



***From digital volatility to  
digital permanence***

***Preserving text  
documents***

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 Den Haag

Tel. 070 888 77 77  
Fax: 070 888 78 88

E-mail [testbed@nationaalarchief.nl](mailto:testbed@nationaalarchief.nl)  
[www.digitaleduurzaamheid.nl](http://www.digitaleduurzaamheid.nl)

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

ISBN 90-807758-1-9

The Hague, December 2003.

© Digital Preservation Testbed, The Hague 2003

All rights reserved. No part of this publication may be published or reproduced by printing, photocopying, microfilm or any other means without the prior permission of the programme office. The use of all or part of this publication to explain or support articles, books and theses and suchlike is permitted, provided that the source is clearly identified.

# Contents

<b>Foreword</b> .....	<b>IV</b>
<b>Reading Guide</b> .....	<b>V</b>
<b>1. The Dutch Digital Government</b> .....	<b>1</b>
1.1 Developments in digital government .....	1
1.2 Working effectively means managing digital longevity .....	2
1.3 Working digitally also means preserving digitally .....	2
1.4 Digital preservation and the law .....	3
1.5 A technical solution on hand? .....	4
1.6 The Digital Preservation Testbed assignment .....	4
<b>2. Digital Records and Authenticity</b> .....	<b>6</b>
2.1 Definition of a digital record .....	6
2.2 The digital record as a combination of hardware, software and computer file .....	6
2.3 Authenticity as a key concept .....	7
2.4 Digital records, digital characteristics .....	8
2.5 Metadata .....	10
<b>3. Preserving text documents in an authentic state</b> .....	<b>12</b>
3.1 The nature and role of text documents .....	12
3.2 The status of digital text documents .....	14
3.3 Characteristics of text documents .....	14
3.4 Creating text documents .....	16
3.5 Authenticity requirements for text documents .....	17
3.6 The digital signature .....	20
3.7 Summary .....	21
<b>4. Three Preservation Strategies Researched</b> .....	<b>22</b>
4.1 Introduction .....	22
4.2 Migration as a preservation strategy .....	22
4.2.1 Backward compatibility .....	23
4.2.2 Interoperability .....	25
4.2.3 Conversion to standards .....	26
4.3 XML as a preservation strategy .....	28
4.4 Emulation as a preservation strategy .....	31
4.4.1 Hardware-emulation .....	32
4.4.2 The Universal Virtual Computer strategy (UVC) .....	34
4.5 Conclusion .....	37
<b>5. Approach to the preservation of text documents</b> .....	<b>38</b>
5.1 Introduction .....	38
5.2 Decision-making table .....	38
<b>6 Concrete Actions</b> .....	<b>47</b>
6.1 Action plan for managers .....	48
6.2 Action plan for records managers .....	50
6.3 Action plan for ICT specialists .....	55
6.4 Action plan for end users .....	46
<b>Glossary</b> .....	<b>53</b>
<b>Bibliography</b> .....	<b>58</b>
<b>Appendix A Settings for conversion to PDF</b> .....	<b>60</b>
<b>Appendix B Preservation Transaction Log</b> .....	<b>65</b>
Technical Metadata .....	65
Preservation action metadata .....	65
Metadata which refer to the access of the records .....	65

# Foreword

In the initial phase of the project the Testbed team needed time to get to know each other's different disciplines. It was sometimes difficult, but ultimately it provided the quality required for this recommendation. 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.

Testbed would not have been able to do its work without the active help and support of not only the enthusiastic team members, but also of many other people at home and abroad. The Ministry of Transport and Communications, the Ministry of Housing, Spatial Planning and the Environment (VROM), the Ministry of Agriculture, Nature Management and Fisheries (LNV) and the Ministry of the Interior and Kingdom Relations have also contributed by providing us with material to experiment with.

Governments who want to manage their digital information responsibly have a great deal to do. The Testbed has attempted to be as specific as possible in indicating which technical and other solutions are the most obvious and which activities the various parties should undertake. I hope that this publication offers what is necessary to take control.

**Jacqueline Slats**  
**Programme Manager**  
**Digital Preservation Testbed**

# Reading Guide

This publication of *From digital volatility to digital permanence* consists of four parts that can be read separately. You are now in possession of part 3, *Preserving text documents*. Part 4 has already been published. Parts 2 and 1 will appear by the end of 2003. The titles of these parts are:

- Part 4: Preserving email
- Part 2: Preserving spreadsheets
- Part 1: Cost and decision models/Functional specifications  
Preserving databases

This publication is written for all those involved in managing and preserving digital information properly for the government. Testbed has tried to avoid the use of jargon as far as possible, or, when it could not be avoided, to explain it. The activities 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.

Part 1 of the series is the final piece of the research Testbed carried out into preserving digital information. This part will appear last and will complete the series, since it contains extra information about all the parts, such as cost and decision models and functional specifications for a preservation system.

This part 3, about preserving text documents, is structured as follows. Chapter 1 is an introductory chapter about the digital government, an outline of 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 record type that is central to this publication, namely text documents. What exactly is a text document and which authenticity requirements are relevant? In other words, what criteria should text documents meet so as 'not to lose their authenticity', so that it is clear to everyone that the text document is what it claims to be.

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

Chapter 5 then looks at the preservation strategy that has emerged from our research as being the most promising for preserving text documents. This 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 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 two appendices with the following subdivisions:

- Appendix A: Settings for the conversion to PDF  
(Acrobat Distiller version 4.0 and 5.0)
- Appendix B: Preservation Transaction Log

# 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 policy, summarised by Minister Remkes of the Ministry of the Interior and Kingdom Relations as 'Better Governance for Citizens and Businesses', stands or falls with the correct application of ICT within the government.

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](http://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](http://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<sup>1</sup>.

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

---

<sup>1</sup> 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 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 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 text documents. These refer to such matters as retaining metadata on the content, form and structure of a record, 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 1995 Archives Act. 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 record, 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 activities, 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 sustainable preservation of different 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, email 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

In this part of the series *From digital volatility to digital permanence* we specifically examine the first three points mentioned above.

## 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 21<sup>st</sup> 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 records or examine them. In the context of Testbed we looked specifically at text documents, databases, email messages and spreadsheets. Multimedia records, 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. Think of text documents which were created 15 years ago using WordPerfect 2.0 running on a DOS computer with an 8086 Intel processor. This software has become obsolete and is no longer being supported on the current Pentium 4 platforms. Accessing these obsolete files with current word processing software is not possible at all or will present unexpected and unreliable results. The management of digital records is still insufficient. An additional problem is that 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.

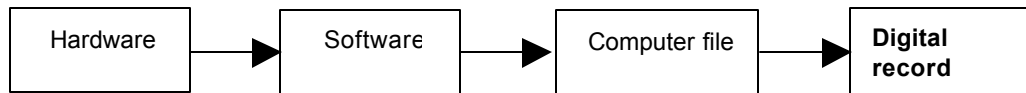
These aspects will be developed further in the sections below.

### 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 such as 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 accessed 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<sup>2</sup>:



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 zeroes and ones. This series of zeroes and ones 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, for instance Appleworks 6 for the Power Mac G4. If the digital record is reproduced in a different computer environment than that in which it was originally made, it may look and behave 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.

---

<sup>2</sup> InterPARES Authenticity Task Force Final Report, [http://www.interpares.org/book/interpares\\_book\\_d\\_part1.pdf](http://www.interpares.org/book/interpares_book_d_part1.pdf)

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 its 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 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.

For the 1995 Archives Act<sup>3</sup>, 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 the long-term 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<sup>4</sup>. 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.

<sup>3</sup> Archives Act 1995, article 1c "Archival records are records, regardless of their form...."

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

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'<sup>5</sup>.

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<sup>6</sup>

'Context' refers to the original environment in which the digital record is made and used. In order to give the record meaning, a certain 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, for example, to 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 others from the same business process, must also be described and preserved. Dossiers are an example of this.

#### Content

The content of a text document is the flat text, independent of the structure (for example, chapters, sections, and paragraphs) and the appearance of the text document (for example, the font, font size, upper and lower margins, colour, position of the page numbering, etc.).

#### 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 structural elements of a text document can, for example, be comprised of a cover page, chapters (subdivided into sections and paragraphs), and a bibliography and/or appendix. It is important that these structural elements are identified correctly, and that the report is displayed in the correct sequence. Moreover it is also important to be cognisant of any other essential structural elements the document might include, such as the presence of footnotes and endnotes. The loss of this structure during a migration could result in the incorrect display and interpretation of the text.

#### 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. The appearance of the digital record can also be influenced by other characteristics, such as page and section breaks or margin width. In addition, colours can also form a visual characteristic; as already mentioned above, the colours can influence the significance or meaning of a record.

---

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

<sup>6</sup> Een uitdijend hee!al? *Context van archiefbescheiden*, H. Hofman, Stichting Archiefpublicaties, Jaarboek 2000.

### Behaviour

The behaviour of a digital record is the most difficult to preserve. 'Behaviour' refers to the interactive characteristics of a record. In some situations the behaviour of the digital record constitutes an essential element of the actual record and consequently needs to be preserved; in such instances, the behaviour is usually derived from the content of the record. One example is a link to an Excel spreadsheet via Object Linking. An amendment made to an Excel spreadsheet is also processed in the table, which is 'pasted' (using Paste special... Paste link) into the relevant text document by means of a link.

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 (email, text documents, spreadsheets, and databases). It can generally be assumed that the appearance of email messages will be of lesser importance, since the display of emails will vary between PCs which use different email programs and have different personal settings. Conversely, for text records the appearance can 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:<sup>7</sup>

- 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.

---

<sup>7</sup>

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

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 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 record is authentic.

### 3. Preserving text documents in an authentic state

*Of the four record-types investigated by Testbed and in use in government agencies, text documents are the most prolific. Our starting point is based on the principle that text document records must be preserved in an authentic state. To this end it is necessary to specify both the essential characteristics of text documents and the authenticity requirements governing their preservation.*

#### 3.1 The nature and role of text documents

Word processors play an important role in the creation of government documents. In the years since personal computers first appeared on the market at the beginning of the 1970's, word processing has grown into the most frequent office-automation application.

Over the course of the years the interface between the word-processing software and the end user has undergone a drastic change. At the beginning of the 1980's documents were produced using a simple non-graphical interface. WordPerfect 4.2 is a well-known example of these early types of applications. Commands were given by using a range of specific key combinations. These were displayed in what was referred to as the 'underwater screen'. The combination of the document's content, appearance and structure was only visible upon printing the document to paper (and in 'Print Preview'). Modern software packages employ WYSIWYG (What You See Is What You Get) and GUIs (Graphical User Interface) for the user's convenience. In addition, modern word-processing software incorporates a continuously-increasing range of advanced features. However, this means that printing the document to paper no longer guarantees that the document can be viewed and used in the manner that was originally intended. One example of this is the tracking of amendments, whereby the date and time at which changes were made and the name of the person who made those changes can be displayed solely onscreen and not in the printed version (see Figure 1).

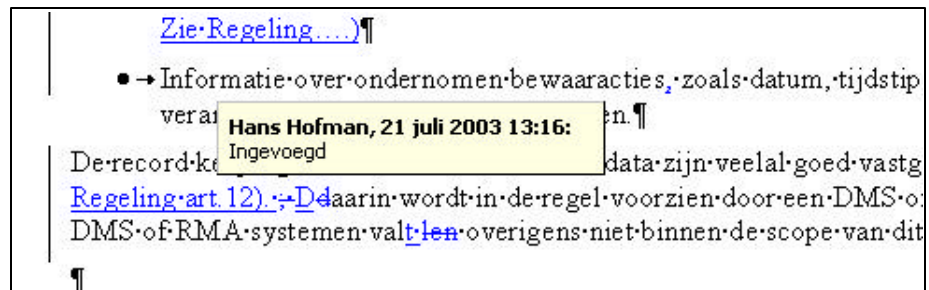


Figure 1: Example of information displayed solely on the screen

Government personnel create text documents using modern word-processing software. On preparing documents, organisations should take account of whether they will need to be preserved and, if so, for how long they will need to be preserved. It must be possible to locate the documents, and users must be confident that the documents are correct and complete. A number of organisational, legal and technical factors involved in this issue are discussed below.

### **Organisation and organisational culture**

Whereas in the recent past records were carefully prepared by the typing pool, nowadays everyone types their own records – often in a personal style (font, etc.). In general, little use is made of the templates that can be used both to comply with the house-style guidelines and to impart the necessary structure to the layout of the document. Not all users of word processors are equally proficient in the use of templates and styles. Experiments carried out by Testbed have demonstrated that poorly-constructed digital records can cause problems for their durable preservation. Cutting/copying and pasting text from old WordPerfect documents into Word documents yielded unexpected results on migration.

Digital processes also enable a number of persons to work on the same text document more or less simultaneously. For this reason version management is an issue which requires separate attention, since the status of a document in relation to the business process is important.

Organisations that create and manage their text documents in a suitable manner are in a better position to share and reuse their knowledge; in addition, they simultaneously lay the foundations for accountability. This is conducive to the transparency of the public administration.

### **Legal aspects**

The existing legislation, such as the *Regulation on the Arrangement and Accessibility of Records*, lays down a framework for the preservation of digital records. Chapter 1 reviewed this Regulation and other relevant legislation. The Regulation prescribes that text documents be durably preserved by converting them into (translated from the original Dutch) “Portable Document Format (PDF), SGML, or XML, accompanied by a stylesheet (XSL, CSS) or TIFF or PDF with the metadata in an XML wrapper”<sup>8</sup>. The specific implementation of such an approach falls outside the scope of the Regulation, and is left to the discretion of the individual government agencies.

### **Technical aspects**

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 text documents. The primary problem encountered with the preservation of text documents is due to the proprietary file format in which the files are usually stored. Most text documents are still created and stored using proprietary word-processing software. The bitstreams of files of this nature can only be interpreted by software that can read the format in which the files are coded. The specifications of these file formats are not usually disclosed by the suppliers, thereby resulting in a dependency on the supplier of the software.

---

<sup>8</sup> Regulation on the Arrangement and Accessibility of Records, February 2002, Article 6b.

### **3.2 The status of digital text documents**

Not all text documents need to be preserved. The choice of the text documents that must be considered for preservation will depend on the selection criteria according to an analysis of the tasks of the organisation that created or received the text document. These are 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 on governmental actions should either be destroyed or transferred to an archive for long term preservation.

Pursuant to the relevant Dutch legislation and regulations, records that were created and used in digital form must also be preserved in digital form. At present, most documentary records are prepared with the intention of printing them onto paper. Consequently insufficient account is taken of the useful digital life of the record at the time it is created. For example, authors of text documents often make use of automatic date fields. The use of fields of this nature causes problems for the long-term preservation of digital text documents. Text document records are increasingly prepared and used in digital form, and consequently will also need to be preserved in digital form. This will require a more meticulous approach to the creation of text document records. The authors of documents, records managers, ICT specialists and (general) managers will need to collaborate in the design of digital information so that it can be preserved in the appropriate manner. The specific issues they will need to address are reviewed in chapter 6 of this publication.

### **3.3 Characteristics of text documents**

A precise review of the characteristics possessed by text document records is required to determine which information is needed to preserve them in the appropriate manner.

As the name implies, a text document is based on text. The text document records most frequently employed by government agencies are in the form of policy papers, memorandums, memos, letters, reports, and agendas and minutes of meetings. However in addition to text, text documents can also contain the following components:

- tables;
- forms;
- headings and paragraphs;
- sections;
- bullets and numbered lists;
- footnotes and endnotes;
- characters and symbols;
- fields and frames;
- automatically updated values;
- graphical/inserted items (images/illustrations), and;
- linked or integrated objects (OLE = Object Linking and Embedding).

On occasion the many features offered by the word processing software can give cause to unexpected consequences. Automatic date and time fields are an example of this.

Experiments carried out within Testbed have demonstrated that the behaviour of automatic date and time fields is difficult to predict, even in a controlled environment. The content of the automatic date field is often updated when it has not been explicitly activated. This can have consequences for the authentic preservation of text document records.

Complex, large documents can also incorporate navigation and structural indicators, such as an index, table of contents and bibliography, which can in turn incorporate active links to selected sections of text. Documents can also incorporate security features, such as password protection, digital signatures, and protected or 'closed off' sections.

The appearance of a document can often result from the use of a variety of fonts and font sizes, styles, colours, and user-defined paragraph and line spacing. In addition, the document can also incorporate invisible style commands for tab formatting, hard returns, and section or column breaks, all of which exert an influence on the manner in which the document and its contents are displayed onscreen.

Most digital text documents possess a page-oriented layout, since most text documents are (still) prepared with the ultimate intention of printing them onto paper. Consequently the layout of the digital text document as displayed onscreen simulates the appearance of the printed document. However, information that used to be distributed in the form of paper (printed) documents is increasingly being published in HTML on websites and intranets that do not simulate the form of the printed page. A different preservation strategy needs to be adopted for the preservation of information published on websites; however, this issue falls outside the scope of the Testbed project.

The structure of the text document is a further characteristic of the record that is perhaps not immediately apparent. Within this context 'structure' refers to the logical relationship between the elements of the record. In contrast to paper documents, the explicit structure of a digital text document can be displayed separately from its appearance. The document can also be created with an implicit structure.

- Explicit structure  
This relates to the structure implemented with the help of a formatting profile, such as chapters and sections with pre-defined headings: heading 1, heading 2, heading 3, and an automatically generated table of contents. These are integrated and rendered directly in the document.
- Implicit structure  
This relates to the structure created without the use of styles and formatting profiles. The structural relationships are manifested solely through the document's appearance. The user manually enters the numbers of the chapters and the font sizes used to indicate chapter headings and sections. The table of contents is also manually created and maintained.

Structure is important. The better a document is constructed, the easier it is to comprehend. Clear structure supports the interpretation of the text; think for example of the format of a letter. Everyone can identify, for example, the sender and their address, due to the position that this information occupies.

Many documents are created with an implicit structure, which means that the structure of the document can only be rendered through the document's appearance. Testbed experiments have revealed that the appearance is one of the most difficult characteristics to retain in preserved records; one example is the loss of the formatting on reading a WordPerfect 4.2 file in MS Word.

The ability to impart an explicit structure to digital records is one of the several benefits offered by working in a digital environment. The creation of different types of text documents based on specific templates (see section 3.4) offers opportunities for the large-scale preservation of government documents in an efficient and controlled manner. By creating documents with an explicit structure, future changes in appearance (such as those that result from preservation action) can be better controlled and limited so that the document is not incorrectly interpreted as a result.

### 3.4 Creating text documents


#### *The use of templates when creating text documents*

Text documents can, if so required, be prepared using templates<sup>9</sup>, an organisation's collection of pre-defined and formatted styles. A template is a specified fixed model to be used when creating a specific type of document, such as a letter or memo. An explicit structure is herein defined. A template can determine and incorporate a number of aspects, such as:

- the house-style rules, such as logo and font;
- metadata, such as the organisation and the author's department;
- (secured) sections, and
- automatic fields.

Nowadays many organisations require the use of templates when new text documents are created. They ensure that text documents are created with a uniform appearance; they also contribute to the recognition and reliability of the document by, for example, including a watermark in the background or by the inclusion of contextual information. Templates contribute to the specification of an explicit structure in the document. However, there are also disadvantages associated with the use of templates. They often contain hidden text providing user instructions for the use or completion of each part or section of the template. Hidden text of this nature can be influenced by different sorts of preservation actions; for example, the instructions could unintentionally be incorporated in the display of the text document. A second disadvantage of templates relates to one of the methods commonly employed in the creation of documents. This is copying or cutting and pasting sections of text (and structure!) from existing documents and saving them directly in the new document. Creating documents in this manner can sometimes have the undesired result of including hidden text from the template of the existing document in the new one. Often, this goes unnoticed. However, if a specific preservation action is later carried out, such as a migration to a higher version, and the existence of the hidden text is still unknown, then it can result in unexpected and undesired additional record content<sup>10</sup>.

---

<sup>9</sup> Strictly speaking, all MS Word documents are based on a template. Starting Word or clicking New  prepares a new blank document based on the Normal template. However, within the context of these recommendations a 'template' is understood as a set of styles compiled by the organisation to characterize a specific type of document. Consequently the universal Normal template is not regarded as a 'template' in this sense.

<sup>10</sup> This can, in particular, be the case when reliance is placed on the use of Viewers to access the document.

### *The creation of text documents without templates*<sup>11</sup>

Many documents are prepared without using a template. It is, for example, possible that a template has not (yet) been defined for that specific type of document, or that the organisation concerned has not yet developed a set of templates.

Such documents can also assume a variety of different formats, such as letters, minutes of meetings, policy documents, or policy proposals. The structure of these documents may or may not be explicit. Formatting styles may be used despite the absence of a pre-defined template. The use of such styles depends on the author's proficiency with the relevant word processing software. In general, it may be expected that documents prepared in this manner contain less complex components than documents based on templates.

The structure and appearance resulting from the execution of a preservation action to such documents is more difficult to predict than with those documents created from a template. This can cause problems during the large-scale preservation of text documents, since records of this nature – each with their own individual implicit or explicit structure – will require more attention during the automated execution and evaluation of a specific preservation strategy on large batches of documents.

In general, the fewer conventions governing the creation of a text document, the higher the probability that the text document will contain unexpected, hidden components that will require additional attention when performing migrations or conversions to XML at a later date.

### **3.5 Authenticity requirements for text documents**

As discussed in chapter 2, the concept of authenticity is highly important in the preservation of information, regardless of whether the information is on paper or in digital form. However, the authenticity requirements for each type of digital record, such as email messages, spreadsheets, databases and text documents, 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*).

Testbed has experimented with text documents and with strategies to ensure their authenticity. The results from these experiments have been used to draw up guidelines specifying a minimum set of authenticity requirements that identify the essential and minimum characteristics of text document records that must be preserved in order for the records to be properly represented.

The requirements specified below are compatible with the characteristics of digital records, i.e. the context, content, structure, appearance, and behaviour. In addition, the organisation can impose supplementary authenticity requirements on the basis of the business process. It can, for example, be necessary to preserve a special colour on a digital map when that colour has a specific meaning that may not be lost.

---

<sup>11</sup> See footnote no. 9

### **Context**

All text documents need to be accompanied by metadata, such as the organisation's name, duties and the business process – or, in other words, the organisational context. In addition, the technical context of the text document – such as the operating system and the application – must also be identified if the record 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 authenticity and ongoing accessibility of the record in the future.

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

*The specification of the organisational context, such as:*

- name of the organisation;
- business process;
- date;
- 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 required for the interpretation of the current file format (for example, a specification of the name of the application program used to prepare the document; a description of the platform, with the name and version of the operating system and the name and type of the hardware);
- Information about the preservation actions that have been undertaken, such as the date, time (for example, using a 'timestamp'), and the person(s) responsible for those actions.

### **Content**

The content of a record is of vital importance: without content there is no record. The content of a text document record can vary greatly, including alphanumeric flat text, illustrations, spreadsheets, columns, tables, etc.

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

*All content, alphanumeric flat text, images or otherwise, forming part of the record must be preserved.*

This means that background information (such as page numbers, headers and footers) and automatically-created content (such as tables of contents and indexes generated for the document) must also be preserved, as well as content which does not constitute part of the wording of the document but does constitute part of the file, such as the document properties or comments (see Figure 1).

*The plain text content of the document must always be legible.*

This is not to say that the text must be of an identical appearance; however, it must always remain accessible and legible.

### **Structure**

The structure of a text document is generally displayed in the appearance and the hierarchy of logical components from which the content is constructed. As a rule, text documents have either an explicit and/or implicit structure. For an explanation of explicit and implicit structure, see section 3.4. An appropriate structure assists in a better understanding of the content and therefore constitutes an essential aspect in the authenticity of a record.

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

*The structure of the content of the original text document must be retained.*

The structure of the text document must be retained such that the logical relationships between the various components of the text document are evident when it is rendered onscreen.

### **Appearance**

The term 'appearance' refers to the manner in which a record is displayed onscreen. The appearance is usually used to communicate a specific significance that cannot be conveyed solely by plain text or a well-defined structure. By underlining certain words or including them in italics the author can emphasise aspects that he or she finds important. The emphasis on pieces of the text is in this instance communicated by the manner in which the record is displayed. A specific appearance can in this way associate additional meaning to the text that cannot be conveyed by the content and structure alone.

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

*The appearance of the preserved text document may deviate from the appearance of the original, provided that the meaning of the record is unchanged.*

Important properties of the appearance that are used to communicate additional meaning, such as font size/style, bold, italics and underlining, must be preserved. General presentation properties, such as the sizes of the margins, the pagination or the default line spacing, that do not convey any additional meaning, do not need to be preserved precisely. A number of small changes are therefore acceptable, provided that they do not affect the way in which the record can be interpreted. Some text documents can possess additional characteristics specific to the business process or the organisation. Any such properties must be recorded in the document's metadata. One example of a specific characteristic of this nature is Agrofont, a font used exclusively by the Ministry of Agriculture, Nature and Food Quality.

### **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 application used to create and manipulate the record, and this functionality is not stored as an integral element of the file contents.

However, some forms of behaviour certainly are part of the record. Examples are automatic date fields, which are updated each time the document is opened, or active hyperlinks to other files or websites. Testbed has experimented with various forms of automatic behaviour in text documents. These experiments revealed that from a records-management perspective these usually result in undesirable changes to the record.

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

*A description of all active links to different documents must be preserved.*

This means that the link does not necessarily need to operate when clicked by the user. Examples of links of this nature in text documents are hyperlinks leading to a website, or links to other files originally stored in the same system. In these instances the preservation of the text of the link will suffice; it is not necessary to preserve the behaviour or the functionality as associated with the link.

There are two reasons for this decision. Firstly, it will be almost impossible to maintain the operability of active links, since the storage location may often continually change and that means that the content of the text document record would be continually changed. Secondly, it is not always permitted to include the content of linked websites in the files for reasons of copyright.

*The active behaviour, i.e. those elements that automatically update the record (such as the automatic date field, etc.), must not be preserved; however a definition of that behaviour must be preserved.*

Proof of so-called 'behaviour-driven content' must be preserved, but that behaviour must no longer be automatically activated. This means that the content that was produced upon the activation of a specific function of the application must be preserved, including information (metadata) about the fact that a function was activated or used to generate the content. It is necessary to prevent the function from being reactivated after the record has been archived, to preclude the possibility that new content is generated when the record is (re)opened. This also means that the content of the field should no longer be automatically updated, even when this was the original setting. In other words, the preservation of the authentic record requires the retention of the content of the text document at the time it played a role in the business process.

### **3.6 The digital signature**

Electronic communication within the government, and also between government, citizens and business, will increasingly be carried out with digital signatures.

This development will be reinforced by legislative actions (Electronic Signatures Act, Electronic Administrative Transactions Act) and also by development of the required technical infrastructure such as PKI in government.

As the use of digital signatures increases, the question of preserving the signatures also comes to the fore. What is the policy on preserving digital signatures?

Digital Preservation Testbed has made an initial exploration of this subject. Further research is needed before the analysis can be completed and policy choices established.

A number of the findings arising from this initial exploration are set out below.

Some of the data on which digital signatures are based and which to a large degree determine the trust that can be placed in a digital signature, is held by the certification service provider (introduced in the Electronic Signatures Act), or a trusted Third Party. This data is mainly data that proves the certification is genuine (data on consulted identity documents, application forms and signed conditions of use) and historical data about cancelled certificates. This data may be highly significant in the event of a dispute about the authenticity and applicability of a digital signature.

Once the certificate has expired, this data must be retained by the certification service provider for a minimum of seven years (according to the Electronic Signatures Act). This minimum seven-year period was selected with non-public transactions in mind, (e-business), although the parties (the user of the digital signature and the certification service provider) are free to agree a longer preservation term if desirable or necessary.

Digital Preservation Testbed recommends that, until further research has taken place, all data about the digital signature and the identity of the signatory, data relating to authentication of the signature, and the accompanying certificate, should be preserved in or with the metadata of the digital record at the moment of signature authentication within the business process.

### **3.7 Summary**

Integrity and verification are crucial in determining the authenticity of digital records:

It is necessary to preserve the characteristics of the digital record according to the set of minimum authenticity requirements. In addition to the aforementioned requirements, each organisation will also need to establish further requirements relating to the essential characteristics of the records they generate in their various business processes. The most important aspects are the content, the structure, the appearance and the behaviour of the digital record, together with the accompanying contextual data.

This contextual data relates to information about the context (such as the business process, why, by whom, etc.) in which the digital record was created 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 digital record in connection with the required management and preservation activities. This information can be used to demonstrate or verify the extent to which the digital record can still be deemed authentic, even when that digital record is no longer exactly the same as the original version.

## 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 text documents.*

### 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 long term preservation of text documents, as described earlier in chapter 3. Based on these considerations, it is decided which is the most suitable strategy for the long-term preservation of text documents.

### 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 text document made with Word 95 can be transferred to Word 2002 or from Word 2002 to Adobe’s PDF 1.4 (Portable Document Format).

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:

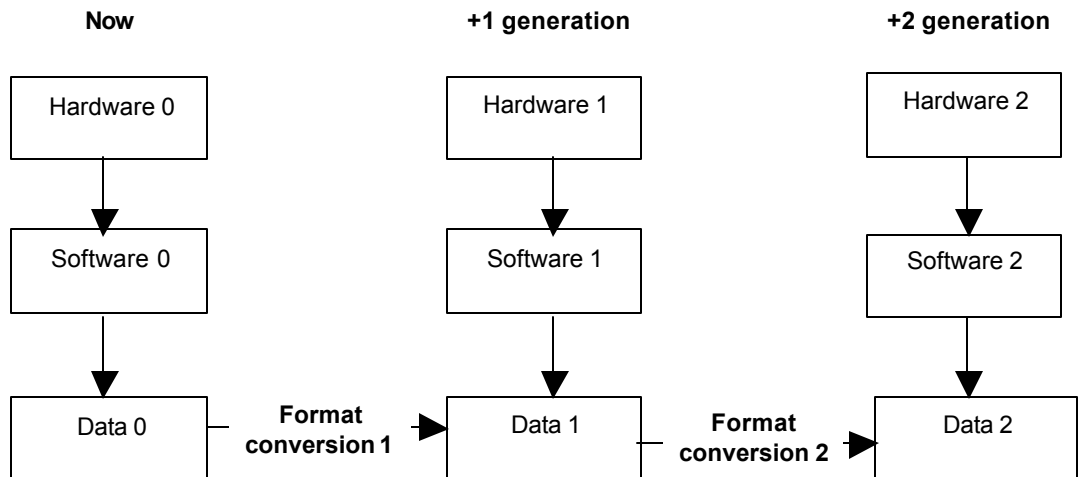


Figure 3 . Basic migration diagram

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, Word 2002 can be used to read files created in Word 95 and saved in the Word 95 format.

It is recommended that digital records preserved using this strategy are saved in the file format of the new application version, since software usually only supports a limited number of older generations of file formats. It is usually necessary to migrate to a new version once every few years. Testbed experiments have revealed that each migration of a digital record can result in changes that, irrespective of how minor those changes might be, are detrimental to the authenticity and integrity of the record. Although backward compatibility can be suitable for short-term preservation, this strategy is less suitable for long-term preservation in view of the potential accumulation of minor errors.

Another disadvantage of backward compatibility as a preservation strategy is that the digital record often continues to be stored in the supplier's own proprietary file format (for example, \*.doc for text documents created with MS Word). From the perspective of digital longevity, this retains an undesirable dependency on the original application software.

A final 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 text documents?**

Backward compatibility is a suitable preservation strategy for text documents 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 the text documents are not placed at risk. Such a migration has been found to be an appropriate manner of preserving the plain text, the appearance, the structure and the (active) behaviour of the document.

In addition, the Testbed studies have revealed that better results may be obtained when the migration bypasses a number of versions rather than including migrations via all intermediate versions. The benefit offered by migration bypassing several versions is the reduced number of migrations required for the preservation of the record, thereby reducing the risk of changes. It should be realised though that every migration is accompanied by the risk of change to the record, no matter how small that change might be. A further benefit is the reduced cost.

However, backward compatibility is not a feasible approach to the long-term preservation of text documents. The problems begin when new software can no longer interpret older files in a reliable manner, a problem which can sometimes arise within as few as three or four generations.

Records-management considerations are also a factor involved in this approach. Opening a file in an interactive word-processing application such as MS Word increases the risk of changes to the content of the file. These changes can be manifested in the form, for example, of an automatic date field that is updated to display the current date although the intention was to display the date on which the record was created. Consequently on occasion this approach is unable to guarantee the content of the record – even though the record is still being used in the original software environment.

In conclusion, the experiments have revealed that backward compatibility is a suitable short-term preservation strategy. However, the use of this strategy is subject to the condition that account is taken of the ability to preserve the text documents in an authentic manner at the time those text documents are created. This will be discussed in more detail in chapter 6.

In view of the disadvantages of backward compatibility as a preservation strategy (storage in the supplier's own proprietary 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, nevertheless, not a realistic approach to the avoidance of long term digital 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 operating systems. Software manufacturers issue versions of applications suitable for each operating system; for example, different versions of MS Word 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 WordPerfect can be read by MS Word and vice versa. Nevertheless, even with simple text documents, this can lead to loss of information.
- 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 MS Word, into an exchange format, such as ASCII (America Standard Code for Data Interchange) or RTF (Rich Text Format), which can then be read into another word processing program such as WordPerfect. Adopting this approach involves a great risk that the essential characteristics of the digital record may be lost, in particular when the text document has a complex layout or a multimedia content.

#### **Is interoperability suitable for preserving text documents?**

Testbed has experimented with interoperability between the two most widely used word processing software packages. It concluded that interoperability can be a successful approach for the short-term preservation of such files. However, relying on interoperability for long term preservation does involve some risks. If the migration is to another proprietary format, then the objection raised in section 4.2.1 remains valid. If the migration involves an application that automatically interacts with the record, then the record may undergo change in its content in the same manner as with backward compatibility. Above all, it is also possible that the new application will interpret and render the record in a manner different from the original application.

The detrimental consequences of an interoperability strategy will increase with the length of time between the creation of the source and target file formats. For example, more problems were encountered with the migration of a text document from WordPerfect 4.2 to Word 10, than with the migration of a text document from WordPerfect 5.2 to Word 10<sup>12</sup>.

---

<sup>12</sup> Testbed experienced several examples: migrating a file created with an old version of WordPerfect to newer versions of Word resulted in changes in the appearance and structure of the record. Manual intervention was required to resolve the problems before the record could be used. Migrating WordPerfect documents to a newer version of WordPerfect and afterwards to Word proved to be a more successful migration pathway.

Consequently, interoperability alone is not a reliable strategy for the preservation of text documents, although it is suitable for use as an interim solution in which obsolescent file formats can remain temporarily accessible whilst a long term solution is sought. Should interoperability be selected as an interim solution then it will be necessary to check the extent to which the source and target formats are interchangeable, so as to ensure for the guaranteed authenticity and integrity of the record.

#### **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 hardware and software.

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)<sup>13</sup>. 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.

#### **Is conversion to standards suitable for preserving text documents?**

Conversion to standards can be a suitable approach to the preservation of text documents. A conversion of this nature will achieve both backward compatibility and interoperability. In this instance backward 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 ministerial *Regulation on the Arrangement and Accessibility of Records* identifies, amongst others, standards for the preservation of text documents and images<sup>14</sup>. Within this context Testbed examined PDF and XML.

<sup>13</sup> XML: de mogelijkheden en valkuilen voor de overheid, W. Thomas, 19 September 2002.

<sup>14</sup> Regeling geordende en toegankelijke staat archiefbescheiden, February 2002.

It has transpired that PDF is a promising file format for the preservation of text documents. Testbed experiments have demonstrated that PDF is suitable for the long-term preservation of content, appearance, implicit structure and some basic behaviour (such as hyperlinks). However, problems were encountered with automatic (date) fields, since some of these fields were updated every time the record was opened. As a result, the incorrect date was assigned to the record every time it was printed to PDF. However, this is a problem originating with the creation of a text document. A text document that has not been created in the appropriate manner will encounter problems with all preservation strategies; it is not an effect caused by the conversion of the document to PDF.

Although PDF is a proprietary format, the specifications are freely available and the file format is not dependent on Adobe software for rendering. Many alternative viewers are available, most of which can be downloaded from the internet free of charge, as is also the case with Adobe Reader.

Further advantages offered by PDF are that it can render a practically identical appearance of the original text document, and that once a record has been converted to PDF it cannot readily be modified<sup>15</sup>. PDF is, in principle, a so-called 'final publication' format that can no longer be changed or edited after it has been saved.

The most significant disadvantage of PDF is undoubtedly the dependence on one specific supplier. Adobe does promise backward compatibility, although this once again relates to only a restricted number of older versions.

It is also possible to include proprietary elements in the PDF document that are not included in the file format specifications, such as compression algorithms<sup>16</sup>. This can cause problems, since the appropriate preservation action will be selected according to current knowledge and understanding of the file format and the specifications of the records to be preserved. In the event that files contain proprietary elements not foreseen in the specifications, the success of the preservation action could be limited. This in turn may result in lengthy and costly recovery programmes.

Finally, Testbed has also experimented with the use of XML. The results were positive, even though the developments relating to XML and the derived standards are still in their infancy. XML is an example of an open standard, independent of specific hardware and software, that is able to preserve not only the content, structure and appearance, but also the context of text document records. An XML approach is also readily able to incorporate metadata about the record in the same file as the text document is saved. This precludes the need to create an extra file to specify the metadata. It is important that a suitable conversion tool is selected, as not all the XML being generated by many commercial and non-commercial conversion tools is suitable for archiving purposes. It will also be necessary to create supplementary XML/XSL files to ensure that the appearance of the text document can be rendered with sufficient accuracy. The following section contains a detailed discussion of the advantages and disadvantages associated with the use of XML as a preservation strategy for digital text document records.

<sup>15</sup> This does not mean that the document cannot be changed once converted to PDF, but the chance of unintentional changes is considerably smaller than when the document is rendered using an active application program.

<sup>16</sup> *Digital Preservation Strategy*, Simon Davis, National Archives of Australia (2002).

### 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 enrichment of 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<sup>17</sup>.

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 technique (see the aforementioned Conversion to standards).

XML is a good preservation 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 independency of 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.

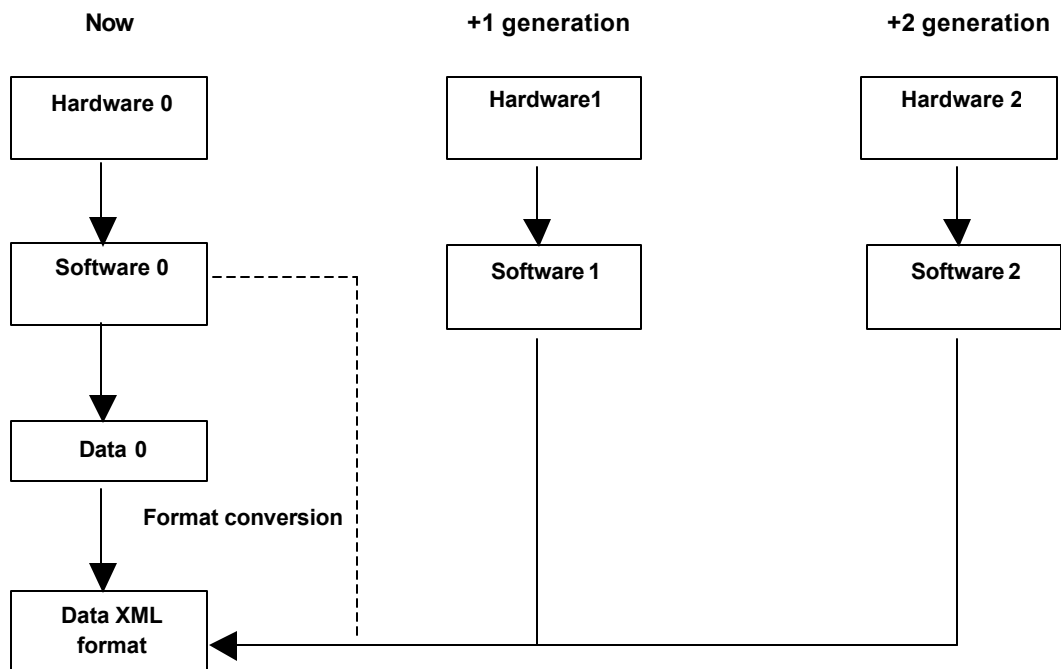


Figure 5: Conversion to XML involves fewer conversions than migration

<sup>17</sup> See <http://www.w3c.org>.

XML is a suitable file format for the specification of metadata and the representation of the five aforementioned characteristics of digital records, i.e. the 'content', 'context', 'structure', 'appearance' and 'behaviour'.

XML can specify the *content* and *context* in a manner suitable for the explicit representation of the content and context. XML can also readily display the structure of a digital record. Moreover it is possible to formulate an explicit specification of the structure of the digital record using an XML schema or DTD.

The structured and consistent design of XML documents renders them extremely easy for computers to read. However, for general users XML is more of a semi-finished product that is in need of a more accessible *appearance* (without the tagging). An appropriate appearance can be generated with the help of a StyleSheet mechanism. XSL (eXtensible StyleSheet Language), a language included in the XML group, can be used to define the appearance of the record when it is displayed. A stylesheet processor is required to transform the XML according to the instructions in a stylesheet. This software is however increasingly incorporated in browsers and other software.

Finally, the behaviour of a digital record can also be represented with XML. Simple behaviour such as hyperlinks and email addresses can be represented by means of tagging. More complex behaviour is more difficult to represent and a further review of this issue is required.

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 Center, for example, regards a wrapper as a piece of software which is used by a 'mediator'.<sup>18</sup> Conversely, the Roquade project uses the term 'container' for the 'packaging' of digital records<sup>19</sup>. A step beyond encapsulation is the additional use of XML as a 'framework' on which to mount (parts of) records in, for example, TIFF or PDF format. In this instance XML forms the backbone of the preserved digital record.

---

<sup>18</sup> "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/nhprcf2k.doc>.

<sup>19</sup> "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 electronic Archive for academic communities* (Dekker, R. *et al*, Nov 2001). The AIP term originates from the Open Archive Information System (OAIS) model.

### *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<sup>20</sup> 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.

### *Case: VERS*

In Australia a pioneering project, the Victorian Electronic Records Strategy (VERS), has been successfully completed. In the final report (from 1999)<sup>21</sup> it was described with fitting pride that it is possible to preserve electronic archival records for the long term. To this end a model is proposed, with the following features:

- The records, context and authenticity information must be encapsulated in a single object and not saved separately.
- The data structure must enable metadata to be added in layers (the 'onion model').
- XML must be used for the coding of the encapsulated archival records.
- Each electronic record must have a digital signature.

In the demonstrator which was delivered as a product of the project, the records themselves were converted to PDF. Because PDF is a binary format and XML is based on text, the PDF files were converted to text files prior to encapsulation in XML<sup>22</sup>.

### **Is XML suitable for preserving text documents?**

XML is an appropriate choice for the long-term preservation of text documents. XML can be used to specify the context, content and structure of text documents. The secondary XML standards, such as XSL and CSS, allow for the precise representation of the record's appearance and the reproduction of specific behaviour using a viewer.

Testbed experiments have demonstrated that the deployment of XML as a file format is a viable approach to the long-term preservation of text documents, in particular when used in combination with a framework approach. The exact strategy will probably differ for existing and new text documents. The deployment of XML as a file format will have the greatest possibility of success when the records possess an explicit structure. However, most existing text document records possess an implicit structure, thereby rendering them less suited to conversion to XML. Templates can be used to force an explicit structure in new documents, although it is possible to bypass the explicit structure specified by templates.

---

<sup>20</sup> See <http://partners.adobe.com/asn/developer/xmp/download/docs/MetadataFramework.pdf>

<sup>21</sup> See <http://www.prov.vic.gov.au/vers/final.htm>.

<sup>22</sup> In this instance use was made of the well-known Base64 standard also used by email systems.

Although a large number of commercial and shareware conversion tools are available, the quality of the XML they produce can vary. A number of obstacles still need to be overcome relating to the accurate representation of a text document that has been converted into XML, in particular with respect to the appearance. However, there are a growing numbers of XML editors on the market offering a WYSIWYG interface that enables users to create records directly in XML. Software suppliers such as Microsoft and Corel are currently working on expansions of their applications which offer an opportunity to generate records directly in XML. This will avoid the conversion problems mentioned above.

This implies that different preservation strategies may be required between new and existing text document records.

#### 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, the focus is 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 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:

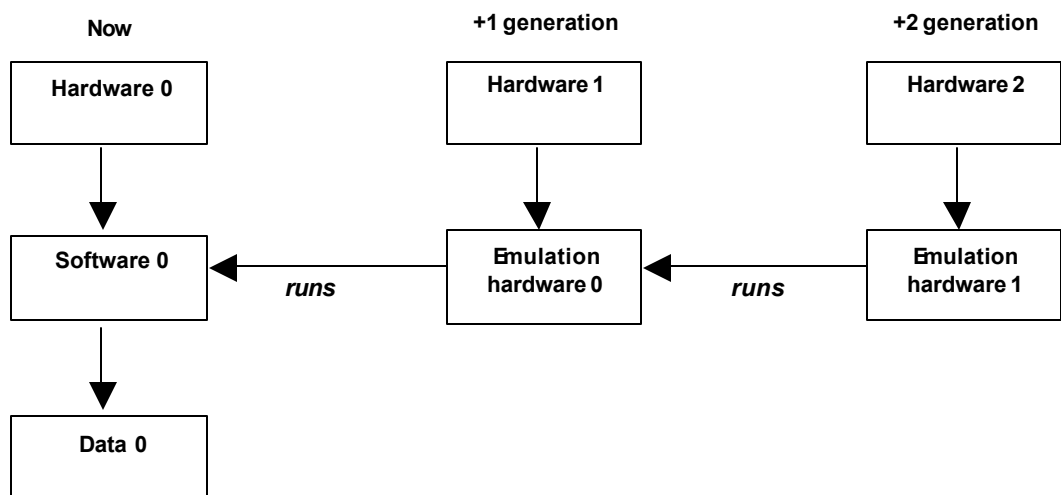


Figure 3: Basic emulation diagram

#### **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<sup>23</sup> 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.

#### **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 can be referred to as '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 emulation' is not really a meaningful preservation approach for preserving digital records either. The idea is to recreate the operating system (OS) that was 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

<sup>23</sup> Reverse engineering – decompilation: the attempt to track down and describe the logic in compiled computer programs, of which the source code has disappeared. In any case, it is a difficult task to perform: you cannot recreate a pig starting with a sausage (Pagrach 1991)

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 text documents?**

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. The disadvantages of this strategy are the technical complexity and time-consuming nature of the design, testing, use and long term 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.

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 requisite 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 advantage of hardware emulation is that the original digital record does not have to be converted or migrated. In view of the complexity of this strategy, hardware emulation will be profitable only if this strategy is not chosen for individual record-types<sup>24</sup> but is employed for all files generated from a specific computer generation.

---

<sup>24</sup> As indicated above, hardware emulation involves the 're-creation' 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 imitate the full hardware 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.

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 key combinations to create and use documents. There were more than forty combinations, and for this reason a card template for the keyboard indicating the combinations was supplied with the software. Testbed staff experienced difficulty working with this old program, just 15 years after it was in daily use – even those who at the time were thoroughly familiar with the application.

#### **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 converts 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 which 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 in its original format – is stored with a program that can extract the data from the bit stream and can specify the data in a simple manner that is independent of a specific technology, so that it can then be processed via a viewer.

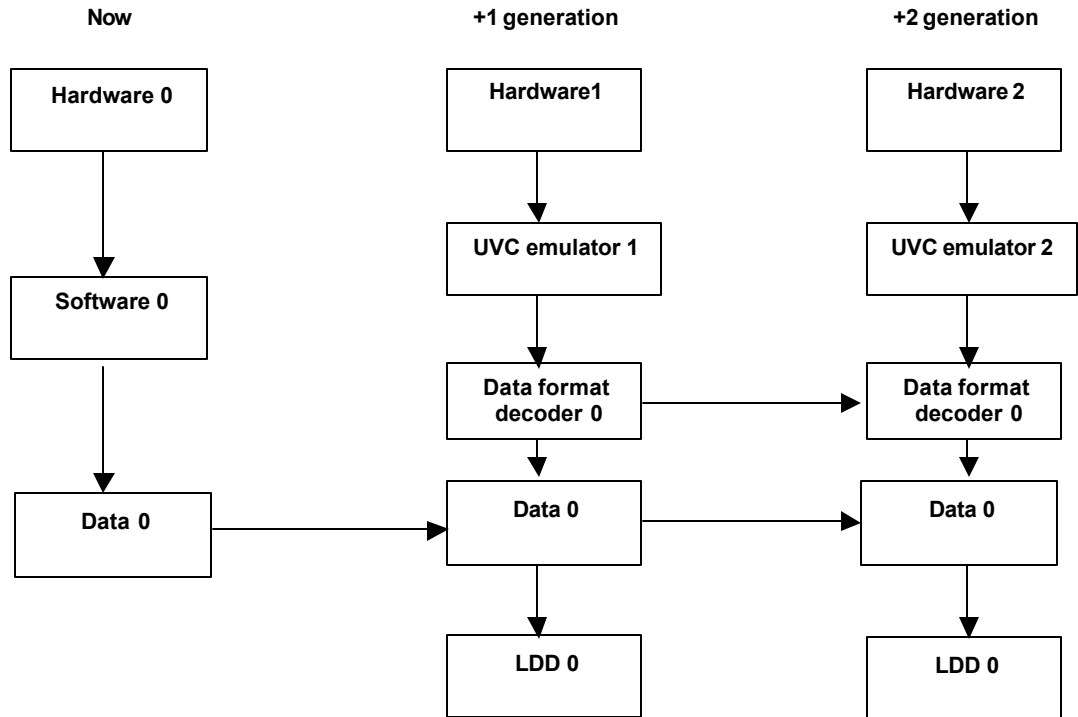


Figure 4: *Diagram of the Universal Virtual Computer*

The original file – for example, a JPEG file – is saved together with the specific UVC data format decoder program for JPEGs. 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 on a future computer platform and displayed via a viewer.

This strategy does not modify the original bit stream, and the new file (the LDD) created when running the UVC JPEG program is not saved. The LDD is 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 Dutch archival institutions will be restricted by the Regulation for the Arrangement and Accessibility of Records.

The next phase of the UVC development will be to classify records 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 the need to write a UVC data format decoder program for each file format (to generate 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 long-term 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 text documents?**

The use of the UVC as a preservation strategy for the long term and authentic preservation of text document records has many attractive properties and much potential. The UVC approach has sufficient basis to evolve into a fully-fledged preservation strategy. The impediment lies in the UVC decoder program, which is not yet able to accurately decode all of the essential attributes in the file. Experiments carried out by Testbed have demonstrated that it is in principle possible to decode a file and re-present it using a viewer. However, all of the relevant information in the file can only be extracted via an intermediate file, that can then be converted into an LDD more easily than the original proprietary file format. It has transpired that the conversion to a logical data description is anything but simple.

Experiments with the preservation of electronic publications carried out jointly by the KB and IBM also made use of an intermediate file of this nature and the information in the LDD was not derived directly from the original file<sup>25</sup>. The original file, in this instance a PDF, was first converted into two other formats, namely JPEG and HTML<sup>26</sup>. The UVC-format decoder programs convert these HTML and JPEG files into logic data descriptions in the UVC environment; these are subsequently processed for the representation of the publication.

To summarise, the UVC approach has potential, but more time and energy must yet be devoted to develop data format converter programs that are capable of converting popular, proprietary file formats. These data format decoder programs only have to be written once and others can then make use of them. Another possibility is that the large software suppliers, when developing new versions of their software, themselves deliver a UVC data format decoder program.

---

<sup>25</sup> The UVC: A Method for preserving digital documents - proof of concept' IBM/KB Long Term Preservation Study, Raymond Lorie (2002).

<sup>26</sup> This transformation was carried out outside the UVC environment, and made use of freely -available conversion tools. The specifications of these two formats have been disclosed.

#### 4.5 Conclusion

The major benefit offered by emulation is the reproduction of the original record in the environment in which it was originally created. This is a particularly attractive prospect with which the so-called 'look and feel' of the digital record can be preserved. The disadvantages of this strategy are 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 game and computer community there are no emulators available for digital preservation.

The two 'proof of concept' explorations carried out by the KB and Testbed have shown that the UVC approach possesses potential, but that it will yet be necessary to devote time and effort to writing the data format decoder programs.

Backward compatibility as a preservation strategy can be a suitable approach to the short-term preservation of digital text document records. 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.

On its own interoperability is not a reliable strategy for the preservation of text document records, although it is suitable for use as an interim solution in which obsolescent file formats can remain temporarily accessible whilst a long term solution is sought.

A combined approach based on the use of PDF and XML constitutes the most effective strategy for the long-term preservation of text document records. The advantages and disadvantages associated with the use of XML for the preservation of text documents are more or less complementary to those associated with PDF. Although it is relatively simple to achieve an accurate representation of the appearance of a text document when using PDF, this is more difficult in XML. Conversely, XML is highly capable of representing context and, in particular, (explicit) structure, which PDF is less suited towards. This strategy can be implemented using a number of different methods. The details of this approach are discussed in the following chapter.

## 5. Approach to the preservation of text documents

*Chapter 4 discussed and compared the various preservation strategies against the record-type 'text document', and came to the conclusion that the recommended preservation strategy involves the use of a combination of PDF and XML. This chapter explains how such a strategy can be implemented.*

### 5.1 Introduction

Although a combination of PDF and XML is the best strategy for long-term preservation, migration is nevertheless an alternative for organisations that create text document records that only need to be preserved for a short period of time. Organisations that create a variety of text some of 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 for all text document records.

In addition to the preservation term, the absence or presence of explicit structure (as discussed in chapter 3) also plays a role in deciding on the most practical and suitable preservation strategy. Conversion to XML serves less purpose when the text document does not possess an explicit structure. The use of an explicit structure can not be enforced for existing text documents, since the decision as to whether or not to implement an explicit structure was taken at the time the document was created. In Testbed's experience many text documents possess an implicit structure. In the creation of new text documents, it is preferable to use a template with specific instructions for users to create new documents with explicit structure.

### 5.2 Decision-making table

The following decision-making table indicates which approach is feasible in a given situation:

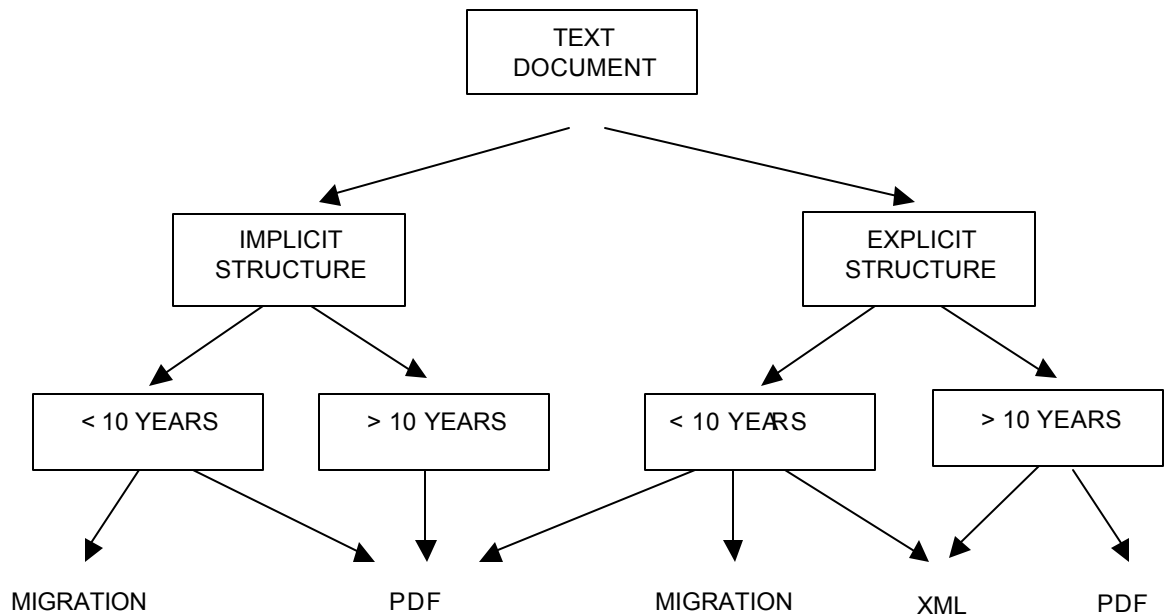


Figure 6: Decision-making table, text documents

The demarcation adopted for this decision-making table is to some extent arbitrary. Irrespective of whether the demarcation is specified as 8, 10 or 12 years, the point is that migration is a reasonable alternative for the short-term preservation of text documents (less than 10 years).

#### DOCUMENTS WITH AN IMPLICIT STRUCTURE + PRESERVATION FOR LESS THAN 10 YEARS

##### **Migration (backward compatibility)**

Migration in the form of backward compatibility is a suitable preservation strategy for text documents with implicit structure and a short preservation term, which will retain their authenticity and integrity and ensure ongoing accessibility to the records during their limited lifespan. Although it is possible to preserve text documents in their original file format, it is nevertheless preferable to upgrade the records to the newer file format since application software is only able to reliably render older file formats for a limited number of generations. However, random visual inspections of the results are necessary.

##### **Conversion to PDF<sup>27</sup>**

PDF can preserve a virtually identical 'image' of the original text document. Conversions to PDF using Adobe Acrobat software employ a process analogous to printing: a virtual PDF printer is added to the list of printers available to the word processor, and using this printer results in the creation of a PDF version of the document. Other Adobe tools are available for the conversion of large numbers of documents.

<sup>27</sup> To accommodate the needs and requirements relating to long term preservation an ISO Working Group is currently defining a new standard based on PDF. This project, PDF for Archiving (or PDF/A) is being furthered by AIIM International (Association for Information and Image Management International) and NPES (The Association for Suppliers of Printing, Publishing and Converting Technologies). At the time this publication is being prepared PDF/A is still in draft form; however, it is expected to become an ISO standard in 2004.

## DOCUMENTS WITH AN IMPLICIT STRUCTURE + PRESERVATION FOR MORE THAN 10 YEARS

### **Conversion to PDF**

PDF can preserve a virtually identical 'image' of the original text document. Conversions to PDF using Adobe Acrobat software employ a process analogous to printing: a virtual PDF printer is added to the list of printers available to the word processor, and using this printer results in the creation of a PDF version of the document. Other Adobe tools are available for the conversion of large numbers of documents.

## DOCUMENTS WITH AN EXPLICIT STRUCTURE + PRESERVATION FOR LESS THAN 10 YEARS

### **Migration (backward compatibility)**

Migration in the form of backward compatibility is a suitable preservation strategy for text documents with implicit structure and a short preservation term, which will retain their authenticity and integrity and ensure ongoing accessibility to the records during their limited lifespan. Although it is possible to preserve text documents in their original file format, it is nevertheless preferable to upgrade the records to the newer file format since application software is only able to reliably render older file formats for a limited number of generations. However, random visual inspections of the results are necessary.

### **Conversion to PDF**

PDF can preserve a virtually identical 'image' of the original text document. Conversions to PDF using Adobe Acrobat software employ a process analogous to printing: a virtual PDF printer is added to the list of printers available to the word processor, and using this printer results in the creation of a PDF version of the document. Other Adobe tools are available for the conversion of large numbers of documents.

### **XML**

Conversion to XML is most purposeful for text documents which possess an explicit structure.

New documents can be saved directly in XML, using software which offers this possibility. The latest versions of MS Word and WordPerfect offer an option for saving files in XML, although research will be required to determine the extent to which the resultant XML complies with the W3C standard. Open Office is an example of an open-source software package that already uses XML as a file format, although Testbed experiments have revealed that the resultant XML is not (yet) entirely independent of Open Office.

## DOCUMENTS WITH AN EXPLICIT STRUCTURE + PRESERVATION FOR MORE THAN 10 YEARS

### **XML**

Conversion to XML is most purposeful for text documents which possess an explicit structure.

New documents can be saved directly in XML, using software which offers this possibility. The latest versions of MS Word and WordPerfect offer an option for saving files in XML, although research will be required to determine the extent to which the resultant XML complies with the W3C standard. Open Office is an example of an open-source software package that already uses XML as a file format, although Testbed experiments have revealed that the resultant XML is not (yet) entirely independent of Open Office.

## **Conversion to PDF**

PDF can preserve a virtually identical 'image' of the original text document. Conversions to PDF using Adobe Acrobat software employ a process analogous to printing: a virtual PDF printer is added to the list of printers available to the word processor, and using this printer results in the creation of a PDF version of the document. Other Adobe tools are available for the conversion of large numbers of documents.

The following sections contain a more detailed description of the procedure used to upgrade text documents and convert them to PDF and XML.

### **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 (5.3.1 the conversion to PDF as *ade facto* standard (5.3.2), and the use of XML as a preservation strategy (5.3.3) for the representation of text documents. 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 10 years) strategy for preserving text documents in an authentic state. Since new versions of software support only a restricted number of older generations of file formats, text documents 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 Word 97 [= version 8] to Word 2000 [= version 9]). Experiments at Testbed have shown that migration over different versions of an application (from Word 95 [= version 7] to Word 2002 [= version 10]) can sometimes deliver better results than migration through each and every successive 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. In addition to procedures for visual inspections of the results, it will also be necessary to select or develop tools that can cater for the automated (and batch-processed) migration of large numbers of text documents to the required version.

#### **5.3.2 PDF**

The software manufacturer Adobe began distributing PDF in mid 1993. PDF is based on PostScript<sup>28</sup>, and it is for this reason that PDF files are page-based. PDF, in analogy with PostScript, is independent of the specific hardware, operating system and software used to create the documents. A viewer is required to open PDF files. Most of the information required to present the documents as they were originally displayed can be preserved in the PDF file, which is beneficial to the record's autonomy. This is achieved by saving the supplementary support data together with the PDF document in the PDF file. One example is the font descriptors for every font used in the document that are stored in the PDF file. PDF's popularity originates from the ease with which text documents can become finalised in an immutable form and exchanged, whilst retaining their original appearance irrespective of a specific resolution. PDF files can be created using one of two different auxiliary programs:

<sup>28</sup> PostScript was created by Adobe, and has been in use since 1985. The specifications of PostScript have been disclosed. PostScript files specify the ultimate appearance of the printed page. PostScript is not based on bit mapping, as a result of which the \*.ps files are resolution-dependent.

➤ Acrobat Distiller

Distiller is the standard program used to create PDF files. Distiller can convert PostScript files into PDF. The Distiller options allow the selection of other fonts for inclusion in the PDF file and the definition of an algorithm with which to compress images.

➤ PDF Writer

PDF Writer is a sort of printer driver program that converts files from other applications directly into PDF. A printer driver usually translates images and text into commands that can be understood by a printer. PDF Writer does not transmit the commands from the application to a printer, but instead converts the commands into PDF operators that are incorporated in a PDF file.

A PDF file can be stored in three different ways: unstructured, structured, or coded (tagged). Tagged PDF files are preferred, rather than (un)structured files. The tags enable other applications to recognise paragraphs, text formatting, bullets and tables and display them in the correct manner. This is not (or much less) the case with (un)structured PDF files. Tagged PDF files yield the best results on conversion to other formats; they are also most reliable for rendition by screen readers. These tags are more comparable with HTML tags than XML tags. Users wishing to preserve a PDF file as a tagged file must check the "Embed tags in PDF" option in the conversion settings (in MS Office). Adobe introduced this functionality in Acrobat Version 5.0.

The standard Windows installation of Acrobat includes a macro (Adobe PDF Maker 5.0) with which PDF files can be quickly and conveniently created from Microsoft Office applications. PDF files created with PDF Maker generate standard coded (tagged) PDF files which retain the hyperlinks, styles and bookmarks contained in the source document (note that PDF Maker uses the Distiller program).

The default setting of PDF Writer and Acrobat Distiller is for file compression. The compression ratios vary from 10:1 for colour images to 2:1 for combinations of text and images. When printing to PDF, the user can select the compression algorithm (for example, automatic, JPEG or ZIP for colour images) and determine the ultimate quality. PDF files are usually smaller than Word or Postscript files. Acrobat Reader decompresses the file. As indicated below, the compression of files is not recommended.

Procedures employed for conversions to PDF require careful specification and control. PDF can represent a continually increasing number of features of, for example, Word; consequently the possible settings that can be used for a conversion to PDF are becoming increasingly complex. Acrobat Distiller contains a number of arrays of predefined settings for the creation of PDF files. These settings have been designed to achieve a balance between the file size and the quality, selection of which depends on how the Adobe PDF file will be used. Distiller always uses the set of command options most recently defined and does not automatically reset to the default settings. The recommended settings for the conversion of text documents to PDF can be found in the Appendix and are based on those in *DEPOT 2000: Functional Design for a digital depot*<sup>29</sup>. Settings that could possibly result in difficulties reading the file in the future should be avoided. Examples of such settings are compression, password protection, and other security measures set by the user. It is also recommended that the fonts are integrated in the PDF file (by checking the "Embed all fonts" option) so as to reduce any dependency on external information.

---

<sup>29</sup> DEPOT 2000 Functional Design for a digital depot, National Council of Archives, The Hague, April 2000.

### 5.3.3 XML

The conversion of text documents into XML is best employed when the document possesses an explicit structure, as explained in chapter 3. The use of predefined styles or formatting profiles such as Header 1, Header 2, Body text, etc., enables the conversion software to record the document's structure. This allows for the explicit representation of the structure of the text document. Chapter 6 explained the best method of implementing an explicit structure when using word-processing software.

XML can specify the structure of groups of documents by use of an XML Schema or DTD<sup>30</sup>. As a result, the conformity of a created document against a specific structure can be verified.

Testbed's White Paper on 'XML and digital preservation' describes how XML offers two possibilities for reproducing appearance, namely Cascading Style Sheets (CSS) and the more complex eXtensible StyleSheet Language (XSL). XSL can itself be further sub-divided into two components, i.e. XSLT (XSL Transformations) and XSL-FO (Formatting Objects). XSLT is a powerful language for defining transformations of XML documents. One application of XSLT is the transformation of XML into a format that can more readily be used to display the document's appearance, such as HTML. XSL-FO is used for purposes such as the definition of a stylesheet. The stylesheet contains instructions for the representation of the document's appearance, including definition of the page layout and the text formatting.

Although CSS is a flexible and powerful language, it is primarily intended for the specification of the appearance of web documents; consequently it is less suitable for page-based text documents.

XSL, one of the newest members of the XML family, was adopted by the W3C in October 2001. As a result, XSL is not yet used or supported on a large scale.

Testbed has investigated a selection of existing tools for the conversion of text documents (primarily in Microsoft Word format) to XML. Although many tools functioned as expected, the results did not entirely comply with the requirements as specified in chapter 3.

The majority of the tools appeared readily able to extract the content and structure of text documents and convert them to XML. However, none of the tools could automatically create an XSL-FO stylesheet with which to represent the text document's original appearance. It was possible to add a separate stylesheet to the XML files. OpenOffice<sup>31</sup>, one of the tools that was investigated, is an open-source software package that appeared capable of reading a Word file, representing the majority of the appearance, and saving the file in its own OpenOffice XML format. This OpenOffice XML file format is a clear and well-structured file format which incorporates instructions for representing the document's appearance, but does not make use of an XSL-FO stylesheet. At present, OpenOffice itself is required to display text documents saved in the OpenOffice XML file format. Although it is an advantage that OpenOffice is open source, that the file format is based on XML, and that the format's specifications have been published, this link between the file format and the specific software is not ideal for long term preservation. For this reason Testbed investigated whether it was possible to convert the instructions for the representation of the appearance included in OpenOffice files into an XSL-FO stylesheet. The resultant stylesheet was then used together with the XML file to display the document.

<sup>30</sup> A DTD can define data types to only a very limited extent, and is not itself XML. For this reason DTD is increasingly making way for another standard, XML Schema. See also [www.w3c.org/TR/xmlschema-2/](http://www.w3c.org/TR/xmlschema-2/)

<sup>31</sup> [www.openoffice.org](http://www.openoffice.org)

This approach rendered the representation of text documents independent of the OpenOffice software. Although the results were successful it would, selfevidently, be preferable for OpenOffice to be fully W3C-compliant.

#### 5.4 Long term preservation of text documents

This section describes the ideal implementation for the long-term preservation of text document records with what is referred to as the 'preservation object'. The ideal is a combined approach of PDF and XML, although in practice this will not always be feasible. Testbed recommends that the original file is saved alongside the PDF and XML file, since it is impossible to predict which opportunities will be available in the future.

The ideal approach is illustrated in the following diagram.

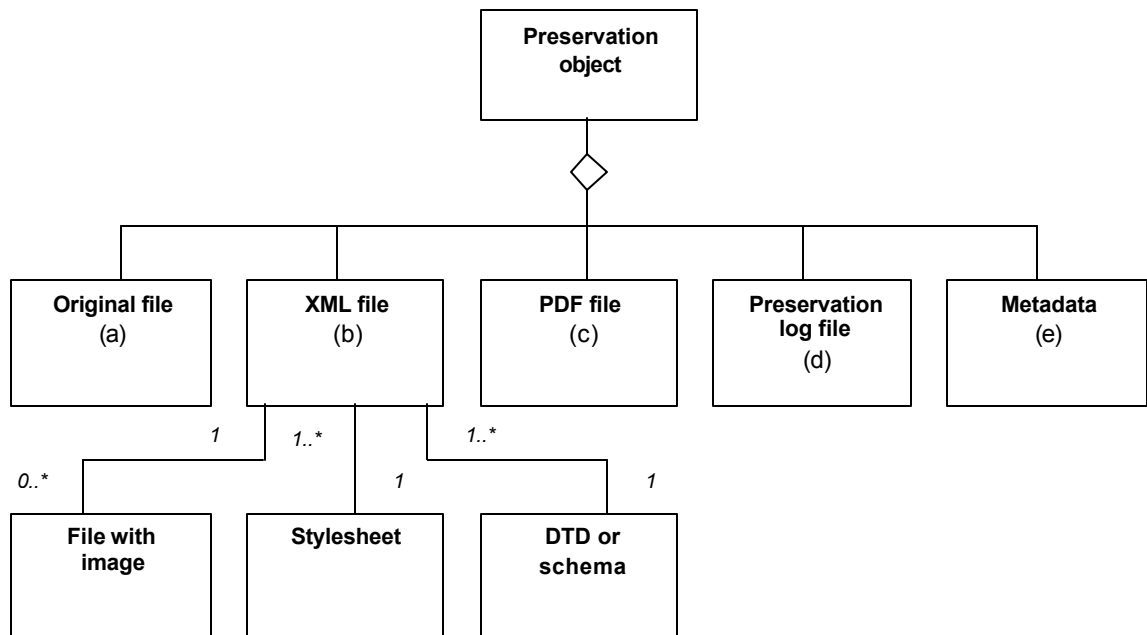


Figure 7: Structure of the preservation object

Notes: the diamond-shaped symbol indicates that the preservation object is comprised of the components to which it is linked. The significance of "0..\*" is "zero or more"; "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 file (a)

Testbed recommends that the original file is preserved alongside the PDF and XML files. This offers maximum flexibility from the perspective of possible future preservation strategies. It also provides for the most authentic representation of the record during the period in which the original software is still available.

**XML file (b)**

The XML file represents the content and the structure. The appearance is represented by means of an XSL stylesheet. A stylesheet can be linked to more than one XML file, as a result of which it is no longer necessary or efficient to save a stylesheet for each individual XML file.

The XML file is linked to a DTD or XML Schema that specifies the structure of the XML file. A schema or DTD can be applicable to a number of records, as a result of which it is no longer necessary or efficient to save a DTD or XML schema for each individual XML file.

Text documents can contain charts, images or other external objects. On conversion to XML, these are stored in separate files. References will be required to specify their relationship with and position in the XML file. The records manager can implement the measures required for the specific file formats. As such, a different preservation approach will be required for each type of digital record.

**PDF file (c)**

The PDF version of the text document is the version which accurately displays the original content and appearance of the record. This PDF file can be regarded as the copy for use.

**Preservation log file (d)**

The preservation log file contains all information about the preservation actions undertaken on the text document. 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 text document to PDF or 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. This information must be followed by data about the conversion, including the conversion tool that was used, the date and time at which the conversion was made, and the new format of the text document record.

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 text document record as a result of preservation actions. Appendix A reviews the possible content of the preservation log file.

**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 a significance 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 records (for example, the arrangement and classification codes for the records). 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 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 register 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 text documents. Now it is up to organisations to make use of this information. Chapter 5 dealt with the implementation of the combined PDF/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 different 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 read separately from the complete publication.

## 6.1 Action plan for managers

### Introduction

In reading the publication *From digital volatility to digital permanence: Preserving text documents* you will have discovered the advantages of working digitally, and also the specific problems that arise in the long-term preservation of digital records in general and text documents in particular. Digital Preservation Testbed has tested preservation strategies for the record-type 'text document'. The best way of preserving text documents at present is to use PDF and XML, but in practise this will not always be possible. The publication also discussed in detail what should be considered and how the proposed approach 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 text document records: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including word processing software. 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 text document records 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 text documents, 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 text documents 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 digital records 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 text documents
- agreements on the management and preservation of text documents

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

*Use templates to create official text documents:*

Using templates is an organised and suitable approach give records a more uniform character and contributes towards confirming the identity and reliability of the text document. A template also makes it possible to incorporate contextual metadata into the content of the record. Finally, the creation of groups of text documents based on the same template makes it easier to preserve text documents on a large scale in a more efficient and controlled way.

*Inform all staff* about the policy and the procedures. Train all staff in the use of the word processing software and when applicable, the use of templates in your organisation. A text document 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 permanence: Preserving text documents* you will have discovered the advantages of working digitally, and also the specific problems that arise in the long-term preservation of digital records in general and text documents in particular. Digital Preservation Testbed has tested preservation strategies for the record-type 'text document'. The best way of preserving text documents at present is to use PDF and XML, but in practise this will not always be possible. The publication also discussed in detail what should be considered and how the proposed approach 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 text documents: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including word processing software. 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 text documents 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 text documents 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 text documents? 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 cooperation 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 files 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 files were created; the dates of changes.
- Any conversion(s)/migrations(s) carried out.
- The hardware used within the organisation.
- The name and version of the operating system (e.g. Windows NT4)

*Establish which text documents constitute archives*

Not all text documents received or created by government agencies are records for the archives in the sense of the 1995 Archives Act. Only text documents which have played a role in a business process are deemed to constitute records for the archives. Consequently letters and other text documents created in connection with the performance of a duty are records that must be preserved in the archives. However, a draft that has been drawn up by a public official for personal use is not a record for the archives – but a draft that has been submitted to the public official's supervisor for comments *is* a record.

*Establish whether the text documents are to be destroyed or preserved*

Establishing that parts of the digital files constitute records does not imply that all these files 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 which 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 text document records 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 text documents*

Files to which data has been neither added nor changed after 1 January 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 text documents 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 floppy disks and CD ROMS, 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 text documents*

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 long-term 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 text documents*

The selection of the most appropriate manner for the storage of text documents 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 Digital Preservation Testbed recommends migration or a PDF/XML approach, depending on how long the text documents 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 text document is necessary to establish its origin, destiny, dates, etc. This metadata is required to determine the authenticity and function of the record. It is necessary to determine which metadata must be registered<sup>32</sup>.

<sup>32</sup> For the determination of metadata see the aforementioned Regulation under Article 12, or *Een uitdijend heelal? Context van archiefbescheiden*, H. Hofman, Stichting Archiefpublicaties, Jaarboek 2000.

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.

*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, 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**

*Encourage the use of templates for the creation of official text documents*

Templates impart text documents with a more uniform character, and contribute to the recognition and reliability of the text document. It is also possible to supplement the text document with contextual information (metadata). Finally, the creation of groups of text documents according to the same template offers opportunities for the large-scale preservation of text documents in an efficient and controlled manner.

*Ensure that digital text documents 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 records. A Records Management Application (RMA) or a Document Management System (DMS) can be of assistance in this respect. Applications of this nature provide for the optimum management of documents and records, including their classification, and can prevent the modification or deletion of stored text documents and records.

*Specify the manner used for classification and filing*

A classification system (as identified above) is used to assign a text document 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 text documents*

The access possibilities are closely related to the selection of the storage format and the quality of the metadata. In general, text documents stored on a central server can be made accessible to all staff. Assign 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 text documents are converted to PDF and/or XML*

At present the best approach for the storage of text documents that must be considered for long term preservation involves the combined use of PDF and/or 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 subject to regular evaluation and/or modification. It is to be expected that better software will be available in the future for the management and long-term 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 permanence: Preserving text documents* you will have discovered the advantages of working digitally, and also the specific problems that arise in the long-term preservation of digital records in general and text documents in particular. Digital Preservation Testbed has tested preservation strategies for the record-type 'text document'. The best way of preserving text documents at present is to use PDF and XML, but in practise this will not always be possible. The publication also discussed in detail what should be considered and how the proposed approach 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 text documents: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including word processing software. 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 text documents 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 for the preservation of digital records, including text documents, 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 text documents eligible for (permanent) preservation, and that agreements have been made within the organisation relating to the creation and use of templates. Besides this, the end users have received adequate training for the word processing software used by your organisation.

#### Concrete actions for ICT specialists

The following paragraphs review the (technical) ICT issues involved in the implementation of a text document preservation strategy. However, it is not possible to specify exactly how the proposed preservation strategy should be implemented. This depends on the existing computer environment and the specific requirements imposed by the relevant organisation, which will vary from case to case. However, the following review addresses the most important requirements, and also provides an overview of a possible system architecture.

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.

**(a) General principles**

*Save the text documents to be preserved in a central management system, not in the computer or the personal folders of the individual users.* This will prevent the accidental or intentional modification of records and the accidental deletion of the documents. Access to the centrally-stored records can be controlled, and preservation actions can be carried out on them. Adopting this approach allows ongoing access to the information for those who require it, and prevents unauthorised access. A central system also provides for the control 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, including text documents.

*Register 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 spreadsheet (such as the dossier to which the spreadsheet belongs) for the arrangement of the stored text documents, particularly in support of search functions.

*Make sure that the preservation system supplements each stored text document 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 word processing program, the version of the preservation system that is used, and an overview of any preservation actions carried out on the text document such as the date and the time at which the text document was accessioned into the preservation system. See Appendix B 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 of text documents is given in chapter 5. A brief summary of this description is provided below, 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 text documents. 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.

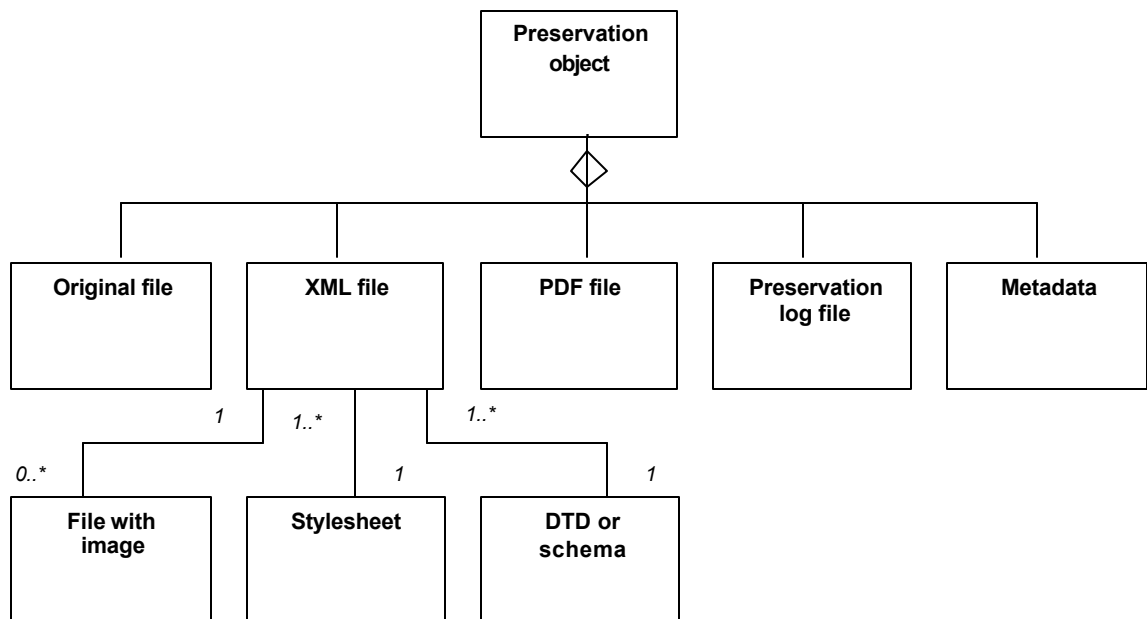


Figure 8: Structure of the Preservation Object

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

The most important requirements to be met by the system for the preservation of text documents 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 PDF and/or 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 such time as they can be transferred to an archival institution.

The European Commission has drawn up guidelines for the required functions of RMA software in the form of the MoReq specifications<sup>33</sup>. Attention is expressly drawn to one element of these specifications, namely the section on the export of digital records from the RMA<sup>34</sup>. 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 another. 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**

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 need of different metadata; moreover the needs of a specific group may 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.

<sup>33</sup> Model Requirements for the Management of Electronic Records, March 2001.

<sup>34</sup> <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/moreq.pdf>

Section 5.3, "Transfer, Export and Destruction".

## 6.4 Action plan for end users

### Introduction

In reading the publication *From digital volatility to digital permanence: Preserving text documents* you will have discovered the advantages of working digitally, and also the specific problems that arise in the long-term preservation of digital records in general and text documents in particular. Digital Preservation Testbed has tested preservation strategies for the record-type 'text document'. The best way of preserving text documents at present is to use PDF and XML, but in practise this will not always be possible. The publication also discussed in detail what should be considered and how the proposed approach 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 text documents: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including word processing software. 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 text documents a success. The concrete steps or actions are preceded by a description of the prior conditions.

### Prior conditions

You are at the start of the chain, at the source, by which we mean that that you create and manage text documents. In so doing, you determine to a great extent whether your organisation is capable of the long-term preservation of text document records. Your organisation will have laid down policy, agreements and procedures governing the creation of text documents, for instance the use of templates to create official documents on a standardised manner.

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 when creating text documents – because our studies have, above all, revealed that the long-term preservation of digital records must begin at the source. And that source is you.

### Concrete actions for end users

Even when you are sufficiently familiar with the relevant word-processing software package, possibly after completing a course (and maybe precisely then) you will certainly need to comply with a number of digital preservation do's and don'ts. Within this scope it is not possible to provide a complete summary, since applications such as Word or WordPerfect offer such a range of different functionalities. Below is a summary of the recommendations for frequently-used functions.

*Make use of the templates* supplied by your organisation. Every Word document is based on a so-called template. A template determines the basic structure of a document and also includes the settings for the document, such as AutoText fragments, fonts, key assignments, macros, menus, the page lay-out, and the special formatting characteristics and profiles. There are two basic categories:

- general templates, and
- document templates.

The 'general templates' contain settings which are available for all documents. The 'Normat.dot' template falls within this category. In contrast, 'document templates' contain settings available only for documents based on that template. Examples of these are the Memo and Fax templates in the 'New' dialog box.

Word contains a wide range of document templates. In general, each organisation will also prepare its own document templates. The advantage of using templates of this nature is that you can incorporate requests for specific metadata in the text document, such as the date, the reference number, and the name of the business process in which the text document concerned plays a role. Moreover, in principle the correct use of templates is conducive to the quality of the text documents created by the user.

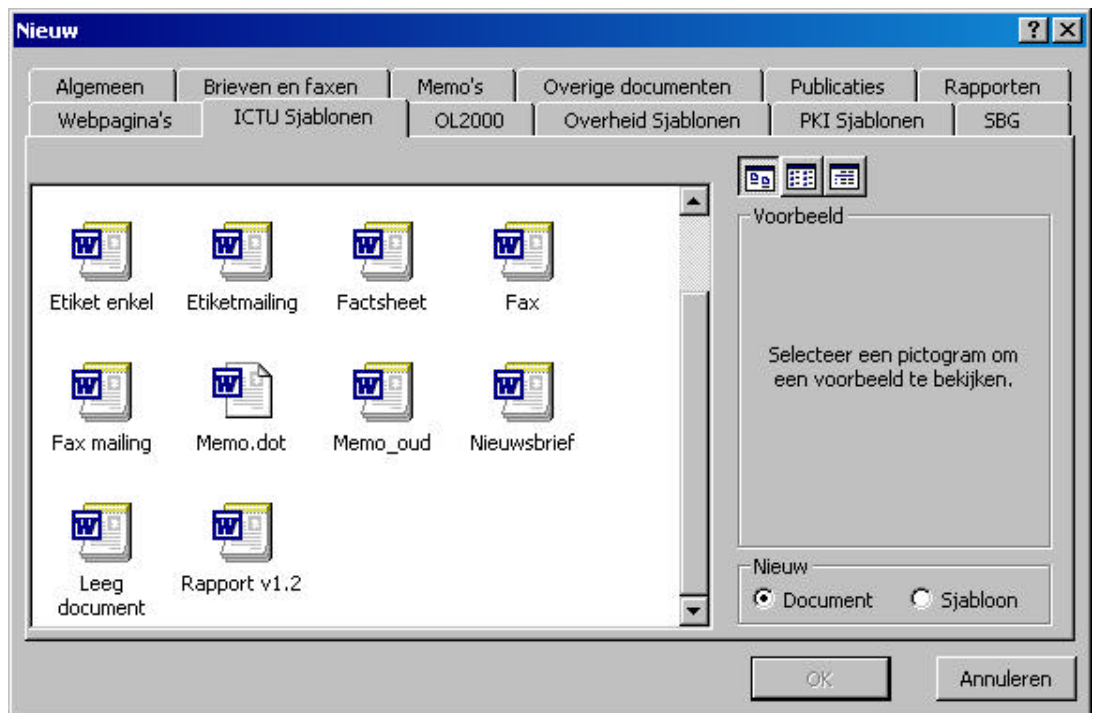


Figure 9: Examples of document templates in use within the ICTU

*Always begin with a blank template!*

It is recommended that when your organisation employs templates, you create a new text document by starting with an empty template– i.e. do *not* create a new document by changing an existing document based on the same template. This is because templates often offer an opportunity to include metadata, as shown in the following example; consequently opting to copy an existing document and modify it as required is accompanied by the risk that not all the relevant data is completed and that specific values, such as the creation date, are not updated as was intended.

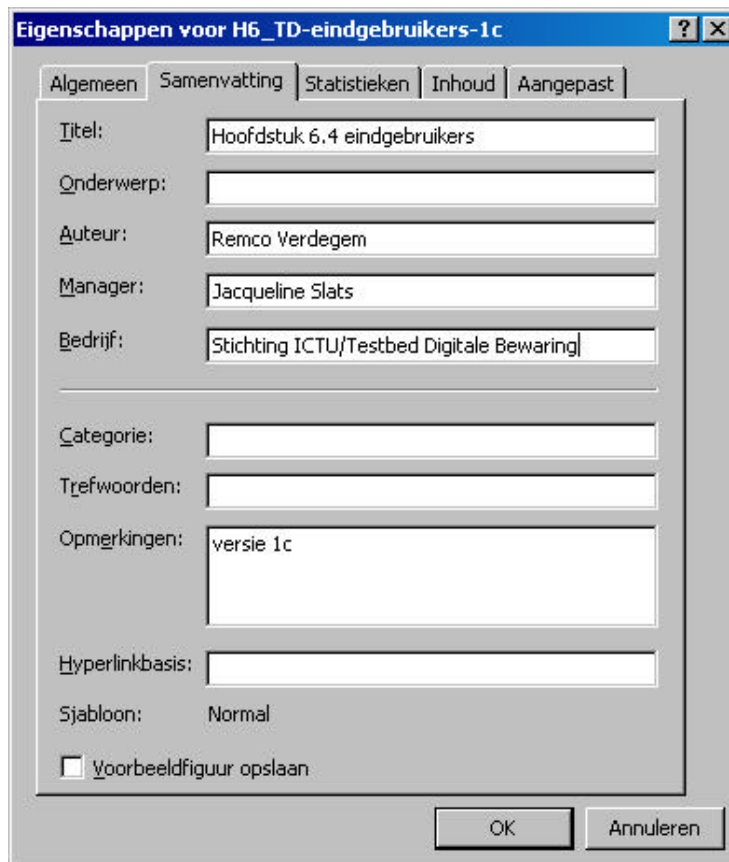
The screenshot shows a window titled "Invoerscherm Brief ICTU" with a close button (X) in the top right corner. The form contains the following fields and controls:

- Bedrijf: [Text input field]
- T.a.v.: [Text input field]
- Adres: [Text input field]
- Postcode: [Text input field]
- Plaats: [Text input field]
- Datum: [Text input field containing "23 mei 2003"]
- Onderwerp: [Text input field]
- Ons kenmerk: [Text input field]
- Uw kenmerk: [Text input field]
- Aantal Bijlagen: [Text input field] Inlichtingen: [Text input field]
- Aanhef: [Dropdown menu] [Toevoegen] [Verwijderen]
- Naam: [Dropdown menu] [Toevoegen] [Verwijderen]
- Afsluiting: [Dropdown menu] [Toevoegen] [Verwijderen]
- Afzender: [Dropdown menu] [Toevoegen] [Verwijderen]
- OK [Button] Annuleren [Button]

Figure 10: *Example of a document template*

*Use styles* to impart structure to text documents (such as Header 1, Header 2, Header 3, etc.). Formatting profiles and styles provide text documents with explicit structure. This not only improves the readability of the text, but also plays a role in the sustainable preservation of text documents.

Verify that the information displayed in the “Properties window” is up to date  
The Summary tab in the Properties window is an excellent aid in the specification of additional (meta) information about the record you have created, such as the subject, author, manager, version, etc.



The image shows a Windows-style dialog box titled "Eigenschappen voor H6\_TD-eindgebruikers-1c". It has a tabbed interface with five tabs: "Algemeen", "Samenvatting", "Statistieken", "Inhoud", and "Aangepast". The "Algemeen" tab is selected. The dialog contains several text input fields and a checkbox. The fields are labeled as follows: "Titel:" (Hoofdstuk 6.4 eindgebruikers), "Onderwerp:" (empty), "Auteur:" (Remco Verdegem), "Manager:" (Jacqueline Slats), "Bedrijf:" (Stichting ICTU/Testbed Digitale Bewaring), "Categorie:" (empty), "Trefwoorden:" (empty), "Opmerkingen:" (versie 1c), and "Hyperlinkbasis:" (empty). Below the fields, there is a "Sjabloon:" dropdown menu set to "Normal" and a checkbox labeled "Voorbeeldfiguur opslaan" which is currently unchecked. At the bottom right, there are "OK" and "Annuleren" buttons.

Figure 11: *Properties window*

The storage of incorrect context data is a real risk. One of the major benefits in using computers is the opportunity to reuse (parts of) existing documents. By always starting with a new document (preferably based on one of the organisation's templates) and then copying any sections of existing documents you require to that document (bearing in mind the provisions below) you can prevent old context information that was retained in the properties window of the original document from being saved with the new document. Accurate records of the document properties must be made when saving that text document.

*Exercise restraint in Copying/Cutting and Pasting* sections of text with different formatting. On pasting text with a formatting style unknown to the target-document template, that style will be added to the template. This can be undesirable, since the template is supplemented with styles that were not originally part of that template. This can cause problems when preservation actions are carried out.

*Do not use passwords* to protect text documents. Word offers a number of options for secure against unauthorised changes by rendering a document less accessible or inaccessible. One option we strongly recommend against is setting a password to open the text document. If you forget the password you will no longer be able to open the protected record, and the data in the record will no longer be accessible. However, and without detriment to the digital sustainability of the document, it is possible to set a password for editing the document. Only users issued with the password can then edit it. All other users can only read the document, and not change it.

*Preference is given to the use of 'standard' fonts*

We recommend you restrict the use of unconventional fonts, since such fonts reduce the probability of the authentic preservation of text document records. Unusual fonts can be lost in a migration.

*Use headers and footers to include suitable (metadata) information.* Headers and footers are ideally suited to use as a means of including (meta)data, such as the name of the file, the document's version number, the logo, etc.

*Avoid date and time insert fields (Insert, Field...)*

The use of automatic date and time fields is very popular. A number of options are available (Create Date, Date, Edit Time, Print Date, Save Date and Time). The Date field, for example, displays today's date and is updated every time the file is opened; this is undesirable from a records-management perspective. Consequently it is recommended that these automatic fields are not used. However, an exception can be made for Print Date, the date on which the document was last printed. If the Print Date insert field is used then the significance of the date must be stated.

*Make consistent use of date and time notations*

Date and time notations can cause a great deal of confusion, especially in an international context. For this reason preference is given to a date notation in which the full name of the month is shown, for example 10 May 2003 rather than 10-05-2003. In the United States 10-05-2003 will be understood as 5 October 2003.

*Insert any images or illustrations in JPEG or TIFF format*

*Do not use text boxes when a table would be more appropriate*

Preference is given to the use of tables or a specific alignment to create columns. Many documents use spaces to vertically align information into columns. However, this is not the correct procedure. On a possible migration this could result in the loss of the layout. Preference is given to the use of a table or otherwise to the insertion of the appropriate alignment (using left, right, or centre alignment, or the decimal tab).

*Use the indent function instead of spaces*

Many documents create indents using spaces, whilst the indent function has been created for this purpose. The indent function is located under **Format**, **Paragraph**, **Indents**. The use of spaces risks loss of the original layout after a migration.

*Object Linking and Embedding*

OLE is a technique used to link information from various sources. You can use an embedded or linked object to add a file or a section of a file created in Office, or in another program that supports linked and embedded objects, into another file. If the file that you wish to use was made using a program that does not support linked or embedded objects, you can still copy and paste the information from that file. The most important difference between linked and embedded objects relates to the location in which the data is stored and to how the object is updated once it has been included in the target file.

### Linked objects

The information contained in linked objects is updated only when changes are made to the source file. Linked data are stored in the source file. The target file only stores information about the location of the source file, and only an image (icon) of the linked file is displayed. Linked objects should only be used when it is necessary to restrict the file size.

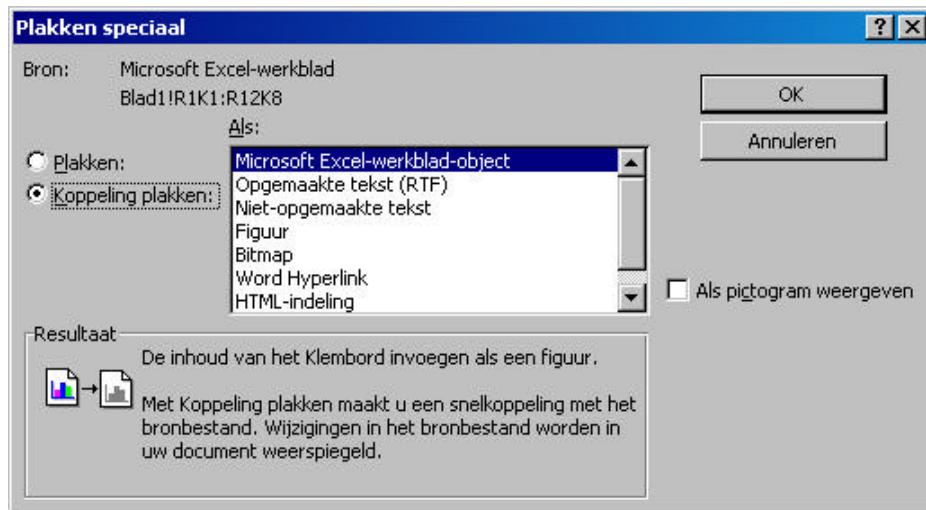


Figure 12: *Paste Special dialog box, with the option to make a link*

If you decide to make use of Linking you must ensure that the link between the source and target files is removed at the time the text document has acquired its definitive form and may no longer be changed (via Edit, Links, Break link).

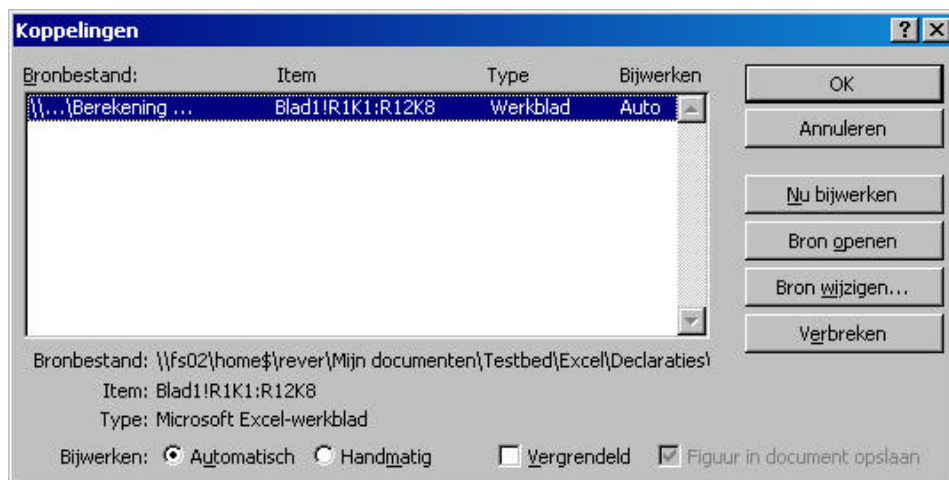


Figure 13: *Links dialog box, with the option to break a previously constructed link*

### Embedded objects

With an embedded object the information in the target file is not changed when the source file is changed. Embedded objects are incorporated in the target file. Objects that have been embedded no longer make use of the source file. Double-clicking on the embedded object will open it in the source application program. Preference is given to embedded objects.

### *Save the text document in the central digital archiving system*

Once the formal text document has been created, it is recommended to transfer it as quickly as possible into a Records Management Application (RMA) or a Document Management System (DMS).

# Glossary

## **Accessibility**

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

## **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

## **Assembler**

A computer program for translating assembly language into object code. This code is directly readable by a microprocessor. An assembly language is a programming language very similar to machine language, but that uses mnemonics in place of numeric values for ease of understanding.

## **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.

## **Basic Selection Document**

A Basic Selection Document is the form whereby a selection list, according to Article 5 of the 1995 Archives Act, is established. A selection list forms the basis for the destruction or transfer for permanent preservation of the produce of business processes from organisations and those under their competencies.

## **Behaviour**

Behaviour is one of the five attributes of digital records, 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 record, for example, by opening an attachment or by activating a hyperlink. The other four attributes are content, context, structure and appearance.

## **Computer file**

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

**Context**

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

**Conversion**

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

**Digital longevity**

The result of safeguarding the authenticity, the accessibility and the readability of digital records for the duration of a given preservation period.

**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.

**Documentary structure plan**

A plan specifying the way in which the accessibility of archival records is organised and the way in which archival records are ordered and classified.

**Emulation**

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

**Font**

A coordinated 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.

**Form**

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

**Formatting profile**

A formatting profile is a specific set of formatting characteristics (or styles) that can be applied to the text in a document and whereby the appearance of a document can quickly be changed.

**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.

**J2EE**

Stands for Java 2 Enterprise Edition. A software development environment that has become an industrial standard for developing large scale Java applications over the last few years.

**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.

**Normal.dot**

A template for general use which you can apply to any type of document. When you start Microsoft Word or click on New, Empty document, a new empty document is created based on the template Normal.dot. You can change this template and in consequence of this adjust the standard appearance or standard content of the document.

**PDF**

Portable Document Format.

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

**PKI**

Public Key Infrastructure

A system of digital certificates, certificate authorities, and Trusted Third parties that can verify the validity of each party in an electronic transaction.

**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.

## **RIO - Institutional Research Report**

Since 1991, selection lists for central government are formulated according to the method of Institutional research. The results of the institutional research are laid out in an Institutional Research Report (RIO). A RIO consists of:

- an overview of the actors who are active in this policy area;
- a historical overview of the policy area;
- an overview of the business processes of government organs in this policy area.

A RIO defines the context in which the archival records in question are created. Because of that, the RIO is the basis for the basic Selection Document.

## **RMA**

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

## **RTF**

Rich Text Format.

Format of a text document including the layout and appearance. A Microsoft protocol for a file format that contains bold, highlighting, underlining, and many other formatting characteristics.

## **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.

## **Template**

A special type of document that forms the basis for the creation and layout of a definitive document or record. Templates can contain the following elements:

- Formatting that is the same in every document based on the template
- A formatting profile, or styles
- Fragments of automatically generated text
- Macro's
- Menu and key changes
- Toolbars

## **URL**

Uniform Resource Locator. An Internet naming convention for resources available via various TCP/IP application protocols. For example:

[HTTP://www.digitalduurzaamheid.nl](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](http://www.w3c.org)

**Wrapper**

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

**WYSIWYG**

“What You See Is What You Get”. MS Word is an example of a ‘WYSIWYG-editor’ which shows on the screen how the text looks like when the text is being printed.

**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/](http://www.w3c.org/Style/XSL/)

## Bibliography

Boudrez, Philip	<XML/> en Digital Archiveren (2002) <a href="http://www.antwerpen.be/david/teksten/xml_digitaalarchiveren.pdf">http://www.antwerpen.be/david/teksten/xml_digitaalarchiveren.pdf</a>
Boudrez, Philip	Standaarden vo or digitale archiefdocumenten (October 2001) <a href="http://www.antwerpen.be/david/teksten/DAVIDbijdragen/Standaarden.pdf">http://www.antwerpen.be/david/teksten/DAVIDbijdragen/Standaarden.pdf</a>
Davis, Simon	Digital Preservation Strategy (2002) <a href="http://www.naa.gov.au/recordkeeping/rkpubs/fora/02nov/digital_preservation.pdf">http://www.naa.gov.au/recordkeeping/rkpubs/fora/02nov/digital_preservation.pdf</a>
Giesbers, Saskia	Records Management Terminologie (6 March 2002) <a href="http://www.rmconventie.nl/publicaties-rm/RecordsManagement_termen_en_definities.pdf">http://www.rmconventie.nl/publicaties-rm/RecordsManagement_termen_en_definities.pdf</a>
Feeney, Mary (Ed)	Digital Culture: Maximising the Nation's Investment (National Preservation Office UK, 1999)
Hofman, Hans	Een uitdijend heela! Context van archiefbescheiden. Stichting Archiefpublicaties, Jaarboek 2002
InterPARES Project	Authenticity Task Force Final Report (2002) <a href="http://www.interpares.org/book/interpares_book_d_part1.pdf">http://www.interpares.org/book/interpares_book_d_part1.pdf</a>
InterPARES Project	Preservation Task Force Final Report (2002) <a href="http://www.interpares.org/book/interpares_book_f_part3.pdf">http://www.interpares.org/book/interpares_book_f_part3.pdf</a>
Lorie, Raymond	A Project on the Preservation of Digital Data <a href="http://www.rlg.org/preserv/diginews/diginews5-3.html">http://www.rlg.org/preserv/diginews/diginews5-3.html</a>
Lorie, Raymond	The UVC: a method for preserving digital documents – a proof of concept (December 2002)
Lourens, Wim, et al	Emulation and Conversion: Organisational and Architectural Overview of an Electronic Archive <a href="http://www.library.tudelft.nl/e-archive/Documenten/Resultaten/reportone13.pdf">http://www.library.tudelft.nl/e-archive/Documenten/Resultaten/reportone13.pdf</a>
Mellor, Paul et al	Migration On Request, a Practical Technique for Preservation (2002) <a href="http://www.si.umich.edu/CAMILEON/reports/migreq.pdf">http://www.si.umich.edu/CAMILEON/reports/migreq.pdf</a>
Ploeg, Dr. F. van der	Regeling geordende en toegankelijke staat archiefbescheiden (February 2002)
Prins, prof. mr. J.E.J. Matthijssen, dr. L.J.	De digitale overheid en de wet; juridische kaders voor gebruik van digitale documenten bij overheden (Programma Digitale Duurzaamheid, The Hague, 2000)
Rothenberg, Jeff & Bikson, Tora	Carrying Authentic, Understandable and Usable Records Through Time (1999) <a href="http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf">http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf</a>
Rijksarchiefinspectie	Wet- en regelgeving <a href="http://www.rijksarchiefinspectie.nl/wetgeving/">www.rijksarchiefinspectie.nl/wetgeving/</a>

Testbed Digitale Bewaring	<i>Migration: Context and current status (December 2001)</i> <a href="http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf">http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf</a>
Testbed Digitale Bewaring	<i>XML and Digital Preservation (September 2002)</i> <a href="http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf">http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf</a>
Testbed Digitale Bewaring	<i>Emulatie: Context and current status (June 2003)</i> <a href="http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf">http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf</a>
Thomas, Wimpe	<i>XML: de mogelijkheden en valkuilen voor de overheid; 19 September 2002.</i>
VERS	<i>Victorian Electronic Records Strategy Final Report</i> <a href="http://www.prov.vic.gov.au/vers/published/final.htm">http://www.prov.vic.gov.au/vers/published/final.htm</a>
Zuurmond, A, Mies, K.	<i>Winst met ICT in uitvoering; Zenc, Den Haag, June 2002.</i>

## Appendix A Settings for conversion to PDF

Based on Adobe Acrobat Distiller 4.0

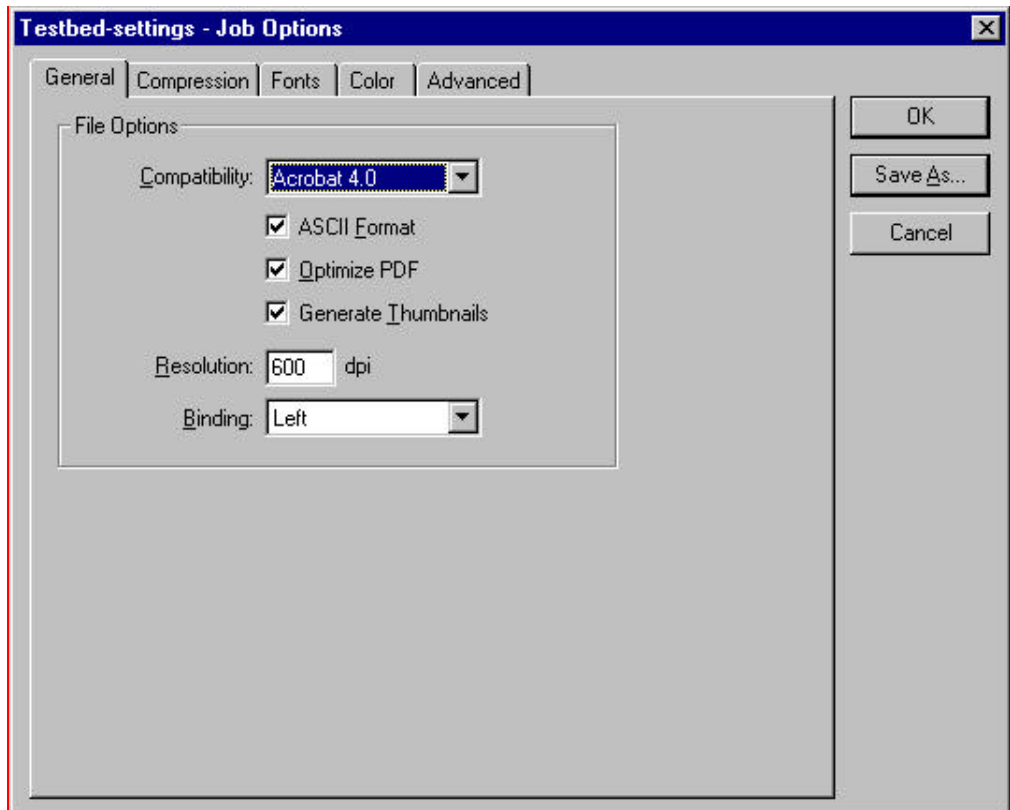


Figure 14: Acrobat Distiller Settings: General

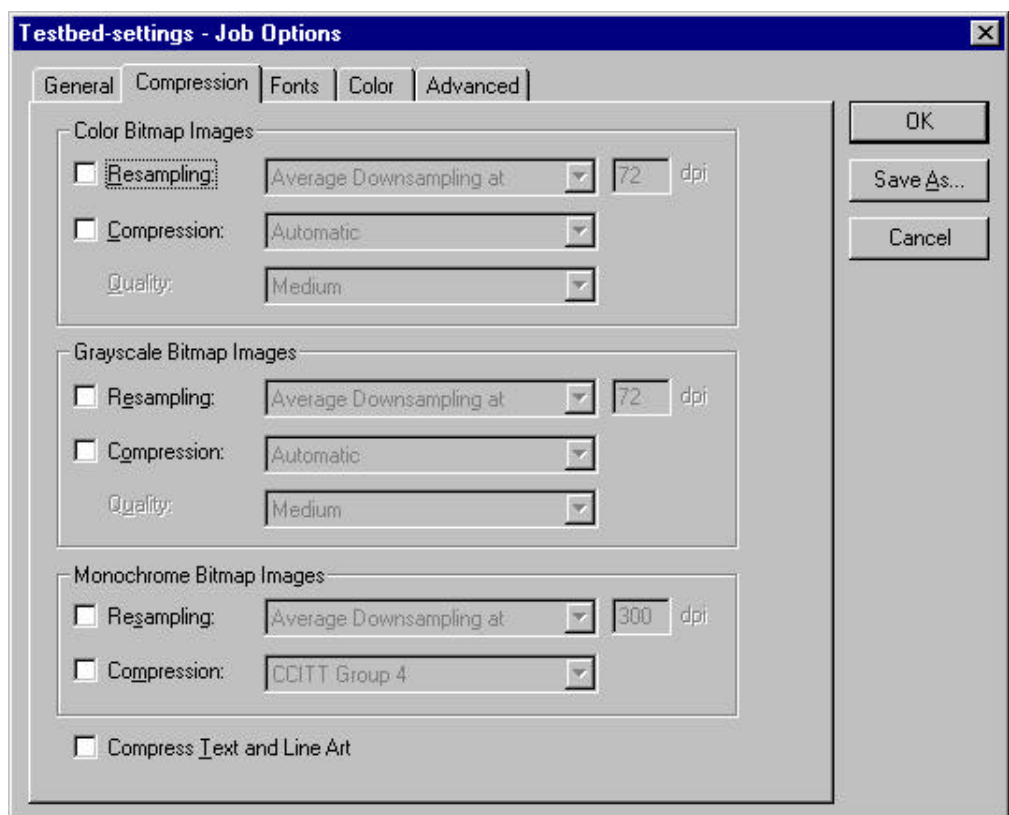


Figure 15: Acrobat Distiller Settings: Compression

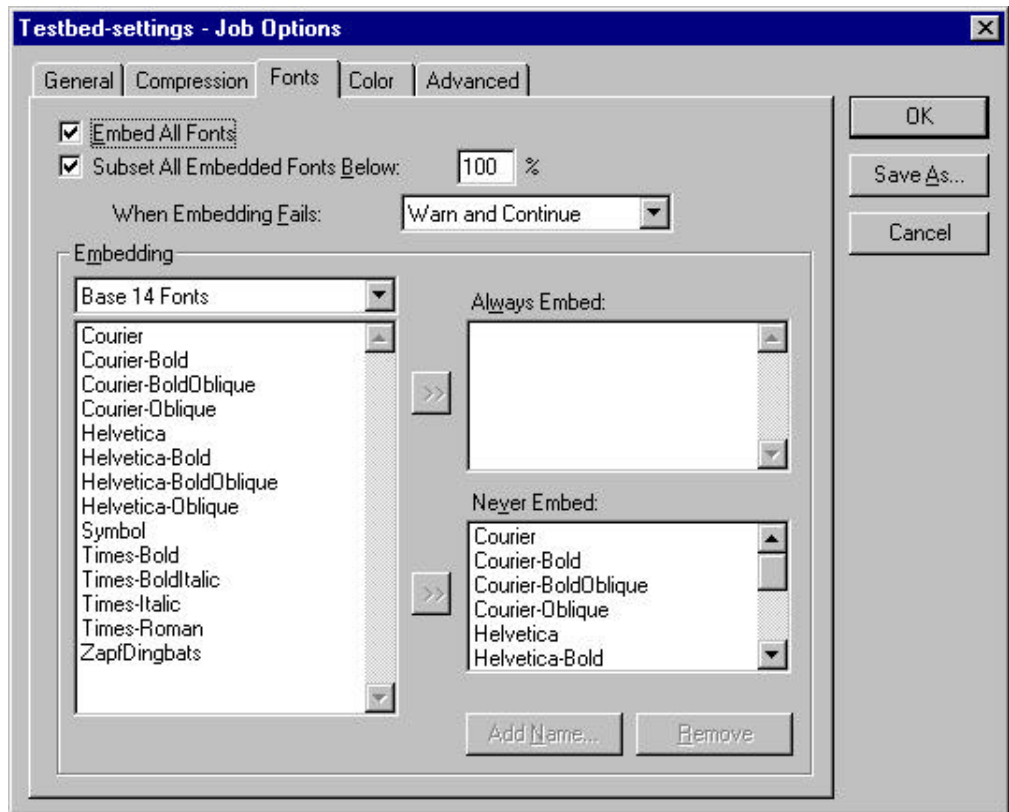


Figure 16: Acrobat Distiller Settings: Fonts

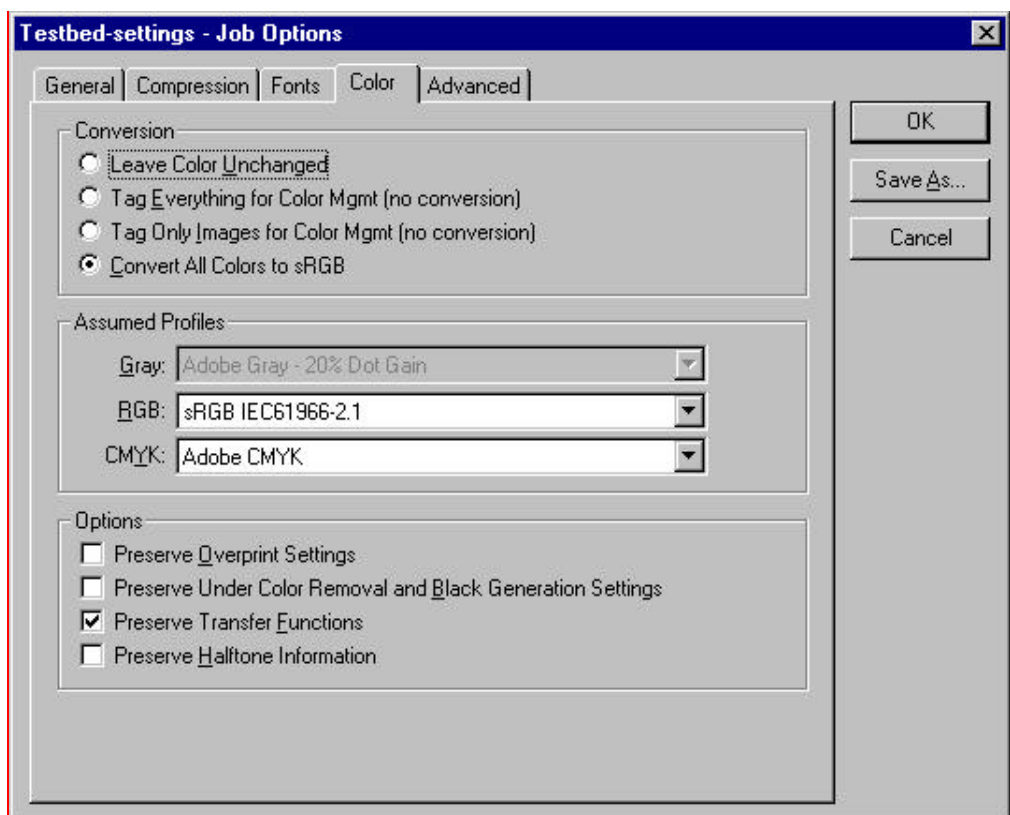


Figure 17: Acrobat Distiller Settings: Colour

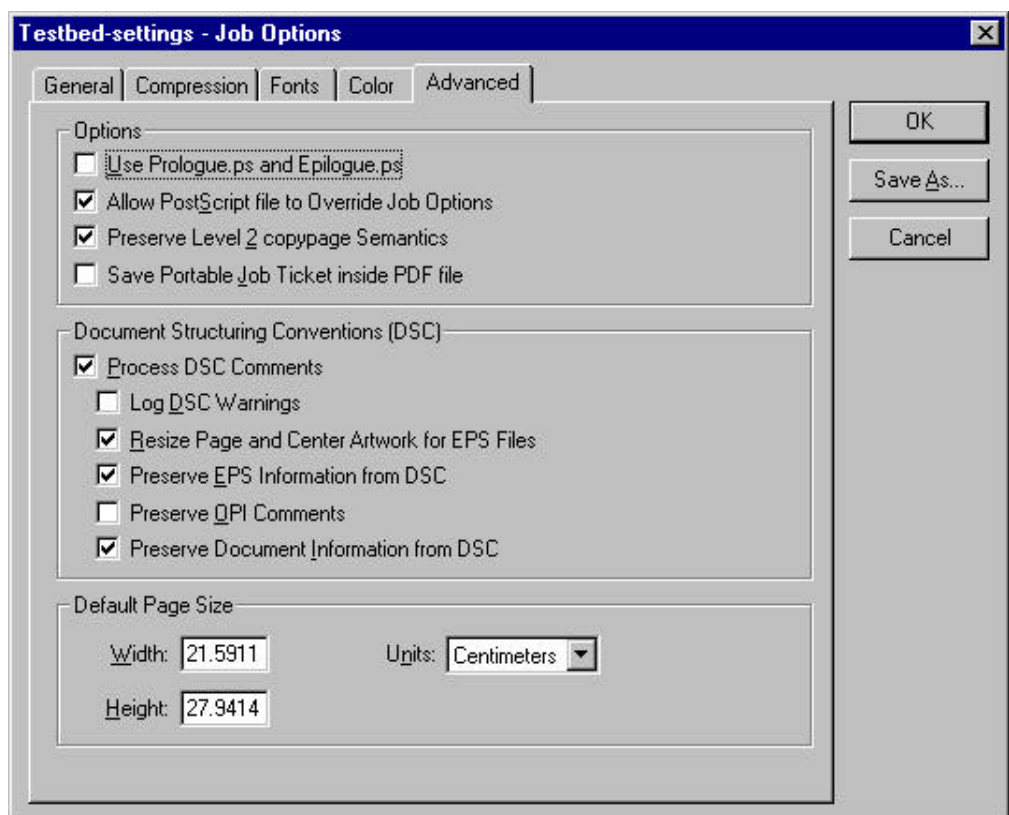


Figure 18 Acrobat Distiller Settings: Advanced

# Appendix B Preservation Transaction Log

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 Word) + hardware environment (e.g. Pentium 4) + operating system (e.g. Windows XP);
- Details of interim formats (e.g. ASCII, RTF);
- 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.