From digital volatility to digital permanence

Preserving databases

The Digital Preservation Testbed is an initiative of the Dutch National Archives and the Dutch Ministry of the Interior and Kingdom Relations. It is a research programme set up to test the practical applicability of various ways of preserving government and other digital information and keeping it accessible for the future. The Digital Preservation Testbed is part of the ICTU foundation, which houses a number of programmes, all of which aim to build the digital government.

ICTU
Nieuwe Duinweg 24-26
2587 AD The Hague
The Netherlands

Tel. +31 (0)70 888 77 77
Fax: +31 (0)70 888 78 88

Email testbed@nationaalarchief.nl
www.digitalduurzaamheid.nl

Digital Preservation Testbed *From digital volatility to digital permanence. Preserving databases* (version 1.0)

ISBN 90-807758-1-9

The Hague, December 2003.

Contents

Foreword

In the initial phase of the project the Testbed team needed time to become familiar with each other's disciplines. Although this sometimes proved difficult, it ultimately provided the quality required for these recommendations. The multi-disciplinary approach is reflected in this publication; after all, different employees with a wide range of backgrounds have to work together in your organisation too.

I would like to take this opportunity to thank everyone who has been involved in Testbed, either for a shorter or a longer period of time, for both their tremendous efforts and their great involvement and enthusiasm – in particular:
Remco Verdegem (National Archives), Tamara van Zwol (Ministry of the Interior and Kingdom Relations), Marjo Barthels (The Utrecht Archives), Maureen Potter (Audata), Unchalee Phimolsathien (ICTU), Carolien Nout (ICTU), Jules Ernst (Pareto), Bill Roberts (Tessella), Chris Rose (Tessella), Ingmar Evers (Tessella), Nancy McGovern (Audata), David Bowen (Audata), Jeff Rothenberg (RAND), Raymond Lorie (IBM), Raymond van Diessen (IBM), Hette Bakker (CGEY), Onno Stegeman (CGEY), James Buckthorpe (Tessella), Bruce Fairley (Tessella), Stuart White (Tessella),Chris Maré (Tessella), Sidney Huiskamp (IBM), and Eric Groen (IBM). I would also like to express my gratitude to Kees Dogterom and his staff at Studio Kader, who were responsible for the design of all our communications.

It would not have been possible to complete Testbed without the active assistance of and support from the enthusiastic members of the team, and from many others both in the Netherlands and abroad. We are also grateful to the Ministry of Transport, Public Works and Water Management, the Ministry of Housing, Spatial Planning and the Environment, the Ministry of Agriculture, Nature and Food Quality, the Ministry of the Interior and Kingdom Relations and the NIWI (the Netherlands Institute for Scientific Information Services) for their contribution of materials for our experiments.
Last but not least, I would also like to thank the following for their contributions:
In the first instance Hans Hofman (National Archives), who provided extremely valuable and detailed comments on our recommendations, as well as Helen Heskes (Public Records Inspectorate), Carolien Schönfelt (Amsterdam Municipal Archives), Jacques Bogaarts (National Archives), Jeroen van Oss (Municipal Archives of the City of Rotterdam), Petra van Santen (Tax Office), Hans Waalwijk (Archiefschool, Netherlands Institute for Archival Education and Research), Hanna Luden (Getronics), Abel Banus (Grafische Bedrijfsfondsen), Henk Duits (Municipal Archives of the city of The Hague), and Albert de Jonker (Amsterdam Municipal Archives).

Our work is complete; now it's *your* turn – whereby local and national authorities wishing to adopt a responsible approach to their digital information will find themselves confronted with a major task. Testbed has endeavoured to be as specific as possible in indicating the most logical (technical) solutions and the measures to be implemented by the various parties. I hope that this publication offers the necessary assistance.

Jacqueline Slats
Program Manager
Digital Preservation Testbed

Reading Guide

This publication of *From digital volatility to digital permanence* is comprised of four separate and self-contained parts. This document is Part 1, *Cost and decision models; Functional specifications; Preserving databases*. The other parts have already been published in the following sequence:
Part 4: Preserving e-mail;
Part 3: Preserving text documents;
Part 2: Preserving spreadsheets.

This publication has been written for all those involved in the appropriate management and preservation of government digital information. Testbed has endeavoured to avoid the use of jargon wherever possible, and where its use *was* unavoidable to at least explain the meaning of the relevant term. The actions that the various people or disciplines in an organisation have to undertake to preserve digital information properly, now and in the future, have been divided up by target group and can easily be found by way of the tab sheets.

This Part 1 of the series forms the concluding part of Testbed's studies into the preservation of digital information. This Part is the last and completes the series, since it contains supplementary information relating to all four parts, such as a cost and decision model and functional specifications for a preservation system.

The layout of this part is structured as follows. Chapter 1 is an introductory chapter to the digital government, the problem of digital preservation, and the assignment given to the Digital Preservation Testbed to decide on the most appropriate preservation strategy through practical experiments.

In chapter 2 you can read about how digital records differ from paper records. We look in detail at the specific properties of digital records, explaining the five main characteristics of a digital record: content, context, structure, appearance, and behaviour.

Chapter 3 discusses the type of record specific to this publication, namely databases. What exactly *is* a database, and which authenticity requirements are relevant? Or, in other words, which criteria need to be met by databases if they are 'not to lose their authenticity' – and thereby assure all those involved that the database is indeed what it purports to be.

Chapter 4 discusses various preservation strategies that are receiving a great deal of worldwide attention. Testbed assesses these strategies in relation to databases.

Chapter 5 then looks at the preservation strategy that has emerged from our research as the most promising for the durable preservation of databases. The chapter also discusses an implementation method.

Chapter 6 contains a concrete plan of action for the various target groups within a (government) organisation, i.e. managers, records managers, ICT specialists, and end users. Each target group has been assigned its own specific responsibilities in this plan and this chapter gives them the information to enable them to contribute to building a reliable digital government.

The publication concludes with a glossary, a bibliography, and the following four appendices:

Appendix A: Preservation transaction log
Appendix B: Decision model
Appendix C: Cost model
Appendix D: Functional requirements for a preservation system.

1. The Dutch Digital Government

Great ambitions have been expressed over the last few years with regard to a better performing government. The digital government is under construction on many fronts and there are wide-ranging initiatives at local, regional and national levels. Digital preservation however, is not always getting the attention it deserves. Action is needed because a digital government cannot exist without digital memory.

1.1 Developments in digital government

The Dutch government is increasingly working with digital records. The second Kok government formulated its aim of having 25% of the transactions between the government and the public take place digitally by 2002, an aim that was then easily achieved. In the meantime, the government has set new targets: by the end of 2006, 65% of all transactions between the government and the public must be dealt with electronically. Meeting this target fits the image of a government that is operating effectively, whereby rules have been simplified, bureaucracy has been reduced to a minimum, and citizens need to submit data only once. This is in accordance with the 'Transfiguring Government' action program that was presented by the Minister for Government Reform and Kingdom Relations, Minister Thom de Graaf, in December 2003.

The advantages of working digitally are, in as far as they are still a topic of discussion, enormous. Firstly, digital information is *more accessible*, to the public, but also to other governments. The World Wide Web, www, is also a significant source of information. Governments can be better controlled if they make their information easily available to, for example, the National Audit Office or Inspectorates. They can in principle produce *better work*, because information is available in a more complete form and can, for example, be used more than once. *Service* to the public can be delivered faster, and *better.* Take, for example, applying for official documents, or identifying hazardous business in a region (as the province of Friesland does on its website www.fryslan.nl), to inform the public and business more adequately. Finally, working digitally not only provides organisational benefits, but also financial ones. Millions of euros can thus be saved [1].

Now, little by little, everyone has become convinced of the advantages of a digital government, but its problems are sometimes difficult to identify or tackle. More digital transactions between the public and the government mean massive changes to the back offices of government organisations, in other words, information management. Besides keeping the back office running well, transparency in its work and the continuing accessibility of information are problems requiring an urgent solution. This last point, the continuing accessibility of digital information, is examined in detail below.

[1] See *Winst met ICT in uitvoering*, A. Zuurmond, K. Mies; Zenc, The Hague, June 2002.

1.2 Working effectively means managing digital longevity

The fact that the government now has to preserve information not only on paper but also digitally is registering with an increasing number of organisations. Durable digital work is the slogan. This means creating, storing, and managing digital records, making them accessible so they are still available for consultation and are authentic even with the passage of time.

Managing digital longevity is not simply a question of technology. Government organisations must (if they are not yet doing so) recognise the problem of digital longevity and be prepared to do something about it. That means making finances available and giving the subject some attention: formulating and implementing policy, regulations and procedures; buying and installing technical and other tools; and training and instructing staff. Individual employees, too, must recognise the need for policy, regulations, and procedures, and must be prepared to observe them. That will only be the case if these things do not or barely hamper them in their normal work and if the supporting technical tools make things easier for them.

Furthermore it is important that government organisations can choose from a wide range of software applications available on the market, applications in which durable preservation of text, images, pictures, sounds and combinations of these is integrated from the outset (in other words as soon as the information is created).

1.3 Working digitally also means preserving digitally

The government has built up several centuries of experience with paper records and registries; it only came into contact with digital records a few decades ago. The specific properties of digital records mean that the procedures for paper cannot be used (this is discussed further in the following chapter).

Digital information differs substantially on certain points from paper information. Digital records do not have a fixed form and are often made by several people. In the past, special archive departments made sure that records were managed in compliance with the law and job responsibilities. Nowadays, because of ICT, government employees have access to many new ways of making records, which vary from text documents and email messages to spreadsheets and databases. Correspondingly, the management of these records is becoming further removed from the supervision of the department responsible for them. Existing procedures and regulations for paper records are not applied to digital records, and they lead a risky existence.

Although this gap in the operation is part of the learning process in the transition from paper to digital records, this development must not continue. Even in the digital age, records must be made that can survive the ravages of time. They must also be managed properly. This is not the case for most of the records made nowadays.

On the one hand therefore, the problem is related to information management in organisations. On the other hand, the problem of preserving digital records lies in the speed of hardware and software obsolescence. If nothing is done, digital information will be lost because it will no longer be readable or accessible. The period we are talking about is short: information may become unavailable after just one or two years.

The consequences of this could be that important information disappears and that it is, no longer possible to reconstruct, for example, a government decision-making process. A recent example of this can be found in the parliamentary inquiry into Srebrenica by the Bakker committee (January 2003). Witness statements were sometimes taken by email, but how were they to be preserved? It is not enough to print them out. After all, an official digital record must be digitally preserved (see also chapter 2 for details).

Another example relates to retrieving information, such as in the question of how many unemployed people an administration agency has helped to find work in the last few decades. This question will not be properly answered if the information management of an organisation is not in good order, or not properly discharged. This subject was the central theme of the symposium that the Digital Longevity project organised together with the *Arbeidsvoorziening* in November 2002. In short, proper preservation (including for the long term), retrieval and re-use of digital data are the keywords.

Government digital services are under construction. The question might yet be asked whether a digital permit issued by a municipality still has exactly the same meaning after five years and three conversions to more modern software.

In short, the examples given above encroach directly on the way the government operates. The continuity of operations, the external responsibility of the government, and future generations studying how the government worked: all this is only possible if there is a good, reliable method for preserving digital information.

1.4 Digital preservation and the law

The government has partly recognised the importance of digital preservation and has changed certain parts of existing legislation to reflect this. A brief summary of these laws and guidelines is set out below.

The 1995 Archives Act
In article 1, part c of the Archives Act, the following definition of archival records is given: “records, *regardless of their form*, received or drawn up by government organisations…”.
Every document, paper or digital, that has a function in the performance of a task is therefore in principle a record or an archival document.

The Regulation on the Arrangement and Accessibility of Records (2002).
The Regulation on the Arrangement and Accessibility of Records is an extension to article 12 of the 1995 Archives Decree. The Regulation states that the most important requirements are that records must be authentic and that records must be readable and retrievable within a reasonable period of time. There are extra requirements for digital records, including databases. These refer to such matters as retaining metadata on the content, form and structure of a document, and technical data on conversion, migration and storage formats.

Open Government Act (WOB) (1998)
When archived records from government organisations are transferred to an archive depository, they are in principle made public by virtue of the Archival Law1995. Whilst records are still stored in government organisations, their public status is organised differently. In these cases, the WOB comes into effect. The WOB gives everyone the right to request information from a government body. In this, as in the Archives Act, no distinction is made between the type of information carrier for the document, whether it is on paper or digital.

Personal Data Protection Act (2001)
The Personal Data Protection Act has also been tightened up to include records in digital form. The same legislation now applies to both paper and digital records.

In summary, it can be said that awareness-raising amongst organisations and their employees is a pre-condition for preserving information properly, particularly in the digital age. A few legislative offerings have already been made. The question now is whether technology can offer a simple solution for effective preservation in both the present and the future.

1.5 A technical solution on hand?

All over the world ICT experts and scientists are busy seeking answers to the question of how digital information can best be preserved. Several existing approaches appear to offer good potential for dealing with the digital outpourings of government actions in a responsible and sustainable manner. We will examine these strategies in detail in chapter 4.

The problem at the moment is that there is no *ready-made* solution for government organisations that really want to start building their digital memory. Which preservation strategy an organisation ought to choose and which facilities ought to be bought are questions to which there is not yet an answer. Additionally, most strategies are, in practice, untested.

To research solutions for this situation, the Ministry of the Interior and Kingdom Relations and the Ministry for Education, Culture and Science, (in this case the National Archives), decided to set up a 'Testbed' to gain knowledge and experience of the sustainable preservation of different types of digital records through experimental research: Digital Preservation Testbed.

The Digital Preservation Testbed was begun in 2000 and carries out experiments defined around a series of solution-oriented research questions, in order to decide which preservation strategy or combination of strategies is most suitable. Testbed focuses on three different, largely theoretical methods for the long term preservation of digital information, namely migration, XML and emulation. Not only are these methods assessed in terms of their effectiveness, but also in terms of their limitations, cost and possibilities for use. As part of its work, Testbed takes account of the legal and policy-induced context outlined above.

The Digital Preservation Testbed team is made up of an international group of experts in the field of archives, ICT, information management and communication.

1.6 The Digital Preservation Testbed assignment

The Testbed team set to work on the assignment from the departments. A unique laboratory environment was built in which to assess and evaluate the approaches, using a system the team designed and built themselves that contains all of the research data. The experiments and tests that are performed are completely reproducible and scientifically sound. The recommendations are freely accessible on the website http://www.digitaleduurzaamheid.nl.

The Testbed project is delivering the following products and services:

- Knowledge and understanding of technical solutions for the long term preservation of digital records
- Advice on how to deal with current digital records
- Well-substantiated strategies for the long term preservation of four types of digital records: text documents, spreadsheets, e-mails and databases
- Functional requirements for a preservation system for digital records: i.e. the functional specifications for building a preservation function into a records system
- Cost models for the different preservation strategies:
  What are the cost indicators when implementing a particular preservation strategy?
- Decision model for preservation strategies (as an aid to determining which preservation strategy is the most suitable, given a particular record type)
- Proposals for altering current legislation and rules.

2. Digital Records and Authenticity

What makes digital records so special? In this chapter we examine the properties and characteristics of digital information. We also look at the key concept of 'authenticity', because it is essential that a record can be guaranteed authentic: once preserved, a record may not be significantly changed.

2.1 Definition of a digital record

Digital records are not simply the 21st century equivalent of traditional paper records. They have other properties, characteristics and applications. However, both digital and paper records must meet the same legal requirements. In practice, this requires a different approach.

Digital records are not tangible objects like a book or a magazine, but a combination of hardware, software and computer files. This combination is necessary to be able to use the documents or examine them. In the context of Testbed we looked specifically at text documents, databases, email messages and spreadsheets. Multimedia documents, digital video and sound can also be digital records, but these remained outside the scope of this study.

An important difference compared to paper records is the greater loss of information that can occur even while the records are being used, or afterwards when the records are being maintained. After all, hard discs and computers are replaced regularly and there are few barriers to destroying computer files. A single click on the 'delete' button and a record can disappear without leaving a trace.

To analyse the problem of technological obsolescence and to test suitable preservation strategies, Testbed makes a distinction between four aspects of digital records:

- The concept of a ‘digital record’ as a combination of hardware, software and computer file;
- The concept of ‘authenticity’ in digital records;
- Digital characteristics;
- Metadata for safeguarding the authenticity of digital records.

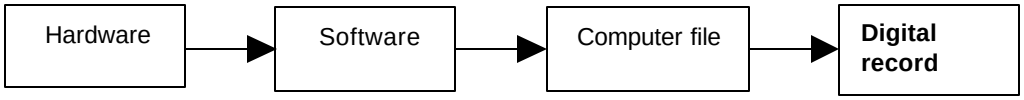
These aspects will be developed further in the sections below.

2.2 The digital record as a combination of hardware, software and computer file

In the paper age, the concept of a 'record' was simple. The record as evidence of a transaction was recorded on a physical entity like parchment or paper, possibly in the form of a charter, a receipt, a letter, a memo or a photograph.

In the digital age a record is not fixed in the same way. Digital files have to be processed technically before the user can read the records and use them for the purpose required. It is this dependence on hardware and software that compels us to think differently about the way we make and use digital records.

The diagram below shows the components needed to reproduce and use the digital record[2]:



A digital record is made using a particular combination of hardware and software and is stored in the form of a code, the computer file. This computer file consists of a series of ones and zeroes. This series of ones and zeroes is read by a certain application and interpreted in a way that is often unique for that application. The result of that interpretation is then shown on the screen and that representation is the digital record.

In most cases the computer file can only be correctly read by way of the above-mentioned combination of hardware and software. If the digital record is reproduced in a different computer environment than that in which it was originally made, it may look and behave entirely differently. If the transition to the other computer environment is not controlled, the authenticity of the digital record may be affected.

2.3 Authenticity as a key concept

Authenticity is a key concept in the preservation of records. Authenticity means that a record is what it says it is. It may not be illicitly changed or corrupted. A decision taken by parliament, for example, is recorded on a paper record that includes the date and the names of the parties involved. These names and dates add value and credence to the record, and nothing may be changed on that parliamentary record once it has been made. If changes are added to this type of record, they can usually be easily identified.

It is less easy to decide whether a digital record is authentic. The problems this can cause must not be underestimated. In September 2001, for example, the Dutch Christian Democrat party (CDA) found itself involved in an internal crisis. A policy official in the CDA parliamentary party in the Lower House played a crucial role by editing a digital report in such a way that it seemed as if an opinion poll had revealed that the parliamentary party leader Mr De Hoop Scheffer had a weak image. The document was passed on to a current affairs column. By the time people discovered that the document was not authentic, the damage could not be repaired, and both the chairman of the party, Mr Van Rij and Mr De Hoop Scheffer resigned. It cost the CDA parliamentary party a great deal of effort to find the culprit. An external IT company had to inspect all the personal computers to trace the culprit, but he was eventually found.

According to the Testbed definition, authenticity is the representation of a record completely and entirely in accordance with the original recording and function that it was intended to fulfil. Authenticity has two central concepts:

Integrity: that the record is intact and not changed or corrupted in such a way that it's meaning is no longer clear. A record has integrity when it is complete and uninterrupted in all essential aspects. Changes are acceptable to a certain extent, as long as they do not affect the original meaning or function of the record. An example of this is the website mentioned above that belongs to the province of Friesland, which

[2] InterPARES Authenticity Task Force Final Report, http://www.interpares.org/book/interpares_book_d_part1.pdf .

has maps showing the position of hazardous businesses indicated in colour. The colours on the map have a significant meaning and must therefore be preserved in their original condition. Converting this record to a higher version of the file format that changes the colours (red becomes green, for example) would affect the integrity of the record.

Verification (or Authentication): that the record is what it says it is. Authentication allows us to confirm that a record, digital or otherwise, is what we think it is and that it was made by a specific organisation or person. Information is required to determine if a record is authentic, concerning both the initial meaning of the record as well as how it has been managed since then. This can be guaranteed by establishing the provenance of the record and ensuring its adequate and uninterrupted management ('unbroken chain of custody').
In general, it will be assumed that the information displayed in a record is authentic; it is primarily a matter of trust. In the event of uncertainty, an investigation (verification) can be carried out to confirm the essence of the information.

Insofar as the 1995 Archives Act is concerned [3], it makes no difference whether a record has a digital or a physical form. The problem that arises with digital records, however, is that due to changing technology, not all aspects of a record can be preserved as precisely as when it was made. This does not mean, though, that sustainable preservation of authentic digital records is impossible.

2.4 Digital records, digital characteristics

In the paper age the characteristics of a record formed a physical entity. The characteristics context, content, structure and appearance make a record authentic. If one property is changed, it has an effect on the others. For instance, the structure of the paper record, such as in the breakdown of a piece of text into chapters, is represented in its appearance. The appearance of the record, for example a complete publication with tab sheets, in turn displays the entire content of the record, comprising many references to the context such as the author's name or the publication date. All these aspects of the paper record, i.e. context, content, structure and appearance, are fixed and can no longer be changed after the record has been published.

Digital records are different. It is true that they still have the four characteristics mentioned above, but they can also have another characteristic: behaviour[4]. In contrast to paper records, however, the characteristics of digital records are not as firmly connected to each other. They are highly dependent on the way in which the software interprets the computer file. This makes them much more susceptible to unwanted changes. Monitoring these characteristics and their relationships thus requires extra measures.

Dutch legislation and regulations refer to context, content, structure and form. The characteristic 'behaviour', which can be important for digital records, is not mentioned. In addition, current regulations define the concept of 'form' as 'the outward appearance in which the structure and layout are visible"[5].

---

3 Archives Act 1995, article 1c "Archival records are records, despite their physical form…".

4 Carrying Authentic, Understandable and Usable Records Through Time, Rothenberg, Jeff & Bikson, Tora, The Hague, 1999.

5 See article 1, section 1, sub o of the Ministerial Regulation on the Arrangement and Accessibility of Records.

For the purpose of its research, Testbed has broken down the characteristic 'form' into two unique attributes, and distinguishes between structure and appearance as separate characteristics of a digital record. The five characteristics of digital records are explained in more detail below.

Context[6]
'Context' here refers to the original environment in which the digital record is made and used. In order to interpret the record and give it meaning, a specific amount of information about its originating context is required. This information relates solely to the record, separate from the medium, and does not necessarily include the technical environment in which the record is made and used. This information relates to the function, the business process and the government body in the context of which the digital record is received or made. In addition, the relationship with other records, including those from the same case file (dossier) and the same business process, has to be described and preserved. Dossiers are an example of this.

Content
Databases can be used for a wide variety of purposes and can thus contain a wide variety of information, ranging from data on births, deaths, & marriages registration, and patient data, to the complete administration of an organisation.

Structure
The structure of a digital record is given shape by the logical hierarchy of and the relationships between the various sections of a record. The structure of a (relational) database relates to the various tables from which the database is constructed, the mutual relationships between the tables, and the construction of the individual tables (comprised of records, or of rows which are in turn comprised of fields). The (partial) loss of this structure from a migration will result in a database that no longer presents all the information in the correct manner.

Appearance
The 'appearance' of a digital record refers to the ultimate presentation of that record, i.e. the form in which the digital record is displayed onscreen. The appearance includes characteristics such as the font, font size, and the use of underlined, bold or italic letters, etc. With a database, the appearance primarily relates to the application that uses a GUI (Graphical User Interface) to access the underlying database. The application enables users to add, amend or delete data. In this instance, 'appearance' refers in particular to the onscreen appearance of the data as presented by the application whilst the user submits queries to and updates the underlying database.

Behaviour
The behaviour of a digital record is the most difficult to preserve. 'Behaviour' refers to the interactive characteristics of a record. In the instance of a database, the behaviour is primarily linked to the application accessed by the user to submit queries to and update the database. Behaviour plays a more important role in databases than in e-mail messages, text documents and spreadsheets.

It should be noted that the importance attached to these characteristics (context, content, structure, appearance, and behaviour) is primarily determined by the relevant business process. However, the importance attached to each characteristic can vary according to the nature of different types of records (e-mails, text documents, spreadsheets, and databases). It can generally be assumed that the appearance of e-mail messages will be of lesser importance, since the display of e-mails will vary between PCs which use different e-mail programs and have different personal

[6] Een uitdijend heelal? Context van archiefbescheiden, H. Hofman, Stichting Archiefpublicaties, Jaarboek 2000.

settings. Conversely, for text document records the appearance will be of essential importance. The five aforementioned characteristics play an important role in the evaluation of the various preservation strategies discussed in Chapter 4.

2.5 Metadata

Metadata is data about data. We add metadata to a digital record to describe extra information about the five characteristics of a record mentioned above so that, among other things, checks can be made on whether the record is what it 'says' it is. At the same time, metadata makes it possible to retrieve and use a particular digital record. Examples of such data are author of the record, subject, business process in which the record was created, and date on which the record was created. But metadata is also important in the context of registering that preservation activities have been carried out.

A distinction can be made between a number of categories of metadata:[7]

- Institutional context
  This category of metadata focuses on contextual data that imparts significance to the digital record: the person or organisation, the function, the mandate, and the business processes.
- Management data
  The management data encompass the intellectual management (for example the arrangement and classification codes for the records), the administrative management (for example, the location, size, frequency of consultation), the technical or physical management (such as processes carried out on the record relating to, for example, conversion or migration, and a description of the result), and the technical context (both the technical environment in which the record was made and that in which it is currently stored).
- Metadata relating to structure, appearance, and behaviour
  This metadata forms the third category and describes the essential (authenticity) characteristics of the digital record, for example the presence of a hyperlink to a specific website.

We can use metadata to create an image of the digital record without actually having to reproduce the record in question. Metadata is part of the digital record and accompanies a digital record throughout its life cycle. It contains information about the creation of the digital record and preservation activities that have been performed. Metadata is therefore vitally important.

Metadata can be used to ensure that the right preservation action is taken. It can be used it to check, for example, whether the essential elements of the digital record are still the same following a migration, and whether the record has or has not been affected. Metadata thus forms part of the evidence that a document is authentic.

---

[7] Blijvend in business, naar een geordende en toegankelijke staat van informatie, *Bijlage 2 Overzicht van metagegevens*, Hans Hofman, The Hague, 2003.

3. Preserving Databases in an authentic state

Databases have assumed an important position in today's government as a result of the opportunities they offer for the processing and management of large quantities of structured data. Our starting point is based on the principle that the databases must be preserved in an authentic state. To this end it is necessary to specify both the essential characteristics of databases and the authenticity requirements governing their use.

3.1 'Will the real digital record please stand up'?

During the initial years of the automated era databases consisted of nothing more than automated card-index boxes. Each card-index box was in the form of a collection of independent files. In the sixties it became possible to link the various files with each other. For example, the Land Registry is not only able to furnish information about Cadastral plots and any buildings on those plots, but can also provide information about the identity of the owner and the price the owner paid for the plot. However, this is not the sole function of a database system; it can also verify that the data are correct and complete, offer the functionality required to maintain the data, and protect the data from unauthorised use.

A database system consists of three components (see Figure 1):

- the database itself (the actual content);
- a DataBase Management System, DBMS (for example, Oracle 9i);
- the database application. This incorporates both the graphical user interface and the functionality the user needs to search through and process the content of the database, as well as programs that function automatically to support the system in processing inputs and outputs.

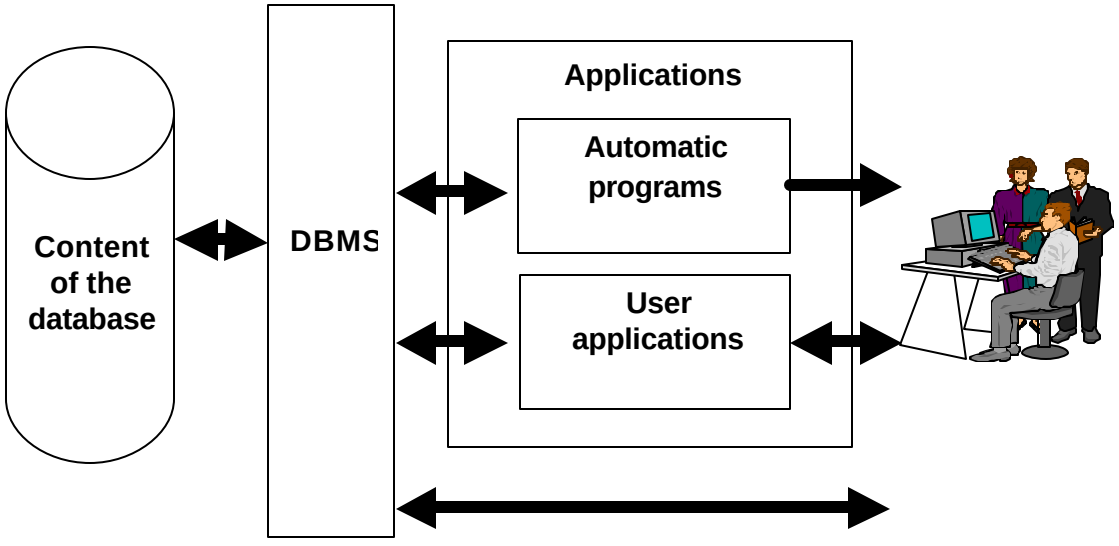


Figure 1: *The components of a database system*

User applications enable users to update the database, whilst automatic programs initiate processing operations in the background. An example of a user application is the processing of the transfer of ownership of a building plot from owner A to owner B. An example of a program running automatically in the background is the maintenance of a log file: which users are logged into the system, when do they log in and out, etc.

The DBMS (DataBase Management System) is the software layer between the physical database and the user. The DBMS processes all user queries relating to the database, including maintenance of information on the physical details of the location of files and file formats, descriptions of the indices, etc. Moreover a DBMS can be designed to comply with the requirement for centralised control of the security and integrity of the data. In addition, the DBMS enables the user to define, create and maintain a database. DBMSs are based on data models. A variety of data models are in use:

- Relational (RDBMS)
- Object-oriented
- Native XML
- Hierarchical
- Network
- Associative

The DBMS market continues to be dominated by relational databases. For this reason, Testbed focused primarily on this type of database (-model).

It is relatively straightforward to identify e-mail messages, spreadsheets, or text documents as digital records. However, this is not the case with a database. There are a number of possibilities:

- the complete database system (database, RDBMS, and application) together constitute the digital record;
- the database is the digital record;
- a single row of data stored in a database table ('tuple') is the digital record;
- data distributed over a number of tables constitutes the digital record;
- information in the database as displayed onscreen by the application forms the digital record.

From the above summary it is evident that it is also impossible to give an unequivocal answer to the question of which information should ultimately be preserved, since databases vary in their nature and each one is different to the next. For practical reasons it was decided that the research into the long term preservation of databases would focus on the storage of the actual database and the user application employed to display the data onscreen.

Organisation and organisational culture

Whilst government bodies have for many decades made use of complex database systems such as Oracle, with the emergence of office automation these systems were supplemented with desktop databases, for example Access.

Complex database systems undergo continual modification to accommodate the new requirements imposed by the organisation, and on average are replaced every five years (i.e. the re-building of an existing system or the construction of a new system). The management, design and construction of such systems is often centrally organised by IT specialists.

Desktop databases are designed, constructed and managed by the user. Moreover the users often then decide what they do and do not wish to preserve: they can exercise their discretion in saving, amending and deleting the data in the database.

The situation that arose on the dissolution of the *Arbeidsvoorziening* (the Dutch Public Employment Service) in 2001 revealed the need for a long term preservation strategy. The *Arbeidsvoorziening* had to tidy up its digital files before they could be transferred to its legal successor(s), in a labour-intensive and expensive operation that cost several million Euros. In the first instance the data amounted to terabytes of information in databases belonging to more than ten different systems. Moreover the need for the preservation of this data was made all the greater by claims the European Social Fund (ESF) had submitted for the reimbursement of millions of Euros.

Legal aspects
The existing legislation, inclusive of the *Regulation on the Arrangement and Accessibility of Records*, offers a framework for the preservation of digital records. These regulations and other legislation were discussed in chapter 1.

Article 6e of the Regulation stipulates that at the time of the transfer to an archival institution the databases shall be stored in the following manner (translated from the original Dutch):
'*in the original storage format or ASCII (flatfile, with field separators) and accompanied by documentation preferably in XML-DTD relating to the structure of the database, and at least extending to a complete logic data model with a specification of the entities; queries shall be specified in the SQL (SQL2) query language.*'

Testbed has assumed that SQL2 was cited in the Regulation, since at the time the Regulation was drawn up SQL2 was the latest version of the SQL standard. The current version is SQL3.

Technical issues
Hardware and software rapidly become obsolete, as a result of which digital files are no longer accessible. Virtually no practical studies have been carried out, either at a national or an international level, into technical approaches to the durable preservation of databases.

Organisations invest a great deal of time and money in the maintenance of their database systems, since they are often used in processes of crucial importance to their operations. Much of this maintenance work is of a technical nature and insufficient account is taken of the need for the long term preservation of databases. Similarly, desktop databases created by the user are not usually designed in a manner which immediately takes account of their long term preservation.

As a result of the complexity of these problems, it is rare that database systems are preserved in an appropriate manner. The solution is to be found in an approach that addresses all the issues (organisational, legal and technical) mentioned above. As such, a technical or legal solution is not enough: it is also necessary to create an awareness of the importance of preservation in an appropriate manner. To develop such a practical approach, first the essential characteristics of databases must be described and the authenticity requirements defined, in other words: which criteria must a well-preserved database meet?

3.2 The status of databases

Not all databases have to be preserved. The choice of which databases must be considered for preservation depends on the selection criteria specified on the basis of an analysis of the tasks of the organisation that owns the database. This is described in an Institutional Research Report (RIO). The Basic Selection Document (BSD) based on this Report forms the foundations for the decisions as to whether records

relating to governmental actions should be destroyed or transferred to an archival institution for long term preservation.

3.3 Characteristics of databases

The relational database model offers a simple conceptual model for data, namely related tables. An example is shown in Figure 2. This simplified example contains three tables, namely: owner, plot and buildings. The data are entered in these tables. Figure 3 shows an example of a summary of entered data. The table entitled 'owner' contains the fields 'owner_ID', 'name' and 'address'. In this example the owner_ID is used as the unique (primary) key for each owner in the database. In this example each owner can own one or more plots, and each plot can be empty or accommodate one or more buildings (houses, apartments, etc.). Conversely, in this example each plot can have only one owner, and a building can stand on only one plot. The relationships between the tables are also referred to as 'constraints'; they define the relationships between the tables.

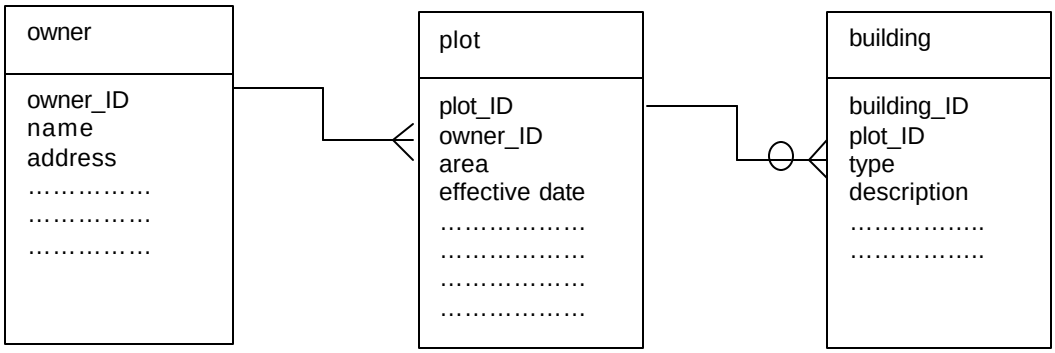


Figure 2:*An example of a relational database structure*

Owner_ID	Name	Plot_ID	Area	Type of building	…….
……..	……..	……	….	……	…….
834456459	De Wilde	153	300	Single-family dwelling	…….
834456459	De Wilde	153	300	Semi-detached house	…….
834456459	De Wilde	201	1000	Villa (30's style)	…….
583490528	Janssen	50	200		…….
……...	……..	……	…..	……	…….

Figure 3:*An example of the registered data*

In order to interact with a relational database, SQL (Standard Query Language) has been designed. The current version is Version 3, also referred to as 'SQL3'. An important part of the SQL specification is the language specification for queries: commands have been defined to retrieve specific information from one or more tables in the database.

Complex database systems often exhibit the following characteristics:

- They are developed using a combination of DBMS for the database and a separate programming environment for the development of the application.
- Their design, construction, and maintenance requires specialist skills.
- It is possible to convert the content of the database to another database format, or it can be migrated to a new version of the same DBMS. Specialised skills are also required for this conversion/migration.

Examples of complex DBMSs are Oracle products, SQL Server, and DB2.

Desktop databases often exhibit the following characteristics:

- The DBMS can be used as a complete solution for both the database itself and the application.
- Software of this nature is intended for end users, and consequently is easier to use than the software designed for the development of complex database systems. Although users need to receive training, this is insignificant in comparison with the amount of specialised training required for complex database systems.
- Users can migrate their database system to a different database system format or a new version of the same DBMS themselves.

Examples of desktop databases are Access and dBase IV.

3.4 Authenticity requirements for databases

As discussed in chapter 2, the concept of authenticity is of great importance to the preservation of information. For each type of digital record, such as spreadsheets, e-mail messages, text documents and databases, the authenticity requirements can differ. These requirements play a crucial role in the selection of a preservation strategy. The requirements are determined by the business process in which the record plays a role, and by the requisite legal context (see *Regulation on the Arrangement and Accessibility of Records*).

The following requirements relate to the characteristics of a digital record: the context, content, structure, appearance, and behaviour. In addition, the organisation can impose supplementary authenticity requirements on the basis of the relevant business process.

As explained in section 3.1, it was decided for practical reasons that the study of the durable preservation of databases would focus on the storage of the actual database itself and the user application.

Testbed has experimented with databases, and with strategies to assure their authenticity. The results from these experiments have been used to draw up guidelines specifying a minimum set of authenticity requirements, identifying the minimum characteristics of a database itself that must be preserved in order for the database to be properly represented. These are supplemented by additional authenticity requirements relating to the user application.

Context
All databases need to be accompanied by metadata, such as the organisation's name, tasks, and the business process – in other words, the institutional or organisational context. In addition, the technical context of the database – such as DBMS – must also be identified if the database is to be preserved in an efficient manner. A further important element of the contextual information is the relationship with other records expressed, for example, in the form of a classification code or dossier. Finally, all preservation actions and their results must be registered so as to ensure the authenticity and permanent accessibility of the database in the future.

Testbed has identified the following set of minimum authenticity requirements for the context:

Database
The specification of the organisational context, such as:
- the organisation's name
- the business process
- the relationship with other files

The maintenance of a preservation logbook that contains at least the following information:
- Information about the original and current file formats
- Information needed for the interpretation of the current file format (for example, the name and version of the DBMS, the name and version of the operating system, and the name and type of the hardware equipment)
- Information about the preservation actions that have been taken, such as the date, time (for example, using a 'timestamp'), and the person or persons responsible.

User application
The maintenance of a preservation logbook that contains at least the following information:
- The name and version of the query languages that are used.

Content
The content of a database is of vital importance. A database can contain numerous sorts of content, such as flat alphanumeric text, Unicode text, and embedded objects.

Testbed has identified the following set of minimum authenticity requirements for the content:

Database
The actual content of the tables must always be preserved.
The content of the various database tables must remain intact and legible.

User application
The content of the database displayed onscreen must be preserved.
SQL is used to retrieve the data from the database. The queries that are used must be preserved so as to be able to represent the required content.

Structure
The structure of a database relates to its composition and the logical hierarchy of the elements of a database. A database consists of one or more interlinked tables. Each table consists of a number of rows, which are in turn constructed from a number of fields (see also section 3.3 for the principle of relational databases). In the absence of structure, a database is nothing more than a disjointed collection of data.

Testbed has identified the following set of minimum authenticity requirements for the structure:

Database
The physical structure must be preserved.
The tables, the relationships between the tables – including the constraints (such as the name, type, columnName, referenceTable [solely for 'foreign keys']) - the views and the field attributes (such as the name, data type and field length) must all be preserved.

The logical structure of the database must be preserved.
The logical data structure of a database can, for example, be presented in the form of a so-called 'ERD' (Entity Relationship Diagram) or an XML schema.

User application
The structural composition of the data as presented onscreen must be preserved.

Appearance
The term 'appearance' refers to the manner in which the content of the database is displayed onscreen. The appearance usually conveys a certain meaning. For example, the use of a currency notation or a date notation with the display of certain figures can impart an additional meaning to those figures. Font sizes and colours can also be used to emphasise certain fields. For this reason a specific appearance can indicate an additional significance that cannot be conveyed solely by the content and the structure.

Testbed has identified the following set of minimum authenticity requirements for the appearance:

Database
None.

User application
The onscreen representation must be preserved.

Behaviour
'Behaviour' is a property possessed solely by digital records and not by their paper counterparts; a paper record does not exhibit an (active) behaviour. Behaviour is often linked to (or made possible by) the user application that is used to generate and manipulate the digital record. Consequently the preservation of the behavioural aspects of databases requires the presence of the user application designed and developed to enter data, submit queries, and manipulate the data in the database.

Testbed has identified the following set of minimum authenticity requirements for the behaviour:

Database
None.

User application
The behaviour of the user application must be preserved.
The behaviour can be preserved in the form of descriptions of system-supported functions, as well as screenshots of the displays used for entering and amending data, generating reports, etc. In this instance the preservation relates to the system documentation and the user's manual. This information can be stored in both paper and digital form.

3.5 Summary

The applied concepts of integrity and verification are essential to establish the authenticity of digital records:

Testbed has specified two sets of minimum authenticity requirements: one for the database itself, and one for the user application (optional).
The characteristics of the database must be preserved in accordance with the minimum set of authenticity requirements for the database. Furthermore, each organisation must establish the essential characteristics of the databases they generate in their various business processes and determine whether the authenticity requirements for user applications are also applicable.

The contextual data relatesto information about, for example, the business process in which the database has been generated and used. This information is necessary to understand the content of a given record and its relationship to other records. The contextual data also contains information about any changes that may have been made in the database in connection with the required management and preservation activities. This information can be used to demonstrate or verify the extent to which a database can still be deemed authentic, even when that database system is no longer exactly the same as the original.

4. Three Preservation Strategies Researched

The most well known strategies for preserving digital information in a sustainable way are migration, XML and emulation. These methods, which have been studied throughout the world, will be discussed here briefly and assessed on their suitability for preserving databases.

4.1 Introduction

Migration, XML and emulation are the three basic approaches most often discussed for preserving digital records. Each preservation strategy has a number of sub-categories, which we will also discuss in this chapter. At the same time, where possible, we will describe how each strategy might be implemented. The advantages and disadvantages of each strategy will be assessed in the light of the specific requirements placed on the long term preservation of databases, as described in chapter 3. Based on these considerations, we will decide which is the most suitable strategy for the long term preservation of databases.

4.2 Migration as a preservation strategy

Digital Preservation Testbed applies the following definition to migration:

“The transfer of records from one hardware/software environment to another”.

Migration is a common way of tackling digital obsolescence. Records created in an old format are transferred to a new format that will run on modern computer platforms. A database can, for example, be transferred from FoxPro version 3.0 to Access 2000.

Every migration requires advance research. After all, the target format must be compatible with the source format so that all the important properties of the digital record are represented in the converted version and the authenticity and integrity of the digital record are safeguarded.

The following diagram shows the relationships between the hardware, software and data when migration is used:

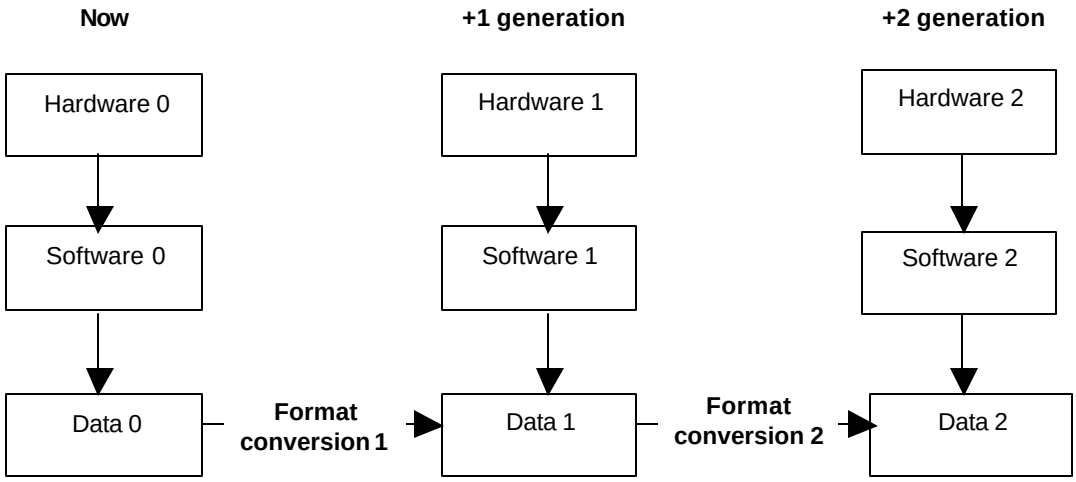


Figure 3. *Basic migration diagram*

Testbed has studied and experimented with the following forms of migration:

- Backward compatibility
- Interoperability
- Conversion to standards

In choosing the most suitable approach, an organisation must first take into account the authenticity requirements of the digital records they are working with. The length of time the digital record has to be preserved is also a determining factor: two years, ten years, twenty years or in perpetuity?

4.2.1 Backward compatibility

Backward compatibility makes it possible to interpret and correctly reproduce a record that was made in an older version of an application, using a later version of that application. Software suppliers often guarantee that new versions of their software are compatible with previous versions. For example, Access 2002 can be used to read files created in Access 97 and saved in the Access 97 format.

With databases both the database itself and the user application can be migrated. A great deal of program code is often involved in the development of complex database systems. This source code is compiled to create the user application, the so-called object code. The same can be said for the program code used to communicate with the DBMS. The SQL as implemented in most DBMSs is supplemented with additional proprietary software, and modified, in comparison with the SQL standard. A migration is therefore normally dependent upon the tools provided by the software supplier.

A disadvantage of backward compatibility as a preservation strategy is that the digital record continues to be stored in the supplier's own file format (for example, *mdb for databases created with Access). From the perspective of digital longevity, this retains an undesirable dependency on the original application software.

A further disadvantage is that migration to a higher version must be repeated every few years, since compatibility is often restricted to only a few generations of the application. Even then, it is still possible that the new version of the software will interpret and display some properties of the record in a different manner.

Is backward compatibility suitable for preserving databases?

Backward compatibility is a suitable preservation strategy for databases that only need to be preserved for the short term. Testbed experiments have demonstrated that such migrations can generally be carried out without significant problems, and that the authenticity and integrity of database systems are not placed at risk. It is preferable to save database systems preserved using this strategy in the new version's format, since the software usually only supports a limited number of older generations. Consequently it will be necessary to migrate to a higher version every few years, although it is possible to skip a number of versions. However the greater the time scale between the source- and target- formats, the greater the risk of data loss and problems during the migration.

In view of the disadvantages of backward compatibility as a preservation strategy (storage in the supplier's file format, the need to repeat the migration every few years, and the risk of adverse effects on the authenticity and integrity of the digital record) backward compatibility is not a realistic approach to the long term avoidance of database obsolescence.

4.2.2 Interoperability

In a technical sense, interoperability tackles the problem of digital obsolescence by reducing or eliminating the dependency of files and records on a particular combination of hardware and software. Interoperability means that a file can be transferred from one platform to another and can then still be reproduced in the same or a similar way:

- A file can be read and processed using different versions of the same application running under different operation systems. Software manufacturers issue versions of applications suitable for each operating system; for example, different versions of Access for use with Windows, Linux or Solaris.

- A further form relates to interoperability between similar software applications. Modern software can always partly interpret files created in a similar software package; for example, files created in dBASE III can be read by Access.

- A last form of interoperability requires the use of an interim conversion program. This involves the conversion of files created in the supplier's own format, such as Access, into an exchange format, such as ASCII, which can then be read into another database program, such as Oracle.

Is interoperability suitable for preserving databases?

Desktop databases, such as Access, frequently have tools to export the databases to other formats such as ASCII or RTF. Conversions of this nature are relatively simple to carry out. It is necessary to check each such conversion to verify that the procedure has been carried out correctly and that no changes have occurred in reading the selected (interim) format into the new environment.

Complex database systems, which usually run on a central server, often incorporate tools for importing and exporting the content from or to other databases. However, specialised knowledge is required for conversions of this type, which can involve a

number of phases. Problems are occasionally encountered whilst carrying out these conversions. In addition, there is also the risk of a loss of data. Modifications must usually be made to the user application's search commands and database transactions. This generally requires the manual rewriting of scripts. In view of the technical difficulties associated with conversions of this nature, they are performed only when there are compulsive reasons for discontinuing the use of a specific database system, for example when the size and the number of users increases to an extent that is detrimental to the performance, or when an organisation decides to switch to another supplier. In such situations it is necessary to export the data to a database system with a larger capacity and/or better performance, or to the database system of the new supplier.

In view of the difficulties associated with the use of interoperability as a preservation strategy, interoperability is not a suitable strategy for the long term preservation of complex database systems. However, interoperability can be used as an interim to provide temporary access to obsolete desktop databases whilst searching for a longer-term solution. Should this approach be selected then it will be necessary to check the extent to which the source- and target- formats are interchangeable, thereby ensuring the guaranteed authenticity and integrity of the desktop database. Once again: the greater the time scale between the target- and source- formats, the greater the potential detrimental consequences.

4.2.3 Conversion to standards

Conversion to standards is in essence migration from a proprietary format (which is often closed) to a format based on a published (non-proprietary, or open) standard. The advantage is that digital records are no longer dependent on the original hardware and software used to create them; consequently they are no longer exposed to the unsustainability risks arising from the obsolescence of the original system.

This method can employ *de jure* or *de facto* standards.

De jure standards are drawn up in a formal and open process involving an officially accredited standardisation organisation (ISO, NEN, W3C), since consensus and participation are important motives for their development. XML is an example of a *de jure* standard.

De facto standards are standards which are in widespread use; a critical mass employs the standard. De facto standards are usually drawn up in closed processes (manufacturer's standards)[8]. PDF is an example of a *de facto* standard.

In general, preference is given to *de jure* standards above manufacturer's *de facto* standards since the maintenance and future development of *de jure* standards does not depend on a single organisation; *de jure* standards are maintained and developed by a broader community. Moreover, in some instances licence fees can also be charged for *de facto* standards.

However, these are not the sole considerations in the selection of a preservation standard: the technical suitability and popularity of the standard are also of importance.

[8] *XML: de mogelijkheden en valkuilen voor de overheid,* W. Thomas, 19 September 2002.

Is conversion to standards suitable for preserving databases?

Conversion to standards can be a suitable approach to the preservation of databases. A conversion of this nature will achieve both backward compatibility and interoperability. In this instance backwards compatibility and interoperability are benefits offered by the strategy rather than the strategy itself. A conversion to a standard offers more benefits than a strategy based solely on backward compatibility or interoperability.

The aforementioned ministerial *Regulation on the Arrangement and Accessibility of Records* cites a number of standards for the durable preservation of digital records[9]. The regulations include stipulations for the preservation of databases, including their storage 'in the original storage format or ASCII (flatfile, with field separators) and accompanied by documentation, preferably in XML-DTD, relating to the structure of the database and at least extending to a complete logic data model with a specification of the entities; queries shall be specified in the SQL (SQL2) query language.' Testbed has assumed that SQL2 was cited in the Regulation, since SQL2 was the latest version of the SQL standard at the time the Regulation was drawn up. The current version is SQL3.

Within this context Testbed has examined ASCII, SQL, and XML.
ASCII has been used as a data storage and exchange format for many years. The ASCII table, a 7-bit character set, was officially adopted by ISO in the 1977 ISO 646 standard. This offers 2^7 (=128) combinations. In view of the restricted opportunities offered by a 7-bit table regarding the inclusion of non English-language characters, the ASCII table was expanded by the use of an 8-bits code table, thereby providing for the specification of 256 different characters (ISO/IEC 8859 standard[10]). ISO/IEC 8859 is comprised of a series of code tables which are each intended for a specific language; for example, ISO 8859-6 is used for Arabic. To resolve the problem of the various national code tables ISO proceeded to the development of one large code table. This 16-bit code table, with a capacity of 6556 characters, was laid down in ISO 10646.

Major computer companies were unable to concur with the ISO 10646 standard; for this reason they joined forces in the Unicode consortium, with the objective of developing a new standardised code table – Unicode – at present a *de facto* standard. At the same time, Unicode representatives joined the committees at work on the preparation of ISO 10646. They succeeded in harmonising both code tables (in 2000, Unicode 3.0 and ISO 10646). In contrast to the ISO standard, the Unicode code table *is* open and freely available. Unicode is more extensive than ISO-10646 because the consortium also investigated ways in which to implement their character set, so that it would run smoothly on different platforms and could be easily exchanged between different software applications.

The content of databases can be converted to ASCII without problem. ASCII files can be read by most software (including database systems). However, when saving the content of databases it is important to use metadata to specify exactly which ASCII code table and, where relevant which version, is used. The major advantage offered by ASCII (and Unicode) files lies in their form of flat text files, which are consequently independent of specific hardware and software. However, they do suffer from the disadvantage that flat text files are not able to represent structure (and appearance). XML, which uses Unicode, can then be an alternative.

---

9 Regulation on the Arrangement and Accessibility of Records, February 2002.

10 http://www.iso.org.

It should be noted that the opportunities offered by ASCII are also offered by XML– whereby XML also offers the advantage of its use of Unicode *and* the possibility to define and specify the structure of a record in a standardised manner. Section 4.3 contains a detailed discussion of the advantages and disadvantages associated with the use of XML as a preservation strategy for databases.

Each DBMS uses a slightly different version of SQL. Over the years the SQL dialects used by the various products have grown apart. This has resulted in a decreased degree of exchangeability. Suppliers have added their own type of functionality, an approach that is detrimental to durability. It is advisable to convert supplier's own SQL to the standard SQL3[11].

4.3 XML as a preservation strategy

XML is an abbreviation of eXtensible Mark-up Language, a mark-up language based on text characters used to enrich data with information about structure and meaning. This language – which can also be used as a file format – is an open standard defined by the World Wide Web Consortium, a non-profit organisation that develops interoperable standards such as the specifications, guidelines, software and tools required for the optimum use of the Internet[12].

XML is not dependent upon a specific platform and can be read by both humans and machines using a simple word processor. For the above reasons XML is suitable for digital preservation. The XML strategy can, depending on the method of its implementation, possibly overlap with other strategies reviewed above. As such, the conversion of files into XML can be regarded as a specific type of migration (see the aforementioned Conversion to standards).

XML is a good storage format since it can be readily processed by computer programs. In the future it will be possible to write relatively simple software capable of processing current XML files.

Files can be converted to XML, or generated directly in XML. XML's independence from a given combination of hardware and software makes the format more durable than many commercial formats. Consequently the number of conversions will be greatly reduced, and therefore so will the risk of adverse effects on the authenticity of the digital record. Moreover it is possible to formulate an explicit specification of the structure of a database using an XML schema or DTD [13].

---

11 ISO and ANSI publish the latest SQL standard. SQL3 is the third and completely version; it replaces SQL2, which was published in 1992 and was in turn the successor to the first version published in 1987.

12 See http://www.w3c.org.

13 XML has inherited the DTD mechanism from SGML. However since a DTD (document type definition) can define data types to only a very limited extent and is not XML it is making way for another standard: W3C Schema. The official name of this standard is XML Schema Definition Language (XSDL), although in practice the names W3C Schema or XML Schema are used.

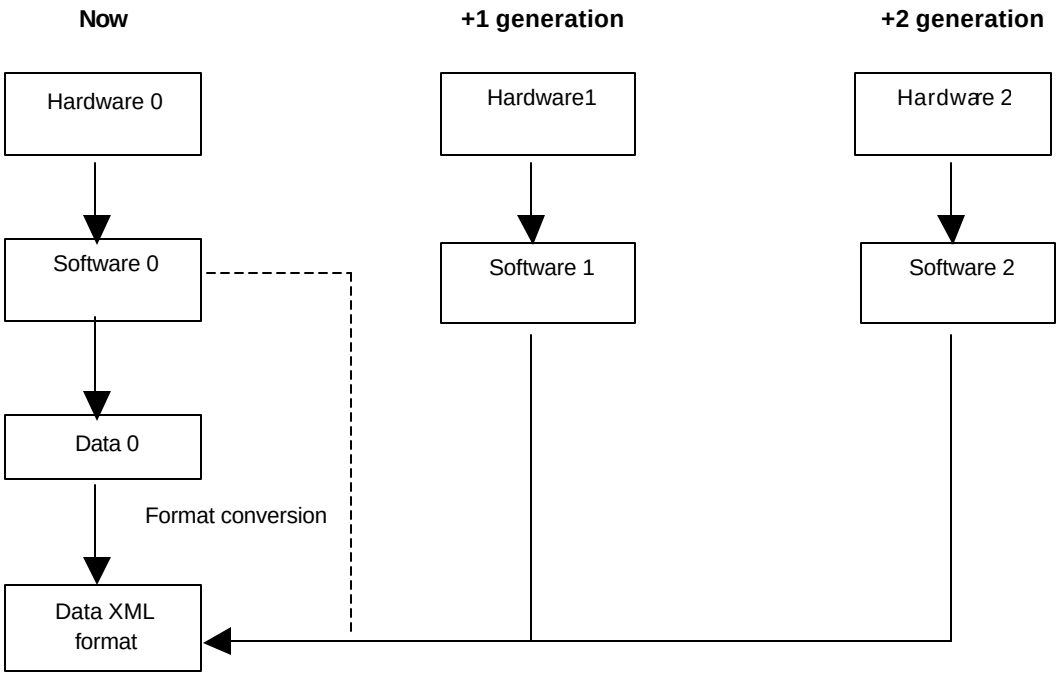


Figure 5:*Conversion to XML involves fewer conversions than migration*

The application of XML as a preservation strategy can be implemented in a number of different ways.

Encapsulation

This approach focuses on the retention of the original format. XML is often referred to as a language that can be used to specify metadata and instructions relating to the object to be preserved. The following sections review a number of terms which are used within this context.

Wrappers, containers, encapsulation and framework

The Regulation refers to an 'XML wrapper' as a means of adding metadata to PDF and TIFF files. Although the term does to some extent suggest the nature of the procedure, the term itself has not (yet) been definitively specified. The San Diego Supercomputer Centre, for example, regards a wrapper as a piece of software which is used by a 'mediator'[14]. Conversely, the Roquade project uses the term 'container' for the 'packaging' of digital records[15]. A step beyond encapsulation is the additional

[14] "A wrapper is a piece of software that acts as a translator between the native format of an information source and a commonly agreed protocol (XML for us). The end-user or application interacts with a piece of software called mediator that collects information from multiple wrappers", page 4 of Methodologies for the Long term Preservation of and Access to Software-Dependent Electronic Records, http://www.sdsc.edu/NHPRC/Pubs/nhprcfeb2k.doc.

[15] "It was decided to work out the idea of XML containers. So the Archival Information Packages (AIP), to be stored in the electronic archive, will be wrapped in XML." *An*

use of XML as a 'framework' on which to mount (parts of) records, for example, in TIFF or PDF format. In this instance XML forms the backbone of the preserved digital record.

Metadata

XML also offers excellent facilities for the specification of metadata, which is the reason why XML is also encountered in other strategies in this respect. With emulation, for example, XML could be the language used to specify the technical metadata. Adobe, the proprietor of PDF, has recently launched its eXtensible Metadata Platform[16] which also uses XML to specify metadata.

Once agreement has been reached regarding a permanent collection of metadata items (which is often much more difficult than the technical implementation!) it is then possible to specify the collection in the form of an XML schema that can again be used as schemas for specific types of records.

Is XML suitable for preserving databases?

XML is an excellent choice of format for the long term preservation of databases. It can be used to specify the context, content and structure of databases.
Testbed has examined three proprietary tools for the conversion of Access databases to XML, namely Access 2002, HiT Software's WinAllora Express, and Data Junction Corporation's XML Junction. All these tools are of good quality, and all have a well-designed user interface. WinAllora and XML Junction were able to convert a wide variety of databases into XML. However, time was required to configure the tools in a manner such that they supplied the desired output. It was possible to gain full control of the final form of the XML. For this reason Testbed developed its own conversion tool that can be used to determine how the data in the database is divided between the XML files, as well as which data will or will not be included. In addition, tests have been carried out with the SIARD application[17] developed by the Swiss Federal Archives. Chapter 5 contains a detailed discussion of the conversion to XML.

4.4 Emulation as a preservation strategy

The term emulation is used in computer science to denote a range of techniques, all of which involve using some device or program in place of a different one to achieve the same effect as using the original. The term "simulation" is often confused with - and sometimes even used as a synonym for - emulation, but we distinguish between the two terms here by noting that a simulation describes what some other thing would do or how it would act, whereas an emulation actually does what that thing would do. For example, an aeroplane simulator does not actually fly. That is, simulation generally involves the use of a model to understand, predict or design the behaviour of a system rather than the practical recreation of that system's capabilities. In contrast, emulation is generally used to create a surrogate for the system being emulated.

For preservation purposes, we focus on emulating older, obsolete computers on future computers. In this context, emulation would enable future computers to "impersonate" any obsolete computer, virtually recreating the obsolete computer and thereby allowing its original, obsolete software to be run in the future. This would allow the

---

electronic Archive for academic communities (Dekker, R. *et al*, Nov 2001). The AIP term originates from the Open Archive Information System (OAIS) model.

[16] See http://partners.adobe.com/asn/developer/xmp/download/docs/MetadataFramework.pdf .

[17] SIARD: Software-invariant Archiving or Relational Databases (Schweizerisches Bundesarchiv.) www.bundesarchiv.ch .

original rendering programs for obsolete digital formats to be run on future computers, under emulation.

The following diagram shows the relations between the hardware, software and data when emulation is employed:

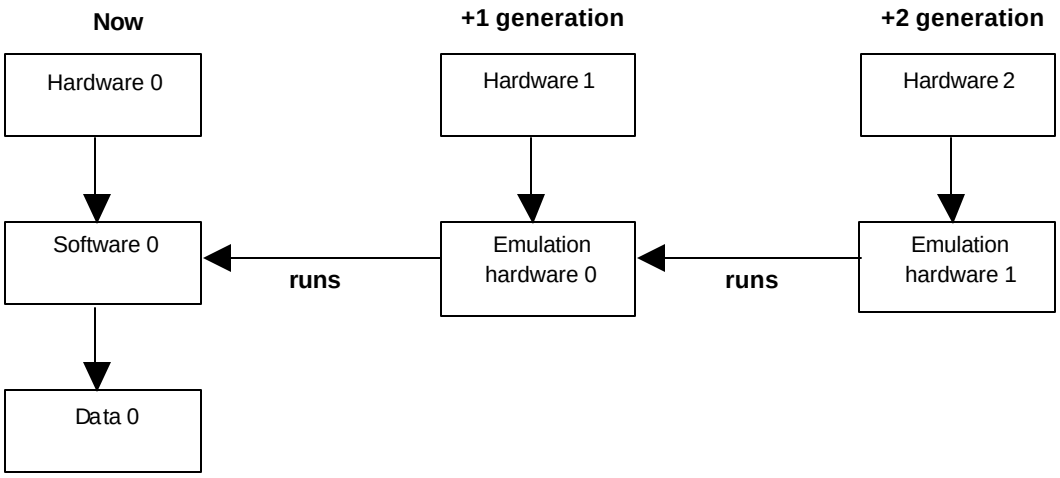


Figure 6. *Basic emulation diagram*

4.4.1 Hardware emulation

Emulation avoids the need to write new software in the future to render obsolete formats. This is a significant advantage, since an obsolete format must be understood in great detail in order to write such rendering programs, which may require extensive research and possible reverse engineering [18] if the format in question is not well documented.

The hardware emulation approach described here is the only way that has so far been proposed to run original software on future computers. This means that the behaviour of that original software will be recreated (within the limits of the emulation approach, as discussed below) without anyone needing to understand or rewrite any of that software. None of the original rendering programs or their original operating system environments need be recreated or modified in any way: they are simply saved and run exactly as they were originally, albeit under emulation on future computers. When this original software is run under emulation in the future, it should be completely unaware that it is running on anything other than its original hardware. Running a digital record's original rendering software in this way should allow preserving and rendering the record in its original format.

The major advantage of hardware emulation is that the original file does not have to be migrated or converted. However, writing an emulator of a given computer system (including its peripherals) is not a trivial undertaking. Yet only one such emulator need ever be written for any given type of computer.

[18] Reverse engineering – decompilation: the attempt to trace and describe the logic of compiled software programs for which the source code has disappeared. This is always a difficult task, akin to attempting to recreate a pig on the basis of a sausage.

Other forms of emulation

The approach discussed here is that of using software to emulate computer hardware, on which original rendering software can then be run: for convenience in this discussion, we will refer to this as the "software-emulation-of-hardware" approach. Sometimes two alternative uses of emulation are discussed, both of which involve emulating software with software and which do not share most of the advantages of the software-emulation-of-hardware approach. These might be called ‘application emulation’ and ‘operating system emulation’.

Application emulation consists of writing one application program to do what another application program does. In the preservation context, this is essentially the "viewer" approach, in which new programs are written in the future to render obsolete digital formats. This is different from the software-emulation-of-hardware approach: instead of writing a single emulator of a hardware platform, the viewer approach requires writing a new program (or adding a significant new piece to an existing viewer program) for every distinct digital format. Because many formats are proprietary, this entails reverse engineering each such digital format. Furthermore, this approach does not allow running a record's original rendering software.

Operating System (OS) emulation is not really a meaningful preservation approach for preserving digital records either. The concept is to recreate operating systems like Windows 98, Windows XP or Linux, that are used by several application programs for different digital formats. This requires a significant amount of reverse engineering effort, but even so, the result is not a program that can run other programs, since this is not what an OS does. An OS merely provides facilities (user interfaces, file systems, interprocess communication, networking, etc.) that are used by programs when they run, and it allows invoking programs to be run (e.g., by double-clicking on their icons). An application program may use these OS facilities to access files, interact with users, or communicate with the network or with other programs, but the application program must always execute on hardware, just as the OS itself does. That is, any program must run on its expected hardware platform, regardless of whether its expected OS is also running on that platform. Computer scientists often say (perhaps confusingly) that an application program "runs on" an OS, but all this means is that it relies on the facilities provided by that OS - it does not mean that the application "runs on the OS" in the same sense that the application runs on hardware. All programs (applications and operating systems alike) must run on hardware. Therefore, implementing an emulator of an OS does not enable us to run application programs, such as rendering programs, without also having the appropriate hardware platform - either as a physical computer or as a software-emulation-of-hardware (which, of course, must itself run on some physical computer)

Is hardware or software emulation suitable for preserving databases?

Hardware emulation offers the advantage of the preservation of the digital record in its original format, since the original application program is also preserved and can be used in the future. Complex database systems will then require a disk image, backup, or export file. With desktop databases a copy on one of the customary storage media will suffice.

However, emulation is an approach which is difficult to implement. The emulator will need to be designed, developed and tested whilst the old computer platform is still available. It will then be necessary to store the emulator together with the operating system, the original application program, and the files created with this application program. Emulation suffers from a number of disadvantages, i.e. the technical complexity and time-consuming nature of the design, testing, use and durable preservation of the emulator. This complexity is primarily due to the following factors:

- the difficulty of defining what precisely must be emulated;
- the complexity of the hardware functions to be emulated.

Copying a complete set of computer hardware is by definition complex. However, an emulator only needs to emulate the specific hardware functions required to enable the stored application programs to run in the required manner. The specification of all the hardware interactions, for example such as those required by an operating system, is difficult since these interactions are often inaccessible to users. In fact, even when the exact specifications of all the requisite hardware functions *are* available the software implementation of those functions to be simulated by the emulator is still a complex and difficult process.

The implementation of emulation as a preservation strategy for complex database systems is more complicated than for text documents or spreadsheets, since in addition to the preservation of the original database - the original operating system, the DBMS and the database application - it is also necessary to preserve the entire user manual and the instructions relating to the configuration of the operating system, the DBMS, the database application, and the import of data in the DBMS.

It should be noted that it will be anything but easy for future users to install and use old software. Future software will probably have a different appearance onscreen, and require a different approach to its use. This is demonstrated by the manner in which applications worked – and documents were prepared – before the emergence of the Graphical User Interface. One example is WordPerfect 4.2, which was very popular in the latter half of the 1980's. This application required the use of a wide variety of keyboard combinations to make and use documents. There were more than forty combinations, and for this reason a card template for the keyboard was supplied with the software that indicated the combinations. Testbed staff experienced difficulty working with this old program, just 15 years after it was in daily use – even those who had previously been thoroughly familiar with the application.

The advantage of hardware emulation is its potential to preserve complex and customised software over the long term. This renders emulation an interesting strategy for databases. However, emulation as a preservation strategy for digital records has not yet been implemented. Further research is first required in the form of a 'proof of concept'. In view of the complexity of this approach, hardware emulation will be profitable only when the strategy is selected as the preservation strategy for *all* categories of digital records from a given computer generation [19].

[19] As indicated above, hardware emulation involves the simulation of a hardware platform. For this reason the strategy is suitable for all categories of records, subject to the proviso that the emulated components are able to simulate the full hardware

4.4.2 The Universal Virtual Computer strategy (UVC)

Emulation using the UVC differs to some extent from the original emulation concept. An emulator must still be written, but in this case it is for a non-existent, virtual computer: the UVC (Universal Virtual Computer).

The UVC has a simple architecture and a simple set of instructions, thereby ensuring that it will be easy to write an emulator at some point in the future. A specific application (a UVC data format decoder program) is run on the UVC that converts the original digital record into a Logical Data Description (LDD). This logical data description is comprised of tags providing information about the content of the digital record. The tagged information is designed in such a manner that in the future it will be possible to interpret the logical data description without additional aids. A future viewer will then process the logical data description and display the digital record.

This strategy is based only in part on emulation, and includes several aspects of the migration strategy. The UVC coverts the original data files into a Logical Data Description (LDD) using a program written in the UVC programming language. This LDD is a stand-alone, self-descriptive and explicitly structured data format that contains all the information required for the future re-assembly of the digital record.

UVC: data preservation

'Data preservation' is the first and simplest form of implementation of the UVC strategy. In this approach the data – the original file it its original format – is stored with a program that can extract the data from the bit stream and can describe the data in a simple manner that is independent of a specific technology, so that it can then be processed via a viewer.

conduct required for all programs used for all categories of records. Much of the potential offered by this strategy will be lost if the hardware emulation is used solely for one category of record. For this reason any implementation of this strategy shall need to be suitable for all categories of records.

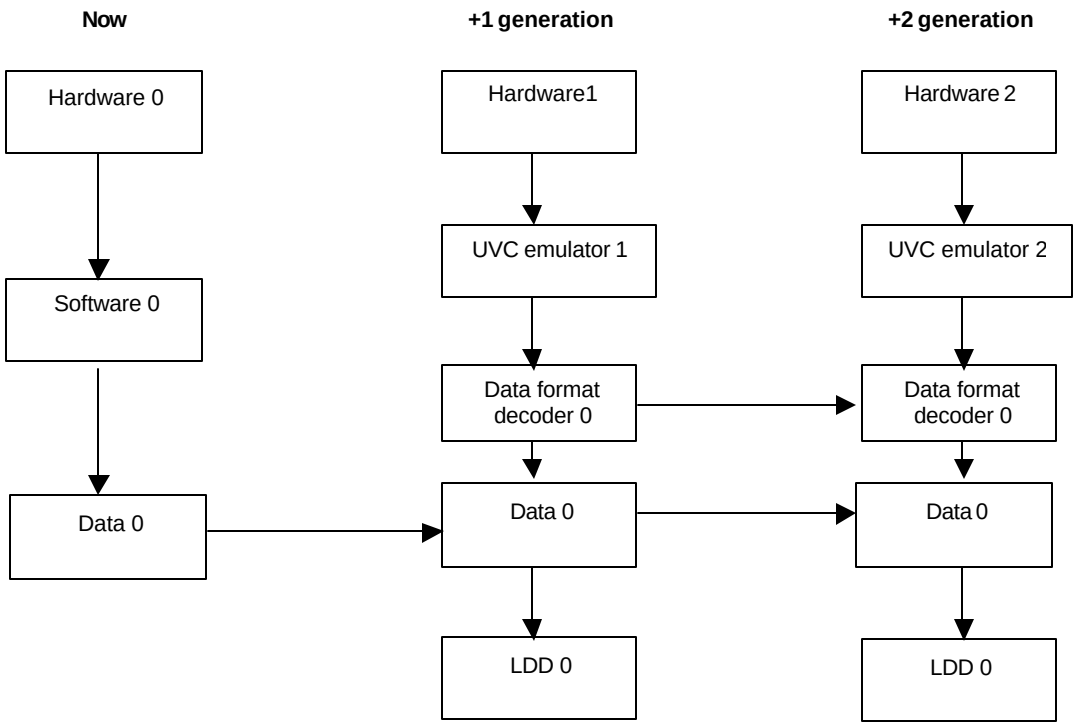


Figure 7:*Diagram of the Universal Virtual Compute r*

The original file – for example, a JPEG file – is saved together with the specific UVC data format decoder program for JPEG's. In the future, this UVC JPEG program will run on the UVC emulator. The UVC JPEG program reads the bit stream of the original file and returns an LDD. This LDD is then processed and displayed on a future computer platform via a viewer.

This strategy does not modify the original bit stream, and the new file (the LDD) created by running the UVC JPEG program is not saved. The LDD is processed and displayed using a viewer. The format and the structure of the Logical Data Description are designed in such a manner that it will be simple to write a viewer at some point in the future. Where necessary, new viewers can be developed for future computer platforms.

At present, a different viewer is required for each category of LDD. As a result, it is possible that hundreds of viewers will be required. However, in practice the number of file formats accepted by the Archives will be restricted by the Regulation for the Arrangement and Accessibility of Records.

The next phase of the UVC development will be to classify files of the same record type into groups of records that function using the same logic. One LDD will be prepared for each specific group (such as the various image file formats), as a result of which it will be necessary to develop only one viewer for that group. Nevertheless, it will still be necessary to develop a separate UVC data format decoder program for each file format in order to convert them to a shared LDD.

One disadvantage of the UVC emulation strategy is therefore the need to write a UVC data format decoder program for each file format (for the generation of the Logical Data Description). It will also be necessary to write a new emulator for each generation of hardware that differs from previous generations to such an extent that the old UVC emulator can no longer run on the hardware with the requisite reliability.

In view of the extremely wide variety of file formats and categories of records, it will be necessary to develop a large number of data format decoder programs if the UVC strategy is to be implemented as a means of providing for the durable preservation of digital records. The ultimate success of the UVC strategy will to some extent depend on the extent to which it is accepted by the software and computer industry. Should software manufacturers themselves develop UVC data format decoder programs for their own applications that are capable of creating Logical Data Descriptions from the original files, then the UVC strategy may experience widespread use.

Other forms of UVC

At present, the UVC program-preservation approach (as opposed to the data preservation approach described above) is still in the design phase, and the viability of the concept will need to be proven in practice. No practical experience with the application of this approach has been acquired to date.

Is UVC data preservation suitable for preserving databases?

UVC data preservation is a highly promising strategy for the durable preservation of digital records. The use of the UVC data -preservation strategy for the preservation of databases is conceptually an attractive approach. The conversion of the content of a database into an XML-like logical data description that can be rendered using an unknown viewer in the future, could offer a feasible long term preservation strategy. The output file can be read by people and computers and is no longer dependent on the original software. However, to date no UVC data format decoder program has been developed and tested for use with databases.

In principle, the UVC data-preservation approach will be suitable for databases only when the data format decoder program is able to accurately decode all the essential aspects of the database without needing to make use of the original application. It should be noted that tests carried out with spreadsheets, for example, have revealed that the extraction of all the required information from a proprietary (and closed) file format is no simple matter.

In summary, it can be concluded that the UVC approach certainly possesses potential, but that it will be necessary to devote time and effort to the development of data format decoder programs for the customary proprietary file formats. These decoder programs will only need to be written once, and then made available to others. Another possibility would be for the major software suppliers to provide a UVC data format decoder program when developing new versions of their software.

4.5 Conclusions

The major benefit offered by emulation is the representation of the original record in the environment in which it was originally created. This is a particularly attractive prospect for the durable preservation of databases, since not only the database but also the DBMS *and* the user application (what is referred to as the 'look and feel' of the record) can be retained. The disadvantages of this strategy are, as mentioned above, the technical complexity and time-consuming nature of the design, testing, use and preservation of the emulator. In spite of the use of emulators by the gaming and computer community there are as yet no emulators in use for digital preservation.

The two 'proof of concept' implementations carried out by the KB and Testbed have shown that the UVC approach possesses potential, but that it will be necessary to devote time and effort to the development of data format decoder programs capable of converting the original files into logical data descriptions.

Interoperability is not a reliable strategy for the durable preservation of databases, although it is suitable for use as an interim solution whereby obsolescent file formats can remain temporarily accessible whilst a long term solution is sought.

Backward compatibility as a preservation strategy can be an approach to the short-term preservation of databases. In view of the disadvantages of backward compatibility as a preservation strategy (storage in the manufacturer's file format, the need to repeat the migration every few years, and the risk of adverse effects on the authenticity and integrity of the digital record) backward compatibility is not a realistic long term approach to the avoidance of digital obsolescence.

At present, XML is the most effective strategy for the durable preservation of databases. XML is highly capable of representing the context, content, and structure of databases. This strategy can implemented using a number of different methods. The details of this approach are discussed in the following chapter.

5. Approach to the preservation of databases

Chapter 4 discussed and compared the various preservation strategies against the record type 'database', and came to the conclusion that the recommended preservation strategy involves the use of XML. This chapter explains how an XML strategy can be implemented.

5.1 Introduction

Although XML is the best strategy for the long term preservation of databases, migration is nevertheless an alternative for organisations that create databases which only need to be preserved for a short period of time. Organisations that create a variety of databases, some off which need short term preservation and others which need preservation for the long term, will need to consider whether they should adopt parallel preservation approaches, or instead opt solely for a long term preservation approach.

5.2 Short-term preservation of databases

Migration in the form of backward compatibility is a suitable preservation strategy to ensure continued short-term (less than ten years) access to databases and user applications without affecting their authenticity and integrity. This ten-year period is to some extent an arbitrary choice; the period could also have been set at eight or twelve years. Although it is possible to preserve databases in their original file format, preference is nevertheless given to the upgrading of the databases to the new file format after a migration since application software is able to read older file formats of only a restricted number of generations with the necessary reliability. However, random visual inspections of the migration results are necessary.

5.3 Conversion and migration procedures

Testbed recommends that any necessary conversions are carried out as quickly as possible. This Section begins with a review of the use of backward compatibility migrations (5.3.1), and the use of XML as a preservation strategy (5.3.2). The preservation of the relevant contextual information is discussed in section 5.4.

5.3.1 Backward compatibility

It has become apparent that backward compatibility is only suitable as a short term (i.e. less than ten years) strategy for preserving databases in an authentic state. Since new versions of software support only a restricted number of older generations of file formats, databases preserved using this strategy should be saved in the new version of the format provided by the new version of the application. Upgrades to new versions of an application normally take place every few years. However, it is not always necessary to upgrade to each and every subsequent version (for example, from Oracle 7.3.3 to Oracle 7.3.4, or from Access 97 to Access 2000). Experiments at Testbed have shown that migration over different versions of an application can sometimes deliver better results than migration through each and every individual version of an application. Random visual inspections will always be needed to verify that the migration has had the required result - or, in other words, whether the organisation's authenticity requirements have still been met.

5.3.2 XML

The conversion of databases to XML is a suitable preservation strategy for databases that need to be preserved in an authentic state for a longer period of time (longer than 10 years).

Testbed has investigated a number of standar d conversion tools in terms of their suitability and limitations to represent the content and structure of databases in XML. The tools Testbed investigated were able to convert a range of different types of database into XML. However, these tools did not offer an opportunity to convert all the database's data tables into an XML file by means of one simple operation. In addition, the XML files and DTDs or XML schemas were not optimum for preserving the information in a structure suitable for digital preservation. Moreover the tools did not offer an opportunity for the assignment of metadata.

For the above reasons Testbed decided to develop a conversion tool based on the following principles:

- Rapid and simple in use;
- Conversion of a group of data tables in one operation;
- Use of an XML file structure readily suited to digital preservation;
- Compatibility with Microsoft Access and Oracle databases.

The tool was written in Java and developed in a short period of time thanks to the availability of database drivers and comprehensive support for programming with XML. The tool was used to develop an assembly of XML files. The tool was designed to generate an overview file to represent the main structure and metadata of the underlying database, and individual XML files to represent the content of each table and view. The tool largely satisfies the requirements for XML files in the preserved record object presented later in this chapter.

Following on from the development and testing of this tool, Testbed received a test version of the application developed by the Swiss Federal Archives for the conversion of databases to XML, SIARD, (as mentioned in chapter 4). Several years of work were involved in the development of SIARD. SIARD is comprised of an advanced collection of applications capable of:

- indicating precisely which components of the database need to be converted,
- the entry of a range of metadata, and
- carrying out a conversion to a group of files comprised of XML, SQL3, and the plain content of the database.

The XML files are comparable to those created using the Testbed tool, although the SIARD files contain more information. SIARD is also capable of converting a stored database back into an Oracle database. SIARD is a highly promising application which, on the launch of the definitive version, will comply with our recommendations for the preservation of databases.

5.4 Long term preservation of databases

This section describes the implementation for the long term preservation of databases with what is referred to as the 'preservation object'.

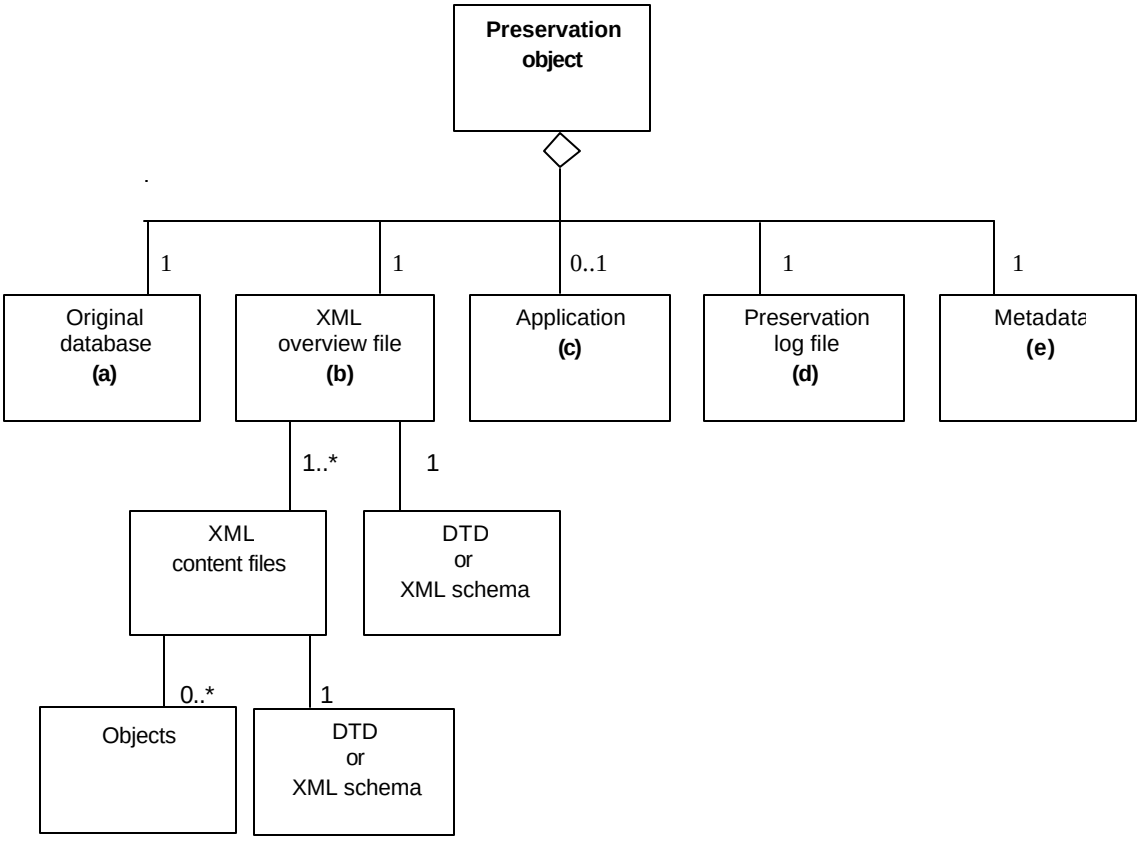


Figure 8*: Structure of the preservation object*

Notes: the diamond-shaped symbol indicates that the preservation object is comprised of the components to which it is linked. The significance of "1" is "1"; "0..*" signifies "zero, or more"; "0..1" signifies "zero or one"; "1..*" signifies "one or more".

The links between the different components can be implemented in a number of ways, for example by means of the 'framework approach' discussed in chapter 4.

Each component is discussed in more detail below.

Original database (a)
The original file of a desktop database is usually directly available, for example in the form of an *.mdb file (Access). With complex database systems it will be necessary to create an export file to obtain the original file. Testbed recommends that the original file also be saved, since this offers maximum flexibility for future preservation strategies. Whilst the original file format is still readily accessible this also provides for the most authentic possible representation of the database, particularly when used in combination with a working user application.

XML overview file (b)
The XML overview file represents of an overview of the tables in the database, the mutual relationship between the tables in the accompanying DTD or XML schema, the content (XML content files), and the structure of the actual tables and views (DTD or XML schema).

The XML overview file must be comprised of the following main elements:

```
<database>
        <schema>
                <table> or <view>
                                <column/>
                                <constraint/>
                <table> or <view>
        <schema>
<database>
```

Multiple schema, table, view, column and constraint XML elements are possible. Each of these elements should contain the following minimum set of information:

- Database – the name of the application, the name of the DBMS product, the version of the database product.
- Schema – the name of the DTD/XML schema.
- Table/view - name, remarks. For views only: the SQL associated with the view.
- Column – the name, data type, not null/nullable, maxLength, remarks.
- Constraints – the name, type, columnName, referenceTable (for foreign keys only).

All tables and views must be supplied with a separate XML content file that specifies the name of the table or view and the column information. The data values should be contained row by row. One general DTD or XML schema is included for all these files. It is neither necessary nor efficient to include a DTD or XML schema for each individual file.

If the database contains content objects, such as images or integrated text document records, then these must be extracted from the database, saved as separate files, and linked to the XML content file corresponding to the table from where they have were extracted. A preservation strategy is also required for these files. The appropriate strategy will depend on the category of record; for example, the recommendations for text document records should be consulted for text documents.

Application (c)
In general, it will not be necessary to preserve the application as a working entity. However, if this is a requirement then hardware emulation is the only way by which to achieve it. In this instance the preservation incorporates, in particular, storage of the system documentation (technical) and the user's manual (functional). In view of the set of minimum authenticity requirements formulated in chapter 3 this documentation will always need to contain information about the functionality (functions) offered by the application, the queries that are used, and the layout and appearance of the application as displayed onscreen. These three components are discussed in more detail below.

Functions

- the names of the system's functions;
- the objective of the functions;
- the conditions governing the functions;
- the activating triggers for the functions.

Queries
The documentation must specify the queries for the entry, modification, deletion and request of data from the database. Preference is given to the specification of these commands in SQL3. Essential blocks of procedural code which form a query, for example Oracle's PL/SQL[20], must also be textually documented. The system documentation must also contain specifications of the version of all code that is used.

Display
The system documentation must specify the layout of the display. This requires a specification of the fields used in the display and their relationship with the columns in the table(s).
Screenshots can be used to illustrate the displayed appearance.

Preservation log file (d)
The preservation log file contains all information about the preservation actions undertaken on the database. Moreover, the preservation log file can also include information about the specific preservation and access requirements.

The preservation log file is created at the time of the first conversion of the database to XML. It is important to ensure that the preservation log file can be updated readily and continuously without overwriting earlier data. A database can be suitable for this purpose; consideration can again be given to the use of XML. The initial content of the preservation log file must be comprised of the data in the original digital record, in this instance the database. This information must be followed by information about the conversion, including the conversion tool that was used, the date and time at which the conversion was made, and the database's new format.

The preservation log file must be updated each time any preservation operations are carried out on the digital record. In addition, the preservation log file must also contain information about any changes that have occurred in the database as a result of the preservation operations. Appendix A reviews the possible content of the preservation log file.

[20] PL/SQL: Procedural Language/SQL.

Metadata (e)
A supplementary metadata file is essential to ensure the authentic preservation of digital records over the long term. This metadata focuses, in particular, on the contextual data that imparts meaning to the digital record: the relevant person(s) or organisation, the function, the mandate, and the business process. The metadata also contains information about the intellectual management of the record (for example, the arrangement and classification codes for the record). This metadata must be collected and saved at the time the record is created, or as soon as possible after its creation, and subsequent updating must be ensured. This metadata must, for as far as is possible, be updated automatically so as to simplify the user's work and to minimise the risk of errors.

Organisations can exercise their discretion in deciding on the exact contents of the metadata file. Many institutions already record and manage metadata, or effect it using a Records Management Application (RMA) or a Document Management System (DMS).

6 Concrete Actions

The previous chapters dealt with the problem of digital obsolescence and proposed the best strategy for preserving databases. Now it is up to organisations to make use of this informati on. Chapter 5 dealt with the implementation of the XML-strategy. The various activities that an organisation has to undertake to successfully achieve this are so specific and different from each other that they justify an approach oriented towards differen t target groups. In that way employees can quickly see which activities they have to initiate. The different target groups are:

- General (line) managers
- Records managers
- ICT specialists and
- End users

Each section is written in such a way that it can be rea d separately from the complete publication.

6.1 Action plan for managers

Introduction
In reading the publication *From digital volatility to digital persistence: Preserving databases* you will have discovered the advantages of working digitally, but also the specific problems that arise in the long term preservation of digital records in general and databases in particular. The Digital Preservation Testbed has tested preservation strategies for the record type 'database'. The best way of preserving databases at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long term preservation of databases: from the line managers, records managers and ICT specialists, to the end users who have desktop databases at their disposal and who make use of complex database systems. The concrete actions listed below are specifically oriented towards:

- General (line) managers
- Records managers
- ICT specialists and
- End users

These four groups each have a specific responsibility in this matter. This final chapter sets out the concrete steps each target group has to take to make the long term preservation of databases a success. The concrete steps or actions are preceded by a description of the prior conditions.

Prior conditions
"You are the inspiration behind improvements in your organisation. You have good contact with the shop floor. Your employees find you approachable. You are prepared to invest time and money in document management to improve the performance of your organisation." It sounds like a recruitment brochure for a management course. Even so, these are the *essential starting points* for giving digital records, in this case databases, a firmly-rooted place in your organisation and for reaping its fruits: accessible, quickly available and reliable information.

Generating awareness among all employees in your organisation that databases are official records, with all the consequences this implies, is a condition for successfully creating an electronic government.

It is also important to take *action quickly*. Examples of cases in which the lack of good preservation of databases was the cause of major problems are increasing in number, because the use of computers has multiplied in the last few years.

Concrete actions for managers
Specify the integral information policy: in your role as manager you are responsible for the specification of the information and archives policy (see also the NEN-ISO standard 15489). This not only contributes to the efficient and effective operations of your organisation, but also forms the basis of your accountability for your actions.

Specify procedures: these must explicitly state who is responsible for what, who can be called to account for which issues, and which staff (positions) should inform each other. The procedures must in any case extend to:

- agreements on the use of databases
- agreements on the management and preservation of databases

Partners in the discussions about these procedures are the records managers, ICT managers, and office managers.

Inform all staff about the policy and the procedures. Train all staff in the use of the database systems. A database which has been created and is maintained in the appropriate manner is one step closer towards durable preservation!

Evaluate the policy and procedures at regular intervals.

6.2 Action plan for records managers

Introduction
In reading the publication *From digital volatility to digital persistence: Preserving databases* you will have discovered the advantages of working digitally, but also the specific problems that arise in the long term preservation of digital records in general and databases in particular. The Digital Preservation Testbed has tested preservation strategies for the record type 'database'. The best way of preserving databases at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long term preservation of databases: from the line managers, records managers, and ICT specialists, to the end users who have desktop databases at their disposal and who make use of complex database systems. The concrete actions listed below are specifically oriented towards:

- General (line) managers
- Records managers
- ICT specialists and
- End users

These four groups each have a specific responsibility in this matter. This final chapter sets out the concrete steps each target group has to take to make the long term preservation of databases a success. The concrete steps or actions are preceded by a description of the prior conditions.

Prior conditions
As records manager you are aware of the various problems that need to be resolved before the management of databases meets the same quality requirements governing the management of paper records. How can you convince the Management to make available the funds and resources that are required for the management and durable preservation of databases? This is not something you will be able to achieve on your own in the organisation; as records manager it is important that you seek co-operation with the line management, with the ICT department, and with the end users.

Concrete actions for records managers
The concrete steps that will need to be taken are:
(a) An analysis of the current situation;
(b) Formulation of the required policy, and;
(c) Establishment of procedures.

(a) Analysis of the current situation

Draw up a description of your organisation's duties or processes, for example on the basis of the Institutional Research Report (RIO). This can be of assistance in locating the relevant digital records. Once you have gained an insight into all the business processes, you will be aware of the operations carried out by your organisation and the (digital) archives that these business processes will generate.
Endeavour to collect as much information as possible about:

- The business processes and the applications that are used (from when).
- The files generated by each business process.

It is also important to establish whether the organisation also out-sources business processes. If this is the case then digital archives could be formed outside the organisation.

Determine which databases are actually present, and where they are stored: on a separate server, on a shared network drive, on an individual section of the network, or on a local hard disk. Endeavour to collect as much information as possible about the following issues. The ICT department can assist you with this task.

- The period in which the database was created; the date of commissioning (and decommissioning) of the database.
- Any conversion(s)/migrations(s) carried out upon the database.
- The hardware used.
- The name and version of the database management system (DBMS).
- The name and version of the user application (the application).
- The name and version of the operating system.

Establish which databases constitute archives
Not all databases are records for the archives in the sense of the 1995 Archives Act. Only databases that have played a role in a business process are deemed to constitute records for the archives.

Establish whether the databases are to be destroyed or preserved
Establishing that a part of the databases constitutes records does not imply that all the databases will need to be preserved. A fixed selection list can be used to distinguish between files that should be preserved and files that should be destroyed. An organisation that does not have a selection list will need to treat all files as records that must be preserved. Until such time as a fixed selection list has been adopted, the organisation will need to ensure that all databases can at least be consulted.
During the storage period of files designated for later destruction, such files, in analogy with files that are to be preserved, will need to be preserved in an appropriate, ordered and accessible manner.

Analysis of the database
Files to which data has been neither added nor changed after January 1st 1996 need comply solely with articles 3, 7 and 9 of the *Regulation on the Arrangement and Accessibility of Records* (February 2002). See the transitional and concluding provisions.

It is then necessary to assess whether the files meet all the requirements. The following points of concern could be encountered:

- The databases contain insufficient metadata.
- The files can no longer be consulted, for example as a result of password protection.
- The carriers used to store them, for example tapes, can no longer be read.

On the completion of the above you will have an overview of all the digital files managed by your organisation, together with an analysis of the files. Moreover you also have an insight into the points of concern relevant to the management of your digital records.

(b) Formulation of the required policy

Make sure that priority is assigned to the successful preservation of databases
Procedures will only have a chance of succeeding when they are based on a policy that has been explicitly conveyed to all those in the organisation. It must be clear what the organisation wishes to achieve with its digital records management, what importance it attaches to this, and how the organisation perceives such developments. This is primarily a line-management duty; however, as records manager you will need to play the role of catalyst and driving force behind the necessary processes.

Establish the required knowledge and expertise in-house
How explicit is the prevailing archives policy with respect to the preservation of digital records? Your department is important in the specification and implementation of that policy. Don't forget that the durable preservation of digital records requires knowledge and skills different to that involved in the preservation of paper records. Make sure your organisation has that knowledge in-house and at its disposal!

Seek partners and interested parties
The formulation of policy is not primarily your responsibility; however, you can play an important role in getting the issue onto the agenda. Whilst doing so, it is also important that you identify other interested parties, such as departmental managers who need specific information for their business operations, the ICT department, and the interests of all users.

Specify the selection criteria
Formulate the selection criteria. In general these will already have been specified in a records structure plan, or a Basic Selection Document (BSD). Ensure selection can be carried out at-source. The formulation and maintenance of a valid selection document may well be the most important step to be taken in this respect.

Retain the authenticity of the databases
The selection of the most appropriate manner for the storage of databases is of essential importance, since this can influence the authenticity. Printing the information out to paper can be detrimental to the authenticity, since some information may be lost. Chapters 4 and 5 of this publication have explained that the Digital Preservation Testbed recommends either migration or an XML approach, depending on how long the databases are to be preserved. Use this information, together with other disciplines in your organisation, to advocate the use of these solutions.

Determine which metadata are required
Specific information about each database is necessary to establish its origin, destiny, dates, etc. This metadata is required to determine the authenticity and function of the database. It is necessary to determine which metadata must be recorded[21].
During this phase, make sure that precise specifications of important metadata are drawn up to ensure that the information can be (re)used and interpreted, and also to ensure that the organisation can be accountable for its actions.

[21] For the determination of metadata see the aforementioned Regulation under Article 12, or *Een uitdijend heelal? Context van archiefbescheiden*, ('An expanding universe? Context of Records') H. Hofman, Stichting Archiefpublicaties, Jaarboek 2000.

Determine the method of arrangement and classification
The objective of arrangement and the subsequent classification of records is to render visible the structure and relationships between records (including databases), and the relationships between records and the processes in which they played a role. This is conducive to their accessibility and provides support for structured searches. Consequently it will be necessary to develop a classification system based on tasks or organisational processes (see also NEN-ISO 15489). Involve the ICT department in the determination of search entries and relationships between records.

Formulate the policy
The performance of the above steps and the choices that were made during those steps must be laid out in a policy document. Specify for each choice what is feasible, and what is ideal. This policy document then serves as the basis for the next phase, which is focused primarily on implementation and during which the actual procedure will be written.

(c) Formulation of procedures

Make sure you are involved in the design process of database systems
It is important that all database systems are of an appropriate design. In addition to the technical issues this relates to the implementation of the business rules, for example in instances when a history will be required and consequently existing data may not be overwritten. When a database system is being designed to replace an existing system then you will need to specify quality requirements governing the conversion process so as to ensure the continued integrity and authenticity of the data in the database. When the logical data model of the new database system differs from the data model, you will need to establish whether the new database system can still account for government actions taken when the old database system was running. If this is not possible with the new database system, then it will be necessary to preserve the old database.

Make sure that databases are preserved
The management of records in a digital environment often takes place out of sight from the responsible records manager. Existing procedures and regulations for paper records are not sufficient for digital records. Procedures are needed to prevent the unintentional or deliberate loss of important databases.

Specify the manner used for classification and filing
A classification system (as identified above) is used to assign a database to a dossier. When the classification system is based on tasks or activities then it is also possible to establish the relationship with the business process when making the classification.

Arrange for the accessibility of the stored databases
The access possibilities are closely related to the selection of the storage format and the quality of the metadata. In general, databases stored on a central server can be made accessible to all staff. Assign the database management authorisations on the basis of the organisation's policy; where relevant, delegate such authorisations to your department. The ICT department is responsible for the actual implementation of this.

Make sure that databases are converted to XML
At present the best approach for the storage of databases that must be considered for long term preservation involves the use of XML.

Make sure that the policy is evaluated at regular intervals
Information technology changes rapidly – and the same is true for organisations. The requirements for digital archiving are also developing. For these reasons the policy must be subjected to regular evaluation and/or modification. It is to be expected that better software will be available in the future for the management and durable preservation of digital records. This is why Testbed also advocates the preservation of the original file.

6.3 Action plan for ICT specialists

Introduction
In reading the publication *From digital volatility to digital persistence: Preserving databases* you will have discovered the advantages of working digitally, but also the specific problems that arise in the long term preservation of digital records in g eneral and databases in particular. The Digital Preservation Testbed has tested preservation strategies for the record type 'database'. The best way of preserving databases at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long term preservation of databases: from the line managers, records managers and ICT specialists, to the end users who have desktop databases at their disposal and who make use of complex database systems. The concrete actions listed below are specifically oriented towards:

- General (line) managers
- Records managers
- ICT specialists and
- End users

These four groups have a specific responsibility in this matter. This final chapter sets out the concrete steps each target group has to take to make the long term preservation of databases a success. The concrete steps or actions are preceded by a description of the prior conditions.

Prior conditions
As an ICT specialist you are indispensable to the preservation of digital records, including databases, in an appropriate manner. Our starting point here is that the required policy has already been formulated for digital archiving, that the records manager has drawn up procedures for the selection of databases eligible for (permanent) preservation, and that agreements have been made within the organisation relating to the creation and use of databases. Besides this, the end users have received adequate training for the desktop database program used by their organisation, or that in the event of a complex database system with a customised user application, the end users have been trained to work with that application.

Concrete actions for ICT specialists

New database systems
Irrespective of whether the new database systems are designed and implemented within the organisation, it will always be necessary to comply with the (industrial) standards for database design and coding standards. These standards also relate to the generation of high-quality system documentation with accurate version management.

This publication focuses on the preservation of relational databases since this is currently the most common database model.

Maintenance of database systems
A qualified database administrator (DBA) must be assigned to all database systems and is responsible for the support and maintenance of the system. This work must be carried out in accordance with the best practices for database-administration.

Preservation of databases
This is the main focus of this section. The concrete actions you need to undertake are related to:
(a) General principles;
(b) Recommendations on the format and possibilities for implementation;
(c) Practical issues.

These aspects are reviewed in the following sections.

(a) General principles

Save the databases that are to be preserved in a centrally-managed system
This prevents the accidental or deliberate deletion of databases. Access to the centrally stored databases can be controlled, to ensure that the information remains accessible to those who need it and to prevent unauthorised access. A central system also provides for the safeguarding and management of the storage media, usually a combination of disks and tapes. This also extends to making copies and backups. It is important to remember that, within the context of digital preservation, there is a world of difference between the storage of backups and the sustainable preservation of digital records such as databases.

Record metadata automatically whenever possible
The importance of metadata for long term preservation has been explained elsewhere in this publication. To ensure maximum simplicity for the users of a preservation system, the metadata should be collected automatically wherever possible. Moreover this minimises the risk of errors during the manual entry of metadata. These measures can also increase the user-friendliness of the preservation system.

However, it is not possible to automatically collect all metadata; consequently the users will need to manually enter some items. This should be made as simple as possible by the development of templates with defaults and drop-down menus from which the appropriate value can be selected. This increases the uniformity of the entered data *and* minimises the risk of errors.

The central preservation system must use metadata on the classification and context of a database (such as the dossier to which the database belongs) for the arrangement of the stored databases, particularly in support of search functions.

Make sure that the preservation system supplements each stored database with a preservation log file (audit-trail information)
A log file of this type must contain metadata about the computer environment, such as the name of the application, the name of the DBMS product, the version of the database product used to create the database system, the version of the preservation system that is used, and an overview of any preservation actions carried out on the database such as the date and the time at which the database was accessioned into the preservation system. See Appendix A for more information about the recommended content of the Preservation Log File.

(b) Recommended format and possibilities for implementation

A detailed description of the strategy Testbed recommends for the preservation strategy is given in Chapter 5. The following brief summary of this description is followed by a number of remarks about the possibilities for implementing this strategy.

Testbed recommends the use of XML as the framework for the preservation of databases. The framework approach indicates the relationship between the different files. See chapter 4 for a detailed review of the 'XML as a framework' approach. The structure of the preservation object is shown in the following diagram, previously discussed in chapter 5.

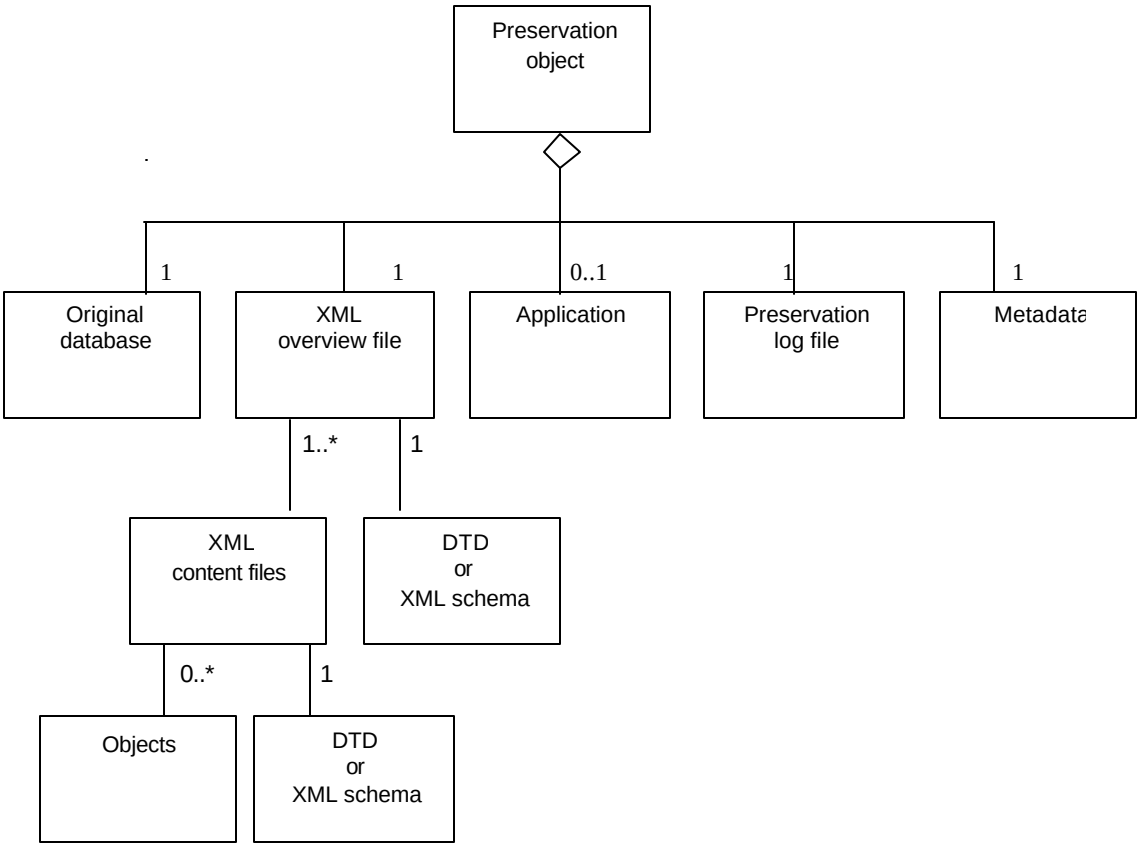


Figure 9 *Structure of the preservation object*

Notes: the diamond-shaped symbol indicates that the preservation object is comprised of the components to which it is linked. The significance of "1" is "1"; "0..*" signifies "zero, or more"; "0..1" signifies "zero or one"; "1..*" signifies "one or more".

The most important requirements to be met by the system for the preservation of databases are:

- the records must be stored in a reliable manner, such that they cannot be lost and cannot be changed subsequent to inclusion in the system;
- the links between the components of the preservation object must be retained;
- when the original file is ingested into the system the XML version must, insofar as is possible, be created automatically;
- if possible the system must automatically collect metadata and provide the user with support during the entry of metadata that cannot be recorded automatically;
- the system must save metadata for preservation and an audit trail (preservation log file).

Many of these functions are included in Records Management Applications (RMAs). Software of this type usually offers opportunities to configure and adapt it so that any of the above recommended functions can be added if they are not already present in the software. Digital records for long term preservation can be stored in the RMA until the time they can be transferred to the archives.

The European Commission has drawn up guidelines concerning the required functions of RMA software in the form of the MoReq specifications [22]. Attention is expressly drawn to one element of these specifications, namely the section on the export of digital records from the RMA[23]. Depending on how long the digital records are preserved by the original organisation before they are transferred to archives, it is possible that the RMA in which the digital records are preserved has already been replaced on one or more occasions. Should this happen then it will be necessary to transfer the digital records from one system to the other. In such a case it is then of essential importance that the digital records in the RMA can be exported in a format independent of the RMA supplier that retains all links between the digital records and between the various components of the preservation object.

(c) Practical issues

When converting the content of a (relational) database to XML it is necessary to take account of the following:

- Does the RDBMS permit the definition of table or column names using characters which are not compatible with the SQL standard?
- Does the RDBMS contain special signs or characters that are not defined in Unicode? It is possible that reserved characters such as '&', '<' and '>' are included in the content of the database. Make sure that these signs are extracted and converted to Unicode in the correct manner.
- Make sure that the format and the content of binary data types or binary objects, for example BLOBs (Binary Large Objects), are extracted and rewritten without change to the encoding.
- Take account of special or user-defined data types that are not specified in the SQL standard. Microsoft Access databases, for example, often make use of automatic numbering or data types with hyperlinks.

[22] Model Requirements for the Management of Electronic Records, March 2001. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/moreq.pdf .

[23] Section 5.3, "Transfer, Export and Destruction".

The design and configuration of the preservation system will need to take account of the following practical issues:

- Security: suitable measures governing access to the central preservation system will need to be implemented to prevent intentional or accidental damage to the stored information (implement an access classification system, see also NEN-ISO 15489).
- Backup: as with every important IT system, it will be necessary to implement a suitable backup strategy that will ensure the ability to restore the system following a system crash, intentional or accidental damage to the system, or a disaster such as a fire or flood.
- Flexibility: each group within an organisation may have a need of different metadata; moreover the needs of a specific group change over the course of time. Consequently it will be advantageous to keep this aspect of the system design as flexible as possible. The records manager will indicate the required flexibility after consultations with the users.
- Response time and reliability: because users may need to access the contents from the preservation system in their everyday work, short response times and reliability are necessary. Two issues are important in this respect: firstly, the user may need to save a file in the system quickly and with ease and, secondly, information already stored in the system must be easy to find and use. It should be noted that the patterns of use in the various business processes can vary greatly.

6.4 Action plan for end users

Introduction
In reading the publication *From digital volatility to digital persistence: Preserving databases* you will have discovered the advantages of working digitally, but also the specific problems that arise in the long term preservation of digital records in general and databases in particular. The Digital Preservation Testbed has tested preservation strategies for the record type 'database'. The best way of preserving databases at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the sustainable preservation of databases: from the line managers, records managers, and ICT specialists, to the end users who have desktop databases at their disposal and who make use of complex database systems. The concrete actions listed below are specifically oriented towards:

- General (line) managers
- Records managers
- ICT specialists and
- End users

These four groups have a specific responsibility in this matter. This final chapter sets out the concrete steps each target group has to take to make the long term preservation of databases a success. The concrete steps or actions are preceded by a description of the prior conditions.

Prior conditions
You are the person using the databases. In making use of them you also largely influence your organisation's ability to preserve its databases in a durable and authentic manner. Your organisation will have laid down policy, agreements and procedures governing the use of databases.
A number of parties play a role in this, such as the general (line) management, the records-management department, the ICT department and yourself, as the end user. The following section describes issues requiring your attention during the use of databases – because our studies have, above all, revealed that the long term preservation of digital records must begin at source. And that source is you.

Concrete actions for end users

Don't forget the GIGO principle

Remember that the only information that can be retrieved from the database is the information that has been entered into the database. This information is governed by the 'GIGO' principle, in other words, 'Garbage In Garbage Out'. If unreliable information has been entered into a database then the output from that database displayed onscreen or printed out in a report will inevitably also be unreliable. Consequently you as a user are largely responsible for determining the reliability of the information in the database, irrespective of whether the database is a desktop database such as MS Access or a complex database system.

Designing a database yourself?
It is possible that you need to design a database yourself, using a desktop database program such as MS Access, and you that are sufficiently familiar with the relevant program (where relevant, after completing a course) to do so. Even so (and, possibly, *precisely* then) you will certainly need to comply with a number of digital preservation do's and don'ts:

Give careful consideration to the design
A database is used to save and use structured data. In general, a database is comprised of a collection of related tables. The design of a suitable structure is anything but simple. One of the important steps in the design process relates to the normalisation of the data and guaranteeing the referential integrity of the data. Normalisation is a technique whereby a collection of data (which must meet your information needs) is divided into groups such that no irregularities can occur during the maintenance of the data.
Referential integrity ensures that the value of a foreign key (the link between two tables) always refers to an existing row (or tuple) in another table. This eliminates the possibility of "dead" references.
During the entry or modification of a row the system then verifies that the foreign keys have a valid value.

Assign meaningful names to tables and columns
In the absence of meaningful names for the tables and columns it will be difficult to determine what information is stored in the database – and certainly after a period of time.

Be consistent in the use of date and time notations
Date and time notations can cause a great deal of confusion, particularly in an international context. For this reason preference is given to a style of date notation in which the full name of the month is shown, for example 10 January 2003 rather than 10-01-2003. In the United States 10-01-2003 will be understood as October 1st 2003.

Avoid the use of the data type "currency"
When working with amounts in a specific currency make use of the numeric data type and include the symbol for the relevant currency in the screen layout. This avoids the problems encountered with the currency data type during migration, whereby the currency symbol can be lost – or, even worse, replaced by an incorrect currency symbol.

Exercise restraint in the use of passwords to protect databases.
Access, for example, offers two options for the protection of databases. You can either set a password to be entered before opening the database, or you can set protection at the user level specifying those elements of the database that can be opened or modified by the user. Another option is to delete the Visual Basic code from database and save this in an *.MDE file; it will then no longer be possible to make changes in the design of the forms, reports and modules.
The simplest approach is to set a password for opening the database. After setting a password, a dialog box will be displayed on opening the database. The password must be entered before it is possible to continue. This option is certainly not advisable. If you forget the password you will no longer be able to open the record, and the data in the database will no longer be accessible.
Protection at the user level is the most flexible and comprehensive form of protecting databases. This form of protection is comparable with the methods used in many other network systems. Authorisations are assigned to groups and users that determine which operations they can perform on the objects in a database. You can further extend the user-level protection by creating groups, assigning authorisations to the groups, and finally allocating users to the groups.

Glossary

Accessibility

The extent to which the authentic reproduction of a document, digital or otherwise, can be consulted without hindrance.

AIP

Archival Information Package. An OAIS term for a file (or group of files) held by an archive to contain, describe and manage an archival data object. The AIP contains metadata about the content, context, provenance, and packaging of the object. It contains the object itself (the record). It contains preservation information, and may contain other descriptive information about the record. See also DIP, SIP, and OAIS.

Archival institution

A location designated under the 1995 Archives Act as an appropriate storage place for the permanent preservation of archival records.

ASCII

American Standard Code for Information Interchange. It is a generally accepted standard established by the American National Standards Institute (ANSI) with the intention to enable the exchange of information between computers. The ASCII-table was registered as an official standard in the ISO-646 norm (1972). The ASCII or ISO-646 character set is 7-bits. This means that 7 bits are used in the creation of 1 character. So there are 2^7 (=128) different combinations. The original ASCII table contains the characters that are required to represent Western languages. Diverse national variations of the ASCII table have been created.

Authenticity

The extent to which the reproduction of a record is complete and totally in accordance with the original recording of the record and, furthermore, the extent to which its function, as intended when it was created, remains intact.

Backward compatibility

This means that software is able to decode or accurately read files made with earlier versions of the same software. Incidentally, most software is only backward compatible to a limited degree.

Behaviour

Behaviour is one of the five attributes of digital documents, described by Jeff Rothenberg and Tora Bikson in “Carrying Authentic, Understandable and Usable Digital Records Through Time”. Behaviour enables the user to interact with the digital document, for example, by opening an attachment or by activating a hyperlink. The other four attributes are content, context, structure and layout.

BLOB

A large single quantity of data, such as an audio file, graphics, an image, or a whole book, that can be saved as a field in a database.

Compiler

In order to execute a computer program written in a specific programming language, the *source code* must first be translated into *object code* (this is machine code). A compiler translates the complete program into object code, which can then be executed.

Computer file

A sequence of bits stored as a single unit conforming to a particular file format.

Context

The administrative, organisational, legal and technical environment, within which the function of the record has to be interpreted in relation to the activities and tasks of the record creator.

Constraint

A condition that restricts, prevents or hinders something.

Conversion

The procedure of converting or transferring data into another storage format.

Digital longevity

The result of safeguarding the authenticity, the ability to consult and the readability of digital records for the duration of the applicable preservation period.

DIP

Dissemination Information Package. An OAIS term for the file or files which are used to share an archival information object (a record) with authorised users. The DIP contains the data and those metadata which are required to give users confidence in the authenticity of the records, and to allow users to access the records. See also AIP, SIP, and OAIS.

DIV

'Documentaire Informatie Voorziening'- Documentary Information Services. The process of communicating by way of documents; this concept thus implies both paper and digital documents, such as textual and financial records, process control data and images.

DMS

Document Management System, also Electronic Document Management System (EDMS). A system that offers functionality for acquiring, storing, archiving and retrieving documents, including their management, whilst implementing, administering, relaying, and authorising users. Document Management Systems monitor access to files and may keep an audit trail of actions and events. They often maintain a version history of their documents.

Emulation

Reconstructing the old hardware using software. Running this software on current and future hardware so that the problem of obsolescence can be avoided.

ERD

Entity Relationship Diagram: describes/outlines the used entities and their mutual relationships in a model.

Font

A co-ordinated set of characters; a complete alphabet in upper- and lower-case letters, numbers and symbols in a specific design. A font is likewise specified through orientation, symbol set, spacing, point size, character type, style, and thickness.

Foreign key

A *foreign key* is the connecting link between two tables, a field in a database record that points to a key field of another database record in another table. Using the value of a row in one table, the right row with the related data in another table can be found. The one table has, so's to speak, the key to the other 'foreign' table.

Form

The outward appearance of a record in which the structure and layout are visible.

GUI

Graphical User Interface. A program that makes the operating system invisible for the user and offers him or her the opportunity to execute different actions by pointing with the mouse. No complicated commands have to be typed in. The most familiar example of a GUI is Windows.

HTML

Hyper Text Mark-up Language. A mark-up language for the creation of hypertext documents. HTML is used to write pages for the World Wide Web.

Integrity

A property of a record when the form, content and structure of a record are the same when the record is consulted as when the record was created.

JPEG

Stands for Joint Pictures Expert Group and is in particular a file format for photos on websites. JPEG divides the image into blocks and only stores the most relevant information in each block.

Mark-up language

Another word for Meta languages, specially intended for adding structure to complex documents. The most well-known variants are HTML and XML

Metadata

Data that describes the context, content, form and structure of digital documents and their management through time.

Migration

The transfer of files from one hardware and/or software environment to another.

OAIS

The Consultative Committee for Space Data Systems (CCSDS) has published a recommendation for a Reference Model for an Open Archival Information System (OAIS). The OAIS model defines an information model for an archival system for digital records and provides a list of responsibilities that the system must meet. This model has been adopted as ISO-norm 14721:2002.

PDF

Portable Document Format.
A file format developed by Adobe Systems Inc. for exchanging documents while retaining their appearance and design.

Platform

All of the hardware and operating software on which the application software runs.

Preservation

Processes and activities relating to ensuring the technical and intellectual conservation of authentic, accessible, and useful records through time.

RDBMS

A Relational DataBase Management System is software that enables the user to implement a database with tables, columns and indexes; to guarantee the referential integrity between rows of different tables, to automatically update indexes; to interpret a SQL search query and combine information from different tables.

RMA

Records Management Application. Application software for ingesting, managing, and making records available.

Script

A programme or subroutine that is saved in a script file and used for a specific purpose, for example to start an Internet session (a login script).

SIP

Submission Information Package. This is an OAIS term for the file or files which are used to submit an information object (a record) to an archive. The SIP will contain the actual record (its content, in its original format), and also appropriate metadata describing its context, provenance, disposition, and access restrictions.

SOP

Standard Operating Procedure. A formal description of the way a process should be carried out. A key part of a formal quality system.

SQL

Structured Query language – query language for a relational database. SQL2 is the ISO-norm 9075 from 1992, the follow-up to the first from 1987; SQL3 norm (with object oriented application) was widely accepted in around 1999.

Storage

Structural retention of digital information, like files and records, on magnetic or optical media.

Structure

The logical connections between the elements of a digital record or of an archive.

Tuple

A row in a table of a (relational) database.

URL

Uniform Resource Locator. An Internet naming convention for resources available via various TCP/IP application protocols. For example:
HTTP ://www.digitalduurzaamheid.nl is the URL for the Digital Longevity programme website.

Viewer

A software application that enables certain files to be looked at but not edited or altered. Works without the original software that was used to create the files .

W3C

The World Wide Web Consortium develops standards for the World Wide Web (WWW), at present the most important application on the Internet. One of W3C's most important domains relates to mark-up languages for defining and structuring web documents. See also www.w3c.org

Wrapper

A term that stands for an approach whereby XML is used as a type of envelope, a casing.

XML

Stands for eXtensible Mark-up Language and is a text-based language for enriching data with information about structure and meaning. It is an open standard, defined by the World Wide Web Consortium and is independent of specific hardware and software.

XSLT

Extensible Stylesheet Language Transformations: a tool for converting XML documents, to HTML for example. See also: www.w3c.org/Style/XSL/

Bibliography

Ashley, Kevin	*Digital Archive Costs: Facts and Fallacies* (1999); http://europa.eu.int/ISPO/dlm/dlm99/Proceed99-down_en.htm DLM Forum 1999 , pp 121 - 128
Besser, Howard et al	*Preserving Digital Materials*, final report of the 'Digital Preservation and Archive Committee', 18 October 2001. Howard Besser, Curtis Fornadley, Anne Gilliland-Swetland, et al.
Boudrez, Philip	*<XML/> en Digital Archiveren* (2002) http://www.antwerpen.be/david/teksten/xml_digitaalarchiveren.pdf
Boudrez, Philip	*Standaarden voor digitale archiefdocumenten* (October 2001) http://www.antwerpen.be/david/teksten/DAVIDbijdragen/Standaarden.pdf
Dekker, R, Dürr, E.H., Slabbertje, M. en Meer, K van der Diessen, Raymond van & Steenbakkers, Johan	*An electronic archive for academic communities* , November 2001 *The Long term Preservation Study of the DNEP project – an overview of the results (December 2002)*
Giesbers, Saskia	*Records Management Terminologie* (6 March 2002) http://www.rmconventie.nl/publicaties-rm/RecordsManagement_termen_en_definities.pdf
Feeney, Mary (Ed)	*Digital Culture: Maximising the Nation's Investment* (National Preservation Office UK, 1999)
Hofman, Hans (redactie)	*Blijvend in business. Naar een geordende en toegankelijke staat van informatie (ex art.12 Archiefbesluit)* VNG Uitgeverij, Den Haag, 2003
InterPARES Project	*Authenticity Task Force Final Report* (2002) http://www.interpares.org/book/interpares_book_d_part1.pdf
InterPARES Project	*Preservation Task Force Final Report* (2002) http://www.interpares.org/book/interpares_book_f_part3.pdf
Jenkins, Clare	*Cedars Guide to Digital Preservation Strategies* , *2 April 2002,* www.leeds.ac.uk/cedars/guideto/dpstrategies
Lorie, Raymond	*A Project on the Preservation of Digital Data* http://www.rlg.org/preserv/diginews/diginews5-3.html
Lourens, Wim, et al	*Emulation and Conversion: Organisational and Architectural Overview of an electronic Archive* http://www.library.tudelft.nl/e-archive/Documenten/Resultaten/reportone13.pdf
Mellor, Paul et al	*Migration On Request, a Practical Technique for Preservation* (2002) http://www.si.umich.edu/CAMILEON/reports/migreq.pdf
Ploeg, Dr. F. van der	*Regeling geordende en toegankelijke staat archiefbescheiden* (February 2002) http://www.nationaalarchief.nl/images/3_2598.doc
Prins, prof. mr. J.E.J. Matthijssen, dr. L.J. Rothenberg, Jeff & Bikson, Tora	*De digitale overheid en de wet; juridische kaders voor gebruik van digitale documenten bij overheden* (Programma Digitale Duurzaamheid, Den Haag, 2000) *Carrying Authentic, Understandable and Usable Records Through Time* (1999) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

Rijksarchiefinspectie	*Wet- en regelgeving* www.rijksarchiefinspectie.nl/wetgeving/
Testbed Digitale Bewaring	*Migratie: Context en huidige stand van zaken* (2001) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_migratie.pdf
Testbed Digitale Bewaring	*XML en digitale bewaring* (2002) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-nl.pdf
Testbed Digitale Bewaring	*Emulatie: Context en huidige stand van zaken* (2003) http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_emulatie-nl.pdf
Thibodeau, Ken	*Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years* http://www.clir.org/pubs/reports/pub107/pub107.pdf
Thomas, Wimpe	*XML: de mogelijkheden en valkuilen voor de overheid* (19 September 2002)
VERS	*Victorian Electronic Records Strategy Final Report* http://www.prov.vic.gov.au/vers/published/final.htm
Waalwijk, Hans et al Zuurmond, A, Mies, K.	*Softwarespecificaties voor Records Management Applicaties voor de Nederlandse overheid*, versie 4.12.9, September 2002 http://www.digitaleduurzaamheid.nl/bibliotheek/docs/remano_versie4_12bis.doc *Winst met ICT in uitvoering, Zenc, Den Haag,* June 2002.

Appendix A Preservation Log File

The exact contents of the Preservation Log File depend on the chosen preservation procedure. At a minimum the log file should contain the following information:

Technical Metadata

- Details of the original computing environment: client software = application (e.g. MS Access) + hardware environment (e.g. Pentium 4) + operating system (e.g. Windows XP);
- Details of interim formats (e.g. ASCII, DIF);
- Details of new computing environment (sufficient details must be recorded to ensure access to the records in their current format).

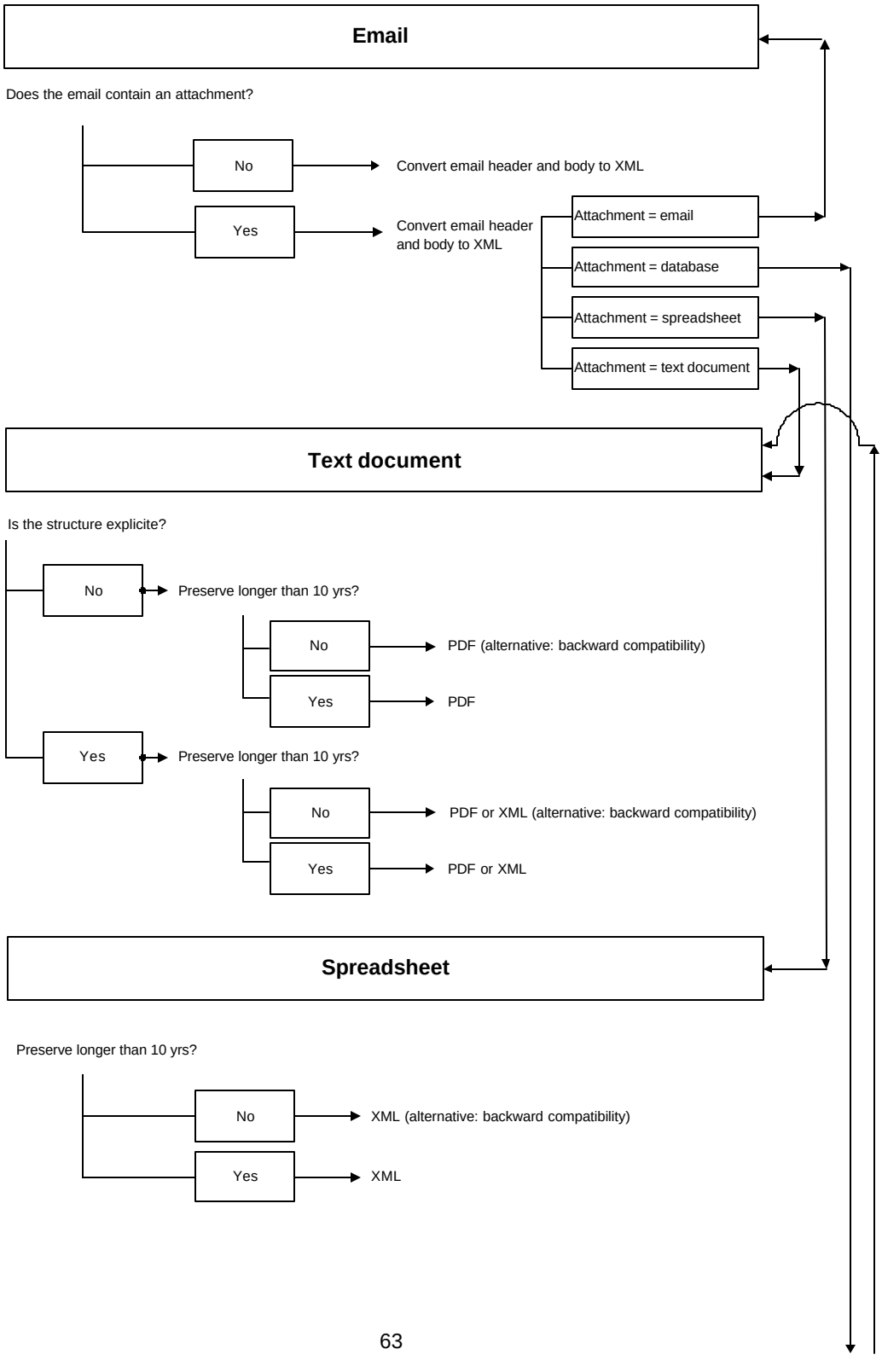
Preservation action metadata

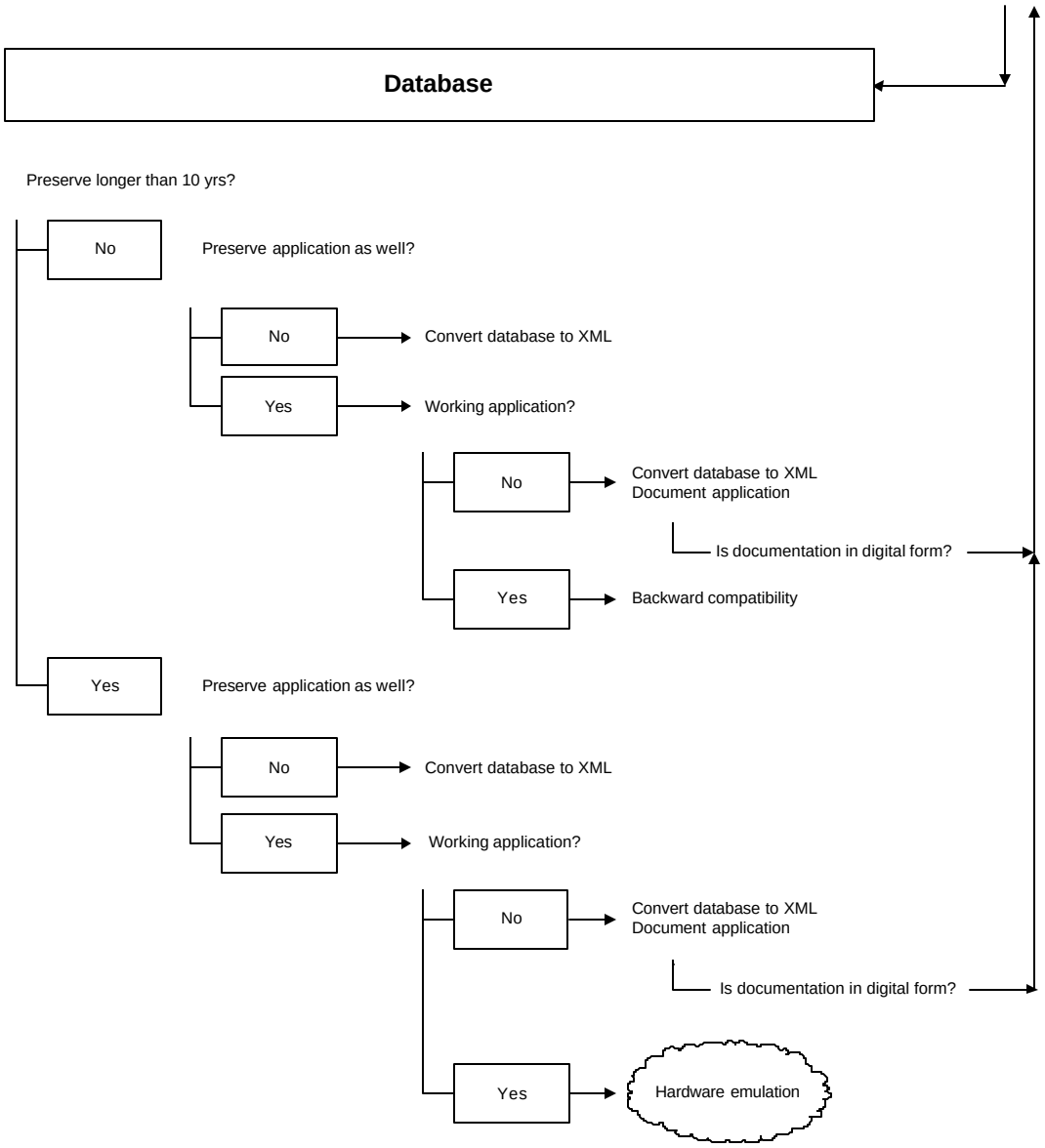
- Date and time of any and all preservation action;
- Person(s) responsible for of any and all preservation action;
- Details of the transformation (conversion) software and;
- Conversion results.

Metadata which refer to the access of the records

- Privileges/rights and;
- History.

Appendix B Decision model




Database
Preserve longer than 10 yrs?
No
Preserve application as well?
No
Convert database to XML
Yes
Working application?
No
Convert database to XML
Document application
Is documentation in digital form?
Yes
Backward compatibility
Yes
Preserve application as well?
No
Convert database to XML
Yes
Working application?
No
Convert database to XML
Document application
Is documentation in digital form?
Yes
Hardware emulation

Appendix C Cost model

Cost indicators and the cost model

Testbed has studied the costs involved in the long term preservation of digital records, drawn up a list of indicators which exert an influence on the total cost of preservation, designed a computational model for the calculation of these costs, and compared the costs involved in the various methods for the creation of digital records and in the various preservation strategies.

These costs are estimates based on Testbed's studies and experience, published information, and information others have supplied to Testbed. These detailed estimates are intended to encourage others to submit their comments on these figures, and to report the costs incurred in practice.

This discussion is, in the first instance, focused on the larger archives; however, it is also applicable to the local storage needs of ministries and other (government) agencies. The costs identified will always be incurred, irrespective of whether the relevant records need to be stored for no more than 10 or 20 years or come into consideration for permanent preservation. Although the scale of the storage system or repository and the relative sizes of the different components of the installation may vary, the following cost factors will in any case need to be taken into consideration.

Although the following list might initially appear to be extremely detailed, it is nevertheless important not to overlook any of these factors. It will, in particular, be necessary to calculate capital and personnel costs. Digital preservation will continue to develop and change. Consequently the functionality for sustainable preservation of digital records will also need to change. The costs incurred in making future changes need to be incorporated in the computational model right from the very beginning.

Summary of the cost indicators

The costs involved in the long term preservation of digital records are influenced by a number of factors. These are summarised below. In the following discussion, a digital archive is included as an element of the costs, as is the storage of the digital records, for example, in an RMA (Records Management Application) or a DMS (a Document Management System). It is often difficult to specify the demarcations between the actual use of the records, their local storage, local preservation, archiving, and long term preservation. The concept of the 'records continuum', which is ideally suited to use in this context, can be defined as:
'a consistent and coherent regime of management processes from the time of the creation of records (and before creation, in the design of record keeping systems), through to the preservation and use of records as archives.'

Consequently digital preservation is not just a necessity for archival repositories, but also for every organisation. Each organisation will need to specify its own requirements, determine its own demarcations, and tailor the cost model discussed in this Chapter to its specific situation.

Testbed makes a distinction between the following cost indicators:

1) The cost of the digital archival system (a digital depot or repository) and functionality for the long term preservation of digital records[24]
2) Personnel costs
3) The cost of the development (or procurement) of software and methods for the preservation of digital records
4) The cost of the actual storage of digital records
5) Other factors that exert an influence on the total cost

1. The cost of a digital repository and functionality for the long term preservation of digital records

The cost of a digital archival repository and functionality for long term preservation is comprised of various components. The cost model (see page 75) indicates the major factors and the minor factors. It also explains which indicators are influenced by the different ways in which records can be created and by the different preservation strategies, and which indicators are not sensitive to (strategic) choices of this nature.

1.1 The physical space
- 1.1.1 Server room, with air-conditioning
- 1.1.2 Sufficient office space
- 1.1.3 Conference room
- 1.1.4 Toilets and kitchen
- 1.1.5 Security

Physical space is required for systems for storage and long term preservation. Servers will be required for the storage of digital records and for the management of long term preservation. It may be advisable to set up separate development, test and production facilities for long term preservation. This can reduce the risks and increase productivity. There will also be a need for offices and conference rooms, for both the staff and visitors.

1.2 Hardware for the digital archival repository
- 1.2.1 Servers for the storage of digital records
- 1.2.2 Disks, tapes, or other storage media
- 1.2.3 Backup equipment
- 1.2.4 Network communications

Hardware is required for the storage of records (in a file system, archival repository, or RMA). It will also be necessary to configure the storage equipment once an impression has been gained of the number of records that will need to be stored.

Appreciable costs can be incurred in the purchase of storage media (tapes, CDs). Make sure an adequate amount is budgeted for effective storage and backup media.

Network facilities may also be important. Archival repositories that receive large numbers of digital records from diverse locations may require a high-speed connection or a flexible connection capable of accommodating varying loads.

[24] See Appendix D: Functional specifications for a preservation system

1.3 Software for the digital archival repository
- 1.3.1 Operating system
- 1.3.2 Security
- 1.3.3 Specific software for archives management
- 1.3.4 Old software applications
- 1.3.5 New (current) software applications
- 1.3.6 Display programs
- 1.3.7 Communications software
- 1.3.8 Database licences

This Section covers issues such as the purchase of operating systems and the standard software for databases. There will also be a need for protection software (against viruses, unauthorised access, and tampering with the archives by unauthorised persons). There may also be a need for specific software for the receipt and storage of authentic digital archival records, such as Depot 2000 or the Digital Archive System of the United Kingdom National Archives [25]. Every organisation which works with digital records will, irrespective of its size, have a need for a DMS (Document Management System) or an RMA (Records Management Application).

The following discussion assumes that the archival repository possesses functionality to provide access to the stored records. Consequently in addition to the customary storage software, there will also be a need for specific applications or display software which enable users to display (or use) the stored records.

Communications software and network and database licences are two other issues that are often overlooked when preparing the budget.

1.4 Hardware for the preservation system
- 1.4.1 Servers for the preparation of software
- 1.4.1 Servers for the testing of software
- 1.4.3 Servers for the storage of records before preservation action
- 1.4.3 Servers for the storage of records subsequent to preservation action
- 1.4.5 Work stations for programming work
- 1.4.6 Disks and tapes
- 1.2.3 Backup equipment
- 1.2.4 Network communications
- 1.4.9 Reading equipment for tapes and disks

The preservation system may require computer systems (servers and storage) of the same type and size as the archival repository. This is necessary so that the preservation system can receive groups of records with a total size in excess of several terabytes. The system will need to store these records in a safe manner ready for, for example, performing preservation operations (such as migration or emulation), assessing the results of preservation actions, and returning the preserved records to the archival repository (digital depot).

In addition more servers and storage equipment may be required for the development and testing of:
- preservation methods,
- software to evaluate the results of preservation operations,
- other tools.

When automated tools are to be developed and tested, the test system may need to test large datasets so as to collect sufficient statistical proof of the tools' success.

[25] http://www.pro.gov.uk/about/preservation/digital/archive/default.htm

The preservation system may need a variety of different types of reading equipment to read the various formats of tapes and/or disks.

1.5 Software preservation system
- 1.5.1 Operating systems
- 1.5.2 Program environments
- 1.5.3 Security
- 1.5.4 Old software applications
- 1.5.5 New (current) software applications
- 1.5.6 Software for the preservation of documents
- 1.5.7 Test and evaluation software
- 1.5.8 Communications software
- 1.5.9 Database licences

Extremely comprehensive software may be required for the sustainable preservation of records. The preservation system may require more than one operating system, since it may be necessary to transfer records from their original operating system to another operating system capable of an improved preservation performance. In addition, there may also be a need for more than one programming environment if the organisation plans to develop in-house software tools or modify third-party tools.

The preservation system will also need to cater for a comprehensive package of software applications. This will allow for research into the different options for preservation and experimentation with a range of records and record-batches.

Automation is a significant factor in controlling the costs of large-scale digital preservation. Manual processing is one of the largest cost items for digital preservation. Consequently, automated preservation actions and automated evaluation (tests) are significant factors in controlling the costs of digital preservation.

2. Personnel costs

This Section reviews the staff duties involved in the operation of a preservation system. The discussion reviews the numbers and types of staff that will be required. The cost model discussed later in this chapter is based on the time that will be needed from such staff, who have various qualifications and skills.

Personnel costs always form a major factor. The necessary staff could be selected or recruited specifically to work upon the digital archive and preservation system. However, in some instances it may be preferable to second staff from other disciplines (such as the records-management department or the ICT department). In yet other situations, preference might be given to the use of temporary employees, or to the use of the services of a specialised company.

2.1 Duties of digital archives personnel
- 2.1.1 Compile requirements
- 2.1.2 Obtain funds and support
- 2.1.3 Design and construct the digital archives

The staff will need to begin by designing and constructing the digital archive. This will require a budget for between one and two man-years; cost calculations should not underestimate this.

Even designs obtained from other organisations or purchased from a commercial supplier will still require modification to meet the needs of the specific organisation.

- 2.1.4 Stocking the digital archive
- 2.1.5 Process management
- 2.1.6 Managing the digital repository
- 2.1.7 Security management
- 2.1.8 Quality control system and documents
- 2.1.9 Standard Operating Procedures (SOPs)
- 2.1.10 User manuals

Once the digital archive has been constructed the next step will be to develop the procedures and commence the management of the digital archive. The internal management encompasses the security and access procedures, a comprehensive quality system (to ensure the authenticity of the records stored in the digital repository), everyday SOPs (Standard Operating Procedures), and user manuals.

Management can also incorporate external activities, such as the identification of records, the arrangement of records, acquisitions, and cataloguing.

2.2 Duties of preservation system personnel
- 2.2.1 Compile requirements
- 2.2.2 Obtain funds and support
- 2.2.3 Design the preservation system
- 2.2.4 Construct the preservation system
- 2.2.5 Process management
- 2.2.6 Management of the preservation system
- 2.2.7 Ongoing security management
- 2.2.8 Quality control and documents
- 2.2.9 SOPs
- 2.2.10 User manuals

The staff responsible for the preservation system will also first need to design and construct the system. They will then need to establish the quality control system, the SOPs, and the procedures. Finally, they will need to begin the development of the preservation methods and evaluation tests, and start work on the sustainable preservation of the records.

Once again, the costs incurred in the development and construction phase are easily underestimated.

2.3 Duties of Public-services staff
- 2.3.1 Access management
- 2.3.2 Training and schooling

3. The cost of the development (or procurement) of software and methods for the preservation of records

3.1 Determine authenticity requirements
3.2 Analyse authenticity requirements

One important Testbed conclusion was that digital preservation is not a question of all or nothing. In many instances the characteristics of records that are essential to the records' integrity and authenticity can be separated from other less important characteristics. Digital preservation activities can then focus on those aspects of essential importance to the integrity and authenticity of the record.

The aforementioned tasks can be carried out solely by the organisation that created the records.
The initial costs incurred in digital preservation relate to issues such as determining the authenticity requirements for each batch of records. In an ideal situation these requirements will be specified by the records managers. However, in some situations it may be necessary for the (authenticity) requirements to be determined by a multidisciplinary team comprised of specialists such as archives-management and IT specialists, whereby every member of the team has some experience in the other specialists' fields.

The cost model assumes that (authenticity) requirements will need to be determined for each batch of records. A batch contains records all made with the same application, the acquisition or preservation of which all takes place at the same time. It will later be shown that the size of the batch is a critical factor in the cost.

3.3 Design preservation approach
3.4 Develop preservation approach
3.5 Preservation software (parser, etc.)
3.6 Viewing software

Once the authenticity specifications have been determined, the next step is to design and develop a suitable preservation approach. Since this is a lengthy process that requires a large number of skills, it is assumed that an international collection of shared preservation strategies will gradually be developed. However, even then it will still be necessary to evaluate these strategies in terms of the specific requirements of the batch of records in question. In some instances it will ultimately be necessary to modify the approach.

3.7 Test the approach
    3.8 If approved, continue
    3.9 If not approved, return to 3.1, 3.2, or 3.3
3.10 Document the approach

Finally, each strategy or approach will need to be tested and documented. All of the IT operations involved in each preservation system will need to comply with the most stringent quality standards. A high level of quality is of essential importance to the authenticity, since the quality systems and the documentation are needed to prove that the preservation actions have achieved the intended results, and that they have had no influence on other records. A high level of quality also increases the probability that the approach will be re-used in this or other preservation systems.

4. Cost of the performance of preservation actions

This Section reviews the costs incurred in the performance of preservation actions on digital records. Within this context the 'performance of preservation actions' can relate to diverse activities:

- The migration of records (transformation)
- The performance of a migration on request
- The use of emulation to retain the accessibility of records

These can be specified with OAIS terminology[26]. A migration or other form of transformation of the records results in changes to the Archival Information Package (AIP) stored in the archives. AIP1 is changed into AIP2. AIP2 serves as the basis for the DIP (Dissemination Information Package) issued to applicants. Migration is one of the preservation strategies examined by Testbed.

The performance of a 'migration on request'[27] has no influence on the AIP. It is possible to produce a DIP which differs from the last DIP produced by the same AIP. It will in any case be necessary to produce a DIP that is both authentic and accessible. Migration on request will require the preparation, testing and issue of an appropriate software tool before the migration (or another form of transformation) is carried out. If a user requests a record, the correct tool is retrieved and the record subsequently transformed. The user receives the transformed copy in the form of a DIP. The transformed copy can be stored, or deleted once the user has finished with it.

Another form of transformation which Testbed examined as a possible approach to the sustainable preservation of records is conversion to XML. The costs for this are included in the migration. The possible cost benefits of XML (because it is an open standard, is expected to have a long and useful life, and can be interpreted by a variety of applications) are explained below. Conversion to XML changes the AIP from AIP1 to AIPX, whereby AIP X is in XML.

Retaining access to digital records through the use of emulation has no influence on the record contained in the AIP. In principle the DIP will also remain unchanged in the future, although in practice it may be necessary to implement a number of small modifications to the DIP to accommodate future technology. In this respect, Testbed has examined the UVC approach formulated by IBM. This approach is based in part on emulation, and in part on migration.

In fact, and as will be revealed by our cost model, the cost of digital preservation activity is only a small fraction of the total cost of the digital archive. The cost of digital preservation also depends on the size of the batch: the cost model will reveal that grouping records in larger batches is a particularly cost-effective approach.

4.1 Determine which digital records will need to be preserved
4.2 Construct the interface with the archives-management system
4.3 Incorporate systems for electronic records management, DMS, RMA
4.4 Receive digital records
4.5 Select the preservation strategy and approach

Records which will need transformation, or which require the development and testing of an emulator, can be identified by an automated process (a technology watch, or a process within an RMA or DMS) or by means of manual identification. Records requiring transformation will first need to be selected and transferred to the preservation system, after which a preservation strategy can be assigned to them.

4.6 Preparing records for transformation
4.7 Supply metadata
4.8 Repair or modify records

In specific circumstances it will be necessary to prepare records for transformation. Records will be in need of 'repair' if metadata are missing, or when the records possess properties that could pose risks with the selected preservation strategy. This can be a slow and labour-intensive process that accounts for the majority of the costs.

---

[26] http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[27] '*Cedars Guide to Digital Preservation Strategies*', Clare Jenkins, April 2002.

Automated methods for the assignment of metadata or the repair of records can greatly reduce the associated costs.

4.9 The transformation of records using the selected method
4.10 Evaluation of the transformation
    4.11 The records are accessible
    4.12 The records are authentic
    4.13 If NO, return to 4.9 or to an earlier step
4.14 Storage of preserved records in the digital archives

The final step is to transform the records. Every transformation will need to be evaluated to demonstrate that the records are still accessible and to ensure their authenticity and integrity after the transformation. Within the context of this discussion 'transformation' refers both to migration and conversion to XML.

Transformed records shall need to be transmitted to the digital archive, where they will be stored until further preservation actions are required and where access to the records can be managed.

Note: it is assumed that hardware emulation is not employed. When hardware emulation *is* used there will be (virtually) no transformation costs, and no recurrent costs for the preparation of records. The cost model reviews the long term cost benefits offered by emulation.

5. Other factors that exert an influence on the total costs

Other factors not mentioned in the above summary are also of relevance. These indirect factors can, however, account for a substantial proportion of the total costs. In addition, they can also have an influence on the impact of a number of the aforementioned factors.

5.1 Public services
    5.1.1 Number of users
    5.1.2 Required training and support tools
    5.1.3 Required maintenance and support

The degree to which users draw upon the services of the archive and preservation systems will have a great influence on the costs; however, the provision of services also offers an opportunity for the generation of income. Kevin Ashley has drawn up a summary of the costs incurred in providing public services[28].

5.2 The time between preservation actions
The time between preservation actions is a critical cost factor. The more preservation actions, the higher the costs. In addition, more preservation actions increases the risk of affecting the authenticity and integrity of the records, and there may also be a need for additional tests.

The costs can be reduced with longer periods of time between the preservation actions. However, preservation actions carried out at excessively great intervals of time can increase the risk of problems with digital preservation and the cost of preservation.

[28] *Digital Archive Costs: Facts and Fallacies* Kevin Ashley, DLM Forum 1999 http://www.europa.eu.int/ISPO/dlm/fulltext/full_ashl_en.htm

5.3 Technology watch – assessing when the hazards increase

A technology watch requires the monitoring of the hardware, software and systems used for the current records. The threatened obsolescence of components on which the digital records are dependent will give cause to the need for an evaluation and implementation of the necessary measures.

5.4 Supplementary storage requirements

Testbed recommends that the original files of the preserved records also be stored. We advise that text document records are stored in both PDF and XML. These recommendations increase the storage space required for each record. In some instances the space can be increased by a factor of between three and five. Although storage is relatively cheap, this will result in additional costs.

5.5 Links to the management systems for electronic records

Testbed has not examined links in DMSs or RMAs. However, it is to be expected that these links will be desirable at some point in the future. Extra costs will be incurred in the construction and maintenance of these links.

5.6 Volume of records

The expected volume of the records to be stored and managed will have substantial consequences for the costs. The storage costs increase linearly with the volume. Moreover the required space will increase even more rapidly when the records need to be stored in a variety of formats (for example, the original file format and two migrated formats).

More expensive servers and storage systems may be required for large volumes of records (more than 500 Terabytes), in particular when there is a need for rapid access to the records.

It should be noted, however, that the cost of digital preservation is influenced more by the variety (diversity) of the records than by the volume of the records. Records that make use of various functions of an application or different application software will generally require different preservation strategies, or at least a variety of tests for the preservation strategies. For this reason it will cost less to preserve a few large batches of records which all use the same application (maybe also the same template) and have the same authenticity requirements, than it will a large number of small but diverse batches that take up the same amount of storage space.

5.7 Requirements for authenticity and reliability

The authenticity requirements for a specific type of records constitute a significant cost factor. Consider, for example, a text document. Preservation of this will be a relatively simple task when only the plain text (the content) needs to be preserved. Highlights can also be preserved, at a slightly increased cost. However the costs will increase if the exact position of each character on the page and the exact colour must to be preserved. This will also complicate the preservation tests for the approach.

For this reason it is important that the authenticity requirements are determined in as comprehensive and realistic manner as possible.

5.8 Preservation of the systems themselves

5.8.1 Preservation of the archival system

5.8.2 Preservation of the preservation system

Finally, it will also be necessary to preserve the systems themselves. These costs will in part be covered by depreciation, as a result of which funds are made available for a three to five-year replacement cycle. However, it is also possible that specific elements of the preservation system form part of the digital record or preservation object[29], as a result of which these will need to be preserved separately.

One example of such an element is the preservation log file, the logbook of earlier preservation operations, which will also need to be preserved. Another example is the emulator, which will need to be preserved to ensure for the continued accessibility of the records and for their possible reuse.

[29] See Chapter 5 with the four recommendations for a further explanation of what is referred to as the 'preservation object'.

Cost model: Results

A cost model has been prepared in the form of an Excel spreadsheet. This computational model is available from the Testbed website (www.digitaleduurzaamheid.nl). The following sections review a number of the most important conclusions, and relate these to other general information about the costs of preservation and of archiving records. The sections are arranged in the lifecycle of a record, i.e. its creation, acquisition (by the archive), transformation and, where applicable, emulation.

1. Assumptions made for the computational model
It is assumed that there are six categories of staff, who are paid four different hourly salaries. It is also assumed that each category of staff works 1620 hours per annum. A cost category is assigned to each category of staff and will be used in later parts of the spreadsheet:

- Administrative support (1), hourly wage EUR 14
- Archivist/Records Manager (2), hourly wage EUR 25
- Supervisor (3), hourly wage EUR 32
- Data-management assistant (4), hourly wage EUR 18
- Programmer (5), hourly wage EUR 25
- Senior IT assistant (6), hourly wage EUR 32.

The above salaries have been assumed for the purpose of calculations with the computational model; they are, of course, open to discussion and/or amendment. Amendments can be made to all calculations in the computational model that make use of the relevant salary.

It is assumed that the archive and preservation system will need four types of space. An estimate has been made of the cost of furnishing each type. A proposal has also been made for the number of staff, based on full-time equivalents (FTEs), to be based in each category of space. Once again, amendments can be made to the computational model that will influence the results from the calculations.

- Space for the digital archive system: 1 FTE, categories 2 and 4; costs EUR 1,897,400 in the first year and EUR 632,403 in every successive year; storage capacity 100 Tb.
- Office space: 1.4 FTE, categories 1 and 3; costs EUR 7,400 in the first year and EUR 2,466 in every successive year. Standard office equipment and furnishings.
- Development and test area: 3 FTEs, categories 4, 5 and 6; costs EUR 50,200 in the first year and EUR 16,731 in every successive year. This room will be equipped with additional software (program environments) and hardware (inclusive of older systems and newer systems). These facilities will be used for the development of digital preservation tools.
- Space for the digital preservation system: 3 FTEs, categories 2, 4 and 5; costs EUR 279,600 in the first year and EUR 93,190 in every successive year. This room is intended for the acceptance, transformation and testing of digital records. The equipment will be capable of storing 10 Terabytes of records each time. If so required, this room can be combined with the space for the digital archive.

The facility specified here should be able to manage a total of 100 Tb of records, and could readily be expanded to 1000 Terabyte (1 Petabyte) or more.

This facility would be able to cater for the annual acquisition and transformation of 40 batches of e-mails (2000 e-mails in each batch), 20 batches of text documents (200 documents in each batch), 20 batches of spreadsheets (20 in each batch), and 20 databases. A batch of 4000 e-mails costs about the same as a batch of 2000 e-mails, provided that they have been created in a durable manner.

Every year this facility would be able to develop 55 ‘preservation approaches’, each for a different batch of records. It would be possible to carry out more work at the same cost in the event that different batches resemble each other. For example, it would be possible to use the same preservation approach for two batches of text documents which differ only with respect to the process in which they were used, but that were created by the same application, possess the same properties, and are governed by the same authenticity requirements. The use of the same preservation approach would result in substantial cost savings.

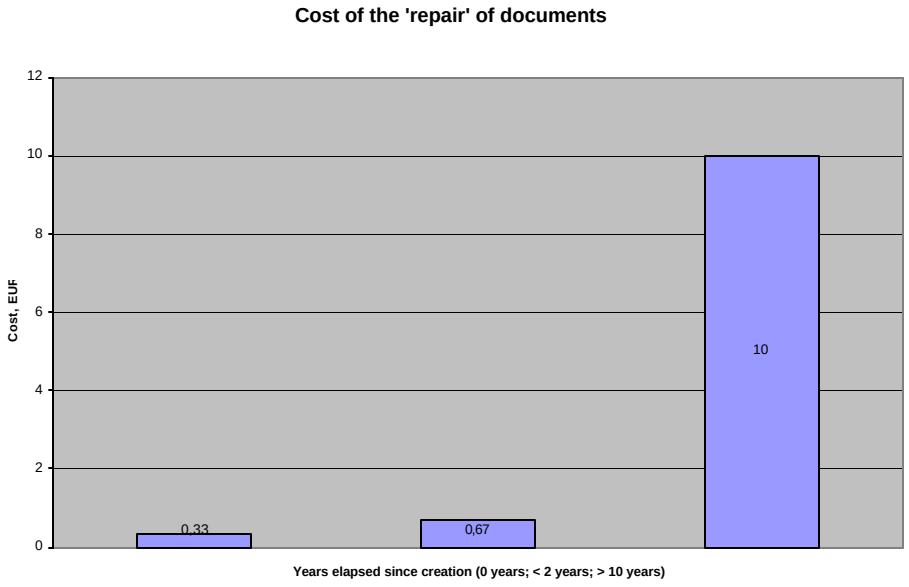
The aforementioned batches would require a total storage capacity of 5 Gb (on the basis of 50 kb per e-mail, 100 kb per text document, 250 kb per spreadsheet, and 2 Mb per database). The discrepancy between these figures and the total storage capacity of the model facility compared on the one hand with the NARA figures for electronic records or on the other hand with the Amsterdam Municipal Archives, indicates two things:

- the estimated number of digital records that can be processed by a team is probably a conservative estimate;
- the automation of the procedures used for the acquisition, inspection, and transformation of digital records will result in tangible benefits.

2. The creation of documents

From the recommendations that Testbed has made for four types of records, it will be clear that digital preservation begins at source, i.e. at the time of the creation of the records. The creation of records in an appropriate manner is a quicker, cheaper and less risky manner of obtaining suitable durable records, compared to the ‘repair’ of those records at a later date.

Testbed has estimated the costs incurred in the creation of durable documents with the appropriate metadata, the addition of metadata and performance of repairs after 2 years or less, and the addition of metadata and performance of repairs after 10 years or less.



Graph 1: Cost of the ‘repair’ of records

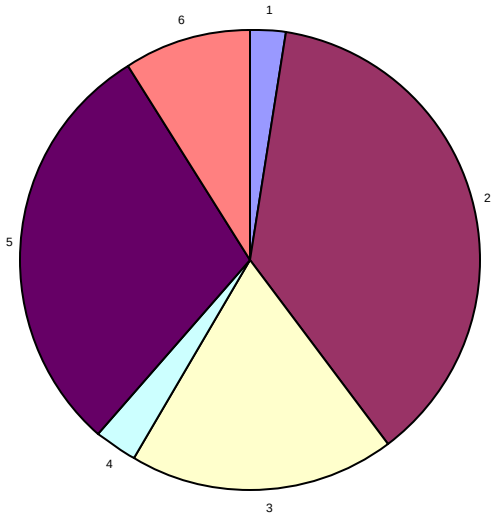
The differences in cost per record become substantial when they are applied to batches of 1000 records.

Approximately EUR 333 must be paid for the creation of a batch of 1000 records in an appropriate manner (the first bar in the graph). This calculation is based on a cost of EUR 0.33 for the creation of a well-constructed record.

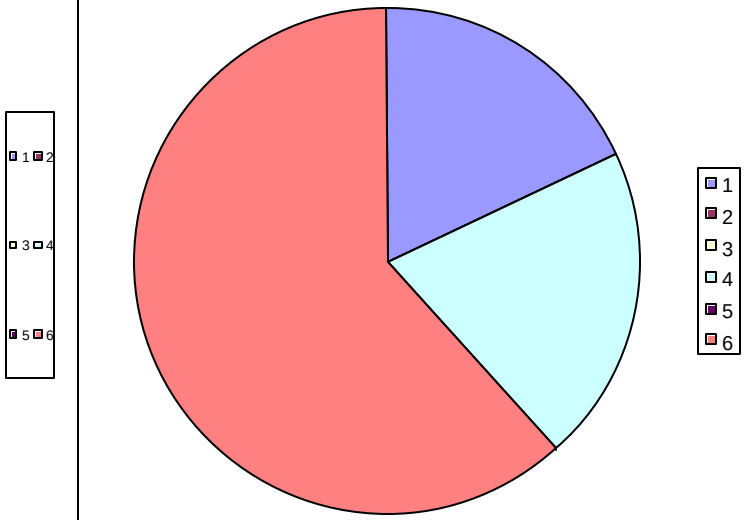
Conversely, it will cost EUR 10,000 (the third bar in the graph) to 'repair' a batch of 1000 badly created records.

Another good example of this effect is the cost difference of emails that have been preserved from a standard email application (such as Outlook), compared to e-mails preserved using a system focused on durability, such as Testbed's XML/e-mail application. The usual cost incurred in the acquisition and input of metadata amounts to EUR 1.41 per normal e-mail, whilst the cost is no more than EUR 0.06 per XML e-mail. The difference is that the XML emails are already equipped with the appropriate metadata and structure.

Graph 2: Cost of acquistion and preservation of existing e-mail messages



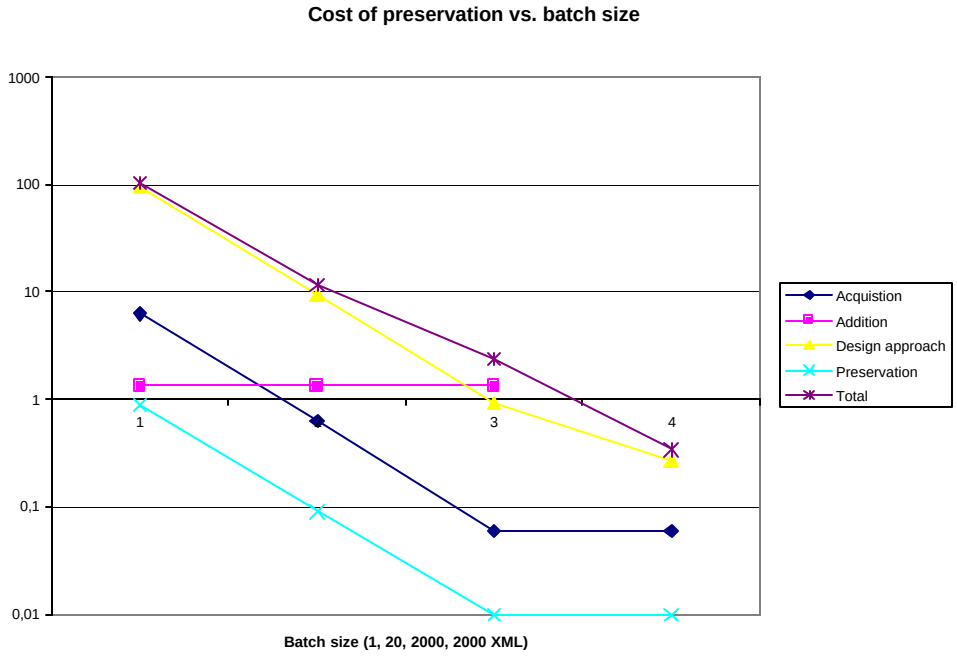
Graph 3: Cost of acquisition and preservation of existing XML e-mail messages



Legend:
1 = Acquire and appraise batch of e-mail,
2 = Necessary metadata,
3 = Repair digital records,
4 = Determine authenticity requirements for batch,
5 = Develop and test the preservation approach,
6 = Perform preservation and assess of e-mails.
The above legend relates to both graphs. In the right-hand graph items 2, 3 and 5 are all 0.

Accession to an Archive

The accession phase shows the effect of batch size on costs. This effect continues into the sustainable preservation phase. At every stage, larger batches of records cost less to manage. The reason for this is that metadata and processes can be added to 10,000 similar records as quickly as they can be added to one record. The time required to process a batch of records does not depend greatly on the size of the batch. For this reason the cost per record is much lower for large batches.
How can records be grouped into batches? A group of records from the same application (such as an e-mail or a word-processing program) that all contain explicit and correct metadata and which can all be processed in the same manner, can be processed together in a batch.



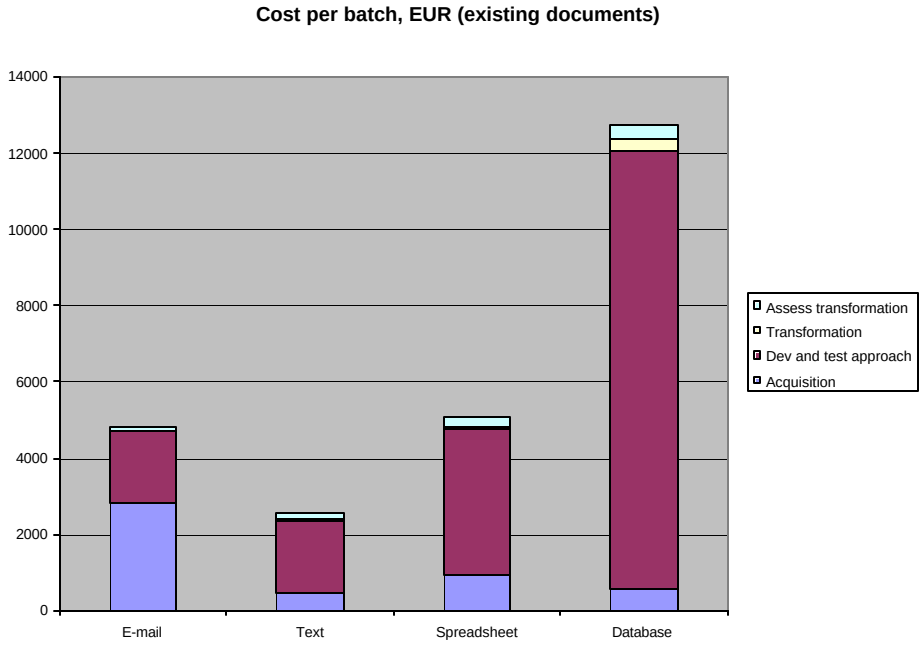
Graph 4: *Cost of preservation vs. batch size*

The graph reveals that the total cost for acquiring and preserving the e-mails decreases from EUR 102 per e-mail to EUR 2.35 per e-mail when the batch size is increased from 20 to 2000 e-mails. Acquiring and preserving e-mails already created in XML (for example, using the application developed by Testbed) costs even less (EUR 0.34 per e-mail for a batch of 2000 e-mails). The development of the approach is the most costly element of the process. If the approach can be used by several batches of records, through careful process design, then a significant saving can be made.

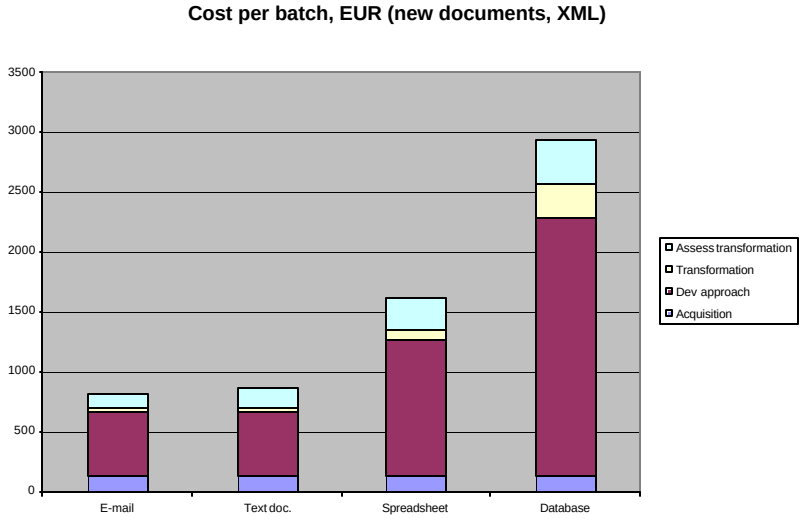
3. Transformation of different types of records
It is difficult to compare the costs of different preservation methods because they depend on so many different factors, the specific values of which will vary between different organisations. The migration costs have been calculated for each of the four record types investigated by Testbed. The costs for transformation do not differ substantially. The majority of the costs relate to the development and testing of the approach, and testing the transformed records. These costs remain constant for the migration from one application (version) to another, or for transformation to a standard format.

As discussed above, the size of the batch is one of the most significant variables for the transformation. For the purposes of this comparison it is assumed that e-mails are supplied in batches of 2000, text documents in batches of 200 and spreadsheets in batches of 20, and that databases are received and managed on an individual basis.

Testbed's experience has also revealed that the amount of work involved in adding metadata to each type of record, and in developing and testing the requisite transformation, varies with the type of record. The following graphs show the costs for existing and new records (such as records created in XML, whereby the correct metadata can be added at the point of creation.).



Graph 5: *Cost per batch, EUR (existing documents)*



Graph 6: *Cost per batch, EUR (new documents)*

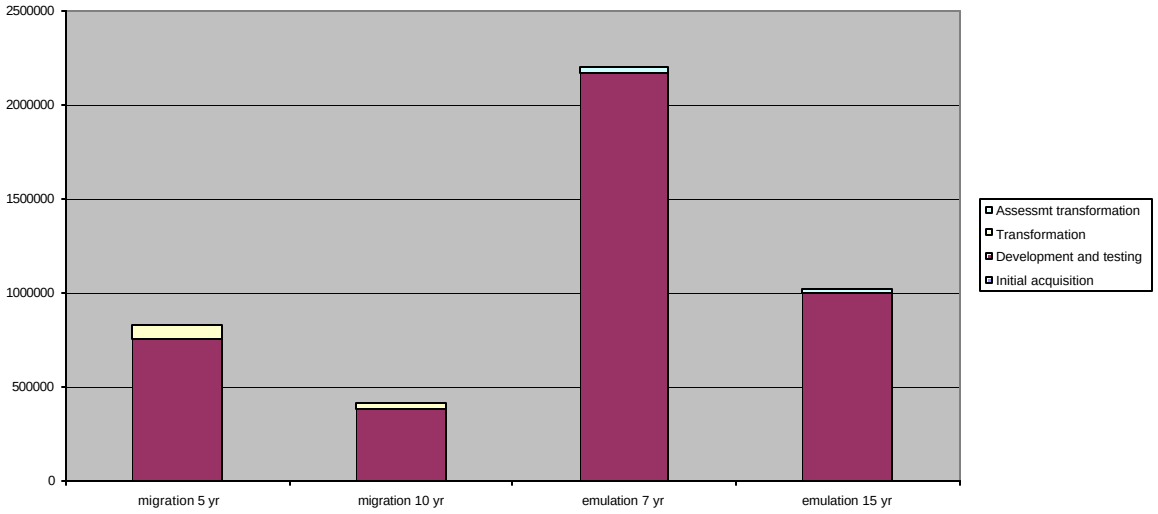
The total acquisition and preservation costs for batches of well-formed records are much lower than the costs for batches of badly created records. Acquisition (which here encompasses the repair of records and the metadata) is a significant cost item for badly formed records. Substantial costs will also be incurred when developing the approach. The development costs per batch can be substantially reduced if the approach can be shared by several batches.

4. Emulation (inclusive of the UVC approach)
Hardware emulation, the only form of emulation included in this review of the costs, offers two benefits when compared to migration as a preservation strategy. The first advantage is that the records themselves do not need to be modified. The second is that one emulator can be used for a variety of record types. Consequently one Pentium-PC emulator would cover text documents (whether prepared in Word, Open Office or Lotus SmartSuite), spreadsheets, and desktop databases.

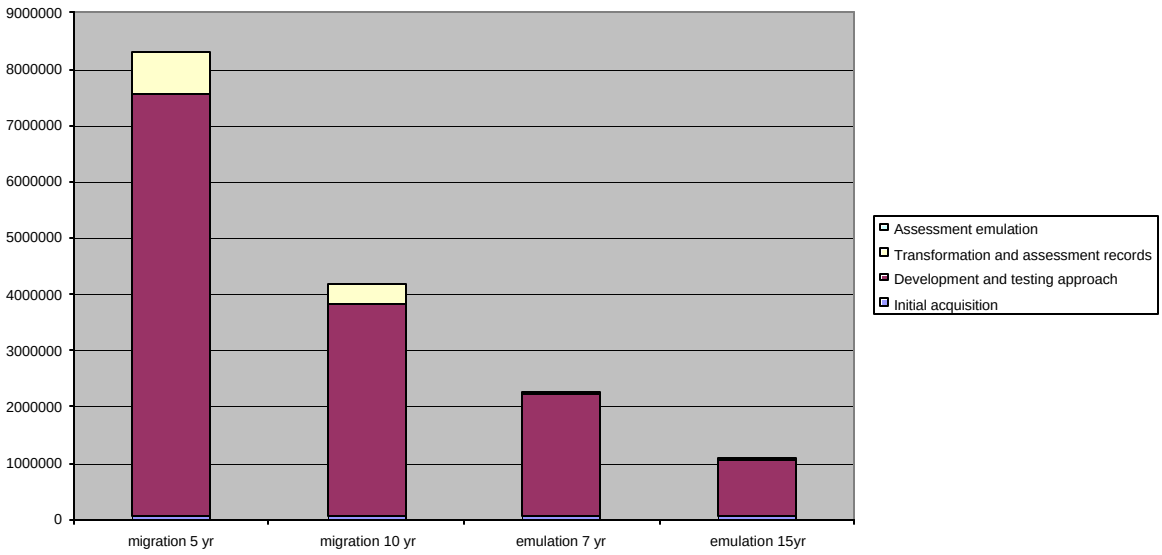
The following graph compares the costs for the maintenance of a collection of records, including e-mails, text documents, spreadsheets and databases, over a 100-year period. This collection resembles a shopping basket: arbitrary, but nevertheless illustrative. It is assumed that:

- the records need to be migrated every 5 or 10 years.
- new emulators are needed every 7 or 15 years.
- once the authenticity requirements for a batch of records have been identified, they do not have to be re-assessed.
- once added, metadata do not need further modification (other than the automatic addition of new preservation metadata about each preservation operation).
- this cost calculation uses a 'basket' of records containing one batch of each type of records, together with a mixture of existing and newly-created (durably-created) records.
- it is estimated that the development of an emulator will require 1.75 years' of work by a team of three persons.

Cost of digital preservation for a 100-year period, EUR



100-year digital preservation: 10x batches



In this second graph it is also assumed that ten batches of each type of records must be preserved and that each type can make use of the same emulator, but that different migration operations are required for each batch, for example as a result of different authenticity requirements.

The advantage of emulation in this example is that only one emulator need be developed for all records created by application software that used to run on the old emulated platform. In contrast, records from different software applications will require different migration approaches, and differences in authenticity requirements between batches from the same application can require different approaches.

5. Cost information from other sources

Testbed has obtained information about the cost issue of digital preservation from a number of other sources. This information can be used for an independent check of the calculations and the cost model.

For the purposes of these cost comparisons it may prove useful to consider the following equivalences:

- 1 metre of records (linear, on a shelf) is approximately equal to
- 0,09 cubic metre (based on A4 pages), which is about
- 3 cubic feet, with about
- 6,500 pages that, in digital form, would occupy about
- 65 Mb of storage space (as flat text).

Delft University of Technology, Utrecht University

The team in Delft and Utrecht has published a cost estimate for medium-term digital preservation that is one of the most detailed and carefully-considered estimates published to date. Dekker, Durr *et. al.*[30] have analysed the costs for the maintenance of records over twenty years.

National Archives of Canada

The National Archives of Canada employs about 661 FTEs and has an annual expenditure of about CAD 49,000,000. This excludes certain services supplied by other agencies, such as accommodation. The total costs of the National Archives amount to about CAD 90,000,000 per annum.

Of the main expenditure, EUR 30,460,000 (at CAD 1 = EUR 0.6217), about 30%, was destined for the acquisition and management of records, 18% for the management of government information, and 26% for services, awareness, and assistance.

The Annual Report of the National Archives of Canada (see www.archives.ca) notes that during his 15-year period of office, the then Prime Minister Mr Pierre Trudeau created some 1.3 Mb of electronic information. The offices of the current Prime Minister, Mr Jean Chretien, create about 1.3 Mb of digital records every day.

The current records stored by the National Archives comprise 111 km of paper government text records, 45 km of paper text documents from private individuals, 3.2 Gb of digital records, 2,500,000 maps and architectural drawings, 345,000 hours of audio, video or film, and 22,734,000 other items.

The cost of managing government records amounts to about CAD 80 per meter shelf, equivalent to about EUR 50 per shelf meter.

[30] An electronic archive for academic communities, November 2001.

The UK National Archives (formerly the Public Record Office)

The UK National Archives (formerly the Public Record Office looks after about 176 km of records at a cost of about GBP 97 per metre (EUR 135) for the selection and preservation of the records. Providing access to the records results in additional costs for each visit, namely GBP 5.59 (EUR 7.80) for on-site access and GBP 0.13 (EUR 0.18) for online access.

The UK National Archives has a workforce of 451 FTEs and an annual budget of about GBP 35,600,000 (= EUR 49,662,000). The annual reports are published on www.nationalarchives.gov.uk and www.pro.gov.uk.

National Archives of Australia

The National Archives of Australia (www.naa.gov.au) has a workforce of about 435, of whom 322 work full-time. The annual budget is about AUD 148,000,000 (EUR 83,428,000 at an exchange rate of AUD 1 = EUR 0.5637) of which AUD 38,274,000 (EUR 21,575,000) is destined for personnel and operational costs. (The remaining costs relate to depreciation, which in Australia also includes changes in the value of the collection itself.)

US National Archives and Records Administration (NARA)

The NARA Performance Report 2002 [31] indicates that the budget for 'space and preservation' amounted to USD 128,000,000 (EUR 109,000,000 at USD 1 = EUR 0.8524), with 338 FTEs. This also includes digital preservation, which is not stated separately in the accounts. NARA reported that in the preceding year the volume of 'logical records' (digital records) increased by 60% and that the volume of digital records during the presidency of President Clinton increased by 1500% (a factor of 15) in comparison with the preceding presidency.

Of these significant electronic holdings, 98% records are accessible, irrespective of their original format. NARA now manages about 4 x 10^9 'logical records'.

It should be noted that the total NARA budget amounted to USD 286,000,000 (EUR 244,000,000), and that in 2002 it had a workforce of 2829 FTEs.

NARA's digital preservation programme has 48.5 FTEs. Of these, 42.5% are active in ingest, i.e. maintaining relations with records producers, receiving records, generating Archival Information Packages (AIPs). A further 20 work on digital archival storage: receiving the AIPs, managing the storage hierarchy, monitoring quality, and generating access copies. 8% are engaged in data management, 7.5% in accounting, and 14% in user services and access. Preservation planning takes up 8% of staff time.

Amsterdam Municipal Archives

In 2001 the Amsterdam Municipal Archives acquired 350 metres (and 16,435 objects) of new archives. The digital archives were also expanded by 185 Gb. The Archives have a workforce of about 160. The reports are available from http://gemeentearchief.amsterdam.nl .

[31] http://www.archives.gov/about_us/strategic_planning_and_reporting/2002_performance_report.html#goal3

Appendix D Functional Requirements for a Preservation System

1 Introduction

1.1 Background and scope

The Digital Preservation Testbed has investigated potential long term preservation strategies for different types of digital records, namely e-mail messages, text documents, spreadsheets and databases. For each of these record types a recommendation has been produced on the most appropriate preservation strategy. These recommendations address the question of preservation mostly at the level of individual records.

The present document discusses the functional requirements for a preservation system to implement the recommendations. Functional requirements relate to the things that the system should do, as opposed to non-functional requirements, such as the necessary capacity of the system, development methods or performance characteristics.

In this document the main features and functions of a preservation system and the choices that must be considered when designing such a system are discussed. Preparing a list of individual, precise, traceable requirements that could be used for building or procuring such a system is beyond the scope of the project.

This document concerns a system for storage and preservation of digital archival records. Much of the document is also relevant to other kinds of preservation system, for example systems for preserving scientific or historical data.

The activities required at each stage of the life of a digital record are considered, from its creation, through its maintenance by the creating department, transfer to a long term archive and long term maintenance of the record in the custody of the long term archive. All stages of this process are essential to the long term preservation of records. Example activities associated with a record are illustrated below in figure 9.

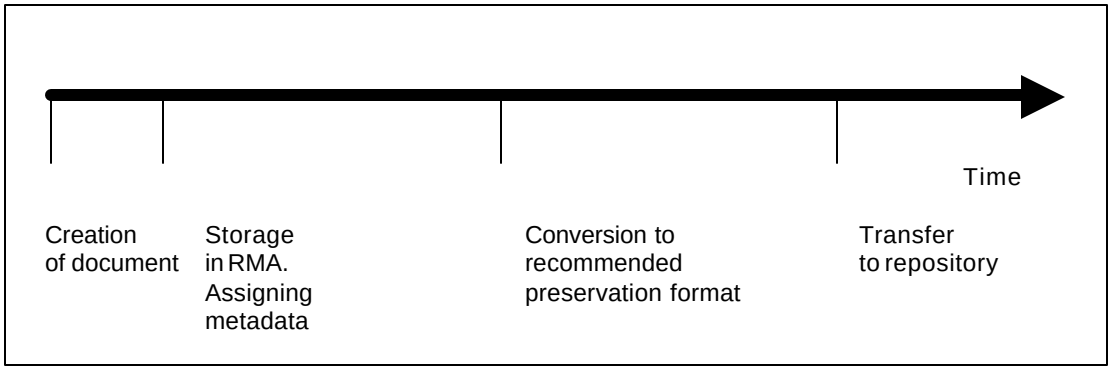


Figure 9 The activity timeline of an example record.

1.2 Context

This document integrates individual requirements associated with a range of long term preservation strategies into a single set of functional requirements. An essential part of a preservation system is the digital archive (repository) where digital records are held and managed. The DEPOT 2000 project produced a "Functional Design for a digital depot"[32], setting out in detail the functions and data structure for such a repository system. The DEPOT 2000 document deliberately sets the choice and implementation of a long term preservation strategy to one side. In section 5, we discuss how the present preservation system requirements relate to the requirements for a digital archive or digital depot. This appendix does not explicitly consider the costs associated with digital preservation. That is addressed in appendix C of this publication[33].

These functional requirements are closely related to the recommendations of the Testbed project on the best approach for the long term preservation of various types of records. Recommendations have been produced for the preservation of e-mail messages, text documents, spreadsheets and databases. The recommendations themselves concentrate on the actions required to preserve each of the four types of records. This document considers the computing environment and systems required to enable these recommendations to be implemented effectively and efficiently.

The Reference Model for an 'Open Archival Information System' (OAIS Model) is generally regarded as the standard work for defining a framework and procedures for the preservation of digital records. This document focuses mainly on the "Preservation Planning" element of the OAIS model, but we also consider other aspects, notably how to deal with the requirement to provide and maintain sufficient representation information. The relationship between the Testbed recommendations and the OAIS model is discussed further in section 5 of this appendix.

Whilst still in the custody of the organisation that created them, digital records are likely to be held in a Records Management Application (RMA). In section 3, we refer to published requirements for such systems and comment on how the need for long term preservation affects these requirements.

1.3 Definitions

The glossary holds a comprehensive list of definitions used in this document. Two further terms are of particular importance and so are defined here, to set the scene for the rest of this discussion.

Digital Archival Record
A digital entity, preserved in the form of a file assembly, that an archival institution receives and preserves. The archival institution preserves the archival records using a strategy dictated by the record type. For simplicity this archived file assembly is referred to as the record in this document, though the files are intrinsically dependent upon suitable applications to read and represent them authentically[34].

Digital Archive System (including a functionality for long term preservation)
A system designed and used to receive and store authentically preserved digital records. The meaning of "long term" refers in this document to a period greater than 10 years.

[32] DEPOT 2000. Functional Design for a Digital Depot. Nico van Egmond, Hans Hofman, Jacqueline Slats, Tamara van Zwol.
[33] See Appendix C Cost indicators and cost model.
[34] See Chapter 2, paragraph 2: "The Digital Record as…".

2 Records continuum

'Record' in this context means the digital document as described above. The concept of the records continuum is very useful for this discussion. It can be defined as:

"...a consistent and coherent regime of management processes from the time of the creation of records (and before creation, in the design of record keeping systems), through to the preservation and use of records as archives."[35]

In this discussion of the preservation of digital records, there are three important phases associated with the continuum, see figure 10:

1. Before transfer to an archival institution
2. The transfer process
3. Long term preservation in the digital archive system

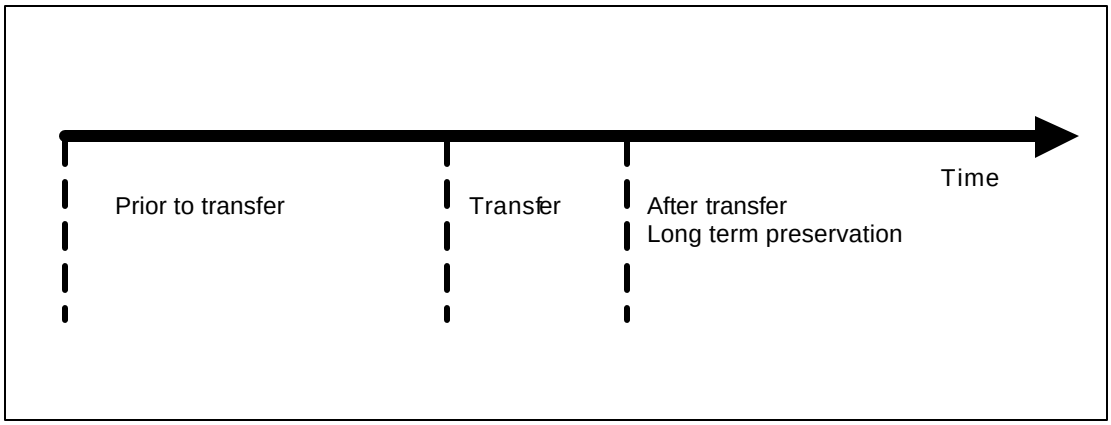


Figure 10 *The 3 phases of the records continuum.*

These three phases will be described in the following three chapters. Each phase of the records continuum is equally important: if a break in the "consistent and coherent regime" occurs before the records are in the care of the long term archive system, then their preservation is put at risk.

Many of the requirements applying to the third phase are also relevant for the first phase. This is because in many cases the digital records will be in the custody of the creating organisation for a period long enough for the need and application of preservation actions to prevent digital obsolescence.

3 Before transfer to an archival institution (phase 1)

It is a requirement that records should be carefully managed according to well-defined procedures. In most cases, a RMA will be a useful tool in achieving this aim. One likely exception to this is in the case of databases. Although the RMA may be a good place to store documentation and metadata about databases that have been identified for preservation, it will usually make sense to store and maintain the database itself outside of the RMA.

[35] Op cit. Australian Standard AS 4390-1996.

3.1 Selecting an appropriate RMA

A number of organisations have published requirements for Records Management Applications:

- European Commission: "Model Requirements for the Management of Electronic Records" (MoReq)[36]
- Ministry of the Interior/Archival School: "Softwarespecificaties voor Records Management Applicaties voor de Nederlandse overheid" (ReMANO)[37]
- UK National Archives[38]
- US Department of Defense[39]

These documents give details of the features which RMA software should have. The UK National Archives and the US Department of Defense each have a programme to evaluate software packages against their requirements specifications. The outcome of this work is lists of compliant software systems[40] [41]. A less formal, but nonetheless useful list of RMA and DMS products and their features is maintained by the Dutch government project Kenniscentrum Elektronische Overheid[42].

3.2 Configuring an RMA

Most off-the-shelf RMA software products are designed to be flexible and highly configurable. Indeed, because every organisation is different, such systems need significant configuration if they are to be able to implement the records policies of each organisation.

The features of an RMA which must be configured to suit an organisation include:

- the records classification scheme based on tasks of business processes (see NEN-ISO 15489-1)
- the set of metadata items which are associated with each record
- retention schedules and disposal actions

In addition, the Testbed recommendations for the preservation of records state further specific requirements for the organisation in relation to the records and associated metadata: in some cases there is a need to transform files from one format to another. Some of these could be implemented within the framework of an RMA, whereas for others, it makes more sense to use external tools. These activities are discussed in more detail in the next section.

---

36 http://www.cornwell.co.uk/moreq

37 Softwarespecificaties voor Records Management Applicaties voor de Nederlandse overheid (Remano). Hans Waalwijk, Geert-Jan van Bussel, Peter Horsman, Archiefschool
Versie 4.12 9 september 2002
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/remano_versie4_12bis.doc. This is based on MoReq, but has been adapted for the specific requirements of the Dutch government.

38 http://www.pro.gov.uk/recordsmanagement/erecords/2002reqs/default.htm

39 http://jitc.fhu.disa.mil/recmgt/standards.htm

40 UK PRO list of approved systems:
http://www.pro.gov.uk/recordsmanagement/erecords/2002reqs/2002listofapprovedsystems.htm,
http://www.pro.gov.uk/recordsmanagement/erecords/1999reqs/1999listofapprovedsystems.htm

41 US Department of Defense list of approved systems: http://jitc.fhu.disa.mil/recmgt/

42 http://www.digitalegereedschapskist.nl/systemen

3.3 Capturing records and preparing them for preservation

3.3.1 Use of a Records Management Application

An important requirement for the reliable preservation of digital records is that, as soon as possible after their creation and their identification as records that should be retained, they should be placed into a well-organised and controlled storage environment, i.e. a records management application. MoReq refers to this as the "capture" of the record in the RMA.

At this point the record should be classified: that is, there must be a defined classification system for the records held by the organisation. On capturing the record in the RMA, the relevant business process, task or dossier where the record belongs should be identified. In some cases, the capture and classification can take place automatically as part of a workflow. If the main workflows of the organisation can be integrated with the RMA, then much of the work of records capture can be automated, reducing the burden on the users and reducing the risk of errors or omissions.

3.3.2 Capturing metadata

On capture of the record in the RMA, the necessary metadata should be added to the record. It is beyond the scope of this document to specify which items of metadata are required. This may depend on the particular requirements of each organisation. However, the previously referenced MoReq and US Department of Defense documents, for example, provide good starting points. As far as possible the RMA should be configured to collect metadata automatically. Items such as the name of the user submitting the record and the date and time can easily be captured by software. If the record is created as part of an automated workflow, then metadata identifying the business process and the place of this record within that process can also be captured by the RMA. Note that different types of record may have different metadata requirements.

An essential requirement for any records management or preservation system is that the association between a record and its metadata must be maintained with 100% reliability.

In addition to the metadata required for record-keeping purposes, there may also be additional items required for technical or preservation reasons, for example relating to the format of the computer file(s) making up the record or describing any transformations that have been carried out on the record. The Testbed recommendations for each record type discuss metadata requirements for preservation in more detail.

3.3.3 Conversion of the file format

It is the responsibility of the creating organisation to conduct file format conversions of records to formats specified in the archival regulation on the Arrangement and Accessibility of Records [43].

Some file formats pose a higher risk for long term preservation than others, so for many records it may be necessary to transform the original computer files into new formats, which have been chosen for their suitability for long term preservation. There are several examples of this in the Testbed recommendations. In general, it is best to carry out this transformation as soon as possible after the record has been created and identified for long term preservation. This will cause additional storage requirements, but we contend that these are outweighed by the increased reliability and trustworthiness of carrying out the transformations whilst the original processing software is still operational and knowledge of the original purpose and context of the record is still available in the creating organisation.

[43] See Article 6 of the Regulation.

Organisations are not obliged to apply conversion to file formats recommended for long term preservation for records with a short retention schedule. Note that, even for records with a short retention schedule, organisations are obliged to keep these records in a good and well-ordered state. There may sometimes be a need to perform preservation actions such as migration to maintain access to the record.

Conversion to the recommended preservation format could be incorporated as a function of an RMA. When a record is first entered in the RMA and associated with a retention schedule involving long term preservation, then automatic conversion tools could be invoked by the RMA to convert the file to the required new format. This process should also involve evaluation of the success of such transformations. These should be automated where possible, but in some cases may also involve an element of manual checking. If a record is initially assigned a short retention schedule, but later it is decided that the record should be transferred to an archival institution, then the transformation to the chosen preservation format should be associated with the act of modifying the retention schedule.
More information about the design and testing of preservation approaches and automated testing procedures is given in section 5.9.

3.4 Transfer from one RMA to another

The current archival regulations in the Netherlands state that records that are not to be destroyed must in principle be transferred to an archival institution within 20 years of their creation. The caretaker can transfer archival records that are not to be destroyed to an archival institution within those 20 years, when, in the opinion of the head of the archival institution, there is immediate cause to do so[44]. It is unlikely that the organisation responsible for the records will use the same RMA throughout this period. Like other areas of information technology, new products and approaches appear frequently and at some point there may be benefits for an organisation to change from one RMA to another.

When this occurs, the records contained in the RMA must be transferred reliably to the new system. In some cases, the system manufacturers may include specific facilities for transferring between specific combinations of systems, but it is not generally safe to rely on such features being available. To ensure that this process of migration between RMAs can be successfully achieved, it should be a requirement that the RMA can export and import records in a vendor neutral interchange format, which maintains the logical structure linking the components of a record and also maintains the links between records (for example grouping records into dossiers), or links between a record with a particular file format and another object containing representation information for that format.

When selecting or building an RMA in the present day, it is not possible to know if the export format of the RMA will match the supported import formats of future RMAs. However, if the system uses a well-documented non-proprietary export format, then it should be possible without excessive effort to provide a matching import facility in the future. XML is likely to be a suitable basis for such an interchange format. MoReq includes a useful section on “Transfer, export and destruction” of records, requiring that an RMA “must provide a well-managed process to transfer records to another system or to a third-party organisation”.

From a technical point of view, the process of transferring records from an RMA to the National Archives has much in common with the transfer between one RMA and another. However, from a regulatory and record-keeping point of view there are additional aspects to consider in the former case and these are discussed in the next section.

---

[44] See the 1995 Archives Act, article 13, sub 1.

4 Accession (phase 2)

4.1 Overview

The accession stage of the records continuum covers the transfer of the records from the creating organisation to an archival institution. There are a number of conditions on how the transfer of records to an archival institution should be organised, including for example the file formats and types of metadata required.

The transfer itself could be carried out over a communications network or on removable media. A group of related records transferred to an archival institution at one time can be defined as an *accession*. The accession will typically involve a large number of records, each with associated metadata. The records will normally be organised into dossiers or other groups and may have other links between them. These are the same issues as discussed in section 3.4. The transfer format must be able to represent the basic information of the records and all the important links between the components.

4.2 Processing the transferred records

Each record must be assigned a unique and final reference on being added to the archive and this becomes the definitive reference. The records may already have a unique reference assigned during the pre-accession stage. This reference should also be retained.

To minimise the human effort required at this stage, the archival institutions - working in consultation with the record-creating organisations – should develop and publish one or more records transfer formats. The transfer format is a specification to allow the grouping of files and metadata together as a package. This can then be implemented as an export format in the RMAs of the record creating organisations, allowing transfer of records to archival institutions to take place as simply as possible.

As the records are imported to the long term archive, any related cataloguing systems should be updated as required, based on the record metadata.

At this stage it may be possible to extract further automatic metadata about the files and records being submitted. This may not be necessary if the creating organisation has already ensured that all required metadata are already available. The digital archive system must verify that all necessary metadata have been provided, according to whatever metadata schema has been agreed between the archival institution and the creating organisation. The contents of each computer file must also be checked for viruses and the result of this check stored in the record metadata.

5 Preservation in digital archive system (phase 3)

5.1 Introduction and background

This chapter discusses the functions and features (properties) that a digital archive system should have to support the preservation process.

The Consultative Committee for Space Data Systems (CCSDS) has published a recommendation for a "Reference Model for an Open Archival Information System (OAIS)".
The OAIS model defines an information model for a long term preservation system and lists a set of responsibilities that the system should fulfill. It has now been adopted as ISO standard 14721:2002. It is therefore a useful basis for our discussion and a few important points from the OAIS model are summarised briefly here. For a full explanation of OAIS, the reader is referred to the OAIS reference model book[45].

5.1.1 The OAIS model

Users

The OAIS Model defines three main types of external users of a digital archive system: producers, consumers and management. The producers are those creating the information to be preserved, the consumers are those making use of the preserved information and the management is responsible for high-level policy making for the digital archive (day-to-day administration of the system is defined as one of the functions of the OAIS).

Data Structure

The OAIS defines an Information Object as a Data Object combined with Representation Information. The Data Object can be thought of as the computer file or files making up the object. Because computer files or bitstreams in isolation cannot be meaningfully interpreted, there must be additional representation information, in the form of documentation or computer software.

There may be several layers of representation information and some items of representation information may be shared by more than one Data Object. For example, one element of representation information is likely to be a specification of the mapping from bits to characters in a particular data object (for example following the Unicode standard). Since a large number of records will have this information in common, it does not make sense to store it separately for every record.

The OAIS Model points out that the amount and type of representation information required depends on the intended users of the Information Object, known in the OAIS Model as the "Designated Community". A Designated Community has an associated Knowledge Base, that is a set of knowledge which a member of that community can be assumed to have. The Representation Information must be sufficient for a member of the Designated Community to be able to understand the Information Object.

[45] OAIS. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

Functions

Figure 11 from the OAIS recommendations, discussed below, illustrates the main groups of functions of an archival information system and we reproduce it here for convenience.

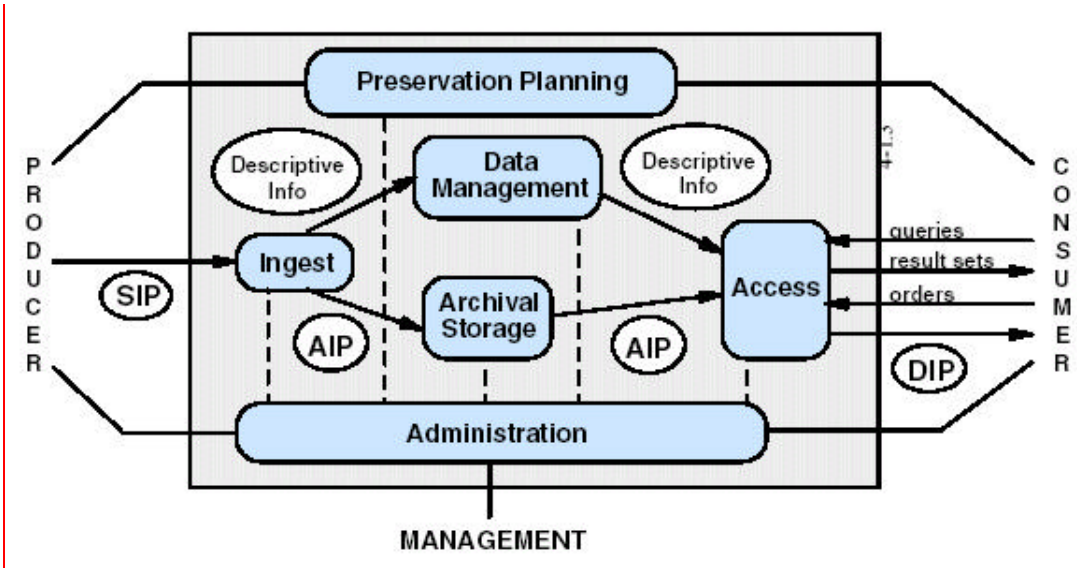


Figure 11: *OAIS Functional Entities, reproduced from the CCSDS Recommendation*

From this it can be seen that the main groups of functions that an archival system for digital records must implement are:

- Ingest
- Data Management
- Archival Storage
- Access
- Administration
- Preservation Planning

Of these function groups, most of the functions are associated with the digital archive system. As explained above, only those aspects with a particular relevance to long term preservation are discussed in this publication. In the following sections, reference will be made to the related OAIS function to show how this publication fits into the OAIS framework.

The abbreviations SIP, AIP and DIP appearing in the diagram above stand for Submission Information Package, Archival Information Package and Dissemination Information Package, and refer to: the way the information is organised during ingest to the archive (SIP); whilst stored in the digital depot/repository (AIP), and; when being distributed to users (DIP). Refer to the OAIS Model for more detailed information.

Thus the OAIS Model provides a framework for long term preservation issues and discusses some of the possible approaches. In the rest of this chapter, we will elaborate on the (functional) requirements for a system to provide long term preservation.

5.1.2 Existing approaches

There are few examples of operational digital archiving systems and even fewer with well-developed preservation strategies. However, a number of organisations have carried out pilot studies, made initial implementations of preservation approaches, or have published information on their preferred future strategies. A selection of these are briefly reviewed in this section.

UK National Archives

The UK National Archives have an operational Digital Archive (since April 2003), storing UK government records and making them available to the public[46]. The focus of this system at present is on a secure repository for digital records (metadata and content files). They have not yet fixed on a particular long term preservation strategy but have ensured that the system is designed to allow future strategies to be effectively implemented.

The policy of the UK National Archives is to accept files in any format. In addition to the core digital archive system itself, they have a database of file format information, known as PRONOM[47]. PRONOM is a system for "managing information about the file formats used to store electronic records, and the software applications needed to render these formats" and forms a component of their technology watch programme.

The digital archive at the UK National Archives includes the concept of multiple "manifestations" of a record, that is different digital representations of the same record, providing a framework to enable possible future migrations.

Public Record Office Victoria (PROV)

The Victorian Electronic Records Strategy (VERS) programme of PROV[48] has developed a strategy for long term preservation based around a choice of a limited number of preservation formats. In contrast to the UK National Archives, they insist that electronic records are submitted to them in one of a small defined set of allowable file types. For "printable" record types, they require that the record-creating organisation submits files to PROV in PDF format. By limiting the file formats in the repository to a small carefully chosen list, PROV aim to simplify future preservation efforts.

Another key element in the VERS approach is to store records as self-contained digital objects. They have defined an XML format which combines their metadata schema with the files themselves. The file (or files) containing the record content are base64 encoded (to convert them from binary format to text format) and are stored as elements in the XML document. This has the benefit of minimising the amount of supporting IT infrastructure needed to allow the record to be correctly reconstructed and interpreted, as no external system is needed to maintain links between record metadata and content files. However, such an external system is still required to support efficient and controlled access to the records.

PROV plan to have an operational digital archive around the end of 2004.

[46] http://www.pro.gov.uk/about/preservation/digital/archive/default.htm
[47] http://www.pro.gov.uk/about/preservation/digital/pronom/default.htm
[48] http://www.prov.vic.gov.au/vers/welcome.htm

NARA

At the time of writing, NARA are involved in the procurement process for their Electronic Records Archives (ERA) project[49]. An important element of their proposed approach for ensuring long term accessibility of digital records is the conversion of files to *persistent formats*. The ERA requirements document explains this as follows:

"A persistent format is one that is supported by a preservation strategy for diminishing the impacts of technological obsolescence, minimising dependence on specific hardware and software and enabling retrieval and output of authentic copies in the future. An ideal persistent format would be self-describing and be able to be validated in accordance with open, non-proprietary standards".

5.1.3 Repository functions versus preservation functions

As discussed above, this document concentrates on the long term preservation functions of a digital archive system. The essential functions of a digital archive system are discussed in detail in other documents, for example the DEPOT 2000 publication of the Rijksarchiefdienst and they are summarised very briefly below.

The basic function of a digital archive or repository is to manage records and their metadata, including relationships between records. It must do this extremely reliably: the system must include comprehensive back-up and recovery systems so that records cannot be lost, even in the case of a disaster.

There must also be systematic monitoring and replacement of storage media, to ensure that data on disks and tapes is not corrupted. Access to the system must be controlled, to regulate what activities can be carried out by which users or types of users.

The digital archive must deal with accession of records and access to records. These aspects are influenced by long term preservation issues and so are discussed in more detail elsewhere in this document: see Chapter 4 and Section 5.11.

5.2 Manifestations

For a migration-based strategy, it is necessary for the preservation system to include the concept of multiple representations or manifestations of a record. As Thibodeau explains[50] it is possible to represent digital records in different ways without losing authenticity: as long as the conceptual record, as presented to the user, is sufficiently similar then it is possible and acceptable to change the file format and rendering software. He gives the example of a text document that can be represented as either an MS Word file or a PDF file.

The OAIS Model refers to various types of migration. Migration including a format conversion is one of these migration types. In the context of OAIS, this type is referred to as a 'transformation'. Transformation of a record (or AIP) leads to a new AIP which is a new version of the previous AIP. The previous version of the AIP will usually be retained, at least temporarily, so that the transformation process can be checked, or as a source format for possible further transformations.

[49] http://www.archives.gov/electronic_records_archives/acquisition/draft_rfp.html

[50] Ken Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years" http://www.clir.org/pubs/reports/pub107/pub107.pdf.

Each manifestation should include or link to all information needed for the record (e.g. shared representation information would typically be linked rather than copied), so that it should not be necessary to access more than one manifestation in order to render the record authentically. A manifestation may involve multiple computer files.

Users of the system would not necessarily have access to all of the retained manifestations. There could be one preferred manifestation for user access, or possibly more than one alternative. This mechanism could also be used for redaction, where original and redacted copies of a record would be different manifestations.

Each time a new manifestation is added, information about the process used to create it must be recorded in an audit trail.

5.3 'Technology watch' and preservation planning

The OAIS Model identifies Preservation Planning as a group of functions that a digital archive system must support. This group is broken down into:

- Monitor Technology
- Monitor Designated Community
- Develop Preservation Strategies and Standards
- Develop Packaging Designs and Migration Plans

The Monitor Technology function is often referred to as a Technology Watch. This is one of the most important functions within preservation planning and the one that we discuss in most detail here. This involves monitoring the various technological components used in the digital archive, to identify as early as possible if a component may be in danger of becoming obsolete. Early identification of potential problems allows action to be taken before access to saved digital records is made difficult or impossible.

One approach to this is to monitor the file formats held in the digital archive system and to maintain information on the status of each file format, for example which application software is able to understand and render the file format, on which hardware platforms the application software is available, whether the application software is still supported by its manufacturer and so on. The UK National Archives have produced a software system called PRONOM intended to help tackle this problem[51].

In the DNEP[52] system of the Koninklijke Bibliotheek, this concept has been formalised as the Preservation Layer Model[53]. Each item in the deposit system will be associated with one or more View Paths, defining the application software, operating system and hardware required to access the item. Note that one record in an archiving system may consist of multiple computer files in multiple formats. Files in different formats will in general have different view paths, so there may be more than one view path per record. This implies that the technology watch must take place at the level of the computer file, not just at the level of the archival records.

When it is identified that a record in the system is in danger of becoming inaccessible, then some kind of preservation action must take place, for example migrating it to a new format, or creating an emulation of a hardware environment where the original software application can continue to run. The steps that will be involved in this are discussed in section 5.10.

---

51 http://www.pro.gov.uk/about/preservation/digital/pronom.htm

52 http://www.kb.nl/kb/resources/frameset_kb.html?/kb/ict/dea/index-en.html

53 "The Long term Preservation Study of the DNEP project – an overview of the results", Raymond J. van Diessen, Johan F. Steenbakkers, IBM Netherlands, Amsterdam, ISBN 90-6259-154-X

5.4 Requirements for e-mail preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving e-mail messages [54]. The recommendations advocate converting an e-mail message to an XML document, linked to separate files containing the message body, in plain text or HTML format or both, and any message attachments. Because e-mail attachments can be any type of file, the e-mail recommendations do not make specific recommendations on how to deal with attachments, beyond noting that an appropriate preservation strategy for that file type should be applied. Using the View Path concept discussed in Section 5.3, each computer file in the record will need its own view path. For example, the XML may be viewed using simple text viewing software.

Processing the e-mail message to transform it to the required preservation format (i.e. this set of linked files plus metadata) should be carried out as soon as possible after the message has been identified as requiring preservation. This could be incorporated as a modification to the e-mail application being used, or it could be implemented as a feature of a Records Management Application.

The Testbed recommendation means that the e-mail record is represented by several files, linked in a particular way. The preservation system must have the ability to make and maintain these links between the files within a record. In the e-mail recommendations it is suggested that this can be done by having an XML file at the core of a record, defining the relationships between the component files and hence the structure of the record. This is one, quite flexible, approach to achieving this objective. Other solutions are also possible, for example using a relational database.

If one or more of the files within a record must be updated as part of its preservation strategy, then this should require creation of a new manifestation of the whole record. This is a general feature for all the record types described in the next few sections.

5.5 Requirements for text document preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving text documents [55]. The Testbed recommendation is to keep the original file, together with a PDF and optionally also a XML representation.

Due to the use of multiple file formats there is a need for multiple representations of same record. The XML approach consists of several associated files including a XML content file, a XSLT (XML Stylesheet Language Transformations) stylesheet and related image files. The XSLT should transform the XML into an XSL-FO (XSL Formatting Objects) file. A suitable XSL-FO processing application is required to render the resulting XSL-FO file. The PDF approach consists of a single binary file that can be viewed using PDF viewing software.

Processing the text document to transform it to the required preservation format (i.e. this set of linked files plus metadata) should be carried out as soon as possible after the text document has been identified as requiring preservation. This could be incorporated at document creation, post creation using a separate application, or it could be implemented as a feature of a Records Management Application.

[54] http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-email-en.pdf

[55] http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-textdocs-en.pdf

As with e-mail messages, there must be a mechanism for representing the relationships between the different files making up the record, which could be the appropriate use of XML or a database.

A technology watch is required to identify major changes in either the PDF format or relevant XML implementations. Such changes will require a re-evaluation of the present strategy.

5.6 Requirements for spreadsheets preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving spreadsheets[56]. The Testbed recommendations advocate keeping the original file as well as converting the spreadsheet to XML.
The XML approach consists of several associated files including a XML content file, optionally an XSLT stylesheet and related image files. The XML files can be viewed in text form using a simple text based application.

Processing the spreadsheet to transform it to the required preservation format (i.e. this set of linked files plus metadata) should be carried out as soon as possible after the spreadsheet has been identified as requiring preservation. This could be incorporated at creation, post creation using a separate application, or it could be implemented as a feature of a Records Management Application.

There must be a mechanism for representing the relationships between the different files making up the preservation object, which could be the appropriate use of XML such as a linking file, or a database.

5.7 Requirements for database preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving databases[57]. The Testbed recommendation is to keep the original database data file(s) or export file(s), an XML representation of the application database tables and associated information in the underlying database and documentation to describe key input and outputs associated with the application. This documentation will contain SQL and code associated with the transactions described.

The XML approach consists of creating several associated files, including a XML overview file and XML content files representing each of the application tables. The application documentation can be preserved and viewed according to the text document recommendations. The XML files can be viewed in text form using a simple text based application. There is even the possibility to conduct a second conversion to transform the XML content files of the database tables into an active database.

Processing the database to transform it to the required preservation format (i.e. this set of linked files plus metadata) should usually be performed when the database ceases to be active or when the database must be transferred for long term preservation. This could be done using a separate application, or it could be implemented as a feature of a Records Management Application.

[56] http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-spreadsheets-en.pdf

[57] http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf

There must be a mechanism for representing the relationships between the different files making up the record, which could be implemented using a range of approaches.
There must be a mechanism for representing the relationships between the different files making up the preservation object. There are several possibilities to implement this, for instance the 'framework approach' described in chapter 4.

5.8 Other record types

The Testbed project carried out research into the long term preservation of e-mail messages, text documents, spreadsheets, and databases. Testbed is not therefore in a position to make recommendations for other types of digital records. However, similar principles will apply. The types of application software available to view or interpret a file format should be considered. Preference should be given to file formats for which the specifications are clearly defined and publicly available (and thus not to 'closed' formats). If the chosen strategy relies on conversion to a format that is deemed more suitable for long term preservation, then the conversion procedure must be fully tested and evaluated to ensure that the essential characteristics of the original record can be retained.

5.9 Preservation actions

The most common type of 'preservation action' is the migration of the file or files that constitute a digital record, into another format. In this case it is advisable to also preserve the original files. If multiple conversions take place over a period of time, then a decision must be taken on whether the intermediate versions are also preserved. The preservation of all versions offers maximum protection against information loss, but with the result of greater storage capacity and more complex management requirements.

When such a file format conversion or other preservation actions are carried out, it is essential to maintain a complete record of the conversion procedure so that the authenticity of the record can be established. It must be possible to determine which file was the predecessor of the newly created version, and to understand the procedure that was applied. If digital signatures or hashing approaches are used to check against corruption of files or metadata, these must be re-applied to the new format once the necessary quality control procedures have been carried out.

If the format of the computer files comprising a record are changed, there is a risk that information is lost or changed, leading to a loss of authenticity. It is therefore important to implement a thorough checking procedure when such a migration takes place. Such checking can be implemented with a separate automatic 'testing module' that can be enhanced in the future to systematically reduce dependency on manual testing.

A separate testing system must be developed to conduct on-going research into preservation approaches, for example, for migrations to new formats or for the emulation of hardware environments in which the original operating and application software can continue to be run. The research must consider and examine relevant new technological developments. The research scope should also cover the development of new automated testing techniques. Promising preservation approaches must be fully tested to ensure that the integrity and authenticity is maintained and can be properly checked. The results of this research are updates to the preservation modules and automated testing modules in the preservation system, and guidelines that specify the most suitable long term preservation approach for a diverse range of digital records.

If a technology watch (as described in section 5.4) indicates that a group of records in the system is in danger of becoming inaccessible, then preservation action must take place. The entire group of records must first be classified by record type, then by subcategories on the basis of corresponding authenticity requirements.
For each records group, the most appropriate preservation approach must be determined one that ensures the formulated authenticity requirements are met. For migration, the

target format is ideally an open standard that is widely used and well documented. The results of research from the testbed research with the test system described above, and the recommendations for long term preservation, can be used to guide this decision.

The selected preservation strategies must be applied for each records group. Such preservation action will result in new manifestations of the records, except for when emulation is applied. The preservation action must be recorded in the digital record's metadata and preservation logbook.

5.10 Representation Information

There are often cases when digital records require the same information, such as an XML schema or reference to a standard, for example Unicode. In such cases the information does not have to be stored explicitly with each record, but rather an appropriate linking approach can be used so that the representation information only has to be stored once. Such links can be implemented in a number of different ways, depending upon the nature of the digital archive system, as long as consideration is given to the way in which links can be maintained should the records be transferred to a new and potentially different digital archive system in the future.

5.11 Security

Security is a very important aspect of a trustworthy digital archive. One of the difficulties of digital record keeping is the ease with which digital records can be changed or deleted. A digital archive must therefore be designed to minimise such risks. The standard approach for IT system security is to assign users particular roles. Different rights are assigned to different roles, which define what users may see, whether they may add new data, and if they may edit or delete existing data files.

The approach of the UK National Archives has been to implement two systems: a closed system that is the definitive version of the archive, and an open system that is accessible by the public and that contains copies of the records held in the closed system. This means that there is no risk of unauthorised access to the closed system, even if the access controls to the open system are compromised. The closed system is separated from the open system by an 'air-gap' and is not linked to any external networks.

In this security model it must be possible for certain documents to be defined as 'public' (accessible to the public) and others to be defined as 'non-public' (not accessible to the public).

To allow verification that the contents of record metadata have not been accidentally or deliberately changed, digital signatures or hash-functions can be used. The VERS approach advocates applying a digital signature to both the record content and the record metadata. If changes or additions to the metadata are made, then the modified version of the metadata must be digitally re-signed.

The system must maintain an audit trail that registers all access, mutations, or deletions, as well as the responsible user and the date and time of the changes.

5.12 Import/export requirements

A digital archive system should be designed to allow the export of records and associated information to users or archivists that require this information. Note that the export consists of a copy of the original digital record. The archive system should further allow for the large-scale transfer of records into another archive. This can be achieved through the use of suitable functionality built in to the archive, combined with a suitable transfer format (for example an open and easily accessible standard, such as XML). The discussion from section 3.4, where the requirements for export facilities are more exhaustively covered, is also relevant here.

5.13 Metadata structure

Metadata can be associated with a record group, with individual records, with manifestations of a digital record, or with computer files. Groups of related records in an archive are often structured on the basis of a dossier or a similar organisational system. Such structure should be defined in the metadata. It is important that metadata functionality is implemented that allows for a flexible digital archive system, which accounts for the possibility that metadata may be altered for example in connection with the addition of new metadata elements.

5.14 Alternative preservation approaches

Testbed has formulated recommendations for four types of records, namely e-mail messages, text documents, spreadsheets, and databases. The recommendations are particularly concerned with variations of migration or conversion to standards. The choice is influenced by what is currently feasible and verifiable. Other preservation approaches will be developed and implemented in the future. For this reason, the preservation system should be designed so that the introduction of a new preservation approach has as minimum an effect as possible on the other parts of the system (as described in section 5.9).

Hardware emulation has also been considered within the framework of Testbed research (more specifically, the 'software-emulation of hardware approach'[58]), but this strategy is not recommended at this stage because hardware emulation remains largely unproven within the framework of digital preservation. If emulation were implemented as a preservation approach in the future, there would be differences in the way that the preservation- and access-modules of the archive operate. It would no longer be necessary to alter the original files associated with a digital record, but more complex information about the required software- and hardware-environments would be needed, as well as operational copies of the necessary application software and operating system.

[58] http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf