is that METS could be used in the role of the OAIS reference model.

Unfortunately many more standards on metadata, repository concepts and architecture, and interoperability between them exist. There is no one-size-fits-all standard, because they all have certain characteristics which are best for one but less for others. While standardization is very important for successful preservation of our intellectual knowledge, it is not succeeded so far [A-11] which is not uncommon in Information Technology.
As a general rule of thumb, it can be stated that people do not like rules and standards. They think on short term and like fancy exotic adventures for which they are willing to step over the border. With this in mind it is very hard to come to a wide accepted and used standard.

### 2.2.2 Co-operation

Despite the complexity of the problem, libraries, archives and other institutes have recognized the struggles and are converging towards a solution for preservation of digitally stored information. It has also gained political support whereas the EU and UNESCO are both setting up guidelines and standards for digital preservation [I-6] [I-7]. Because of the magnitude of the problem and the diversity of the stakeholders, co-operation on international and cross-sectored fields is of crucial importance [A-3]. Not only between the libraries and archives, but also between IT-companies which should take care of the actual design and implementation of preservation repositories.

## 2.3 Which digital repositories are there?

Different libraries, archives and other institutes in the world are currently active in the field of digital preservation. Many projects are running or have recently run, most of them in good co-operation to refine the ideas on this topic. In this section some important projects on repository development are outlined, followed by a couple of organizations who have implemented a preservation repository or are currently occupied with it. The list below is not intended to be complete, but serves to give a representation of the ongoing efforts on digital preservation worldwide.

### 2.3.1 Projects on architecture

#### NEDLIB

During 1998 until 2000 a group of eight European national libraries, together with three publishers started the project Networked European Deposit Library, in short NEDLIB. The project was led by the Royal Library of the Netherlands (Koninklijke Bibliotheek) and financially supported by the European Commission. Their main objective was to ensure that electronic publications are preserved well now and in the future. This resulted in the NEDLIB report series [A-9]. One of the most important outcomes is the agreement on the Open Archives Information System (OAIS) confirming ISO 14721:2003 [I-15]. The OAIS model is originally developed by the Consultative Committee for Space Data Systems of the NASA. It is a reference model for archiving information in both digital and physical form (figure 2-2). The model describes the whole process of acceptance (called ingestion) of information objects, storage and retrieval. When an information object is in its ingestion process, it is wrapped into a Submission Information Package (SIP). Next the process of preservation follows in which the SIPs are converted into Archival Information Packages (AIPs). Finally, if the information objects need to be accessed, they can be retrieved from the OAIS model following the dissemination process whereby the AIPs are converted into Dissemination Information Packages (DIPs). The OAIS reference model leaves the possibility open to extend the framework with a certain preservation sub system (upper part of the model), which makes it able to take precautions for long-term access and interpretation of the preserved data. Besides that, the OAIS is an open model that allows interaction with any other system, depending on its implementation [A-13].

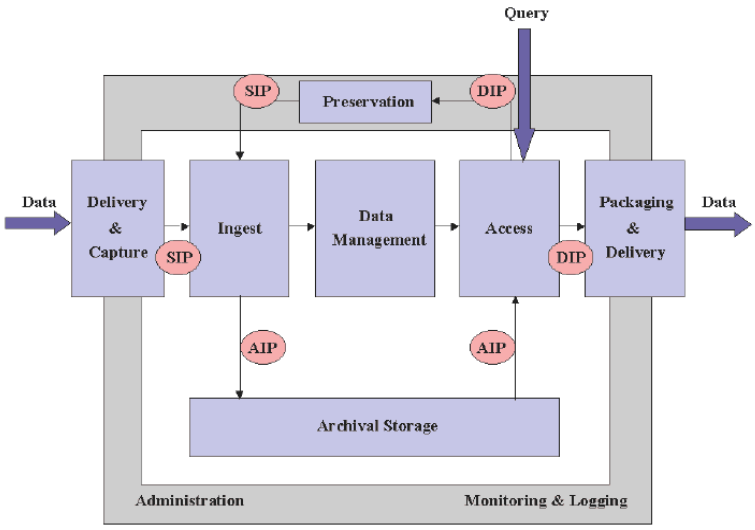



**Figure 2-2: OAIS model overview [A-13]**

#### LOCKSS

The shift from printed material to digital documents has inspired the founders of the LOCKSS project [I-28]. LOCKSS stands for 'Lots Of Copies Keep Stuff Safe' which praises the theory that the chance of information loss will be less if more copies are preserved. They state that publications are more often released in electronic form, soon superseding printed paper.

Therefore they proclaim that libraries and archives should react by acceptance of web journal publications as authentic items to be preserved, and abandon paper versions.
The LOCKSS program, formed by Stanford University Libraries, has developed tools which use local, library controlled computers to safeguard reader's long-term access to web based journals. The crux of these tools lies in the way ingested e-journals are stored. Using a network of computers, all ingested e-journals are stored in a number of caches on different systems. If one falls out, the document still remains available because other copies exist. In this, LOCKSS acts as a selective web cache. Although LOCKSS takes care of preservation and should be able to return the stored information as it was ingested, it does not provide a mechanism to interpret this information. To their opinion there is no need for it. If information in certain format is widely used and preserved, it will always remain accessible and understood. History has showed us however, that although the widely used hieroglyphics of the Egyptians was well preserved, it could not be understood until Napoleon's army found the stone of Rosetta.
In April 2004 the official release of the open source system based on the LOCKSS architecture became available, already participated by a great number of libraries, archives, universities (like Dutch Universities of Amsterdam and Maastricht) and other institutes.

### 2.3.2 Digital preservation repositories

#### National Library of Australia (NLA)

The NLA was one of the first actuators on the topic of digital preservation. They perform a leading role in the development of digital preservation repositories for electronic records in the Australian libraries and archives, and toke initiative to preserve certain information of the World Wide Web as well. To overcome the impact of preserving all information that exists, they follow an approach of selective ingestion, which means that they only preserve information that meets their protocol and standards. Different projects are funded like Pandora which stands for Preserving and Accessing Networked Documentary Resources of Australia. Pandora harvests and preserves information from networks, mainly the Internet, done by their Pandora Digital Archiving System (PANDAS) [I-9]. Other important projects are Picture Australia for preservation of pictures and its metadata, and PADI which stands for Preserving Access to Digital Information. PADI is an important platform that facilitates the development of strategies and guidelines for long-term access to digital information [I-10].

#### DNEP

In 1999 the Royal Library of the Netherlands (Koninklijke Bibliotheek, KB) started the project Deposit for Dutch Electronic Publications, in short DNEP (Dutch: Depot voor Nederlandse Electronische Publicaties). The aim of the project was to acquire a full-scale deposit system which offers large-scale, high quality storage and digital preservation functionality. The system requirements were based on the outcomes of the NEDLIB, such as the OAIS architecture [A-3] [I-4].
In 2000 the KB contracted the development of the deposit system to IBM Netherlands N.V. which took two years to build. In October 2002 the system was delivered to the KB, named Digital Information Archiving System (DIAS) composed from many off-the-shelf components, like DB2 database, Tivoli Storage Manager, Websphere and Content Manager. Together with a newly developed infrastructure by the KB, the integral solution is called e-Depot with DIAS at heart. It is now capable of storing 60.000 articles each day and has a scalable capacity over more than 500 Terabytes of data. In 2002 and 2003 KB has contracted the publishers Elsevier and Kluwer to preserve their electronic publications. Elsevier alone holds more than 5 million e-journal articles which are currently being ingested in the e-Depot [A-3] [I-11], today containing already about 2.9 million articles [I-12].
In parallel with the development of the e-Depot, IBM and the KB worked on a second project known as the Long-Term Preservation (LTP) project. The LTP project researched the possibilities to ensure accessibility of digital documents in the future. In 2002 the study has

finished and resulted in a report series of six documents [I-4] and in April 2004 a first implementation has been presented, based on the Universal Virtual Computer of IBM.

### DSpace

DSpace is the outcome of work done by MIT libraries and Hewlett-Packard (HP). It is a digital institutional repository that captures, stores, indexes, preserves and redistributes the intellectual output of university's research faculty in digital formats [I-18]. By C.A. Lynch [I-19], a digital institutional repository is defined as "A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members."
On that basis, DSpace can be classified as being a repository suitable for the academic world, to preserve and share information much easier. DSpace is freely available, is open source and serves a web-based interface to submit and retrieve information in many different file formats. More than ten universities across the world are using DSpace as their digital repository.

### Fedora project

In 2002 the Fedora project was funded by the Andrew W. Mellon Foundation to build an open-source digital object repository management system based on the Flexible Extensible Digital Object and Repository Architecture (Fedora). It is jointly developed by the University of Virginia and Cornell University. The latest available version (1.2.1 2004) uses web-based technologies, mainly leaning on XML and web services for ingest and export. Basic metadata is stored in Dublin Core format and communicated via the OAI Protocol for Metadata Harvesting. The repository also serves a migration utility to convert large amounts of data to a newer format. Today Fedora is deployed by different universities and libraries all over the world.

# 3 Permanent access

In 1086 the famous Domesday book was written to keep track of statistical information about Great Britain. In 1984 the British Broadcasting Corporation (BBC) started to digitize this book to let younger generations learn about its content. The project used state of the art technology as BBC master microcomputers and LV-ROM video disks for storage, and a total cost of £ 2.5 million. But during the 90s this virtual book has become inaccessible due to media deterioration, old fashioned microcomputers and inappropriate device readers, while the original book is still legible [A-6]. Thanks to the effort of the CAMiLEON project (which is discussed later in this document) the work and investments of the BBC are not completely lost [I-16].
This is just one occurrence involving accessibility to digitally stored information. Another example concerns the Viking Mars expedition in 1976 of which twenty percent of the data collected has become inaccessible. Or a magnetic tape containing photographs of parts of the Brazilian Amazon cannot be accessed anymore [R-2]. These are only a few examples of accessibility failures, but will not be the only ones. Most people are still not aware of the inaccessibility of their digitally stored information, until they need it. Who succeeded to preserve his digital documents for more than fifteen years should praise himself lucky if they can still be opened.

Saving digital objects in a repository is a first step but will not be enough. We have to ensure that preserved digital objects maintain both accessible and understandable over time. This is often called permanent access, but different denominators exist like continuous rendering or long-term access (this also depicts the newness of this topic). It means that it should be able to access, interpret and understand a preserved digital document not only at preservation time, but also over at least fifty or more years from now or until the document become obsolete. Different definitions of long-term exist. The Royal Library of the Netherlands for example uses hundred years as target period for their digital preservation [I-22]. Permanent access implies that appropriate action must be taken at preservation time. This requirement can be divided in two questions [A-7]:

1. What representation of the digital document should we save?

2. How can we ensure proper interpretation of that representation in the future?

## 3.1 What representation should we save?

Because of rapid IT developments of hard- and software, the format in which a digital document is stored becomes obsolete. Obsolescence means that the software needed to interpret the logical format of a static digital object cannot be executed anymore. This may occur because the format has become unpopular or is superseded by a new version of that format which holds a different paradigm. Another reason is that a logical format is proprietary to some company that has gone out of business [A-7].

During the last decades an enormous number of new software applications and a wide variety of formats have been developed. Approximately every two or three year new releases follow up older applications and formats and these developments will not slow down in coming years [I-2]. Standardization of formats could tackle the problem, but ICT standards have often proven to be easily forgotten by developers. It is more a weak spot instead of being a solution [A-2]. Besides that, standardization can limit innovation, because hard- and software developers have to confirm to globally decided rules instead of optimizing their solution for each particular problem.

Therefore, to answer the first question we have to predict which representation will remain useful in the future. This can only be answered when we know the answer of the second question.

## 3.2 How can we ensure proper interpretation?

To prevent inaccessibility of digital documents because of obsolescence, we have to ensure that documents are preserved in such a way that we can always decipher what the meaning was. Over the last years, permanent access has gained growing attention. Different approaches about how to maintain access to digitally preserved information are openly discussed under the term Permanent Access Technology (PAT), which will be outlined shortly. In the chapter about *preservation strategies worldwide* the most promising approaches will be discussed in more detail.

Of course we could just sit back and wait until current digital documents are not accessible anymore. We could hope that in the far future new techniques will be developed which solves the accessibility problem we are concerned of today, but that assumption is very risky. Fortunately six other approaches are currently being discussed [A-7].

### 3.2.1 Technology preservation

To view a digitally stored document in its original way requires the viewing software as well as hardware. Preserving not only the document, but also the needed hard- and software keeps the document accessible over time. Although this seems to be an easy and effective solution, it is not very realistic. Saving and maintaining all software applications and hardware devices ever build is a very costly and complex undertaking. Software support and replacement parts for defect hardware will eventually cease, while reproduction is difficult and costly. Besides that, skills are needed concerning how these programs and hardware work, which will become scarce over time.

### 3.2.2 Saving the hard copy

Libraries and archives have proven to be successful in preserving physical artifacts like books and papers. From this point of view it seems a logical step to transform digitally stored information back to analogues form. Although this is possible without loss of quality for flat data like text and images, it loses functionality and behavioral aspects for other types of documents [I-8]. Spreadsheets for example contain embedded formulas which are not saved when printing the sheet on paper or microfilm.

### 3.2.3 Encapsulation

Saving the interpretation together with the original document can prevent inaccessibility and is called encapsulation [A-1]. With the interpretation by hand we should be able to “read” the bit stream of a document in all detail at anytime. This will certainly work for simple file structures such as plain text documents written in Unicode format. But for more complex formats which embed dynamic, active or interactive behavior, encapsulation does not make it easier to interpret. To reproduce a representation of the document which is understandable for humans, decryption must be done and specific knowledge and skills are needed to retrieve the information from the document. Besides that, interpreting large quantities of digitally stored documents by hand is a time consuming process and error-prone.

### 3.2.4 Migration

The best known and widely applied preservation strategy is migration [A-8]. By the definition of the Digital Preservation Testbed [A-7], migration, also called conversion or transformation, is defined as “the transfer of files from one hardware configuration or software application to another configuration or application.” A simple example of migration is the conversion of a WordPerfect document to a Microsoft Word 2000 format, earlier discussed in chapter two.

Migration forms the centre of debates. Advocates of migration say migration is the only serious candidate thus far for digital preservation of large scale archives [I-30], while opponents like Jeff Rothenberg find migration error-prone, expensive and time-consuming [A-4]. Migration is discussed in further detail in the next chapter.

### 3.2.5 XML

XML stands for eXtensible Markup Language. XML is developed by the World Wide Web Consortium (W3C) [I-32] and designed to describe data in a simple, flexible way. It is a self-descriptive language with a clear separation of content, structure and layout. Each of these aspects is defined in different files. XML is based on its predecessor SGML, a standard from 1986. The content itself is defined in an XML file, structured by element tags which look like HTML (HyperText Markup Language). In figure 3-1 an example piece of code written in XML shows what this looks like.

```
<Report>
	<Title>Permanent Access Technology for the virtual heritage</Title>
	<Author>J.R. van der Hoeven</Author>
	<Abstract>...</Abstract>
	<Content>
		<Chapter>
			<Chaptername>Permanent Access</Chaptername>
			<Section>…</Section>
		</Chapter>
	</Content>
</Report>
```

**Figure 3-1: Example of XML code**

Element tags are defined with “< >”. Each tag has an opening and closing tag, containing the data. To know which element tags can be used and in which order, the structure of an XML file has to be defined. This is done using a Document Type Definition (DTD). In figure 3-2 a simple piece of DTD is shown for the XML code above.

```
<!ELEMENT Report (Title, Author, Abstract, Content)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Author (#PCDATA)>
<!ELEMENT Abstract (#PCDATA)>
<!ELEMENT Content (Chapter+)>
<!ELEMENT Chapter (Chaptername, Section+)>
<!ELEMENT Chaptername (#PCDATA)>
<!ELEMENT Section (#PCDATA)>
```

**Figure 3-2: Piece of DTD code**

Remarkable is that XML together with a DTD is directly usable for computer processing, while it is still human readable without the need of any special purpose application. However it is still flat text without any layout. Therefore a third component is needed to create a correct representation out of this semi-finished product, which could be done using one of the two following approaches: CSS or XSL.

CSS (Cascading Style Sheets) is developed out of the shortcomings of HTML. In 1996 W3C recognized that HTML has its limitations and developed CSS as a supplement on HTML. Today, it is also used to define the layout of XML. See figure 3-3 for an example piece of code.

```
BODY {
       font-family: Arial, Helvetica, sans-serif;
       margin-left : 0px;
       margin-top : 0px;
       margin-bottom : 0px;
       margin-right : 0px;
}

.result {
       background-color : #FFFFFF;
}

.resultheader {
       background-color : #BCBCBC;
}

.inactivelink {
       color : Gray;
}

.big_black {
       COLOR: black;
       FONT-FAMILY: Arial, Helvetica, sans-serif;
       FONT-SIZE: large
}
```

**Figure 3-3: CSS example code**

Another approach is to use XSL (eXtensible Stylesheet Language). XSL can be seen as an improved version of CSS, with extended features like transformation to HTML and filtering.
A detailed explanation of XML can be found at the W3C [I-32] and W3Schools [I-33].

In preservation terms, XML is a new, but potential player in two ways [A-22]. Firstly, it is platform and application independent, which clears the problem of incompatibility of digital documents. As shown by practice XML is perfectly suited to be used on the World Wide Web. Secondly, many organizations use XML to communicate with other organizations or processes, like web services. XML is becoming a standard language, forming a stable basis for XML as preservation strategy.

### 3.2.6 Emulation

As migration changes the digital object each time its logical form is near obsolescence, it abandons the original bit stream. And although XML seems to be a promising approach for newly generated information, it does not reveal an applicable solution for all data that is already created. In this, emulation could be of important significance, as ratified by Jeff Rothenberg, a great profounder of emulation [A-4]. The theory behind emulation is that it ensures the authenticity and integrity of the digital object over the long term and continues to provide access to it in its original environment. By definition of the Digital Preservation Testbed [A-7] an emulator “is a program that runs on one computer (the emulator’s ‘host’ system) and thereby virtually recreates a different computer (the emulator’s ‘target’ system). In this definition the word virtual denotes that the emulator functions like it is the original computer, but physically is not.
The term emulation is often confused with simulation, although these differ significantly. Simulation acts like what it should do for real, whereas emulation really does it. Think of a Formula 1 racing simulator, which gives you the experience that you are really driving, but is actually bounded to a mockup on a motion base and a world model created in software. An

emulator in this case would let you race an F1 car too, but you may actually drive a Lada, which forms a surrogate of reality.

In current debates emulation forms the counterpart of migration. For example Rothenberg proclaims that emulation is the only way to ensure document authenticity [A-4]. Other sounds have more skeptics to emulation, which of they think is too complicated and too great potential for error [A-1].

Considering these different approaches, migration, XML and emulation seem to be the promising strategies. In the following chapter migration, XML and emulation are discussed in further detail, together with different variations on these topics.

# 4 Preservation strategies worldwide

Having heard about permanent access and related technology, most libraries and archives are currently unsure which path should be walked to safeguard access to their virtual heritage in the future. As with all innovative developments, it first has to be proven that a certain approach really works and is safe to apply, before the majority is convinced of the benefits. To reach that point a process of experimenting and evaluating needs to be carried out, mostly done by a small group of enthusiasts. In this chapter more information will be given about different projects and platforms which are currently active in the field of preservation strategies. After that, today's practices on permanent access technology are considered.

## 4.1 Who cares?

During the last couple of years organizations seems to become aware of the fact that digital objects do not have an infinite lifetime, although they can be duplicated endlessly without loss of quality. In chapter two some activities on preservation strategy were issued shortly, like IBM, KB and PADI from the NLA. In this section other important projects and platforms will be outlined. Although the number of active projects is limited, it may occur that this list is not fully complete. In the first place it is meant to offer an overview of the most valuable initiatives.

### 4.1.1 CAMiLEON

The CAMiLEON project stands for Creative Archiving at Michigan & Leeds: Emulating the Old on the New. Not surprisingly they are solely focused on developing and evaluating a range of technical strategies for the long-term preservation of digital objects. It is a co-operative project of the Universities of Michigan and Leeds and funded by JISC and NSF.
The project has chosen emulation as their basis of long-term access strategy and faced their three main objectives [I-16]:

- To explore the options for long-term retention of the original functionality and 'look and feel' of digital objects.
- To investigate technology emulation as a long-term strategy for long-term preservation and access to digital objects.
- To consider where and how emulation fits into a suite of digital preservation strategies.

The project was started in 1999 and ended in 2003. In the duration of this project some remarkable outcomes were delivered. As later stated in this report, the CAMiLEON project was able to recover the results of the BBC Domesday project, which created a computer representation of the original Domesday Book but became inaccessible due to new technology developments. Also later discussed, CAMiLEON designed and implemented a new approach based on migration at retrieval time. Furthermore, they did some tests concerning the comparisons and differences between migration and emulation, and published several articles on this topic.

### 4.1.2 Cedars

The CURL Exemplars in Digital Archives [I-35], in short Cedars project, ran from 1998 until 2002. It consisted of three CURL institutions, the universities of Leeds, Oxford and Cambridge and was funded by JISC (the Joint Information Systems Committee of the UK higher education funding councils). The work Cedars focused on consists of five areas concerning:

- Preservation metadata
- Intellectual property rights

- Collection management
- Technical strategies
- Digital archiving prototype

The results on these areas were very meaningful, especially on collection management and costs, as presented in a series of documents. Cedars way of working is based on the OAIS model, building on the outcomes of NEDLIB. Furthermore, Cedars presented new insight in technical strategies, undertaken in co-operation with the CAMiLEON project. Interesting is that Cedars praises the preservation of the original bit stream, instead of converting periodically to newer formats. Together with CAMiLEON they suggested a new approach, called 'migration on request'. This will be outlined in more detail later in this report.

### 4.1.3 e-archive

The e-archive project [A-24] was executed by the university libraries of the TU Delft, Utrecht and Maastricht, together with the Royal Library of the Netherlands and the Netherlands Institute for scientific information (in Dutch: NIWI). The project can be seen as an extension to of the Cedars project. The main target of the e-archive project was to develop an architecture for academic archives preserving publications. Thereby a couple of important preservation questions were to be answered:

- What is the structure of data objects?
- Which preservation strategy to follow?
- What does the cost model look like?

The first question was answered by choosing XML as digital data format. They preferred XML because of its self-descriptive aspect, but stated that it is not said that XML will be the standard forever. The second question was investigated by looking at both emulation and migration as permanent access technology. Consideration of the last question delivered a business and cost model. The e-archive project emphasized that knowledge and appliance of rules and standards are of great importance to make preservation, interoperability and understanding of digital documents successful. In 2002 the project ended.

### 4.1.4 DARE

As a continuation of the outcomes of the e-archive project, DARE [I-36] started in 2003 as a new program under the Dutch partnership organization SURF to focus on the further development of academic repositories. DARE stands for Digital Academic Repositories and is a joint initiative of the Dutch universities, together with other organizations (KB, KNAW, NWO) to make all their research results digitally accessible. DARE aims on two main goals:

- Implementing the basic infrastructure by setting up and linking the repositories as agreed upon by different participating institutes.
- Starting and promoting the submission of scientific content to the repositories.

To guarantee access to the repository, DARE works together with the Royal Library of the Netherlands on long-term storage. Although results are not visible yet, the program seems to be promising. DARE will run until 2006.

### 4.1.5 PATCH

To guarantee future accessibility to digital repositories, a group of national libraries, archives, universities, research institutions and ICT companies united their forces to accelerate the development of tools for accessibility. They formed the PATCH project, which is the abbreviation for Permanent Access Toolbox for the digital Cultural Heritage [A-3]. The aim of PATCH is to create a technological framework, a so called 'PATbox', that will promote

continuous development of solutions for future accessibility. Furthermore a range of tools for permanent access will be developed and tested. PATCH proposes a new approach to develop an iterative process that creates both concrete tools as well as a general framework (the PATbox). In addition this PATbox should be able to connect with the OAIS standard. Tools will be created for four types of digital objects [A-19]: fixed-format digital items (static digital objects), scientific data sets, web resources, and applications (dynamic digital objects).
In April 2003 the PATCH-consortium submitted this innovative proposal to the European Commission. Although the EC described the PATCH proposal as good and innovative, they did not reward it with the requested funding [I-4]. In the official statement of the EC it was rejected because of lack of commercial sponsoring.

### 4.1.6 Digital Preservation Testbed

During the period from October 2001 to October 2003 a project group, labeled Digital Preservation Testbed (in Dutch: Testbed Digitale Bewaring) was active in the field of long-term access of digital documents, in particular documents created and maintained by the Dutch government [I-17]. The project was focused on fulfilling four primary goals. They wanted to develop:

- Concrete recommendations on how different types of digital files may be preserved so they can be retrieved and remain readable.
- Technical and functional requirements for a preservation system.
- Cost models for the different preservation approaches.
- Recommendations for further legislation and regulations concerning the management of digital archival records.

The project proclaims three main approaches for permanent access: migration, emulation and XML. They have been experimenting with different types of digital files and records in order to determine the best approaches to long-term preservation. E-mails, text documents, spreadsheets and databases are examples of digital objects which were taken into account. The results were described in a series of white papers, which will be discussed later on in this report.

### 4.1.7 PREMIS

Around 2000 many institutions recognized the importance of metadata, with some of them using in house developed schemes to record it. But little of these schemes were interoperable. With this in mind the Online Computer Library Center (OCLC) [I-23] and RLG [I-24] jointly sponsored the creation of two working groups in 2001 [I-26]. The first group was charged with identifying key attributes and responsibilities of trusted digital repositories. The second working group was concerned with identifying and describing metadata necessary to support the digital preservation process. In May 2002 both working groups ended and delivered a description of a metadata framework to support the long-term preservation of digital objects, as one of its important outcomes.

Although the results were satisfying, the metadata framework stopped short because of limits of consensus about how it should be applied in a production environment. Therefore OCLC and RLG sponsored a new working group: the PREservation Metadata: Implementation Strategies (PREMIS) [I-25]. This group started in June 2003 and used the metadata framework as starting point of their work to develop guidelines and recommendations for implementing metadata in support of the long-term retention of digital objects [I-26]. To succeed this primary objective, the group was split into two subgroups. The first concerned with defining a core set of preservation metadata elements. The second charged with identifying and evaluating alternative strategies for encoding, storing, managing and exchanging preservation metadata. The

identification is done by holding a survey under the general community of digital preservation. At this moment PREMIS is finalizing and evaluating its results.

### 4.1.8 InterPARES

The International Research on Permanent Authentic Records in Electronic Systems, in short InterPARES [I-45], focuses on the development of theoretical and methodological knowledge, essential to long-term preservation of authentic records in digital form. The InterPARES project is split up in two sequential sub projects. InterPARES 1 and 2.
InterPARES 1 ran from 1999 until 2001 and was focused on the authenticity of records, e.g. text documents. InterPARES 1 was based on the outcomes of an earlier project: the preservation of the integrity of electronic records, which led to the DoD standard 5015.2 named earlier in this report. One of the components of InterPARES 1 was dedicated to the exploration of issues related to permanent access to digital sound, a quite unexplored field of digital preservation. The outcomes of this sub project were reported in a book and used for the follow-up sub project.
InterPARES 2 started in 2002 and will run until 2006. This project aims to develop and articulate the concepts, principles, criteria and methods that can ensure the creation and maintenance of accurate and reliable records on the long-term. Results are yet to come.

## 4.2 Concepts and practices: what do we have?

The experiences from each project mentioned above have led to more insight about the consequences of each strategy. The knowledge that has been gathered so far is important to libraries and archives, because it clears the fog over different preservation strategies. These findings will be described in the following sections and will be compared to each other at the end by considering their advantages and disadvantages, cost, timescale, suitability, ease of appliance, and future success.

### 4.2.1 Migration

Migration is a denominator for various different approaches. The Task Force on Archiving of Digital Information proposed five basic ways of applying a migration strategy [A-17]:

- Hardware migration: change the media on which the information is stored
- Software migration: change the logical format of the information
- Incorporate standards into a preservation strategy
- Build migration paths
- Use processing centers to do it for you

Transferring information stored on a particular medium to another type of medium is called hardware migration. It is commonly used e.g. copying data from floppy disk to a newer media type like CD-ROM or printing attachments of an email to paper. Migration is something different than refreshing, which transfers data whereby the source and destination medium are of the same type.
Software migration has taken place when a digital document has converted from one logical format to another. For example a Microsoft Word 97 format can be converted to Word 2000. The converted file has now the logical form defined by Word 2000 and is thus migrated to this newer format. New releases of existing applications often support older formats of the same application, although not always visible to the user. Migration of digital objects is sometimes done without notice. When for example importing e-mail files to a newer e-mail management application, normally the content is silently migrated to a newer format. Furthermore formats of other applications are sometimes also supported, like Microsoft Excel supports Lotus 1-2-3 files. This is a pro for migration because no extra effort is needed to guarantee support for older and different formats, although file formats are most of the time only supported for two or three generations and it is not a guarantee for all file formats.
To overcome the diversity of versions and formats, it may be attractive to migrate to a certain standard. Standards, if well defined and accepted, improve the chances of surviving the deterioration of digital objects and give more grip on future accessibility.
Another issue stated by the Task Force is the use of migration paths. A migration path can be denoted as a provision for preservation as an integral part of the process or system that generates digital information [A-17]. It is important not only to migrate to a particular format, but also to keep track of which other formats are supported, by which companies and institutes, and what the expectations are in the future. The path which will be walked should be outlined before any action will be taken.
Finally, the Task Force suggests that it could be useful to develop specialized processing centers which are able to translate one format into another, using their knowledge and practices. This offers the benefit of scale and maximizes the use of uncommon technical expertise.

In the case of preservation of digital documents, software migration forms the centre of debates. Although hardware migration is taken care of with great success, migrating software has both its advantages and disadvantages.

### Considerations

Migration offers a strategy which is applicable to most common formats right away. The support of older formats by software applications creates the possibility to convert documents without having to develop other techniques. It is often assumed that software vendors of these applications offer the best conversion tools, because they fully understand the format. Although they are capable of offering backward compatibility, in practice this holds only for a limited number of generations and is not proven to be lossless. Moreover, it depends on a format which is still proprietary to a company, which holds a certain risk of continuity.

Furthermore migration does not guarantee that the authenticity of the converted digital object will not change. This is especially the case with interoperability [A-22], whereby digital aging is tackled by making digital documents less dependent on hard- and software. In this way a document could be converted from one application running on Windows to a same sort of application depending on Linux. To guarantee the authenticity remains the same, each aspect of a migrated digital document needs to be checked for changes, because alteration in the object can result in a different interpretation. Verifying all different logical transformations of a file format reveals an enormous effort.
The Digital Preservation Testbed [I-17] has performed some tests with migration and authenticity. One of these tests considered the authenticity of spreadsheet documents created with Lotus 1-2-3 and migrated to Microsoft Excel and vice versa. The results showed that although simple data was converted correctly, more complex sheets using different formulas and styles were severely damaged [A-22].

Third, migration is not a single event. It will probably be a periodic process, because migrated digital objects will eventually become in danger again if technological development moves forward. This means that migration of a digital document must be performed several times during its lifetime, making migration only an intermediate solution. To reduce labor this process can best be automated, requiring specific migration control applications. Multiple migration steps can also cause a loss of quality due to error propagation. Any errors or omissions from a transformation step will be propagated and be present in all following migrated versions [A-18], as depicted in figure 4-1.

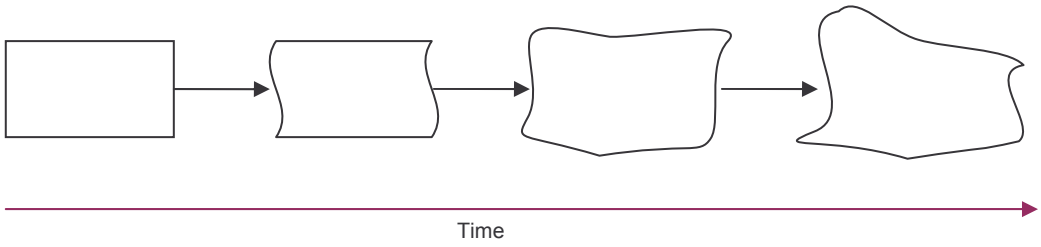



**Figure 4-1: Error propagation through each migration step [A-18]**

Therefore it is important that the number of migration steps is reduced. This means that migration does not necessarily have to be applied to every new version of a format, but only when critical changes in format occur [A-22]. A second prevention is to apply a certain risk assessment [I-42]. With risk assessment it is possible to track changes of a digital document over time. Although this limits errors and loss of information, slight mutations can still occur over the long-term.

Fourth, it is difficult to predict when migration is needed and to which format migration should take place. The success or failure of a target format depends on many factors which are probably not all known at the moment of migration. Of course it should be technologically safe, but other aspects as IT development of alternative formats and standards on preservation could have important influence to what might look like the right choice for migration [A-1].
As stated before different forms of standards on content exist [A-7]. Formal standards, like for digital images as TIFF (Tagged Image File Format) or JPEG (Joined Photographic Experts

Group) can be applied. Or proprietary standards, like MS Word which is very popular and offers lots of functionality. 'De facto' standards such as Adobe's PDF (Portable Document File), or still evolving 'de jure' standards like XML. Each of these standards has its own benefits and drawbacks, which makes it hard to justify choosing only one specific form of standard.

Migration reveals another uncertainty. It is impossible to know what document characteristics are important in the future. Migrating to a newer format can lose aspects which may be found important over fifty years. This requires risk management to investigate which aspects of digitally stored information have to be kept safe and which can be left out during a migration step.

Finally migration is only attractive if there are options to migrate to. If no new conversion tools exist, which could be the situation for unfamiliar formats, migration requires the development of such new conversion tools. This is very costly, technologically difficult and error-prone when not every aspect of a specific format is known. In that case, leaving the digital object in its original format without migration could be an option, but relying on backward compatibility is also very risky.

Considering these issues mentioned above, migration of digital documents is suitable for common formats, whereby the impact on authenticity of the document can be traced and limited. However, it may be risky to classify migration as a long-term preservation strategy because the influence and continuity of multiple migration cycles can not be fully overcome.

### 4.2.2 Emulation

Different types of emulation exist, which are categorized to hardware, operating system and application level [A-4]. First, emulation can be applied at application level (like emulating MS Word). This is emulation at the highest level, because it is still bounded to a specific operating system and hardware. It requires knowledge of how that application works, which in most cases is a complicated task and not always possible because applications are mostly proprietary to software vendors. Besides that an emulator should be written for each type of document viewer.
A second approach is emulating the operating system. This enables the possibility to run all original software once written for this platform. However, as pointed out by the Digital Preservation Testbed [A-7], operating systems (OS) cannot be emulated in the same way as one can do for applications. An OS itself does not control the applications. It offers services and enables user interaction flow, but does by no means control the applications running.
Finally, emulation can be done to mimic hardware by software, which is called software-emulation-of-hardware. In this way computer hardware like a processor is emulated by a software surrogate, as depicted in figure 4-2. The power of this form of emulation lies in its low-levelness. Understanding the way how the instructions of a processor work at logical level is relatively easy and better to validate than emulation at higher levels. It is therefore that software-emulation-of-hardware is preferred as the best way to apply emulation. However, the original digital records, software applications and operating systems need to be preserved.

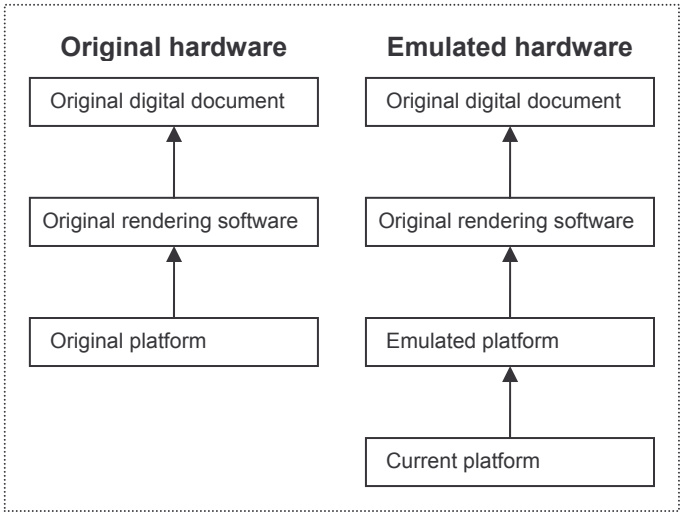



**Figure 4-2: Hardware emulation**

Emulation is not new. It has been applied for years already and for many purposes, e.g. computer Company Apple created an emulator for their new PowerPC processors. To continue supporting the older Motorola 68000 processor, they emulated this processor on hardware on PowerPC [A-7]. They did the same for running the Macintosh Operating System on Intel based machines. Instead of the specific PowerPC chipset on which Mac OS normally runs, Apple created an emulator running on an Intel chipset offering a PowerPC interface to the operating system.
Gaming is another important field in which emulation plays a prominent role. Emulation forms the key to run computer games written for an older, abandoned platform on a current machine. Numerous emulators are written and can be found on the Internet [I-31].

### Considerations

In preservation terms, emulation makes it possible to view any digital object while it maintains its original form. This is a great pro for emulation, because it does not thorn at the authenticity of a digital document, while migration entails new validation. Neither does it require changes or rewriting of the original software needed to interpret the logical form and to view the document. Instead the original program should be able to run in executable form on the emulated environment as it did when running on the original hardware. This makes the knowledge about how this software was build superfluous, which reduces labor and costs.
As a second, no periodic transformations have to be made to archived information, leaving error propagation to a minimum and reducing costs as well. In this case, a single emulator could serve as platform to run many original applications, which enables to view even more types of original documents.

A disadvantage is that an emulator must be developed, which is able to run on future machines and emulate preserved digital objects in their (virtual) original environment. This embarks an important question: how can we ensure that the emulator will be able to run correctly on future computers, without knowing which platform will be available at that time?

In a paper on digital preservation by the CAMiLEON project by S. Granger [A-21], this question is partly answered by linking back to the theory behind Turing machines. Invented by Alan Turing back in around 1930, a Turing machine is defined as one of the simplest computers one could think of. The theory Turing described, can be visualized by thinking of a large linear tape marked into squares and both sides of infinite length (figure 4-3). Each square can be blank or hold a symbol out of a finite set of symbols. The Turing machine itself can be seen as a device equipped with a read and write head, sliding over the tape square by square, and a control unit. At each square, the machine scans the block beneath it and reads the content of it (blank or a symbol). Then the control unit can perform the following actions based on the symbol read in the square:

- Halt the computation
- Move one square to the right
- Move one square to the left
- Write a symbol in the square

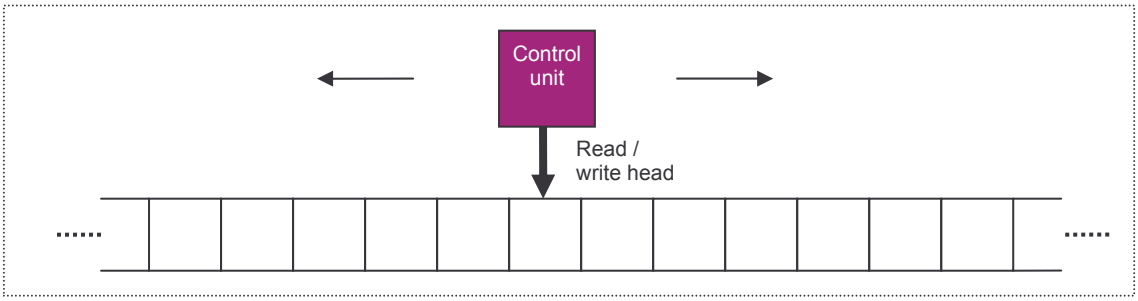



**Figure 4-3: Visualization of a Turing machine**

Based on this instruction set, programs can be written which describe what the Turing machine should do in each transition. See [I-27] for a demonstration of a Turing machine implementation.
Although a Turing machine seems trivial, it is capable of solving any effectively solvable algorithmic problem [A-21]. In general, as defined by Harel [A-12]:

> "…any algorithmic problem for which we can find an algorithm that can be programmed in some programming language, any language, running on some computer, any computer, even one that has not been built yet but can be built, and even one that will require unbounded amounts of time and memory space for ever larger inputs, is also solvable by a Turing machine."

With this theory in mind, Granger of the CAMiLEON project maps this argument back on emulation: "In effect then, the most powerful super-computer can only solve the same class of problems as the simplest home computer. Non-computable (undecidable) problems are solvable on neither, and the computable (decidable) problems are solvable on both."
Thus an emulator in itself can be seen as a basic Turing machine. If Turing machines can solve any algorithmic problem, a future emulator on any future super-computer could also do it. Because computer programs are not more than logical algorithms, this ensures that future computers would always be able to run any of these computer programs.

Although this answers the question by means of computability, it leaves issues as performance, peripheral devices, I/O interfaces and thus development cost uncertain. These issues are hard to predict. However, CAMiLEON points out that creating a certain form of abstraction could help writing emulators more cost-effective. Although emulation offers the ability to run a great amount of digital objects developed for this particular platform, the emulator is useless if a new platform has been developed. As depicted in figure 4-4 abstract interfaces could be created between digital object and the emulator (orange spot) at the one hand and between the emulator and host platform (green spot) at the other. When the host platform is replaced by a different one, the abstract emulation platform needs to be changed, but the upper interface (orange spot) remains the same, accelerating development.

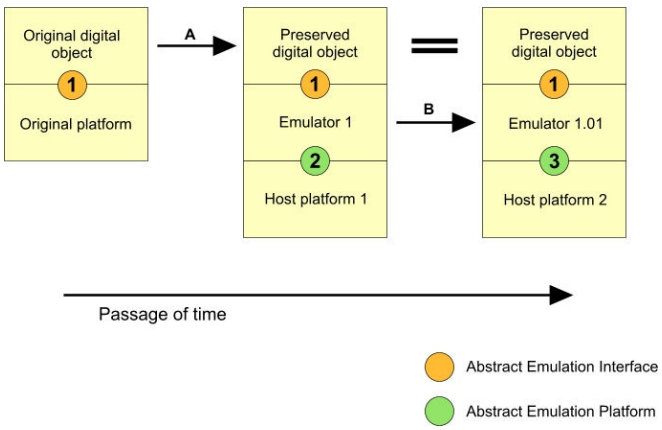



**Figure 4-4: Abstract interfaces [A-21]**

Jeff Rothenberg, one of the greatest supporters of emulation, proposed a slightly different approach. He thinks of an 'emulation virtual machine' as depicted in figure 4-5. As stated by Rothenberg a virtual machine is "something that serves the role of some computer that does not actually exist. Virtual machines are typically generalized or simplified analogues of real computers, implemented by software."
This suggests that a virtual machine could operate as an intermediate layer between the physical hardware and the emulator. The advantage of this approach is that different emulators for different platforms could run all via the same virtual machine. Also, if new platforms replace

older ones, the emulators once written need not to be rewritten because the interface of the virtual machine will remain the same, in contrary with the CAMiLEON approach. However, a difficulty still remains: how to build such an emulation virtual machine which supports all requirements of multiple emulators? This question still remains unanswered.

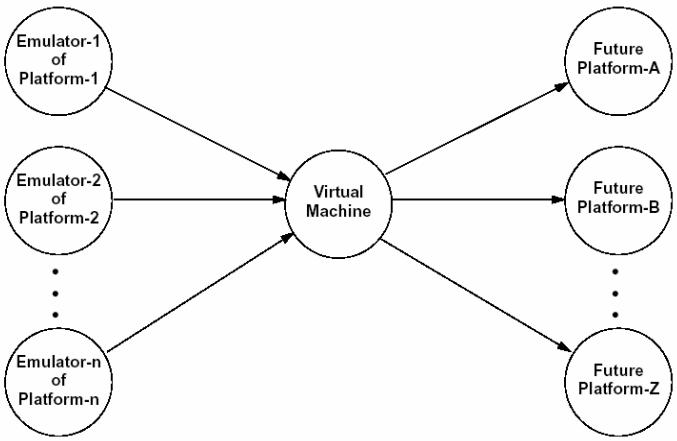



**Figure 4-5: Emulator Virtual Machine [R-3]**

Furthermore, criticism is heard from David Bearman, quoted in the article of Granger, as he states that a big problem of emulation is that emulation breaks with the right protected formats and software, which are proprietary to software producers.
Another statement in Grangers article is that specific knowledge is needed about how to control the emulator, software applications, formats and computer peripherals. It may seem to be common knowledge but how often have we been frustrated by installing new hardware or software products?

## Migration versus emulation: a computer game

Migration and emulation both seem to have their benefits and drawbacks. To get a better insight it is useful to make a comparison and put them to the test. In 2001 a test between these two methods of preservation had been performed as part of the CAMiLEON project, conducted jointly at the University of Michigan and the University of Leeds in the United Kingdom. At the end of 2001 the test results were published in the article “Emulation vs. Migration: do users care?” [I-29].

Their goal is to assess the needs of users and requirements for preserved information. They state that user testing can play an important role in understanding user requirements by assessing user needs in relation to archived digital objects that are preserved via different methods. As a test bed they had chosen a computer game called Chuckie Egg (figure 4-6), which was a popular game in the United Kingdom around the mid-eighties, running on an IBM microcomputer. Three different versions were available: the original game, a migrated version and an emulated one. The first one was installed on a BBC Microcomputer with all original equipment set up. The two others were installed on modern machines.


**Figure 4-6: Chuckie Egg**

A group of participants was asked to play the game on the original machine for one hour. After that, the group was split into two smaller groups. The first group played the migrated game, the

second the emulated one. Eventually their opinions were asked about aspects such as satisfaction, ease of use, performance and differences between both played versions.
The results were interesting. As small differences were noticed between the original and both migrated and emulated version, no statistically significant advantage or disadvantage was stated between emulation and migration. But perceived differences, one of the aspects that were measured, showed a lot of small variances between emulation and migration. Things like “The character in the game can jump further.” and “Sounds are a bit sloppier.” were noticed by users. But also hardware changes were marked like softer keys or no blinking screen. This states that look-and-feel does not stop by software but also hardware should be taken into account.

Despite the fact that no significant differences between migration and emulation were encountered in this test, the circumstances were ideal. Being in the position of having a program that is both available in emulated and migrated version is rather unique. If we are not able to run the original program to see what it looks like, we would not be able to build a migrated version of it. In such a situation emulation would be our only option.

### 4.2.3 XML as strategy

In the introduction to eXtensible Markup Language (XML) described in the chapter *Permanent Access* platform independence of XML is marked as a great advantage. This independence is created by the uniform character of XML and its separation of content, structure and layout. Other pros for XML named earlier, are its simplicity to process by computers, its readability for humans and its growing use worldwide.

The Digital Preservation Testbed subscribes these benefits in their white papers on digital preservation. Furthermore, they tested the usability aspects of XML for preservation of text documents, e-mail and spreadsheets, to advice the Dutch government in choosing an appropriate preservation strategy. The Dutch government is required by rule to preserve archival pieces for about hundred years and should be able to access and view these documents at any time of preservation. In comparison with XML the Testbed tested migration and emulation as other approaches.
The results were published in a series of advisory documents to the government, but also publicly available [I-17]. In general, the Testbed qualified XML as the best preservation strategy to remain access on the long-term. They pointed out that XML can be applied by two ways: the application directly stores the data in XML format (preferred), or the data is later converted to XML. This is depicted in figure 4-7. The great benefit is that future applications should be able to render XML without any extra needs. Condition is of course that XML will be stable and widely supported by future software applications all time.

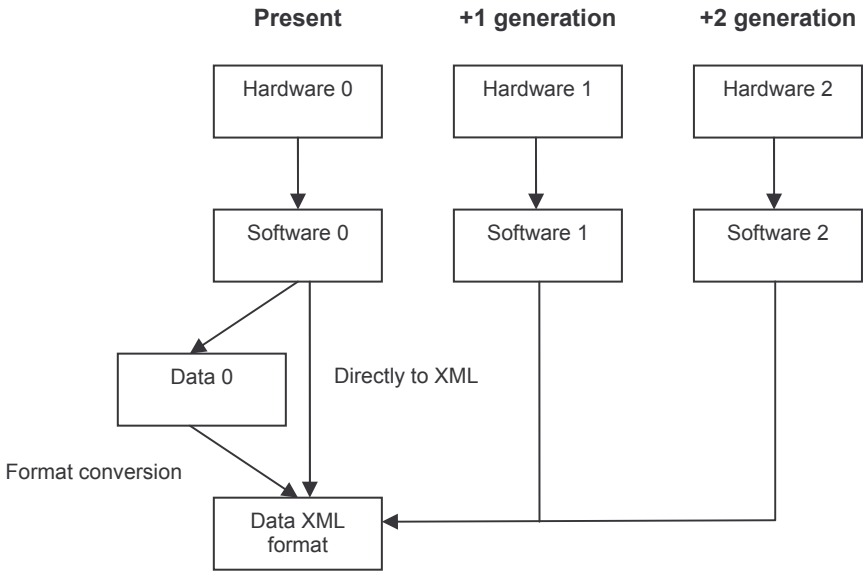



**Figure 4-7: Conversion to XML, now and in the future [A-22]**

For e-mail to XML [A-25] they defined two approaches: post-use and pre-use. Post-use is meant for already existing and received e-mail messages and pre-use for messages newly created and sent. With pre-use e-mail messages that are sent (for now only using MS Outlook) are passed to a web service which converts this message to XML. Then the XML is stored in a central repository. Besides an HTML version of the e-mail is sent back to the mail client (MS Outlook) and ultimately forwarded to the SMTP server responsible for sending. Furthermore, the user has to enter some metadata which is stored together with the e-mail message. Post-use follows a similar approach, only at receive time.
Attachments are treated differently. They are both stored in encoded form together with the XML message, as well as in decoded form (a stream of bits). Although the message and original binary of the attachment is safeguarded, following this approach gives no guarantee if the attachments will be accessible and understandable in the future [A-2]. This suggests that other

preservation techniques should be considered as well. Relying on migration or emulation is an option.

For preservation of text documents [A-26] over the long-term, both XML and PDF formats were designated as good candidates. For short-term and medium-term preservation (less than ten years) migration is also a candidate, but the disadvantages of migration (loss of quality after each migration step, uncertain target format, etc.) make preservation on the long-term less suitable. Also emulation is mentioned, but is not found suitable yet for preservation by the Testbed, because of its complexity and ample design, development and testing time and effort.
If the structure of a text document is well defined (by using chapters, paragraphs, sections and more) the document can best be stored in XML format. A pro is that XML tools are widely available to convert a document into XML format. For example, Open Office already uses XML as its official format to store documents. For text documents which are not neatly structured, the Testbed recommends the PDF file format as the best option to guarantee long-term access. Adobe, the founder of PDF in 1993, supports a large list of PDF viewers, creation and manipulation tools. Besides that, the specification of PDF is now publicly available which has lead to even more PDF specific tools. By using PDF a text document can be stored in almost identical representation as the file initially looked like.

Preserving spreadsheets, the last of the three tested formats, can best be done in XML as stated by the Testbed. Migration tests have shown that conversion of spreadsheet formats lead to significant loss of information. Emulation is for the same reasons as for text documents not considered a practical solution and conversion to PDF means loss of embedded information, like hidden formula's and multiple spreadsheet fields. Based on these outcomes and on the advantages of XML, this format is pointed out as the best way to preserve spreadsheets for a long period of time.

The Testbed noted that the ideal and safest situation for preservation of text documents and spreadsheets, is betting on multiple horses. In this, they suggest preservation of the original document, together with a PDF version and an XML version of it.

### Considerations

Although the Digital Preservation Testbed prefers XML for e-mail, text documents and spreadsheets, it leaves other types of documents unsure. For example, images could also be converted to XML. But if file size is taken into consideration an unwanted effect occurs. Preserving an image in XML with well defined tags explosively grows in size [I-37]. A small image (less than 100 by 100 pixels) can easily consume about one megabyte of storage. If images with print quality (about 300 dpi) are stored this could soon exceed 1 gigabyte of storage! In appendix A of this report the XML output is shown for a 2 by 2 pixel image, containing two pages of code already. Of course tag names could be shortened, saving some bytes, but lowers the readability of the code. Other multimedia formats like sound, video and 3D simulation models (VRML) will even be more difficult to manage regarding to size.

As a second, XML gives no guarantee on a well structure and interoperability. Even XML could be badly defined. Even if its human readable, it does not automatically mean that it is human understandable. Furthermore XML processors could interpret the style sheets, defined by an XSL or CSS, differently resulting in different layouts. This influences the authenticity of a document.

Finally, an important question remains: will XML stay? In general it is hard to predict if XML will remain for a long time. As with many other formats, they come and go. In particular XML is a very young standard and is still in motion. The old DTD format is slowly being replaced by the more flexible XML Schema of W3C. Furthermore, XSL and CSS are still both working standards on layout, but it is expected that one of these will win (and the other lose).

### 4.2.4 Digital Rosetta Stone

Without the discovery of the Rosetta Stone the new world would not have been able to understand the ancient Egyptian scripts which have been preserved on numerous artifacts. The Rosetta Stone formed the key to decipher the hieroglyphics, because it holds an inscription of a decree issued in 196 BC by Ptolemy V Epiphanies. This inscription was repeated two times in two languages written on the same stone: Greek and Egyptian. Fortunately enough of the ancient Greek culture is still understood, making it possible to decrypt the Egyptian language on that stone. Today, because of the Rosetta Stone, we are able to understand the valuable inscriptions on Egyptian artifacts.

In this context, Heminger and Robertson [A-27] made a reflection to the digital world. Many parallels can be drawn between the ancient periods of which many artifacts have been conserved, and the modern world in which information in digital form reveals much of our culture today. But just like with the Egyptian scripts, we have to understand the concepts of information in digital form. To overcome this problem Heminger and Robertson think of a so called Digital Rosetta Stone (DRS), a key to forgotten information in digital form.

The DRS is a model for maintaining long-term access to digital documents. It contains multiple levels of knowledge about specifications and processes by which information is stored on various types of storage media. Strengthened by the words of Rothenberg [A-4] that if an information system is sufficiently described, then future generations should be able to recreate that system's behavior and bring the old digital documents back to life. Although Rothenberg also states that such a description cannot sufficiently described, Heminger and Robertson are positive that a DRS model can succeed.

The DRS consists of three processes: knowledge preservations, data recovery, and document reconstruction, as depicted in figure 4-8. With these three processes each digital document should be reconstructed and interpreted. Each process supports the subsequent one.

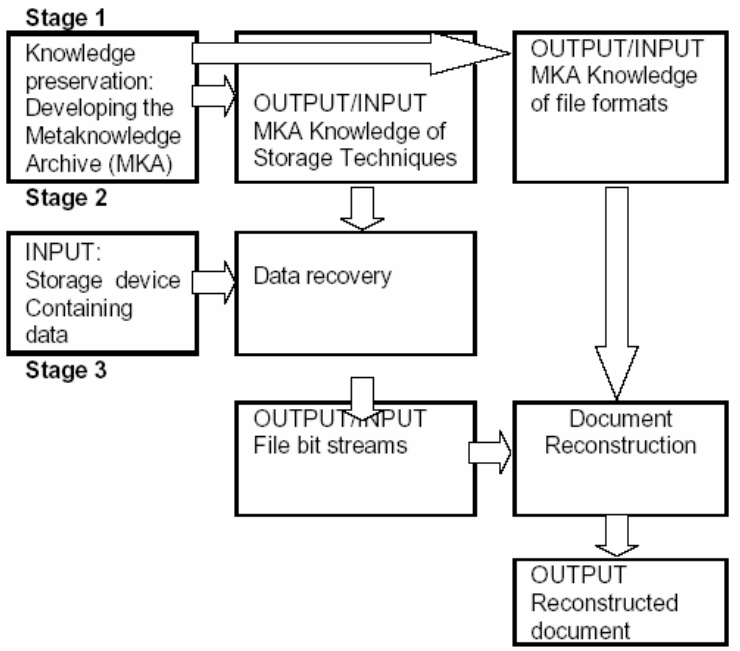



**Figure 4-8: Digital Rosetta Stone model**

The first process, knowledge preservation, is needed to recover digital data and reconstruct the document. This knowledge forms the basis of the system and will be preserved in a metaknowledge archive (MKA). All facts, heuristics, rules, and the context in which it is placed

need to be preserved in two key areas: media storage techniques and file formats. With these two key areas enough information should be available to support the next processes.

The second process, data recovery, is initiated when it is not economically feasible to maintain antiquated hardware systems. Data recovery forms the key to recover data from superseded media. Based on the information gathered during the knowledge preservation process, obsolete media could be reconstructed and ported to newer media. Thereby it is not always necessary to rebuild old hardware devices, such as punched-card readers. Heminger and Robertson demonstrated that punched-cards could efficiently be read not using a reconstructed card reader, but by simply scanning the cards and developing a software application that could interpret the scans. Today's devices can be used to interpret media from the past.

The final process, document reconstruction, builds upon the outcomes of knowledge preservation and data recovery (figure 4-8). With the recovery of hardware media and the gathered information about the format in which the digital data is stored, it should be possible to recreate all the aspects that a digital document once had.

### Considerations

Ideally, the DRS gives a clear and stepwise approach of how digitally stored information could be recovered. However, recovery means that we first lose something. This makes the DRS a risky undertaking, because if recovery fails as result of shortage by one of its three processes, accessibility is at stake. This could for instance be the case if an old medium is not accessible anymore, because of magnetic or chemical failure [A-10]. Magnetic disks can lose its magnetism over time and tapes could fall apart. If something like this happens, nothing could be done at it.

Furthermore, depending on one institution which holds the magic key to historic digitally stored information is also dangerous. What if an earthquake destroyed the Rosetta Stone?
However, developing a DRS would be important. Even if only the first process is carried out it would help other initiatives on preservation strategies, like emulation. But this step alone requires a great amount of effort and time [A-27].

### 4.2.5 Migration on request

It has been stressed that migration is a practical solution for common static digital objects, although it has quite a lot of disadvantages as denoted in the previous section. It is only an intermediate solution and not cost effective for uncommon file formats. One of these drawbacks which happen during multiple migration steps over time, is error propagation. When a digital document is near to become obsolete, it can be migrated to a newer format. Although this transformation should be done with care, differences between the original authentic document and the target document can occur. During following migration cycles, needed if subsequent formats become obsolete too, these differences are propagated and possibly amplified throughout the life-cycle of the migrated digital document, as depicted earlier in figure 4-1.

The CAMiLEON project [I-16] has investigated the possibility of applying migration in a more sensible and effective way and came up with a new approach named 'Migration on Request' [A-18]. Instead of migrating a digital document each time the format is in danger, with migration on request the original bit stream is preserved and migration is only applied at retrieval time. Therefore a special tool needs to be developed which allows the original document to be converted to the most current suitable format. With this approach error propagation is eliminated, because migration is done only once, while 'traditional migration' iterates over time, like shown in figure 4-9 (left part).

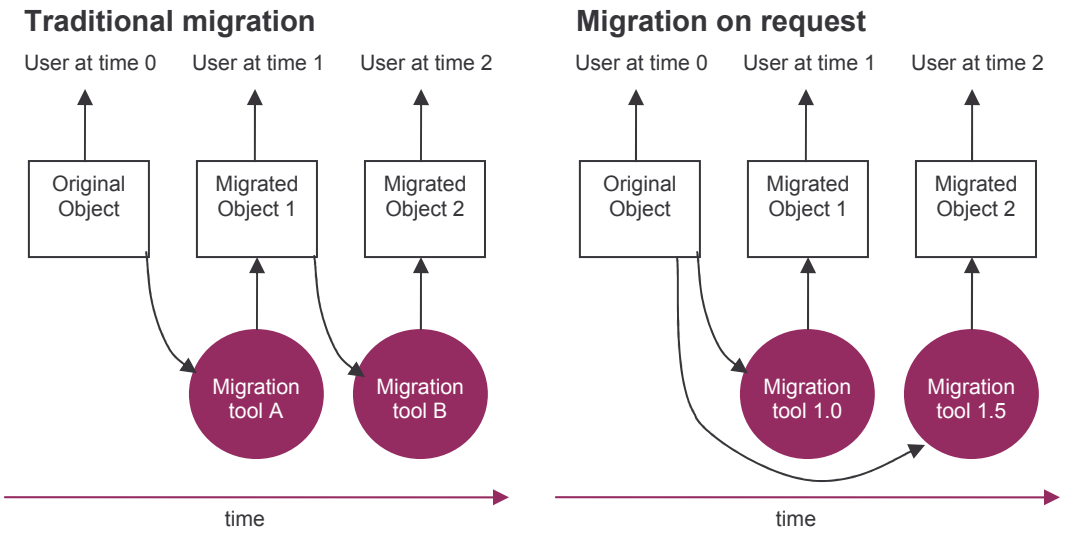



**Figure 4-9: Difference between traditional migration and migration on request [A-18]**

To overcome the necessity of multiple migration steps, CAMiLEON took a new approach in the design of a migration on request tool. They introduced a modular framework with an intermediate layer at heart and the original format and the migrated format as input and output modules respectively. The modular design allows us to add multiple input and output modules for different formats, all using the same internal intermediate representation. This design is also illustrated in figure 4-9 (right part).

At retrieval time, the migration on request tool should operate like the schema shown in figure 4-10. The original digital document constitutes the input defined in a specific logical format. This format is then translated via input module 'format A' into an 'intermediate format'. The intermediate format should be able to represent all different types of formats of a file type. Depending on the format that should be outputted, the intermediate format is translated to a particular 'format C' by the output module for that format.

As a proof of concept, the CAMiLEON project has taken this approach into practice using three different types of vector image formats: WMF (Windows Meta File) by Microsoft, Draw by Acorn Computers, and SVG (Scalable Vector Graphics) defined as new XML based format by the World Wide Web Consortium (W3C). As a basis for the intermediate representation SVG was chosen. SVG was denoted by CAMiLEON to be a well defined structure.

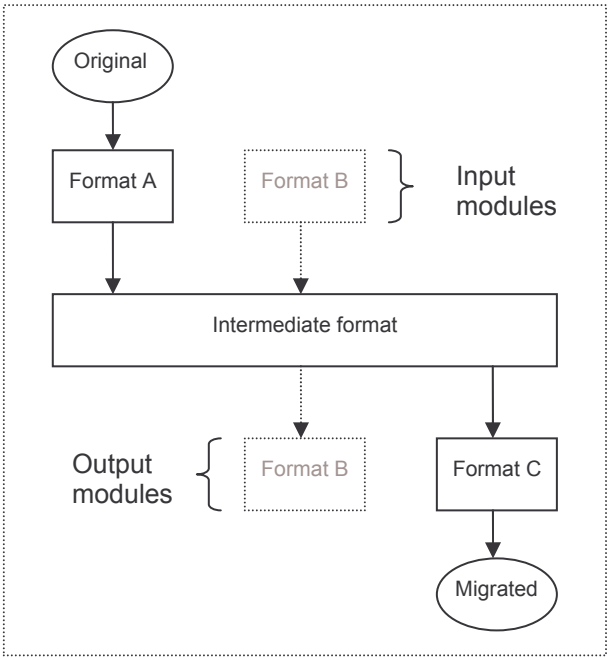



**Figure 4-10: Modular design of migration on request**

To encompass all the features of other vector formats, the intermediate format needs to be adjusted to prevent degradation, resulting in some duplication in format description. All three formats were decrypted by specifications on which algorithms were developed to convert the file specific format into the intermediate format. To verify that the migration on request is performed without error and loss of quality, the Consultative Committee for Space Data Systems OAIS Reference model recalls that migration should be reversible. Therefore a test was done converting a Draw image as source to the intermediate format and then reconverting it to a Draw format again. Both original and migrated results were visually compared to each other, and appeared to be similar. Thus the concept had been proven to be successful.

## Consideration

The CAMiLEON project pointed out that the migration on request approach offers several advantages above 'traditional ways of migration'. These are:

- It lowers the error rate because only one migration step is required instead of multiple steps over time
- The code which reads in and interprets a particular file format needs to be implemented only once.
- Authenticity aspects are simplified because the digital object is preserved in its original form.
- Migration on request spares the need for storing all migrated versions over time.
- Reversible migration is easier to implement using the modular design.
- Migration on request is cost effective because no migration has to be done over time.

Although these facts are great pros for migration on request, several disadvantages are also implied. The modular design allows different output modules to be connected to the intermediate layer. But as technology development moves ahead, new modules should still be developed, porting the intermediate file representation to the latest file formats. This requires a continuing development of modules and an understanding about the intermediate format. Although the costs will be lower than total migration from time to time, development should continue which requires a certain amount of effort.
Secondly, the migration on request tool is platform dependent. It is created as a software application that is aging just like all other digital objects. CAMiLEON assumes that the migration tool, which was written in C, will last but the idea of written once and usable forever is not guaranteed. The migration tool should be migrated or rewritten over time when the current executable is becoming obsolete. As a side effect authenticity remains a hot topic because migration on input and output modules can influence the content, structure, appearance and interactivity of the document.

### 4.2.6 Universal Virtual Computer

In literature the Universal Virtual Computer (UVC) is commonly discussed. It is often depicted as one of the few practical solutions to successfully guarantee permanent access while the authentication of the document will not be affected. The UVC-based preservation method is invented by Raymond Lorie from IBM Almaden Research Center and is further developed by IBM for the Royal Library of the Netherlands (Koninklijke Bibliotheek) as part of the e-Depot [A-14]. The UVC-based preservation method allows static digital objects to be reconstructed in its original appearance anytime in the future. Because the Royal Library of the Netherlands uses Portable Document Format (PDF) as its main logical format for electronic publications stored in their digital repository, a long-term preservation study has been done to the appliance of a UVC for PDF in particular.

#### How does it work?

The central idea of the UVC-based preservation method is that static digital objects preserved in an archive can be reconstructed anytime in the future without loss of the meaning of that object. The UVC concept is based on a combination of emulation and migration on demand and basically consists of four different components. These are:

- Universal Virtual Computer (UVC)
- Logical Data Schema (LDS) with format explanation
- UVC program (decoder)
- Restore Program

Together with the original data it is possible to reconstruct the meaning of each particular digital document. The UVC can be seen as an emulator, not really existing as hardware, but as software application running on future hardware. Because we do not know at this time which hardware is available in the future, the UVC must be created at the time we want to access a particular document from the repository. The UVC serves as a platform on which programs specifically written for the UVC can run. Such a UVC program is needed to decode the file format of a digital document. In turn, it retrieves element tags which hold specific information about the content of the data. These elements form the Logical Data View (LDV) of the data and look like XML. The LDV is an instantiation of the Logical Data Schema (LDS), which describes the elementary parts of a specific logical format as a blueprint.

All this is controlled by a Restore Program. This program also runs on future hardware and therefore need to be created at that time too. The Restore Program starts the UVC and feeds it with data of the digital document and a decoder program specifically developed for the UVC. It retrieves an LDV from the UVC and reconstructs a specific representation of the original objects meaning. This process is depicted in figure 4-11.

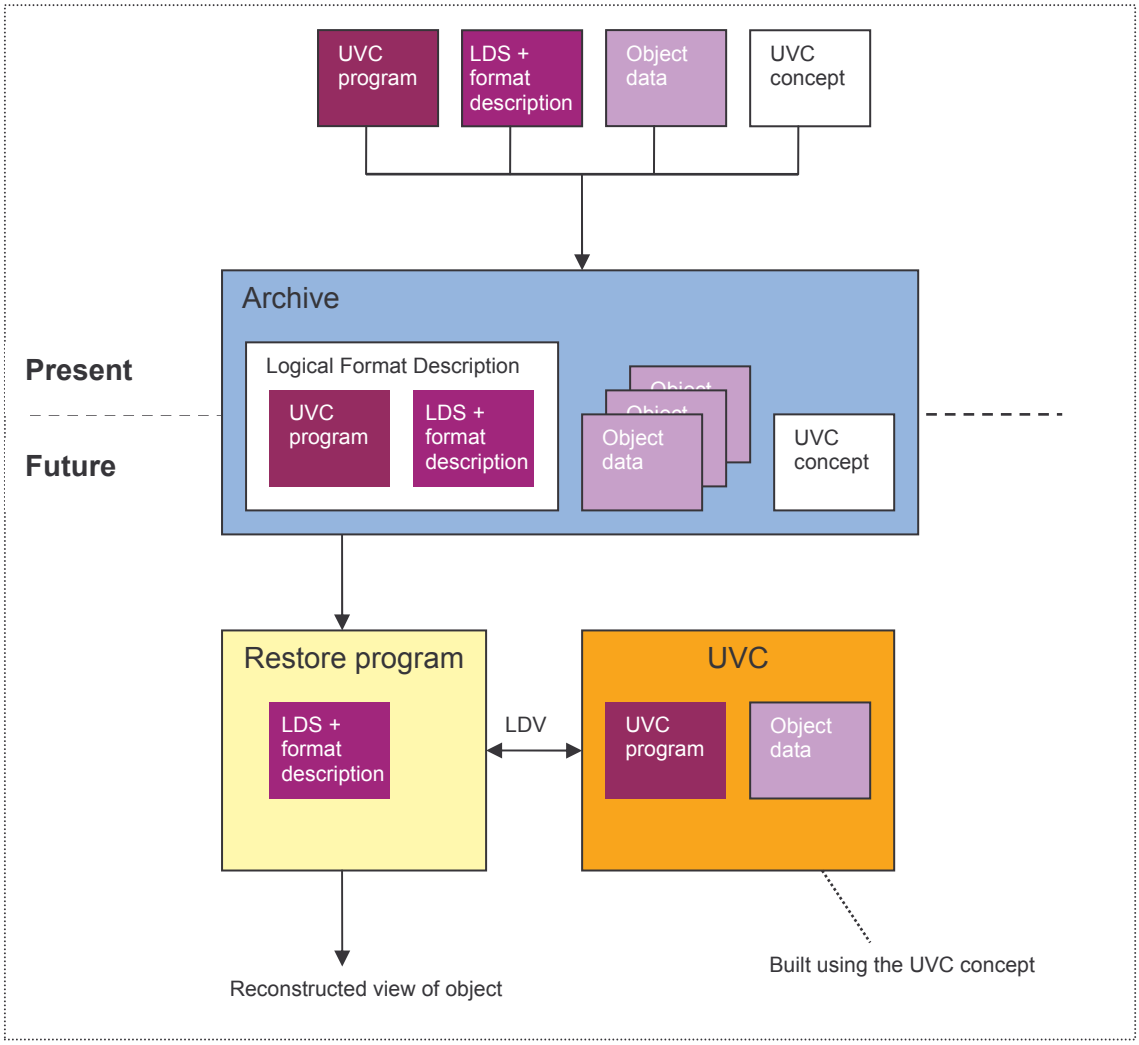



**Figure 4-11: UVC-based preservation method**

In the figure a distinction is made between preservation time (present) and retrieval time (future). Different steps must be taken during both present and future, which are:

### At preservation time (present)

**Step 1** – To view a digital document in the future, we must understand the structure (logical form) of it. Therefore a detailed description needs to be developed which states how the logical view should look like and what it means. This logical view is returned by the UVC in the future and needs to be interpreted by a restore program; therefore it must be understood by future developers. For instance, a logical view for a raster based image, consisting of a set of pixels with for each pixel different values for red, green and blue components, could be schematically depicted as in figure 4-12.

In this schema an image can be decomposed in a number of elements. The image is seen as a number of scan lines, representing the horizontal lines of an image. It can contain one or more of these scan lines, which is pointed out by a "+" sign. Each scan line in turn contains one or more pixels. Each pixel has a position and a color (defined by the three colors red, green and blue).

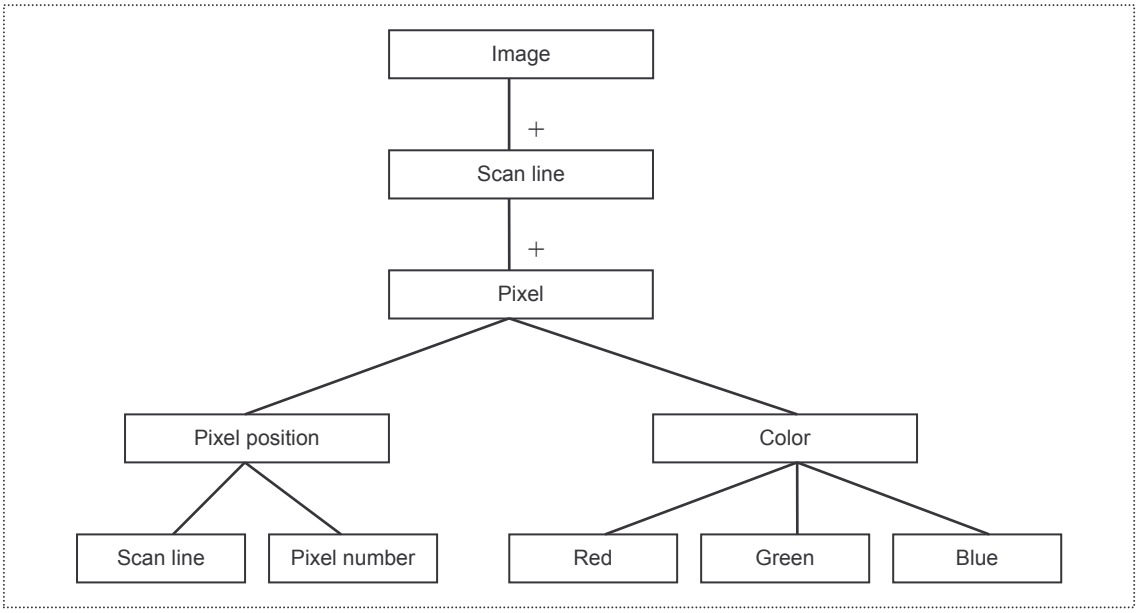



**Figure 4-12: Simplified schematic view of an image**

This information describes the structure of the object's content in detail. As mentioned above, executing the UVC will return this information in elementary tags. All tags are defined in a blueprint called the Logical Data Schema (LDS). The LDS explains which tags are retrieved from the UVC for a particular file format. But knowing how the LDV is retrieved is not enough. To understand what they mean, the LDS provides a description of the meaning of important elements. Think of an image from which each pixel is described by the colors red, green and blue. For each type of color a code value is returned, like red: 230, green: 17 and blue: 0. Without understanding the scale and spectrum of these values, colors can not be reconstructed to its authentic color. A simplified LDS for an image is schematically depicted in figure 4-13. See appendix B for a detailed LDS for image formats with description of the meaning.

```
ELEMENT 1 [Image] (10, 24+)

ELEMENT 10 [Image Size] (11, 12)
ELEMENT 11 [Number Scan Lines]
ELEMENT 12 [Number Pixels Per Scan Line]

ELEMENT 24 [Pixel] (25, 29)

ELEMENT 25 [Color] (26, 27, 28)
ELEMENT 26 [Red]
ELEMENT 27 [Green]
ELEMENT 28 [Blue]

ELEMENT 29 [Pixel Position] (30, 31)
ELEMENT 30 [Scan Line]
ELEMENT 31 [Pixel Number]
```

**Figure 4-13: Simple Logical Data Schema (LDS) for a raster based image**

**Step 2** – Having a digital document and a description of the elements returned by the UVC, is still not enough to reconstruct the object's meaning. The UVC has to know how it should decipher the logical format of a digital object. Therefore a UVC program has to be written which can decode the format and transforms it into a logical data view, using the elements defined by the LDS. It is important that this UVC program is written at preservation time,

because waiting can eventually lead to a misunderstanding of the format due to its obsolescence. For each format a decoder has to be written, demanding a lot of effort. But once a decoder is available it can be applied to every document of that same type.

**Step 3** – Finally future developers have to know how they could construct a UVC, which can execute the UVC decoder program for a particular document format. Developing one now will not guarantee that it is still operational in the future. This implies that it has to be made understandable how software developers in the distant future can create a new one by themselves. The UVC is designed to be a general-purpose computer, running on future hardware. The architecture conforms to the current Von Neumann architecture, but is very flexible. For instance, it assumes it has an unlimited amount of virtual memory and has no fixed bit-size, which differ from today's computers. To reproduce an UVC in the future, a description of this concept needs to be carefully preserved. This could be done as a document in a digital repository, but also as hardcopy on paper and microfilm.

### At retrieval time (future)

**Step 1** – If the digital documents, LDS descriptions, decoder programs and UVC concept description all have been archived successfully, we are now able to reconstruct any digital document for which the UVC fulfills a decoding process. First thing is that a UVC has to be created on a current platform. Because of the simplicity of the UVC concept, it should cost a limited amount of energy for skilled software developers to construct a UVC for a platform. If this succeeds, it is fairly easy to create one for other platforms, because of its general characteristics.

**Step 2** – Assuming that a proper working UVC exists, a simple application program has still to be developed. This application program, called a restore program has to control the UVC and all input / output interaction between it. How this restore program must be implemented is left to the developer. It needs to run on future hardware and can therefore not be specified at preservation time. The restore program has to run the UVC, send the encoded data and decoder program to it and receive logical data view tags back from the UVC. With these tags and the explanation defined in the LDS, a representation of the original document can be created (figure 4-14).

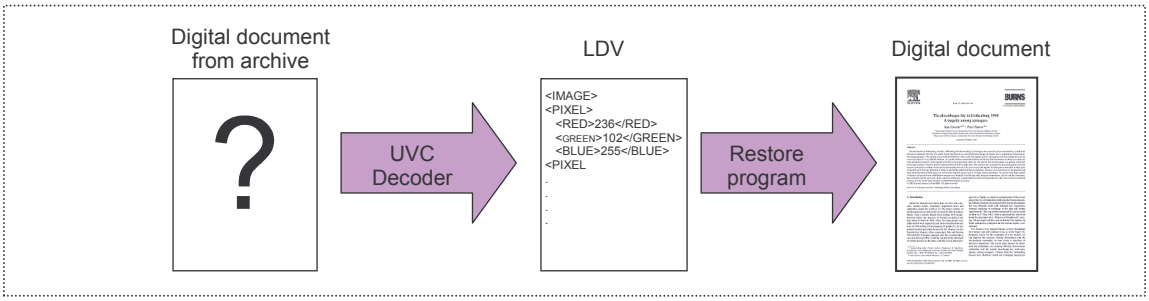



**Figure 4-14: reconstruction of a digital document**

### Considerations

The UVC-based preservation method promises to be a practical strategy to retain access and understanding to digital documents. The concept is well defined although an official UVC description document is not available yet. The combination of emulation (the UVC) and migration on demand (conversion of original document to a Logical Data View) seems to combine the advantages of both approaches. The power of the UVC approach lies in its scalability. Once a UVC has been successfully built, it can be used for all UVC decoder programs. And once an original digital document can be viewed, it can be used for all other documents sharing a similar format.

Although this approach seems to be promising, some uncertainties and disadvantages still exist. Firstly, the concept depends on a number of key factors: the UVC concept description, LDS descriptions and UVC decoder programs. If one of these fail to be available or is not well defined, all or a great amount of digitally stored information solely depending on the UVC is lost from accessibility [A-7]. Therefore the UVC description must be unambiguous and preserved with maximum effort. For each logical format, a Logical Data Schema has to be defined. Just like the UVC description, these must be preserved well. Besides that, another danger exists. The LDS could be misinterpreted by future users if the Logical Data View returned by the UVC decoder is not well understood, as previously said. Therefore the LDS should be self explanatory and well tested.

Secondly, the UVC is based on the Von Neumann architecture. It assumes that this architecture will remain the heart of computers also in the future. Of course, no guarantee can be given to prove this. As today's computer technology is driven to the maximum, like circuitry at atom level, new technologies will probably be endeavored. Quantum computing is one of those expeditions which could become the computers of the future.

Finally, the UVC-based preservation method is only applicable to static digital objects. Although digital documents form an important amount of the digitally stored information, dynamic digital objects are part of the preservation problem too.

Nevertheless IBM and KB together made a good attempt to cope the problem of inaccessibility to digitally stored information on the long-term. At this moment its success depends mainly on the willingness of software suppliers to remark the usefulness of the UVC and the development of logical data schemes and decoders for it.

## 4.3 Overview of strategies

Having considered the different strategies concerning digital preservation, it has become clear that each approach has its own characteristics, benefits and drawbacks. These are summarized in table 4-1.

| | *Migration* | *Emulation* | *XML* | *DRS* | *Migration on request* | *UVC* |
|---|---|---|---|---|---|---|
| **Advantages** | - often easy to apply<br>- often supported by software vendors.<br>- practical results | - authenticity guaranteed<br>- original bits<br>- low error rate<br>- low storage capacity required<br>- no software rewrite<br>- no periodic conversion of documents | - separation of content, structure and layout<br>- easy to understand<br>- universal language<br>- wide appliance | - integral approach<br>- centralized way of preserving all hard- and software specifications | - lower error rate<br>- intermediate layer<br>- authenticity guaranteed<br>- no periodic conversion of documents<br>- modular design<br>- low storage capacity required | - perm. access guaranteed<br>- authenticity guaranteed<br>- low error rate<br>- low storage capacity required |
| **Disadvantages** | - continuity of conversion tools<br>- no guarantee on authenticity<br>- periodic event<br>- rewrite of software may be needed<br>- errors (mutations)<br>- loss of information<br>- storage | - complexity<br>- initial effort<br>- peripherals<br>- formats proprietary to software vendors<br>- knowledge of how old applications work is needed | - not suitable for all formats<br>- standard on structure definition unsure<br>- standard on layout definition unsure<br>- ambiguity of layout processor<br>- storage | - highly theoretical (maybe not applicable for all media and formats)<br>- dependant on one organization<br>- initial effort | - development of input / output modules<br>- platform dependent (risking authenticity) | - interpretation of UVC, LDS unsure<br>- writing decoders<br>- format specifications may be proprietary to software vendors |
| **Cost** | - Periodic (variable cost)<br>- linear with growth of scale | - high initial cost<br>- low periodic cost<br>- insensitive for scale | - medium initial cost<br>- no periodic cost | - high initial cost<br>- continuing cost for updating processes | - each time for new format<br>- each time new platform is widely used | - initial for UVC<br>- high cost for decoder / LDS development |
| **Best appropriate timescale** | Short and Intermediate term | Long term | Long term (if XML remains a widely applied standard) | Long term | Intermediate / long term | Long term |
| **Suitable for** | All common widely used digital documents | All documents | Not all documents (because of size) | All documents | All common widely used digital documents | All documents |
| **Ease of appliance** | - Easy for common file formats (conversion tools available).<br>- Difficult for unfamiliar formats | - Difficult to set up an emulator. But once working, it could run many different applications.<br>- also useful for unfamiliar formats | - XML is easy to apply, but care needs to be taken to ensure interpretation is correct. | - Requires a complete new infrastructure and much information has to be gathered. | - For each format an intermediate level and input / output modules have to be developed. | - Concept sounds difficult<br>- depends on implementation of UVC.<br>- Once working, a lot of benefits. |
| **Future success** | Already widely used. Will still be in the future for common documents wherefore authenticity does not have a high priority. | Experiences with gaming consoles have shown the power of it. If a certain level of abstraction on I/O and peripherals can be reached, emulation is a good candidate. | XML is already widely used and will probably grow further. But as with every format, it may also be a one day fly. | Implementation of this approach requires a lot of resources and effort, but may be taken into practice because it offers a centralized depot for hard- and software specs. | Forms a good strategy for commonly used formats. Is less attractive when target format is not available (output module) | Good chances to become a widely used strategy. It depends on the support of other software houses to make this approach a success. |

**Table 4-1: Overview of the preservation strategies in this report**

## 4.4 Which strategy to choose?

To preserve digital documents over the long-term, retain access and to understand the information kept inside, a preservation strategy has to be chosen. The basic belief that one strategy fits all has shown to be a misunderstanding. Different types of documents exist with different types of purposes which have to be considered before a decision is made about the strategy to follow. In most cases multiple strategies can be applied, leaving the responsible party with the question which strategy suites best? Risk management could help to make this decision easier.

To point out which preservation strategy gives the best assurance for permanent access, the dangers and opportunities of digital documents itself have to be quantified. Knowing the strengths and weaknesses of each type of document creates a better feeling for what is appropriate to choose.
Compliant to this, John Bennett described the use of a Preservation Complexity Scorecard (PCS) [A-28]. A PCS is a matrix which denotes the risk of each digital object. This is done by identifying each risk in four distinct levels:

- type of material
- type of file format
- type of media
- type of platform / operating system

For each level, a scorecard has to be created, such as shown in table 4-2 for some type of materials. Each material is assigned a base score (1 being the least complex to preserve, and 5 the most complex). Hereby is added a complexity factor, triggered by some functionality feature that adds cost and effort ("difficulty") when permanent access has to be safeguarded. Finally, the risks are named for each particular type. Quantifying all four levels gives a complex combined scorecard which denotes the cumulative risks and scores of a digital object. This makes it possible to mark endangered formats and take appropriate action based on its characteristics. Note that the scorecard is not a static matrix. Over time the Scorecard calibration will change as new digital technologies are used to access and preserve digital material.

| Material | Base Score | Complexity Factors (add to the base score) | Risk |
|---|---|---|---|
| Text / Document | 1 | Functionality (+1), Macros (+1), Templates (+1) | Loss of format |
| Spreadsheets | 1 | Formatting (+1) | Loss of format. Loss of meaning |
| Multiple Spreadsheets | 2 | Linkages (+1), Macros (+1) | Loss of external data |
| "Office Suite" documents | 2 | Links, Views, Indexes are standard | Loss of access to all items. Loss of meaning |
| Database records | 3 | Structures and rules (+1) | Loss of meaning |

**Table 4-2: Scorecard for type of material**

Although the scorecard can help identifying risks, it does not give an answer to the question which preservation strategy suites best. It strongly depends on the type of document (characteristics), application software, hardware and operating system, as well as the importance of authenticity and costs.

# 5 Conclusions & recommendations

The virtual environment we live in today is an important part of the cultural heritage for tomorrow. Therefore, aspects of the virtual heritage should be preserved for future generations to learn about today's life and build on the knowledge we created. To do so, digital preservation repositories have to be designed, implemented, used and maintained. But preservation is only the first step. To ensure that the information we stored today remains accessible and understandable in the far future, permanent access has to be guaranteed.
We started this report with the following research question:
"Which of the current strategies regarding permanent access technology taken worldwide ensure accessibility over the long-term?"
To answer this question different related sub questions have been answered.

## 5.1 How to preserve the virtual heritage?

The virtual heritage is formed by all things that happen in a virtual environment, including digital documents. Preservation of digital documents requires a different approach than traditional documents, because of the aspects of digitally stored information. A document in digital form is stored as a sequence of bits on a storage medium. The meaning of a digital document depends on the logical form the bits are stored and on the rendering software needed to view the document. Furthermore, a digital document can contain additional information enclosed as metadata.

### 5.1.1 What makes digital preservation difficult?

In this report a couple of difficulties have been addressed concerning preservation of digital documents.

- **Scale:** an enormous amount of digital content is available today and is still growing rapidly. Each year about 5 exabytes of data has been stored on print, film, magnetic and optical storage media, of which 92% on magnetic disks. It is impossible to preserve all digitally stored information in an organized way. Criteria have to be developed to decide which information is valuable for the future and how this can be preserved.
- **IT developments:** during the twentieth century, the Information Technology sector (IT) has grown enormously. As a result developments in hard- and software have increased explosively. Today many different hardware devices and software applications are in use and replaced by new releases and versions about every two or three years. Preserving digital documents today may be inaccessible over three years already, due to the dependence on particular hard- and software.
- **Authenticity:** a digital document can only remain authentic if its integrity is safeguarded and if it can be verified as 'the real one'. This is hard to satisfy because a digital document highly depends on its environment (hard- and software). Therefore this environment should be exactly recreated or the digital document should be transformed to newer formats without loss of its intrinsic value.

### 5.1.2 How to design a preservation repository?

Libraries and archives have a lot of experience with preservation of traditional documents, but are new to digital preservation. The issues mentioned above form a challenge for libraries and archives. During design of a digital preservation repository these issues should be well considered.

An important question about permanent access is: how can we ensure that preserved digital documents can be properly interpreted in the future?

Ideally success is only guaranteed if repositories and preservation process is independent of computing platform, media technology and format paradigms. This implies that standards have to be developed, widely used, maintained and general concepts for information value including selection and authenticity need to be defined. But these requirements are hard to meet.

### 5.1.3 Which preservation strategies are there?

Different preservation strategies are considered worldwide to guarantee permanent access. In short these are: technology preservation, saving the hard copy, encapsulation, migration, migration on request, emulation (including virtual machine approach), XML, Digital Rosetta Stone, and Universal Virtual Computer.
All of these approaches have their own advantages and disadvantages, which are discussed in the chapters three and four. The first three approaches (technology preservation, saving the hard copy, and encapsulation) are less suitable than the others, because it is practically impossible or loss of information is inevitable. The other approaches are considered in more detail and the results are denoted in table 4-1 of chapter 4.

## 5.2 Conclusions

In general it can be stated that there is a growing attention on permanent access. Many organizations are developing (or planning) digital preservation repositories and are becoming aware of the difficulties of preservation of digital documents. Frontrunners are exploring the possibilities of different preservation strategies and many libraries and archives are watching the outcomes closely. In this report the most important preservation strategies have been discussed to find an answer on the main question of this research.
Based on these outcomes it seems clear that no one-size-fits-all solution is possible. Digital documents differ from each other in too many ways and are used for many different purposes by many different users. Organizations that are waiting for “the” solution will not be successful in preservation of digital documents. Risk management should be applied to find out which strategy is most appropriate for each type of document. Thereby considering how important the authenticity of a document is.

Considering the six approaches that are discussed in detail, migration seems to be suitable for common document formats which are widely supported while authenticity has not the highest priority. Emulation can be seen as a last resort for uncommon file formats, whereby authenticity of a document is important and initial costs are not an issue. XML is different in its kind because it tends to a uniform standard used worldwide. Despite the history of standardization (standards come and go), XML can become the standard of standards if it stays in business. It seems to be very suitable for preservation of e-mail, spreadsheets and text documents, although less for other document formats, e.g. image files. The Digital Rosetta Stone (DRS) is a good theory, but a complete implementation of the model seems far away. Instead, migration on request is very practical and already tested with success for image formats. A disadvantage is that it is platform dependent. This does not hold for the Universal Virtual Computer (UVC) based approach. This strategy is the only one that is platform independent, applicable for all digital documents while offering maximum authenticity. But to make the UVC approach successful, it requires decoders and Logical Data Schema’s to be developed at preservation time. This demands a lot of effort. Therefore more experience with the UVC should be gained to convince others of its potency and gain more support in development.

## 5.3 Recommendations

Although we are heading the right way, more work has still to be done in the field of digital preservation. First of all, more understanding is needed on preservation strategies. Especially emulation and the UVC need further development to convince organizations of their power, overall “seeing is believing”. Because none of the strategies, discussed in this report, offers a one-for-all solution, more awareness is needed on combined strategies and tools. In this, the proposal of the PATCH project should gain more attention to support the development of different approaches altogether.

Besides this, the core of the problem should not be forgotten. The creation of so many file formats depending on all kinds of hard- and software over the last decades leaves us with the preservation struggles today. We are now in the position to solve this problem. This can be done by setting up design criteria for file formats such as a separation of content, structure and layout which is done by XML. Furthermore, we can further develop and use standards on architecture, content and access. A possible approach is to invent rules which force software developers to support older formats when they introduce new ones. Or offer at least well-tested conversion tools to migrate to newer formats.
No matter which actions are required, most important is that valuable information will remain valuable, accessible and understandable for future generations, helping our civilization forward.

# 6 Reference list

## 6.1 Books & Articles

[A-1] *Migration: context and current status*, Digital Preservation Testbed, The Hague, The Netherlands, 2001

[A-2] Meer, van der, K., Dondorp, F.P.A., *Design criteria for preservation repositories*, Delft University of Technology, DIOSE Betake Research Group, Delft, The Netherlands, 2004

[A-3] Steenbakkers, J.F., *Permanent archiving of electronic publications: research & practice*, Koninklijke Bibliotheek, The Hague, The Netherlands, 2003

[A-4] Rothenberg, J., *Using Emulation to preserve digital documents*, Koninklijke Bibliotheek, The Hague, The Netherlands, 2000

[A-5] Dürr, E., lourens, W., *Programs for ever*, NDDL 2002, Ciudad Real, 2002, p. 63-79.

[A-6] Wilkins, J., *Preservation strategies for electronic documents*, 2004

[A-7] *Emulation: context and current status,* Digital Preservation Testbed, The Hague, The Netherlands, 2003

[A-8] Cloonan, M.V., Sanett, S., Survey of Preservation Practices and Plans, Preservation strategies for electronic records, round 1 (2000-2001), InterPARES, 2001

[A-9] Steenbakkers, J., *The NEDLIB guidelines, NEDLIB report series number 5*, Koninklijke Bibliotheek, NEDLIB Consortium, 2000

[A-10] Rothenberg, J., *Ensuring the longevity of digital information*, RAND, Santa Monica, 1999 (revisioned version of 1995)

[A-11] Lorist, H.H.J., Meer, van der, K., *Standards for digital libraries and archives: digital longevity*, Delft University of Technology, Delft, The Netherlands, 2002

[A-12] Harel, D., *Algorithmics: the spirit of computing*, Addison Wesley, 1987, p. 221

[A-13] Diessen, van, R.J., *Preservation requirements in a deposit system*, IBM / KB Long-term preservation study, report series nr 3, IBM, Amsterdam, The Netherlands, 2002

[A-14] Lorie, R., *The UVC: a method for preserving digital documents - Proof of concept*, IBM / KB Long-term preservation study, report series nr 4, IBM, Amsterdam, The Netherlands, 2002

[A-15] Wijngaarden, van, H., Oltmans, E., *Digital Preservation and permanent access: The UVC for images*, Koninklijke Bibliotheek, The Hague, The Netherlands, 2004

[A-16] Diessen, van, R.J., Oltmans, E., Wijngaarden, van, H., *Preservation Functionality in a digital archive*, in proceedings of the Joint Conference on Digital Libraries 2004, Tucson, Arizona

[A-17] Andre, P.Q.C., et al., *Preserving Digital Information – Final Report*, Task Force on archiving of digital information, 1996

[A-18] Mellor, P., Wheatley, P., Sergeant, D., *Migration on request, a practical technique for preservation*, CAMiLEON project, Edward Boyle Library, University of Leeds, Leeds, 2002

[A-19] Steenbakkers, J.F., *The PATCH project: Integrating Research & Standardisation of Digital Preservation*, Koninklijke Bibliotheek, The Hague, The Netherlands, 2003

[A-20] *Design criteria standard for electronic records management software applications*, Department of Defence, 2002

[A-21] Granger, S., *Digital Preservation & emulation: from theory to practice*, CAMiLEON, Leeds

[A-22] *Bewaren van spreadsheets*, Digital Preservation Testbed, The Hague, The Netherlands, 2003 (Dutch)

[A-23] Brodie, N., *Authenticity, preservation and access in digital collections*, conference paper for 'Preservation 2000', December 2000, York, England

[A-24] Dekker, R., Dürr, E.H., Slabbertje, M., Meer, van der, K., *An electronic archive for academic communities*, NDDL 2002, Ciudad Real, 2002, p. 1-12.

[A-25] *Bewaren van e-mail*, Digital Preservation Testbed, The Hague, The Netherlands, 2002 (Dutch)

[A-26] *Bewaren van tekstdocumenten*, Digital Preservation Testbed, The Hague, The Netherlands, 2003 (Dutch)

[A-27] Heminger, A.R., Robertson, S.B., *Digital Rosetta Stone: a conceptual model for maintaining long-term access to digital documents*, 1998

[A-28] Bennett, J.C., *A framework of data types and formats, and issues affecting the long term preservation of digital material*, British Library and Innovation Centre, 1997

## 6.2 Internet

All URL's are checked and valid in May 2004.

[I-1] Varian, H., Lyman, P., *how-much-information? 2003*, University of California, Berkeley, 2003
http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

[I-2] Hodge, G.M., *Best Practices for Digital Archiving*, D-lib Magazine, vol. 6, nr 1, 2000
http://www.dlib.org/dlib/january00/01hodge.html

[I-3] Webb, C., *Saving Digital Heritage – A UNESCO Campaign*, RLG Diginews vol 7, nr 3, 2003
http://www.rlg.org/preserv/diginews/v7_n3_feature3.html

[I-4] Royal Library of the Netherlands (Koninklijke Bibliotheek)
http://www.kb.nl

[I-6] Digicult Report, Technological landscapes for tomorrow's cultural economy, EC, 2002
http://digicult.salzburgresearch.at

[I-7] Preliminary draft charter on the preservations of the digital heritage, UNESCO, 2003
http://unesdoc.unesco.org/images/0013/001311/131178e.pdf

[I-8] *Draft Research Agenda*, National Library of Australia, Australia, 1998
http://www.nla.gov.au/policy/rsagenda.html

[I-9] Phillips, M., *Australia's Web Archive, and the Digital Archiving System that Supports it*, DigiCULT.info, Issue 6, December 2003
http://www.nla.gov.au/nla/staffpaper/2003/mphillips1.html#1

[I-10] Preserving Access to Digital Information, PADI
http://www.nla.gov.au/padi

[I-11] Digital Information Archiving System, DIAS
http://www.ibm.com/nl/dias

[I-12] Steenbakkers, J.F., Treasuring *the digital records of science: Archiving E-Journals at the Koninklijke Bibliotheek*, RLG Diginews, vol. 8, nr 2, 2004
http://www.rlg.org/en/page.php?Page_ID=17068

[I-13] Urban Legends: Gates Memory
http://www.urbanlegends.com/celebrities/bill.gates/gates_memory.html

[I-14] WPDOS under Windows NT, 2000 and XP
http://www.columbia.edu/~em36/wpdos/windowsxp.html#installguide

[I-15] ISO Archiving Standards – OAIS reference model
http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49

[I-16] CAMiLEON project
http://www.si.umich.edu/CAMILEON/

[I-17] Digital Preservation Testbed (Testbed Digitale Bewaring)
http://www.digitaleduurzaamheid.nl

[I-18] DSpace project
http://www.dspace.org

[I-19] Lynch, C.A., *Institutional Repositories: essential infrastructure for scholarship in the digital age*, ARL, nr. 226 (February 2003), 1-7
http://www.arl.org/newsltr/226/ir.html

[I-20] Australian standard for records management – AS ISO 15489 (and AS 4390)
http://www.naa.gov.au/recordkeeping/rkpubs/advices/advice58.html

[I-21] The unofficial TIFF homepage
http://home.earthlink.net/~ritter/tiff/

[I-22] Dutch National Archives Regulations, article 11, 1995 (Dutch)
http://www.rijksarchiefinspectie.nl/wetgeving/arch_archiefwet1995.html

[I-23] Online Computer Library Center, OCLC
http://www.oclc.org

[I-24] RLG
http://www.rlg.org

[I-25] PREMIS work group
http://www.oclc.org/research/projects/pmwg/

[I-26] Lavoie, B.F., *Implementing metadata in digital preservation systems*, PREMIS, D-Lib Magazine, Volume 10, nr. 4 (April 2004)
http://www.dlib.org/dlib/april04/lavoie/04lavoie.html

[I-27] Turing machine simulator using Java
http://www.igs.net/~tril/tm/tm.html

[I-28] Lots Of Copies Keep Stuff Safe, LOCKSS
http://lockss.stanford.edu

[I-29] Hedstrom, M., Lampe, C., *Emulation vs. migration: do users care?,* RLG Diginews vol 5, nr 6, 2001
http://www.rlg.org/preserv/diginews/diginews5-6.html

[I-30] Granger, S., *Emulation as a digital preservation strategy*, CAMiLEON, D-Lib Magazine, Volume 10, nr. 6 (October 2000)

http://www.dlib.org/dlib/october00/granger/10granger.html

[I-31] Computer game emulation, CAMiLEON research
http://www.si.umich.edu/CAMILEON/research/research.html

[I-32] World Wide Web Consortium, W3C, XML
http://www.w3.org/XML

[I-33] W3Schools
http://www.w3schools.com/xml

[I-34] Encoded Archival Description, version 2002
http://www.loc.gov/ead

[I-35] Cedars project
http://www.leeds.ac.uk/cedars

[I-36] Digital Academic Repositories, DARE
http://www.darenet.nl

[I-37] Aschenbrenner, A., The bits and bites of data formats – stainless design for digital endurance, Erpanet, RLG Diginews vol 8, nr 1, 2004
http://www.rlg.org/preserv/diginews/v8_n1_feature3.html

[I-38] Joint Photographic Experts Group, JPEG
http://www.jpeg.org

[I-39] Adobe Systems Incorporated, PDF
http://www.adobe.com/products/acrobat/acro_ad.html

[I-40] Microsoft Corporation
http://www.microsoft.com

[I-41] Machine-Readable Cataloguing, MARC
http://www.loc.gov/marc

[I-42] Lawrence, G.W., et al, *Risk management of digital information*, CLIR, 2000
http://www.clir.org/pubs/reports/pub93/contents.html

[I-43] Dublin Core Metadata Initiative, DC
http://www.dublincore.org

[I-44] Metadata Encoding & Transmission Standard, METS
http://www.loc.gov/standards/mets

[I-45] InterPARES project
http://www.interpares.org

## 6.3 Folders, slides, etc.

[R-1] WordPerfect 5.1 document test, performed on an AMD Athlon 700Mhz computer running Windows XP Professional Edition SP1. Test performed by J.R. van der Hoeven, 2004

[R-2] Lawrence, H.A., *Digital insurance for information at risk*, Eastman Kodak Company, 2000

[R-3] Rothenberg, J., Long-term preservation of digital information: challenges and possible technical solutions, Slides of December 2000

[R-4] PDF/A, Frequently Asked Questions, Adobe Systems Incorporated, Slides 2003

[R-5] PDF/A, SMP and a metadata registry use case, Adobe Systems Incorporated, slides 2004

# 7 Glossary

Supported by: Wikipedia – The free encyclopedia – www.wikipedia.org

**Artefact** : any object or process resulting from human activity

**Backward compatibility** : a system is backward compatible if it is compatible with earlier versions of itself, or sometimes other earlier systems, particularly systems it intends to supplant. That is, other systems or objects that interoperate with the old version of the system should continue to interoperate with the new version.

**Bit stream** : a bitstream or bit stream is a time series of bits.

**Born digital** : information directly created in digital form.

**Continuous rendering** : see permanent access.

**Controller card** : a device that attempts to control states or outputs of a dynamic system.

**Dissemination** : the broadcasting of something.

**HTML** : HyperText Markup Language (HTML) is a markup language designed for creating web pages, that is, information presented on the Internet.

**Ingestion** : the intake of something.

**Interface** : a standard specifying a set of functional characteristics, common physical interconnection characteristics, and signal characteristics for the exchange of data or signals.

**Internet** : the Internet is the publicly available internationally interconnected system of computers.

**Interoperability** : the ability of systems, units, or forces to provide services to and accept services from other systems, units or forces and to use the services so exchanged to enable them to operate effectively together.

**Long-term** : a period of hundred years or longer.

**Long-term access** : see permanent access.

**Microfilm** : an analog storage medium for books, periodicals and engineering drawings.

**Motherboard** : also known as mainboard or systemboard is the central or primary circuit board making up a computer system or other complex electronic system.

**Operating system** : the system software responsible for the direct control and management of hardware and basic system operations, as well as running application software.

**Peripheral device** : any part of a computer other than the CPU or working memory, i.e. disks, keyboards, monitors, mice, etc.

**Permanent access** : the guarantee that access to information and understanding of that information on the long-term is safeguarded.

**Permanent access technology** : technology to ensure permanent access.

**Platform (computing)** : describes some sort of framework, either in hardware or software, which allows software to run.

**Preservation strategy** : a strategy to ensure proper preservation of and permanent access to digitally stored information.

**Punched card** : a medium for holding information for use by automated data processing machines. Made of stiff cardboard, the punch card represents information by the presence or absence of holes in predefined positions on the card.

**Quantum computer** : a device that computes using superpositions and entanglement of quantum states. Simple quantum computers have recently been built, and progress is continuing.

**Repository** : a central place where data is stored and maintained.

**SGML** : The Standard Generalized Markup Language (SGML) is a metalanguage in which one can define markup languages for documents.

**SMTP** : Simple Mail Transfer Protocol (SMTP) is the de facto standard for email transmission across the internet.

**Von Neumann architecture** : the so-called von Neumann architecture is a model for a computing machine that uses a single storage structure to hold both the set of instructions on how to perform the computation and the data required or generated by the computation. Such machines are also known as stored-program computers. The separation of storage from the processing unit is implicit in this model.

**Web service** : a collection of protocols and standards used for exchanging data between applications.

**World Wide Web** : see Internet.

# Appendices

## Appendix A. XML document for defining a 2 by 2 pixel image

```
<?xml version="1.0" encoding="UTF-8"?>

<image>
	<numFrames>
		1
	</numFrames>
	<frame>
		<imageSize>
			<numScanLines>
				2
			</numScanLines>
			<pixelsPerScanLine>
				2
			</pixelsPerScanLine>
		</imageSize>
		<pixel>
			<colour>
				<red>
					170
				</red>
				<green>
					129
				</green>
				<blue>
					87
				</blue>
			</colour>
			<pixelPosition>
				<scanLine>
					0
				</scanLine>
				<pixelNum>
					0
				</pixelNum>
			</pixelPosition>
		</pixel>
		<pixel>
			<colour>
				<red>
					155
				</red>
				<green>
					114
				</green>
				<blue>
					72
				</blue>
			</colour>
			<pixelPosition>
				<scanLine>
					0
				</scanLine>
				<pixelNum>
					1
				</pixelNum>
			</pixelPosition>
		</pixel>
		<pixel>
			<colour>
```

```
                        <red>
                                255
                        </red>
                        <green>
                                217
                        </green>
                        <blue>
                                175
                        </blue>
                </colour>
                <pixelPosition>
                        <scanLine>
                                1
                        </scanLine>
                        <pixelNum>
                                0
                        </pixelNum>
                </pixelPosition>
        </pixel>
        <pixel>
                <colour>
                        <red>
                                206
                        </red>
                        <green>
                                165
                        </green>
                        <blue>
                                123
                        </blue>
                </colour>
                <pixelPosition>
                        <scanLine>
                                1
                        </scanLine>
                        <pixelNum>
                                1
                        </pixelNum>
                </pixelPosition>
        </pixel>
</frame>
<endOfData>
</endOfData>
</image>
```

Appendix B. LDS for raster images

# Initial design of the logical data view for bit map images

/* **Main definition of the data elements** */
ELEMENT 1 [Image] (2, 4?, 8, 9+, 255)

ELEMENT 4 [Spatial Resolution] (5?, 6, 7)
ELEMENT 5 [Metric Unit] (CHAR)
ELEMENT 6 [Width] (100)
ELEMENT 7 [Height] (100)

ELEMENT 8 [Number of Frames] Integer32
ELEMENT 9 [Frame] (10, 13?, 24+)

ELEMENT 10 [Image Size] (11, 12)
ELEMENT 11 [Number Scan Lines] Integer32
ELEMENT 12 [Number Pixels Per Scan Line] Integer32

ELEMENT 13 [CIE Conversion] (14, 15,16, 20)

ELEMENT 14 [CIE Version] (CHAR)
ELEMENT 15 [Observer] (CHAR)
ELEMENT 16 [White Reference] (17, 18, 19)

ELEMENT 17 [Red CIE Primary] (100)
ELEMENT 18 [Green CIE Primary] (100)
ELEMENT 19 [Blue / Luminance CIE Primary] (100)

ELEMENT 20 [CIE Conversion Matrix] (21, 22, 23)

ELEMENT 21 [Red Conversion] (17)
ELEMENT 22 [Green Conversion] (18)
ELEMENT 23 [Blue / Luminance Conversion] (19)

ELEMENT 24 [Pixel] (25, 29)

ELEMENT 25 [Colour] (26, 27, 28)
ELEMENT 26 [Red] Integer8
ELEMENT 27 [Green] Integer8
ELEMENT 28 [Blue] Integer8

ELEMENT 29 [Pixel Position] (30, 31)
ELEMENT 30 [Scan Line] Integer32
ELEMENT 31 [Pixel Number] Integer32

/* **Reserved for potential metadata associated with the digital object** */
ELEMENT 2 [Metadata] (3)

ELEMENT 3 [File Name] (CHAR)

/* **Generally used items appearing in most logical data schemas** */
ELEMENT 100 [Reference Floating Point] (101, 102, 103, 104)
ELEMENT 101 [Quotient] Integer32
ELEMENT 102 [Remainder] Integer32
ELEMENT 103 [Exponent] Integer32
ELEMENT 104 [Sign] Bit1

ELEMENT 120 [Reference Date] (121, 122, 123)
ELEMENT 121 [Day] (CHAR)
ELEMENT 122 [Month] (CHAR)
ELEMENT 123 [Year] (CHAR)

ELEMENT 130 [Reference Time Format] (131, 132, 133)
ELEMENT 131 [Hour] (CHAR)
ELEMENT 132 [Minute] (CHAR)
ELEMENT 133 [Second] (CHAR)

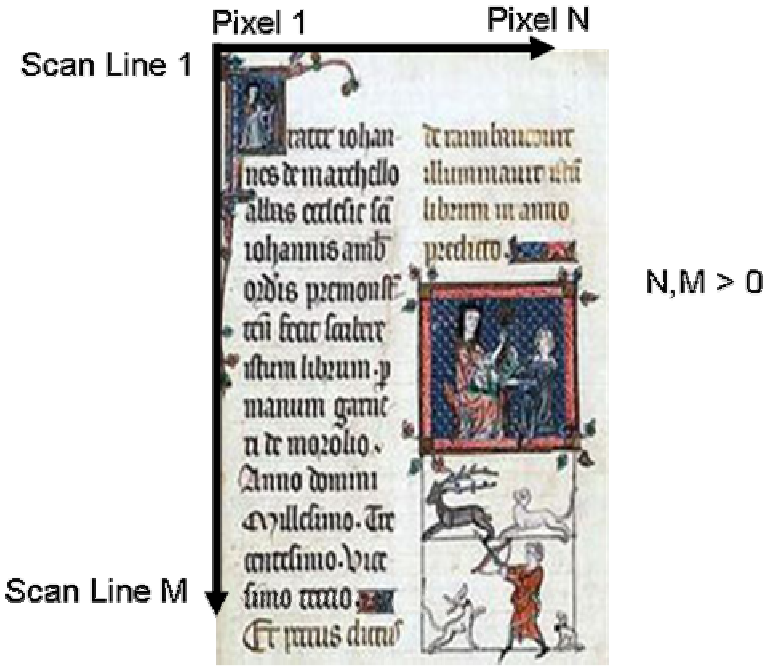
ELEMENT 255 [End of Data]

# Enumerated data types

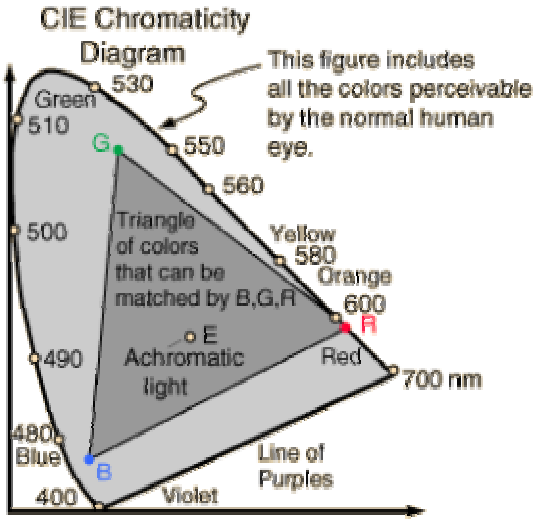
| Attribute | Value: Meaning |
|---|---|
| Metric Unit | Inches, Centimetres |
| CIE Version | 1931, 1960, 1976 |
| Observer | 2 degrees, 10 degrees |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |

# Picture Composition

A picture is build from left to right and from the top to the bottom, see figure.




# Colour Reference




The CIE system characterizes colors by a luminance parameter Y and two color coordinates x and y which specify the point on the chromaticity diagram. This system offers more precision in color measurement than do the Munsell and Ostwald systems because the parameters are based on the spectral power distribution (SPD) of the light emitted from a colored object and are factored by sensitivity curves which have been measured for the human eye.

Based on the fact that the human eye has three different types of color sensitive cones, the response of the eye is best described in terms of three "tristimulus values ". However, once this is accomplished, it is found that any color can be expressed in terms of the two color coordinates x and y.

The colors which can be matched by combining a given set of three primary colors (such as the blue, green, and red of a color television screen) are represented on the chromaticity diagram by a triangle joining the coordinates for the three colors.

The exact formulae to convert RGB to CIE tristimulus values  is explained in a separate PDF RGB_to_CIE.pdf.