From digital volatility to digital permanence

Preserving spreadsheets

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

Email testbed@nationaalarchief.nl
www.digitaleduurzaamheid.nl

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

ISBN 90-807758-1-9

The Hague, November 2003.

Contents

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 2, *Preserving spreadsheets.* Parts 3 and 4 have already been published. Part 1 will appear by the end of 2003. The titles of these parts are:

Part 4: Preserving email
Part 3: Preserving text documents
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 p arts, such as cost and decision models and functional specifications for a preservation system.

This part 2, about preserving spreadsheets, 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 specific to this publication, namely spreadsheets. What exactly is a spreadsheet and which authenticity requirements are relevant? In other words, what criteria should spreadsheets meet so as 'not to lose their authenticity', so that it is clear to everyone that the spreadsheet 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 spreadsheets.

Chapter 5 then looks at the preservation strategy that has emerged from our research as being the most promising for preserving spreadsheets. 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 specific responsibilities in this plan and this chapter gives them the information to enable them to contribute to building a reliable digital government.

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

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 w ith 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 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, 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 a vailable in a more complete form and can, for example, be used more than once. *Service* to the public can be delivered faster, and *better.* Take, for example, applying for official documents, or identifying hazardous business in a region (as the province of Friesland does on its website www.fryslan.nl), to inform the public and business more adequately. Finally, working digitally not only provides organisational benefits, but also financial ones. Millions of euros can thus be saved [1].

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

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

1.2 Working effectively means managing digital longevity

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

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

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

1.3 Working digitally also means preserving digitally

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

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

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

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

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

Another example relates to retrieving information, such as in the question of how many unemployed people an administration agency has helped to find work in the last few decades. This question will not be properly answered if the information management of an organisation is not in good order, or not properly discharged. This subject was the central theme of the symposium that the Digital Longevity project organised together with the *Arbeidsvoorziening* in November 2002. In short, proper preservation (including 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 1995 Archives Act, the following definition of archival records is given: “records, regardless of their form, received or drawn up by government organisations… ”.
Every document, paper or digital, that has a function in the performance of a task is therefore in principle a record or an archival document.

The Regulation on the Arrangement and Accessibility of Records (2002).
The Regulation on the Arrangement and Accessibility of Records is an extension to article 12 of the 1995 Archives Decree. The Regulation states that the most important requirements are that records must be authentic and that records must be readable and retrievable within a reasonable period of time. There are extra requirements for digital records, including spreadsheets. 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 archival institution, 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 21st century equivalent of traditional paper records. They have other properties, characteristics and applications. However, both digital and paper records must meet the same legal requirements. In practice, this requires a different approach.

Digital records are not tangible objects like a book or a magazine, but a combination of hardware, software and computer files. This combination is necessary to be able to use the 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 for example of spreadsheets that were created 15 years ago using Quattro Pro running on a DOS computer with a 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 spreadsheet software is not at all possible or presents 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.

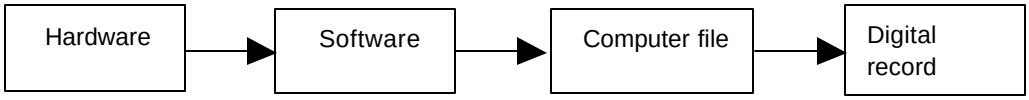
These aspects will be developed further in the sections below.

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 'fixed' in the same way. Digital files have to be processed technically before the user can read the records and use them for the purpose required. It is this dependence on hardware and software that compels us to think differently about the way we make and use digital records.

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



A digital record is made using a particular combination of hardware and software and is stored in the form of a code, the computer file. This computer file consists of a series of 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.

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.

---

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

Authenticity has two central concepts:

Integrity: that the record is intact and not changed or corrupted in such a way that it's meaning is no longer clear. A record has integrity when it is complete and uninterrupted in all essential aspects. Changes are acceptable to a certain extent, as long as they do not affect the original meaning or function of the record. An example of this is the website mentioned above that belongs to the province of Friesland, which 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[3], it makes no difference whether a record has a digital or a physical form. The problem that arises with digital records, however, is that due to changing technology, not all aspects of a record can be preserved as precisely as when it was made. This does not mean, though, that sustainable preservation of authentic digital records is impossible.

2.4 Digital records, digital characteristics

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

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

Dutch legislation and regulations refer to context, content, structure and form. The characteristic 'behaviour', which can be important for digital records, is not mentioned.

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

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

In addition, current regulations define the concept of 'form' as 'the outward appearance in which the structure and layout are visible"[5].

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

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

Content
Spreadsheets can possess a variety of types of content, such as the content of the cells, formulae, mathematical functions and comments, irrespective of structure (rows and columns) and appearance (for example, the font, font size, colour, 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 spreadsheet can, for example, be comprised of the worksheets in a workbook, and the rows, columns and cells. It is important that these structural elements are identified correctly, and that the spreadsheet is displayed in the correct manner. Moreover it is also important to be cognisant of any other essential structural elements, such as the presence of footnotes and endnotes. The loss of parts of this structure during a migration could result in the incorrect display and interpretation of the spreadsheet.

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. Colours can also form a visual characteristic; as already mentioned above, the colours can influence the significance or meaning of a record.

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. Examples of this behaviour are hyperlinks, and what are referred to as "what if queries" in spreadsheets.

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

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

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 will be of essential importance. The five aforementioned characteristics play an important role in the evaluation of the various preservation strategies discussed in chapter 4.

2.5 Metadata

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

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

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

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

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

---

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

3. Preserving Spreadsheets in an authentic state

Spreadsheets have assumed an important position in today's government as a means of working with figures. Our starting point is based on the principle that spreadsheets must be preserved in an authentic state. To this end it is necessary to specify both the essential characteristics of spreadsheets and the authenticity requirements governing their preservation.

3.1 The use of spreadsheets: problems encountered during preservation

Bookkeepers were using spreadsheets long before automated versions became available. Richard Mattessich (University of California, Berkeley) developed the first automated spreadsheet between 1961-1964. In the then absence of personal computers this application did not find worldwide use. However, personal computers had become available by the time Dan Brinklin and Bob Frankston, based in Boston (Massachusetts), developed VisiCalc in 1978-1979 – the first spreadsheet to become known to the general public and to be used on a large scale. VisiCalc was regarded as the first 'real PC application'.

In the years since then spreadsheets had become a standard component of PC software. Everyone who uses a PC will on occasion work with a spreadsheet. Virtually all companies and government agencies use spreadsheets in the management of their time, funds, and assets, and for the organisation of a great variety of other data. The response to a request for a summary of an organisation's projects will often be in the form of a spreadsheet. Timesheets are also often based on a spreadsheet.

Organisations will need to preserve these spreadsheets and ensure their continued accessibility, because they could be needed, for example, for management purposes, for operational accountability, or as a source of information.

When creating records, 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 records, and users must be confident that they are correct and complete. All the problems of relevance to the appropriate and sustainable preservation of digital records in general are also applicable to spreadsheets. A number of organisational, legal and technical factors are involved; these are discussed below.

Organisation and organisational culture
Government agencies have now been using spreadsheets for several years; nevertheless, to date little experience has been acquired with the long-term preservation of spreadsheets. People can create spreadsheets according to their own wishes, which can make spreadsheets very difficult to manage. Moreover, staff often decide what they do and do not wish to preserve: the storage, modification, and deletion of spreadsheets can take place at their own discretion. In fact, spreadsheets are often unjustifiably regarded as part of the personal work domain.

Digital processes also enable a number of people to work on one document more or less simultaneously. Version management requires separate attention, since the status of a record is important in relation to the business process.

Organisations that create and manage their records 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 and regulations, 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. However, they do not prescribe a specific format for the long-term preservation of spreadsheets. In addition, it is often difficult to interpret these regulations in practical terms.

Technical issues
Hardware and software rapidly become obsolete, as a result of which digital files are no longer accessible. Virtually no practical studies have been carried out, either at a national or an international level, into technical approaches to the long-term preservation of spreadsheets. The primary problems encountered with the preservation of spreadsheets are due to the proprietary file formats in which spreadsheets are often saved, and to the mathematical functions and formulae incorporated in them. The presence of these functions and formulae means that alternative requirements must be specified for the preservation of spreadsheets than, for example, text documents.

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

3.2 The status of spreadsheets

Not all spreadsheets need to be preserved. The selection of spreadsheets that must be considered for preservation depends on the selection criteria specified according to an analysis of the tasks of the organisation that created or received the spreadsheet. 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 archival institution 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. Some organisations print their spreadsheets and preserve the printout. However, this is not an appropriate strategy since information will always be lost.

3.3 Characteristics of spreadsheets

To preserve spreadsheets for the long-term, it is necessary to review their characteristics precisely.

Spreadsheets can be used as an aid for planning and control, for reporting purposes, as support for computational work, and as 'scrap paper'.

All spreadsheets contain one or more worksheets. Each worksheet is comprised of rows and columns, and each cell has a unique designation based on the point of intersection between column and row.
Charts can also be inserted in worksheets. These display data from selected cells in the relevant worksheet in a graphical form. Other types of objects can also be inserted in worksheets.

Each cell can possess various attributes, such as:

- the cell content, i.e. the value displayed in the cell;
- one of a series of cell format categories, such as number, currency, accounting, date, time, percentage, or fraction formats. Changing the cell format category may also change the display of the content of the cell;
- a formula, whereby the content of a cell is created on the basis of the content of one or more other cells;
- a comment appended to the cell;
- appearance characteristics, such as alignment, font (including font style and font size), borders, and pattern (background).

The standard display of the spreadsheet onscreen shows the content of the cells within the associated cell format. The content can be comprised either of manually entered data or of the results of calculations using formulae. The actual composition of these formulae (which formula is used and the cells to which it refers, in short how the calculation is designed)) can be displayed by positioning the cursor on the relevant cell or using the relevant option to 'display all formulae'. Whilst this option is activated, the results calculated using the formulae will not be displayed. For this reason a distinction can be made between two layers within spreadsheets: the standard-display layer, and the underlying layer containing the formulae and cell references.

Spreadsheets can be linked to each other, such that the result from one spreadsheet can be used as the input for another spreadsheet.

3.4 Authenticity requirements for spreadsheets

As discussed in chapter 2, the concept of authenticity is of great importance in the preservation of information. However, the authenticity requirements for each type of digital record, such as spreadsheets, emails, text documents and databases, 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 spreadsheets, 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 spreadsheets that must be preserved in order for the records to be properly represented.

The following requirements relate to the characteristics of a digital record: the context, content, structure, appearance, and behaviour. In addition, the organisation can impose supplementary authenticity requirements on the basis of the business process. It can, for example, be necessary to preserve a special colour on the worksheet, because that colour has a specific meaning that may not be lost.

Context

All spreadsheets need to be accompanied by metadata, such as the organisation's name, duties and the business process – or, in other words, the institutional or organisational context. In addition, the technical context of the file – 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 spreadsheets:

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, 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 spreadsheets can vary greatly; they can contain figures, letters, formulae, comments, charts, inserted objects, and even complete maps.

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

The actual content must always be preserved.

The content of a preserved spreadsheet must be as clearly and legibly displayed as in the original digital record, including the content of any possible charts and other inserted objects.

The content of both 'layers' of spreadsheets must be preserved.

Both the standard display of the worksheets (the entered values *and* the results of calculations using formulae) and the underlying layer containing formulae and inserted functions must be represented in the correct manner.

Structure

The structure of a spreadsheet relates to its composition and the logical hierarchy of the elements from which the content is constructed. A spreadsheet (a workbook) contains one or more worksheets. Each worksheet is in turn comprised of a series or rows and columns; each cell in the sheet has a unique designation based on the point of intersection between the relevant column and row. Worksheets can also contain one or more inserted charts or other objects. Cell references are also an important component of a spreadsheet's structure.

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

All worksheets in the spreadsheet must be preserved.
All worksheets must be preserved in the correct sequence, and with the correct names.

The structure as established in rows and columns must be preserved.
The rows and columns in the worksheet must be represented in the same manner as in the original file, thereby ensuring that each cell retains its unique designation as based on the intersection of the relevant column and row. For example, subsequent to a migration operation the coordinates of a cell may not be changed from A3 to B3 as this could cause the formulae to produce a result differing from the original value.

Inserted charts and/or other objects must be preserved.
The presence and position of inserted objects must be represented in the correct manner.

The cell references must be preserved correctly.
If cells in a spreadsheet refer to cells outside the spreadsheet then the other spreadsheet(s) must also be preserved, together with the mutual relationships.

The relationship between both 'layers' of spreadsheets must be preserved.
The link between the results of the calculations in the display layer and the formulae in the underlying layer must be preserved.

Appearance

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

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

The appearance of the spreadsheet may deviate from the original appearance.
The appearance of the preserved spreadsheet does not have to be identical to the original version provided that the meaning of the digital record is unchanged. The appearance characteristics that are used to impart additional meaning must be preserved.

The appearance of charts may deviate from the original chart.
The appearance of the preserved chart does not have to be identical to the original version, provided that the meaning of the chart is unchanged.

The appearance of inserted objects must be preserved.
The appearance of inserted objects, for example bitmap images, objects created using Microsoft Equation Editor or organograms, must be displayed in the same manner.

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 digital record.

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

The behaviour whereby formulae are executed to carry out a (re)calculation should not be preserved.
It is necessary to prevent mathematical formulae from being reactivated subsequent to the archiving of the spreadsheet, thereby precluding the possibility of the generation of new content on opening the record. Testbed experiments have revealed that newer versions of spreadsheet applications sometimes use mathematical formulae in a manner different to the older version, thereby yielding a different result. In other words, the preservation of the authentic record requires the retention of the content of the record at the time it played a role in the business process.

3.5 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?

The 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 of great importance 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.

3.6 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.

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 spreadsheets.

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 spreadsheets, 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 spreadsheets.

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 spreadsheet made with Excel 95 can be transferred to Excel 2002 or from Lotus 1-2-3 version 5 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:

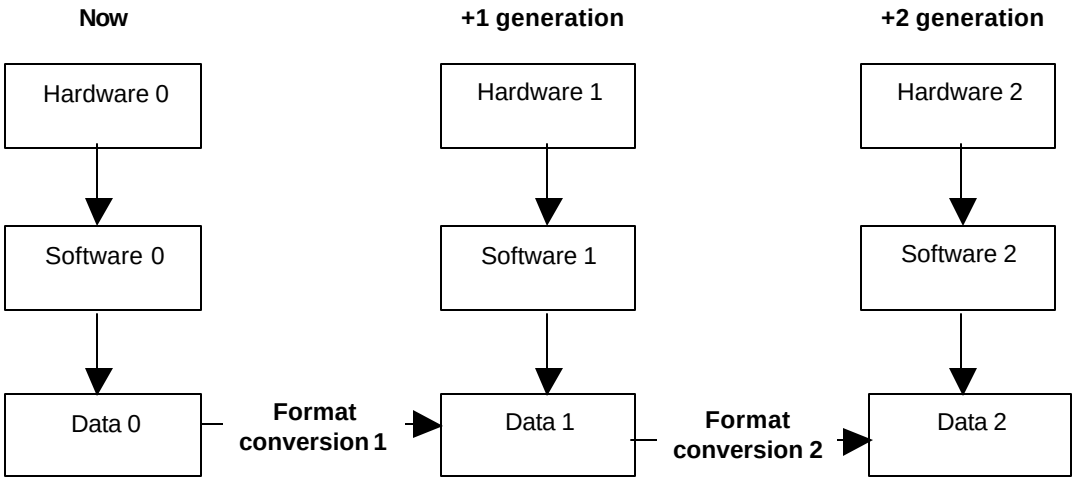


Figure 3. *Basic migration diagram*

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

It is recommended that digital records preserved using this strategy are saved in the file format of the new 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, may be 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, *.xls for spreadsheets created with Excel). 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 spreadsheets?

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

In addition, Testbed studies have revealed that better results may be obtained when the migration bypasses a number of versions rather than including 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 spreadsheets. The problems begin when new software can no longer interpret older files in a reliable manner, 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 aninteractive application such as Excel 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 spreadsheets in an authentic manner at the time those spreadsheets 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 not considered 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 Excel 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 Excel can be read by Lotus 1-2-3 and vice versa. Nevertheless, even with simple spreadsheets, 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 Excel, into an exchange format, such as DIF (Data Interchange Format) or CSV (Comma-Separated Values), which can then be read into another spreadsheet program, such as Lotus 1-2-3. Adopting this approach involves a great risk that the essential characteristics of the digital record may be lost, particularly when the spreadsheet has complex formatting or makes use of formulae or mathematical functions.

Is interoperability suitable for preserving spreadsheets?

Testbed has carried out experiments with interoperability, in particular between Lotus 1-2-3 and Excel. These experiments revealed that neither Lotus 1-2-3 nor Excel were able to represent all the elements of the migrated spreadsheet in an authentic manner. Some of the more complex mathematical functions were no longer operable, and cells that originally displayed the result from such functions now displayed an error alert. Also, many of the fonts were changed. The content, appearance and behaviour of the spreadsheets were all damaged.

Consequently interoperability is not considered a reliable preservation strategy for spreadsheets.

4.2.3 Conversion to standards

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

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

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

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

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

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

Is conversion to standards suitable for preserving spreadsheets?

Conversion to standards can be a suitable approach to the preservation of spreadsheets. 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 aforementioned ministerial *Regulation on the Arrangement and Accessibility of Records* cites a number of standards for the sustainable preservation of digital records[9], but does not state any specific format for the preservation of spreadsheets. Testbed has examined PDF and XML. XML is reviewed in the following section.

The conversion of spreadsheets to PDF format is inadvisable. The standard display of the spreadsheet onscreen shows the content of the cells and the associated cell format. The content can be comprised either of manually entered data or of the results of calculations using formulae. The composition of these formulae (the formula that is used and the cell to which it refers – or, in other words, the design of the calculation) can be displayed by positioning the cursor on the relevant cell or using the relevant option to 'display all formulae'. Whilst this option is activated the results calculated using the formulae will not be displayed. For this reason a distinction can be made between two layers within spreadsheets: the standard-display layer and the underlying layer containing the formulae and cell references. Although PDF is able to represent both 'layers' of spreadsheets, it is unable to specify the relationships between the two layers.

4.3 XML as a preservation strategy

XML is an abbreviation of eXtensible Mark-up Language, a mark-up language based on text characters used to enrich data with information about structure and meaning. This language – which can also be used as a file format – is an open standard defined by the World Wide Web Consortium, a non-profit organisation that develops interoperable standards such as the specifications, guidelines, software and tools required for the optimum use of the Internet[10].

---

8 *XML: de mogelijkheden en valkuilen voor de overheid* ('XML: the opportunities for the authorities, and the pitfalls') W. Thomas, 19 September 2002.

9 *Regeling geordende en toegankelijke staat archiefbescheiden* (Ministerial Regulation on the Arrangement and Accessibility of Records) February 2002.

10 See http://www.w3c.org

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.

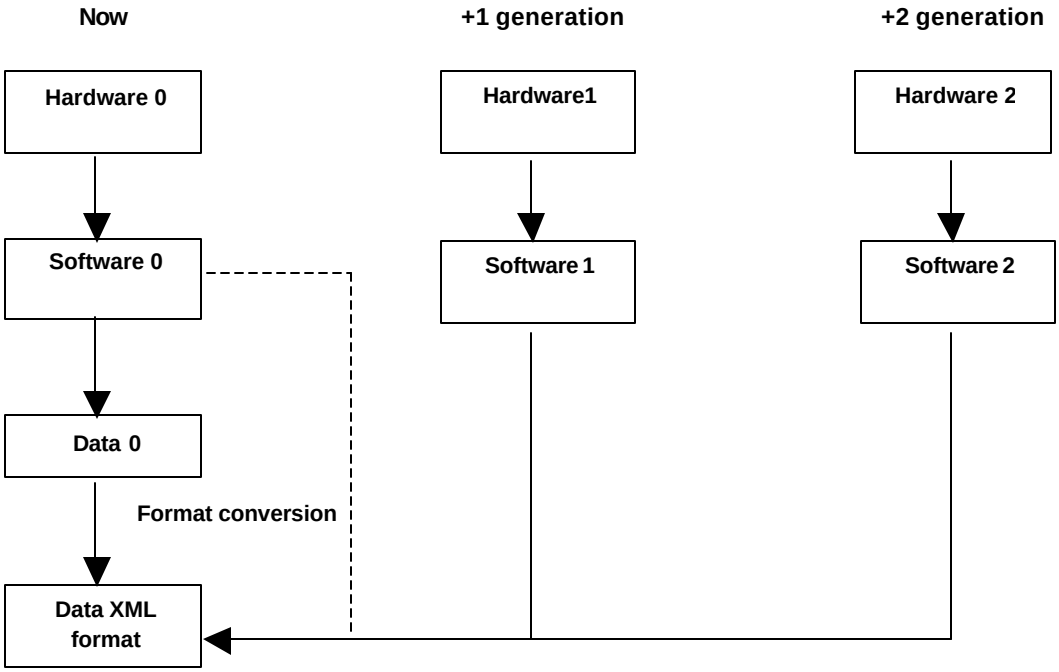


Figure 5:*Conversion to XML involves fewer conversions than migration*

XML is a suitable format for the specification of metadata and the reproduction 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 reproduction of the content and context. XML can also readily render 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 [11].

[11] XML has inherited the DTD mechanism from SGML. However since a DTD (Document Type Definition) can define data types to only a very limited extent and is not XML, it is making way for another standard: W3C Schema. The official name of

The structured and consistent design of XML documents means they are 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 reproduced with XML. Simple behaviour such as hyperlinks and email addresses can be reproduced by means of tagging. More complex behaviour is more difficult to reproduce 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'[12]. Conversely, the Roquade project uses the term 'container' for the 'packaging' of digital records[13]. 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.

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[14] which also uses XML to specify metadata.

---

this standard is XML Schema Definition Language (XSDL), although in practice the names W3C Schema or XML Schema are used.

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

13 "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.

14 See http://partners.adobe.com/asn/developer/xmp/download/docs/MetadataFramework.pdf

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

Is XML suitable for preserving spreadsheets?

XML is an appropriate choice for the long-term preservation of spreadsheets. XML can be used to specify the context, content and structure of spreadsheets. Both the display and the formulae layers are retained, together with the mutual relationships. The secondary XML standards, such as XSL and CSS, make it possible to precisely represent the appearance. Conversion to XML also prevents the activation of behaviour whereby formulae are recalculated.

Testbed experiments have demonstrated that the deployment of XML as a file format is a viable approach to the durable preservation of spreadsheets, in particular when used in combination with a framework approach.

Although a large number of commercial and shareware conversion tools are available, the quality of the XML they produce can vary from tool to tool. 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 create records directly in XML, although at present spreadsheets stored in the XML file format used by Microsoft and Corel still require these applications for their display.

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:

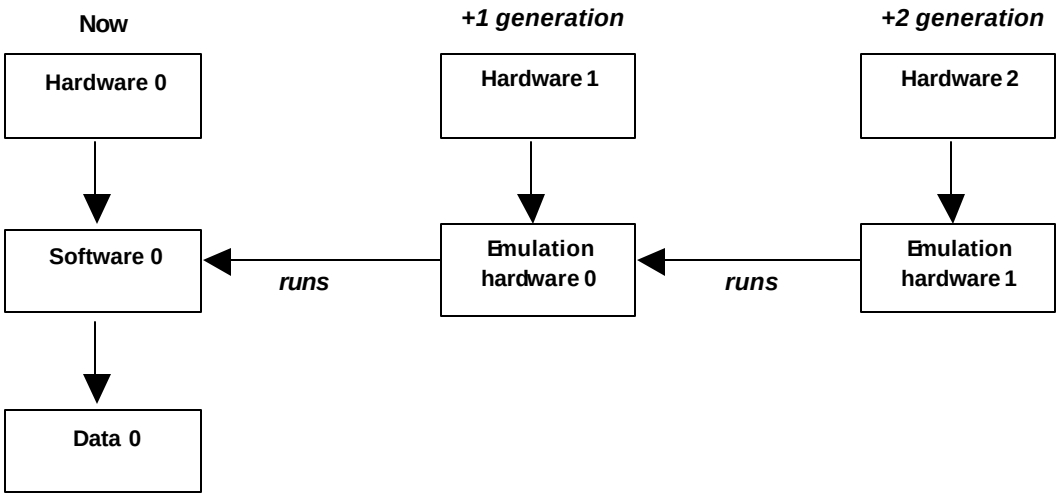


Figure 3: *Basic emulation diagram*

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 [15] 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 theoriginal 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.

[15] Reverse engineering – decompilation: the attempt to track 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)

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, communication between processes, networking, etc.) that are used by programs when they run, and it allows invoking programs to be run (e.g., by double-clicking on their icons). An application program may use these OS facilities to access files, interact with users, or communicate with the network or with other programs, but the application program must always execute on hardware, just as the OS itself does. That is, any program must run on its expected hardware platform, regardless of whether its expected OS is also running on that platform. Computer scientists often say (perhaps confusingly) that an application program "runs on" an OS, but all this means is that it relies on the facilities provided by that OS. It does not mean that the application "runs on the OS" in the same sense that the application runs on hardware. All programs (applications and operating systems alike) must run on hardware. Therefore, implementing an emulator of an OS does not enable us to run application programs, such as rendering programs, without also having the appropriate hardware platform - either as a physical computer or as a software-emulation-of-hardware (which, of course, must itself run on some physical computer).

Is hardware or software emulation suitable for preserving spreadsheets?

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 preserve the emulator together with the operating sys tem, the original application program, and the files created with this application program. The disadvantages of this strategy are the technical complexity and the 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 and that the original 'look and feel' will remain intact.

Taking the complexity of this strategy into consideration, hardware emulation is profitable only if this strategy is not chosen for individual record types[16], but for all files generated from a specific computer generation.

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 make 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 coverts the original data files into a Logical Data Description (LDD) using a program written in the UVC programming language. This LDD is a stand-alone, self-descriptive and explicitly structured data format which contains all the information required for the future re-assembly of the digital record.

---

16 As indicated above, hardware emulation involves the imitation 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.

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 describe the data in a simple manner that is independent of a specific technology, so that it can then be processed via a viewer.

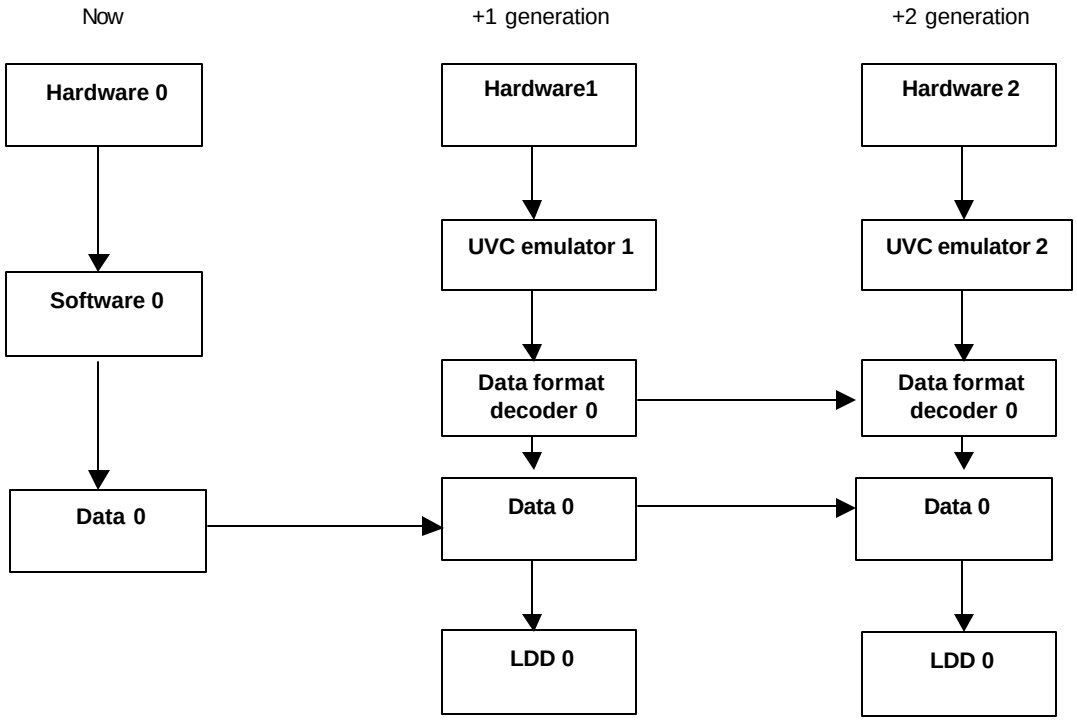


Figure 4:*Diagram of the Universal Virtual Computer*

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 by running the UVC JPEG program is not saved. The LDD is rendered using a viewer. The format and the structure of the Logical Data Description are designed in such a manner that it will be simple to write a viewer at some point in the future. Where necessary, new viewers can be developed for future computer platforms.

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

The next phase of the UVC development will be to classify 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 (for the generation of the Logical Data Description). It will also be necessary to write a new emulator for each generation of hardware that differs from previous generations to such an extent that the old UVC emulator can no longer run on the hardware with the requisite reliability.

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

Other forms of UVC

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

Is UVC data preservation suitable for preserving spreadsheets?

In conceptual terms the application of the UVC data preservation strategy to the preservation of spreadsheets is appealing. The UVC would be able to preserve the content of the cells, the formulae, the comments, and the formatting – and, if so required, even the behaviour.

The complete process is shown in diagrammatic form in the following figure. An intermediate format is generated from Excel and Lotus spreadsheets that contains the relevant information extracted from spreadsheet's original programming environment (through LotusScript or Visual Basic).

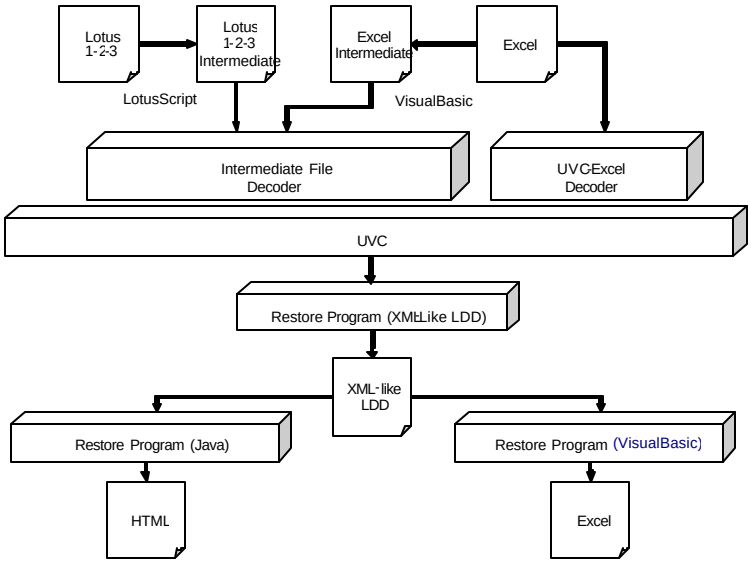


Figure 5:*Basic diagram of the UVC approach for spreadsheets*

UVC data preservation is still in an experimental phase, and the development tools for the implementation of the data format decoder programs are still very basic: at present, the programs are still developed in assembly language! Ultimately it will be possible to implement all the logic for a data format decoder program written entirely in UVC code. A first step in this direction is the UVC Excel-format decoder program that directly implements a subset of the Logical Data Description (LDD) for Excel.

The Intermediate File Decoder and the UVC -Excel decoder generate the LDD for Lotus 1-2-3 and Excel spreadsheets. The LDD is comprised of a hierarchy of labelled elements in which the labels and their position in the hierarchy specify all the relevant information in the original spreadsheet. Appendix B contains a detailed technical specification of the operation of the UVC.

The results from the proof of concept Testbed carried out for the UVC and spreadsheets showed potential. A successful design of an LDD schema was made which specified virtually all the essential characteristics of a spreadsheet. However, the proof of concept also revealed that the UVC is not yet complete, and that further development will be required. One aspect in need of further development is the reverse engineering of the original file format, which is required for the development of a data format decoder program capable of retrieving all the essential elements from the relevant file.

4.3 Conclusions

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 as yet no emulators available for digital preservation.

The two 'proofs of concept' 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 the development of data format decoder programs capable of converting the original files into logical data descriptions.

Interoperability is not a reliable strategy for preserving spreadsheets.

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

At present, XML is the most effective strategy for the durable preservation of spreadsheets. XML is highly capable of representing the context, content, appearance and structure. 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 spreadsheets

Chapter 4 discussed and compared the various preservation strategies against the record-type 'spreadsheet'. It was concluded that the recommended preservation strategy involves the use of XML. This chapter explains how an XML strategy can be implemented.

5.1 Introduction

Although XML is the best strategy for the long-term preservation of spreadsheets, migration is nevertheless an alternative for organisations that create spreadsheets which only need to be preserved for a short period of time. Organisations that create a variety of spreadsheets, 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 spreadsheet records.

5.2 Short-term preservation of spreadsheets

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

5.3 Conversion and migration procedures

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

5.3.1 Backward compatibility

It has become apparent that backward compatibility is only suitable as a short-term (i.e. less than ten years) strategy for preserving spreadsheets in an authentic state. Since new versions of software support only a restricted number of older generations of file formats, spreadsheets 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 Excel 97 to Excel 2000). Experiments at Testbed have shown that migration over different versions of an application can sometimes deliver better results than migration through each and every 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 spreadsheets to the required version.

5.3.2 XML

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

For groups of spreadsheets based on a template, for example estimates or monthly reports, it would be worthwhile using an XML Schema or a DTD[17] to specify their structure. It would then be possible to verify that the created records comply with the specified structure.

The appearance can be represented using an XSL stylesheet, supplementary XML elements or attributes in the XML file itself, or by means of an image. XML elements or attributes can be appended to, for example, the spreadsheet's cells. These elements or attributes refer to the colour of the font that is used, the cell's background colour, and the alignment etc.
Moreover, it is also possible to create an image of a worksheet.

Finally, XSL can be further sub-divided into two components: 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 be more readily used to render 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 representing the document's appearance including definition of the page layout and the text format. Alongside XSL, XML also offers the possibility to make use of Cascading StyleSheets (CSS). CSS is primarily intended for the specification of the appearance of web documents. XSL, one of the newest members of the XML family, was adopted by W3C in October 2001. As a result, XSL is not yet used or supported on a large scale.

Charts can be preserved by recording an image of each chart. It is advisable to preserve the chart's properties by means of extra elements in the XML file. For example, XML elements could be used to specify the chart type, the subtype, the title, the names of the X-axis and the Y-axis, and the source data used to construct the chart. The XML schema must specify explicitly the worksheet(s) to which the charts belong. Images and other inserted objects can be preserved using a similar approach.

Although a large number of commercial and shareware conversion tools are available, the quality of the XML they produce can vary. However, there are a growing number 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 create records directly in XML, although spreadsheets stored in the XML file format used by Microsoft and Corel still require these applications for their display. The same restriction is also applicable to Open Office.

[17] A DTD can define data types to only a very limited extent, and is not itself XML. For that reason the DTD is more and more being superseded by another standard: XML Schema. See www.w3c.org/TR/xmlschema-2/

5.4 Long-term preservation of spreadsheets

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

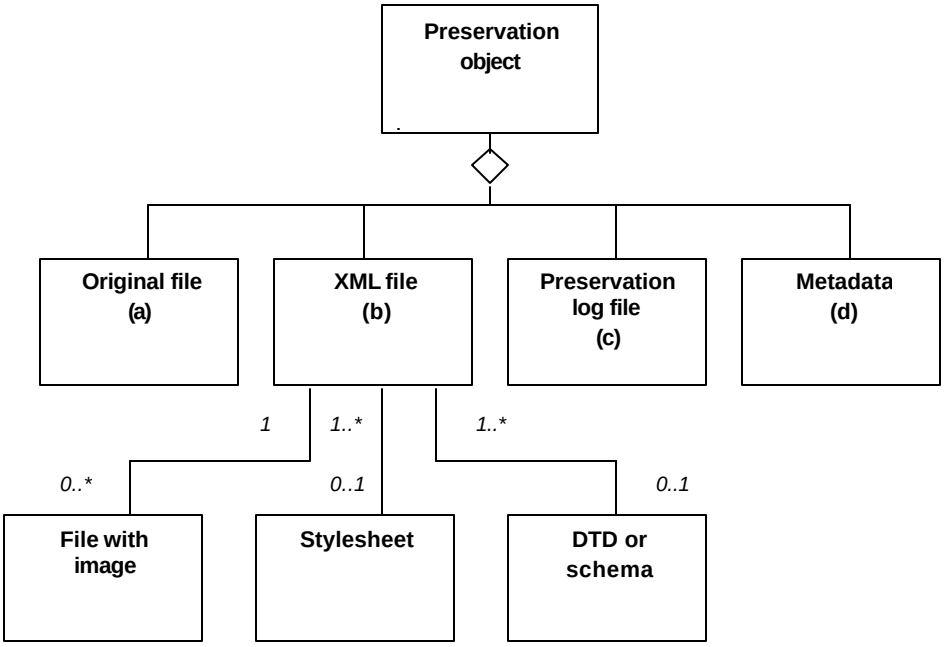


Figure 6: *Structure of the preservation object*

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

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

Each component is discussed in more detail below.

Original file (a)
Testbed recommends that the original file is preserved alongside the XML file. 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 can still be used.

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 neither necessary nor efficient to save a stylesheet for each individual XML file.

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

Spreadsheets can incorporate charts and other inserted objects, etc. On the conversion of the spreadsheet into XML these are stored in separate files, for example as PDF or JPEG files. References will be required to specify their relationships with and positions in the XML file. The records manager can implement the measures required for such images according to the specific file format, whereby different preservation requirements will govern each type of digital record.

Preservation log file (c)
The preservation log file contains all information about the preservation actions undertaken on the spreadsheet. 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 spreadsheet to XML. It is important to ensure that the preservation log file can be updated readily and continuously without overwriting earlier data. A database can be suitable for this purpose; consideration can again be given to the use of XML. The initial content of the preservation log file must be comprised of the data in the original digital record, in this instance the spreadsheet. 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 spreadsheet's new format.

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

Metadata (d)
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 spreadsheets. Now it is up to organisations to make use of this information. Chapter 5 dealt with the implementation of the XML-strategy. The various activities that an organisation has to undertake to successfully achieve this are so specific and different from each other that they justify an approach oriented towards 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 spreadsheets* 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 spreadsheets in particular. Digital Preservation Testbed has tested preservation strategies for the record type 'spreadsheet'. The best way of preserving spreadsheets at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long-term preservation of spreadsheets: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including spreadsheet programs. 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 spreadsheets 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 spreadsheets, 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 spreadsheets 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 spreadsheets
- agreements on the management and preservation of spreadsheets

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

Inform all staff about the policy and the procedures. Train all staff in the use of the spreadsheet program and when applicable, the use of templates in your organisation. A spreadsheet 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 spreadsheets* 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 spreadsheets in particular. Digital Preservation Testbed has tested preservation strategies for the record type 'spreadsheet'. The best way of preserving spreadsheets at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long-term preservation of spreadsheets: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including spreadsheet programs. 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 spreadsheets 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 spreadsheets 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 long-term preservation of spreadsheets? 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 spreadsheets constitute archival records
Not all spreadsheets received or created by government agencies are records for the archives in the sense of the 1995 Archives Act. Only spreadsheets which have played a role in a business process are deemed to constitute records for the archives. Consequently spreadsheets 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 spreadsheets 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 spreadsheets 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.

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

Determine which metadata are required
Specific information about each spreadsheet 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[18].

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.

[18] 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.

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

Make sure that spreadsheets 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 spreadsheets.

Specify the manner used for classification and filing
A classification system (as identified above) is used to assign a spreadsheet 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 spreadsheets
The access possibilities are closely related to the selection of the storage format and the quality of the metadata. In general, spreadsheets 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 spreadsheets are converted to XML
At present the best approach for the storage of spreadsheets that must be considered for long-term preservation involves the use of XML.

Make sure that the policy is evaluated at regular intervals
Information technology changes rapidly – and the same is true for organisations. The requirements for digital archiving are also developing. For these reasons the policy must be 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 durable preservation of digital records. This is why Testbed also advocates the preservation of the original file.

6.3 Action plan for ICT specialists

Introduction
In reading the publication *From digital volatility to digital permanence: Preserving spreadsheets* 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 spreadsheets in particular. Digital Preservation Testbed has tested preservation strategies for the record type 'spreadsheet'. The best way of preserving spreadsheets at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long-term preservation of spreadsheets: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including spreadsheet programs. 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 spreadsheets 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 spreadsheets, 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 spreadsheets 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 spreadsheet program used by your organisation.

Concrete actions for ICT specialists
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 spreadsheets that must be preserved in a centrally managed system, not on the computer or the personal folders of the individual users. This prevents the accidental or deliberate deletion of spreadsheets. Access to the centrally-stored spreadsheets can be controlled to ensure that the information remains accessible to those who need it, and to prevent unauthorised access. A central system also provides for the safeguarding and management of the storage media, usually a combination of disks and tapes. This also extends to making copies and backups. It is important to remember that, within the context of digital preservation, there is a world of difference between the storage of backups and the sustainable preservation of digital records, including spreadsheets.

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 spreadsheets, particularly in support of search functions.

Make sure that the preservation system supplements each stored spreadsheet 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 and version of the spreadsheet program, the version of the preservation system that is used, and an overview of any preservation actions carried out on the spreadsheet such as the date and the time at which the spreadsheet was accessioned into the preservation system. See Appendix A for more information about the recommended content of the Preservation Log File.

(b) Recommended format and possibilities for implementation
A detailed description of the strategy Testbed recommends for the preservation of spreadsheets 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 spreadsheets. 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

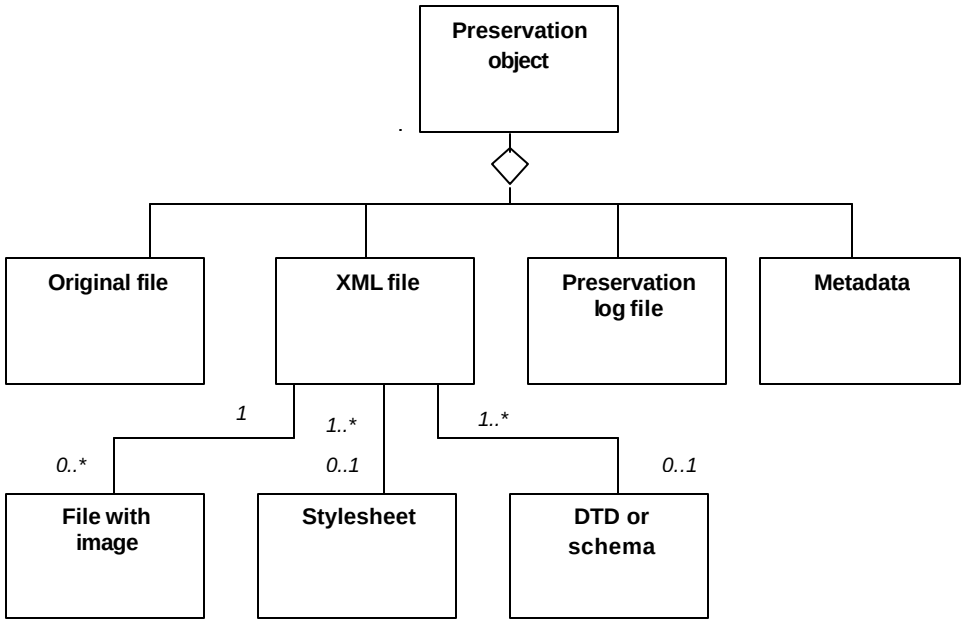


Figure 7 *Structure of the preservation object*

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

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

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

Many of these functions are included in Records Management Applications (RMAs). Software of this type usually offers opportunities to configure and adapt the RMA 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 [19]. Attention is expressly drawn to one element of these specifications, namely the section on the export of digital records from the RMA[20]. Depending on how long the digital records are preserved by the original organisation before they are transferred to the 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.

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

[20] 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 spreadsheets* 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 spreadsheets in particular. Digital Preservation Testbed has tested preservation strategies for the record type 'spreadsheet'. The best way of preserving spreadsheets at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the long-term preservation of spreadsheets: from the line managers, records managers and ICT specialists to the end users who have office applications at their disposal, including spreadsheet programs. 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 spreadsheets 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 spreadsheets. In so doing, you determine to a great extent whether your organisation is capable of the long-term preservation of spreadsheets. Your organisation will have laid down policy, agreements and procedures governing the creation of spreadsheets, for instance on the use of templates to create official records in 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 spreadsheets – 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 spreadsheet program, 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 Excel or Lotus 1-2-3 offer such a range of different functionalities.

Where possible, make use of templates or worksheet templates such as those supplied by your organisation. Excel contains a wide range of templates. In general, each organisation will prepare its own templates. The major benefit offered by the use of well-designed templates is the creation of uniform spreadsheets that, for example, use meaningful names for rows and columns which ensure for the ability to understand the spreadsheets both now *and* in 10 to 50 years' time. Furthermore, templates can also ensure metadata is incorporated in the spreadsheet. Finally, the use of templates is also beneficial to the quality of the XML generated on the conversion of spreadsheets into this more durable file format.

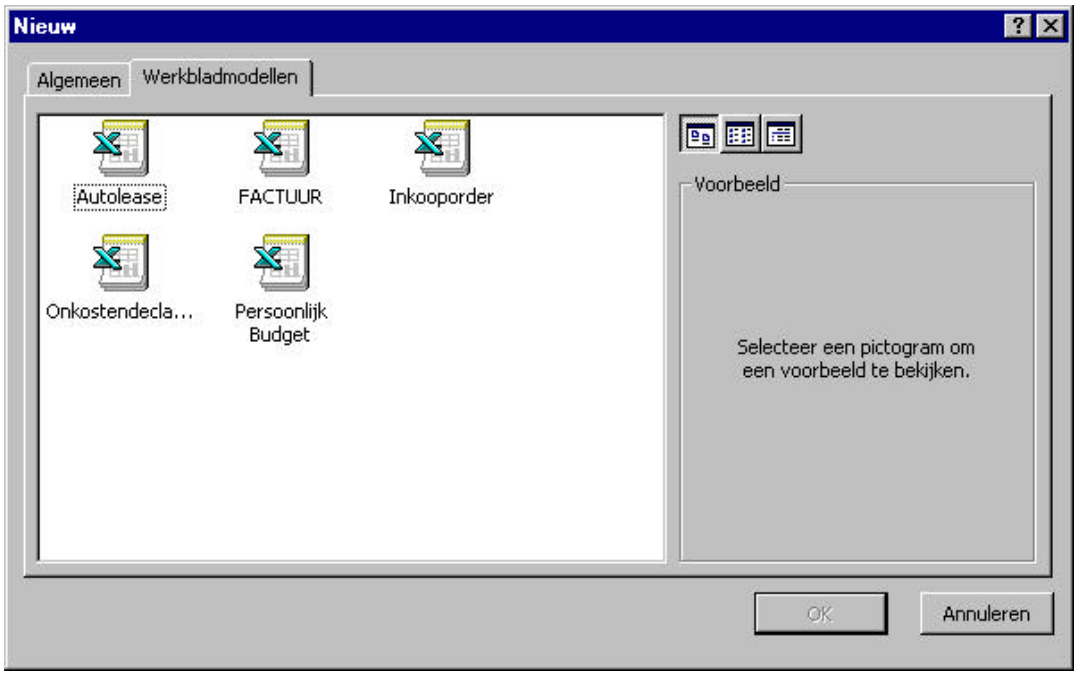
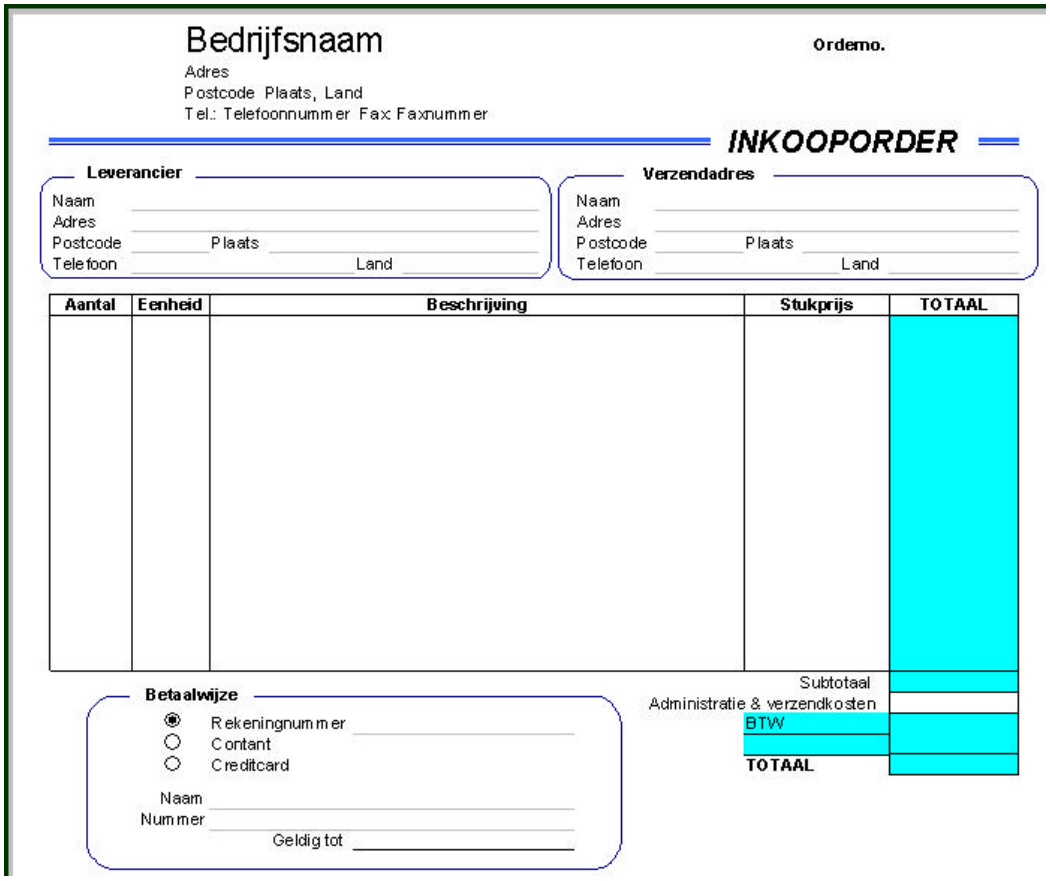


Figure 8: *Examples of Excel worksheet templates*



Bedrijfsnaam

Adres
Postcode Plaats, Land
Tel.: Telefoonnummer Fax Faxnummer

Ordernr.

INKOOPORDER

Leverancier

Naam
Adres
Postcode Plaats
Telefoon Land

Verzendadres

Naam
Adres
Postcode Plaats
Telefoon Land

Aantal	Eenheid	Beschrijving	Stukprijs	TOTAAL
			Subtotaal	
			Administratie & verzendkosten	
			BTW	
			TOTAAL	

Betaalwijze

- Rekeningnummer
- Contant
- Creditcard

Naam
Nummer
Geldig tot

Figure 9: *Example of the Purchase Order worksheet template*

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, etc.

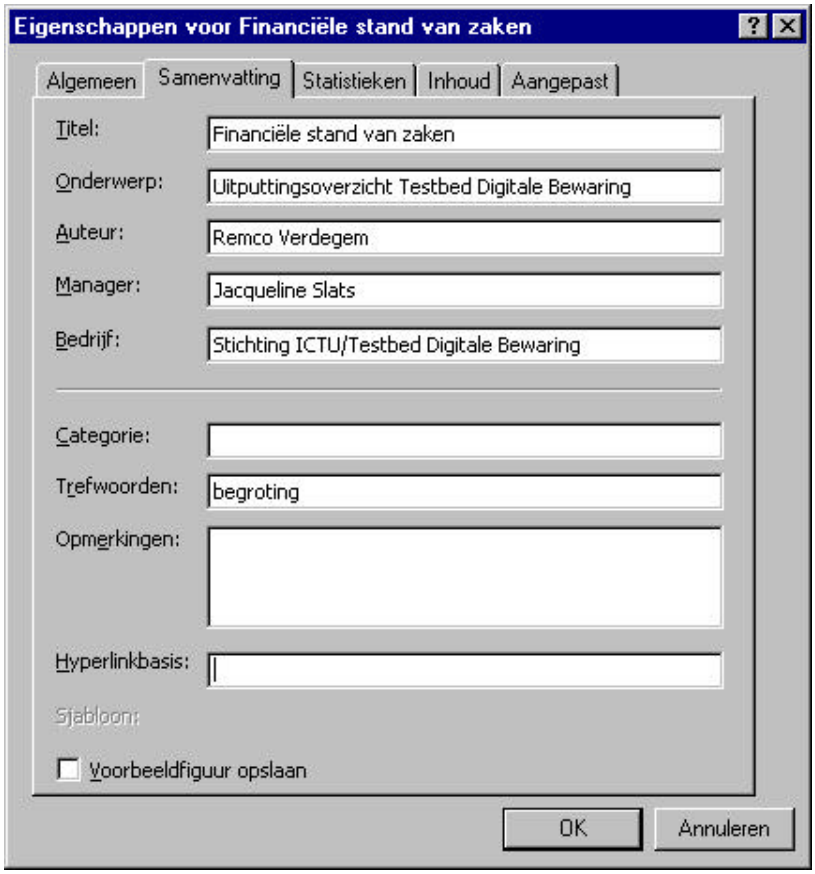


Figure 10: *Properties window*

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. However, this is accompanied by the risk that old context information retained in the properties window of the original record will be saved as part of the new record and, in so doing, becoming an official component of that record. Accurate records of the spreadsheet properties must be made when saving that spreadsheet, especially when making use of existing spreadsheets.

Do not use passwords to protect spreadsheets. Excel offers a number of options for the protection of workbooks and worksheets: you can restrict access to individual worksheets; you can protect entire workbooks from changes; you can protect shared workbooks and secure the change history (a separate workbook that contains information about changes that have been made in a shared workbook and then stored, such as who made the changes, when and where the changes were made, and which data was deleted or replaced), and you can restrict access to a workbook by setting a password. This password will then need to be entered before the workbook can be opened and saved.

Setting a password governing the opening of a spreadsheet is a most inadvisable option. 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 it is possible to set a password for editing the document without affecting the digital durability of the record. Only users issued with the password can then implement changes to the spreadsheet. All other users can only read the spreadsheet, and not change it.

Preference is given to the use of 'standard' fonts
Avoid the use of unconventional fonts, since fonts of this nature reduce the probability of the successful preservation of spreadsheets. Unusual fonts can be lost in a migration.

Use headers and footers to include (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.

Assign meaningful names to rows and columns (titles or labels)
In the absence of meaningful names for the rows and columns it will be difficult to determine what information is presented in the spreadsheet – and certainly after a period of time.

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 the date notation for the cell format in which the full name of the month is shown, for example 10 January 2003 rather than 10-01-2003. In the United States 10-01-2003 will be understood as 1 October 2003.

Avoid automatic date and time functions, such as "=NOW()"
The use of automatic date and time functions is very popular. The result of the NOW() function is the current date and time in the form of a serial number that is automatically recalculated each time the file is opened. This is not desirable from a record-management perspective, and consequently it is recommended that these automatic fields are not used.

State the currency in a separate cell, and in full
When making use of currency amounts, state the relevant currency in the title or name of the column, as shown in the second column in the following column. An integrated currency symbol as used in the last two columns (note the difference in format!) can be lost on migration or – and even worse– replaced by a different currency symbol.

date	amount, Euros	amount	amount
1 January 2001	1000	€ 1.000,00	€ 1.000,00
1 February 2001	910	€ 910,00	€ 910,00
		note: via currency symbol on toolbar	note: via cell properties, currency

Save the spreadsheet in the central preservation system
Once the formal spreadsheet has been created, it is recommended to transfer as quickly as possible into a Records Management Application (RMA) or a Document Management System (DMS). Applications of this nature provide for the optimum management of documents and records, including their classification – and can prevent the unintentional or intentional modification or, even worse, deletion, of official spreadsheets.

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 characterset 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. A Basic Selection Document can consist of one or more selection lists.

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 records, for example, by opening an attachment or by activating a hyperlink. The other four attributes are content, context, structure and appearance.

Cell formatting profile

Collection of formatting properties such as character size, font, and alignment, that you can define and save as a group.

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.

CSV

Comma-Separated Values
File format in which the different fields are divided by comma's.

DIF

Data Interchange Format
File format for data interchange (database or spreadsheet).

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.

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.

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.

Integrity

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

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.

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 records and their management through time.

Migration

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

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, default

The template that is used as the basis for an empty workbook and that is opened when you start Excel and click 'New' on the 'standard' toolbar or via File/New and select the template 'workbook' in the dialogue box.

Template

A workbook created and used as the basis for other, similar workbooks. You can make templates for workbooks and worksheets. The settings that you save in a template will determine the following characteristics that are used in subsequent workbooks and worksheets based on the template:

- The number and type of sheets in a workbook
- Cell and sheet layout, according to settings selected in the Format menu.
- Cell formatting profiles
- Page format and settings for the print area of each sheet
- Text that you want repeated in each new workbook or worksheet, such as page headings, and column and row labels
- Data, formulae, images or other information that you want to insert in each new workbook or worksheet.

URL

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

Viewer

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

W3C

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

Wrapper

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

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/

Bibliography

Boudrez, Philip	*<XML/> en Digital Archiveren (2002)* http://www.antwerpen.be/david/teksten/xml_digitaalarchiveren.pdf
Boudrez, Philip	*Standaarden voor digitale archiefdocumenten (October 2001)* http://www.antwerpen.be/david/teksten/DAVIDbijdragen/Standaarden.pdf
Giesbers, Saskia Feeney, Mary (Ed)	*Records Management Terminologie (6 March 2002)* http://www.rmconventie.nl/publicaties-rm/RecordsManagement_termen_en_definities.pdf Digital Culture: Maximising the Nation's Investment (National Preservation Office UK, 1999)
Hofman, Hans	Een uitdijend heelal? Context van archiefbescheiden. Stichting Archiefpublicaties, Jaarboek 2002.
Hofman, Hans (redactie)	Blijvend in business, Naar een geordende en toegankelijke staat archiefbescheiden (ex art. 12 Archiefbesluit) (Den Haag 2003).
InterPARES Project	Authenticity Task Force Final Report (2002) http://www.interpares.org/book/interpares_book_d_part1.pdf
InterPARES Project	Preservation Task Force Final Report (2002) http://www.interpares.org/book/interpares_book_f_part3.pdf
Lorie, Raymond	*A Project on the Preservation of Digital Data* http://www.rlg.org/preserv/diginews/diginews5-3.html
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* http://www.library.tudelft.nl/e-archive/Documenten/Resultaten/reportone13.pdf
Mellor, Paul et al	*Migration On Request, a Practical Technique for Preservation (2002)* http://www.si.umich.edu/CAMILEON/reports/migreq.pdf
Ploeg, Dr. F. van der	*Regeling geordende en toegankelijke staat archiefbescheiden (February 2002)*
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, Den Haag, 2000)*

Rothenberg, Jeff & Bikson, Tora	*Carrying Authentic, Understandable and Usable Records Through Time (1999)* http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf
Rijksarchiefinspectie	*Wet- en regelgeving* www.rijksarchiefinspectie.nl/wetgeving/
Testbed Digitale Bewaring	*Migration: Context and current status (December 2001)* http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf
Testbed Digitale Bewaring	*XML and Digital Preservation (September 2002)* http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf
Testbed Digitale Bewaring Thomas, Wimpe	*Emulatie: Context and current status (June 2003)* http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf *XML: de mogelijkheden en valkuilen voor de overheid; 19 september 2002.*
VERS	*Victorian Electronic Records Strategy Final Report* http://www.prov.vic.gov.au/vers/published/final.htm
Zuurmond, A, Mies, K.	*Winst met ICT in uitvoering; Zenc, Den Haag, June 2002.*

Appendix A 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. Excel) + hardware environment (e.g. Pentium 4) + operating system (e.g. Windows XP);
- Details of interim formats (e.g. DIF, CSV);
- Details of new computing environment (sufficient details must be recorded to ensure access to the records in their current format).

Preservation action metadata

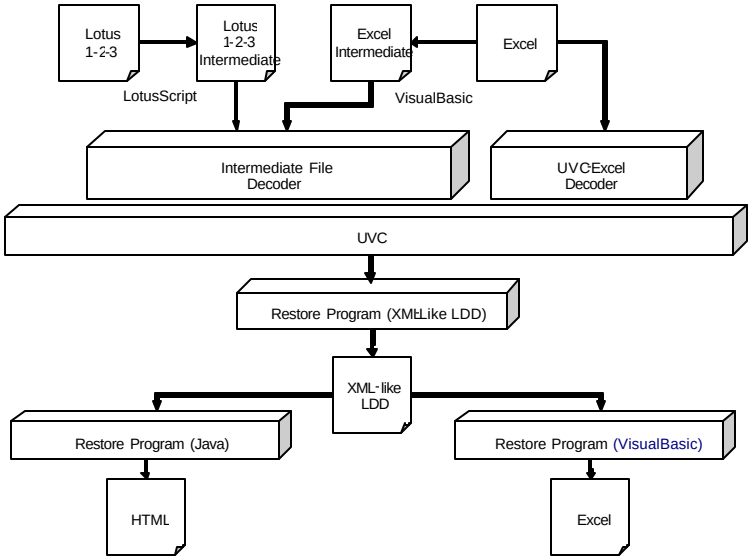
- Date and time of any and all preservation action;
- Person(s) responsible for of any and all preservation action;
- Details of the transformation (conversion) software and;
- Conversion results.

Metadata which refer to the access of the records

- Privileges/rights and;
- History.

Appendix B UVC technical specifications

Overview of UVC data preservation
The UVC data preservation process for spreadsheets is shown in the following diagram. An intermediate format is generated from Excel and Lotus spreadsheets that contains relevant information extracted from spreadsheet's original programming environment (via LotusScript or Visual Basic). The UVC data preservation approach is still in an experimental phase, and the development tools for the implementation of the data format decoder programs are still very basic: at present, the programs are still developed in assembly language! Ultimately it will be possible to implement all the logic for a data format decoder program that is written entirely in UVC code. A first step in this direction is the UVC Excel-format decoder program that directly implements a subset of the Logical Data Description (LDD) for Excel, and consequently without the use of the aforementioned intermediate file.



The Intermediate File Decoder and the UVC -Excel decoder generate the LDD for Lotus 1-2-3 and Excel spreadsheets. The LDD is comprised of a hierarchy of labelled elements in which the labels and their position in the hierarchy specify all the relevant information in the original spreadsheet. The following description specifies the top layers of the LDD spreadsheet hierarchy. The initial element is a spreadsheet (label 10), which indicates that the spreadsheet has a name (label 15) and is comprised of one or more worksheets (label 20), a set of labelled cells, and the metadata associated with the spreadsheet. The entire listing of the LDD used for spreadsheets is enclosed in Appendix C.

ELEMENT 10 [Spreadsheet] (15, 20+, 30+, 40?)

ELEMENT 15 [SpreadsheetName] (CHAR)

ELEMENT 20 [Worksheet] (50, 80+)
ELEMENT 30 [LabelledCell] (90, 50, 170)
ELEMENT 40 [Metadata] (350?, 360?)

The interface of the UVC passes the various labelled elements, in a standard format, on to the Restore program via the memory interface. Each element is comprised of three sections: the label code (tag), the length of the value string (in number of characters), and the value string.

TAG ELEMENT (16 bits == 2 bytes)	**NUMBER OF CHARACTERS (32 bits == 4 bytes)**	**CHARACTERS (16 bits per char == 1 byte)**

The tag and length of the string are represented by numbers in 16 and 32 bits. The value string is coded in Unicode UTF-16 format. The 'little endian' sequence is employed for all bytes; i.e. the most significant (or highest) byte of a number or UTF-16 character code is transmitted first, followed by the remaining bytes, in sequence.

A basic restore program then translates this data into an XML-like structure, whereby each label is assigned a beginning tag and a closing tag comparable to XML. The entire process is shown in the following example:

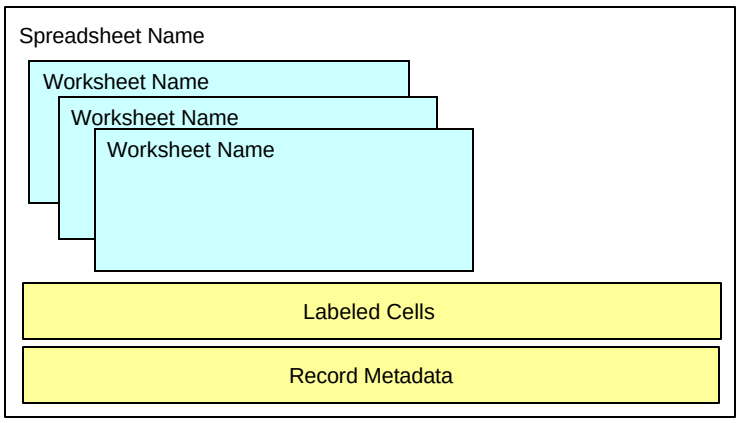
10 3 ABC
20 4 DEFG

which the basic restore program transforms into

<TAG1>ABC
    <TAG2>DEFG</TAG2>
</TAG1>

using the following LDD definitions:
ELEMENT 10 [TAG1] (20+)
ELEMENT 20 [TAG2]

The XML-like translation of the LDD contains, in schematic form, the element categories as shown in the following diagram. This XML-like representation of the spreadsheet is then used either for the conversion into Excel or for the representation of the spreadsheet data in an HTML browser.



The UVC data preservation approach is based on the principle th at the LDD that is generated will be independent of future prevailing *de facto* technology standards. The only technology-dependent component is the UVC itself, which will need to be implemented on each platform. The UVC is the invariant that ensures that i t will always be possible to use the data format decoder program. On its own, the LDD is a

sufficient interpretation of the relevant information from original format, independent of specific technology. Consequently the last steps (translation to XML-like Excel and HTML formats) are not essential for the durable preservation of digital records; self-evidently it will also be possible to convert to the *de facto* standard formats required for future prevailing platforms.

Interpretation of the Logical Data Description

It is essential that conversions to logical descriptions result in descriptions that can readily be interpreted in a unique manner. This is of vital importance to both UVC data preservation and XML-driven solutions employed in the long-term preservation of digital records. Often insufficient attention is devoted to the specific preconditions that need to be satisfied if an interpretation is to be successful.

The preconditions to be satisfied for the interpretation of a LDD for spreadsheets are identified below, on the basis of the above description of the entire UVC data preservation process.

Natural language

The relevant information of the format is displayed in natural language, independent of the technology, such that the comprehension of the natural language is as such a basic requirement. However, this is applicable to the provision of all information. The majority of LDDs are written in English.

Logical Data Schema

The Logical Data Schema (LDS) defines the correct construction of LDDs for spreadsheets. A number of options are available for the preservation of the LDS:

1. Make a printout of the LDS, and store the printout;
2. Use a generic LDD for all LDS specifications, as a result of which it will be necessary to preserve one unique LDS for all the other LDS specifications.

Without the LDS it will be impossible to interpret the LDD generated by the data format decoder program, since this returns solely numbers for the label codes/tags.

Unicode UTF-16

The format decoder returns all values associated with a label in Unicode (UTF-16) format. UTF-16 decoding is currently the worldwide standard for the electronic specification of characters. This standard has global support, and the specification is widely accessible. However from a long-term perspective nothing is eternal; consequently it will be necessary to ensure, should another character format become the *de facto* standard, that the UTF-16 specification is preserved and remains accessible.

Big versus Little Endian

The storage of a data element in a byte-oriented computer is always accompanied by the problem of the sequence to be employed for elements longer than one byte. For example, the decimal number 369 is represented in the form of 0000000101110001 when stated as a 16-bit binary number (2 bytes). Two options are available for the sequence of the bytes:

1. Big Endian
   The bytes are transmitted from left to right, i.e. beginning with byte (00000001) and followed by (01110001).
2. Little Endian
   The bytes are transmitted from right to left, i.e. beginning with byte (01110001) and followed by (00000001).

It will be self-evident that the incorrect interpretation of the sequence will yield a completely different result: what should have represented as 369 will, in the event that the bytes are exchanged, by interpreted as 28929 (0111000100000001). The UVC spreadsheet decoder uses little endian.

Layout References
References will also need to be incorporated in the LDS for any layout information that is included in the data to be preserved. In the instance of spreadsheets, the LDS also incorporates layout elements such as the font, cell borders, background colours, and patterns. Consequently these are present as specific elements in the LDD, such as Times-Roman font. However, these references are always specific to the application program, such as the lines round spreadsheet cells, or to the operating system, such as fonts. It is virtually certain that it will be impossible to use the current Excel program on the prevailing infrastructure in 2025, and consequently it will then be impossible to interpret layout references of this nature incorporated in the LDD. One possible way to avoid this problem would be to preserve images (bitmaps) of the current layout characteristics indexed by the possible values of the corresponding items of the LDS. This will ensure for the continued availability of an example of the appearance of a specific line pattern used in Excel 2003.

Functions in Formulae
Spreadsheets use formulae to calculate cell values, as well as standard computational functions enclosed with the application, such as accounting, mathematical and statistical functions. The LDD for spreadsheets also includes specifications of the formulae and the call procedure used for the functions incorporated in the program.

From the above summary it will be evident that the durable preservation of a digital object involves more than merely defining an LDS and implementing a specific UVC data format decoder program. It is also necessary to make a thorough review of the additional preconditions governing the interpretation of the LDD. The extent to which this will be necessary will depend on the estimation of the risks and the importance of the preserved objects. It is possible that many organisations will be of the opinion that, for example, their duties do not extend to the preservation of UTF-16 format definitions, and that it is unlikely that ICT specialists will be unable to interpret them at some point further in the future – a standpoint based on the wide support for Unicode. Although this can be a totally legitimate standpoint any decision of this nature must be taken after careful consideration of the situation – and not as a result of a lack of an insight into the relevant preconditions.

Appendix C Schema Logical Data Description

ELEMENT 10 [Spreadsheet] (15, 20+, 30+, 40?)

ELEMENT 15 [SpreadsheetName] (CHAR)

ELEMENT 20 [Worksheet] (50, 80+)

ELEMENT 30 [LabelledCell] (90, 50, 170)
ELEMENT 40 [Metadata] (350?, 360?)

ELEMENT 50 [WorksheetName] (CHAR)

ELEMENT 80 [Cell] (130, 150, 160, 170, 180?, 190?, 200?, 210?)

ELEMENT 160 [CellValue] (235?, 240?, 250?, 255?)

ELEMENT 235 [Error] /* Currently empty tag indicating some error in formulae evaluations */
ELEMENT 240 [Float] (260, 257)
ELEMENT 250 [Text] (260+)
ELEMENT 255 [Date] (260, 258)

ELEMENT 257 [ExactFloatValue] (600, 610, 620, 630)
ELEMENT 258 [ReferenceDateFormat] (640,650, 660)

ELEMENT 260 [FormattedString] (270, 280?, 290?, 300?, 310?, 530?, 330?)
ELEMENT 270 [String] (CHAR)
ELEMENT 280 [FontType] (CHAR)
ELEMENT 290 [FontSize] (CHAR)
ELEMENT 300 [Underlined] (CHAR)
ELEMENT 310 [Style] (CHAR)
ELEMENT 320 [RotationAngle] (CHAR)
ELEMENT 325 [Orientation] (CHAR)
ELEMENT 330 [Strikethrough] (CHAR)

ELEMENT 90 [Label] (CHAR)
ELEMENT 170 [CellPosition] (120, 140)

/* The document metadata contained in the spreadsheet */
ELEMENT 350 [DocumentMetadata] (370?, 380?, 390?)
ELEMENT 370 [DocumentCreator] (CHAR)
ELEMENT 380 [DocumentCreationDate] (258, 690)
ELEMENT 390 [DocumentModificationDate] (258, 690)

/*The record metadata supplied in an additional file */
ELEMENT 360 [RecordMetadata] ……..

ELEMENT 210 [CellFormat] (400, 410, 420, 320?, 325?)
ELEMENT 400 [CellShading] (530, 440)
ELEMENT 410 [Alignment] (670, 680, 685)

ELEMENT 200 [Formula] (CHAR)

/* Border characteristics of the cell */
ELEMENT 420 [Border] (450?, 460?, 470?, 480?, 490?, 500?)
ELEMENT 450 [Left] (510)
ELEMENT 460 [Right] (510)
ELEMENT 470 [Top] (510)
ELEMENT 480 [Bottom] (510)
ELEMENT 490 [BottomLeftDiagonal] (510)
ELEMENT 500 [BottomRightD iagonal] (510)
ELEMENT 510 [Line] (520, 530)
ELEMENT 520 [LineStyle] (CHAR)

/* Color coding into 8 bits (0..255) RGB values */
ELEMENT 530 [Color] (540, 550, 560)
ELEMENT 540 [Red] (CHAR)
ELEMENT 550 [Green] (CHAR)
ELEMENT 560 [Blue] (CHAR)

ELEMENT 440 [Pattern] (CHAR)

/* Cells can be merged into a rectangle*/
ELEMENT 180 [CellSpanned] (570, 580)
ELEMENT 570 [NumberOfColumns] (CHAR)
ELEMENT 580 [NumberOfRows] (CHAR)

ELEMENT 190 [Comment] (590, 260+)
ELEMENT 590 [Author] (CHAR)

ELEMENT 130 [Width] (257)
ELEMENT 150 [Height] (257)

ELEMENT 120 [ColumnNumber] (CHAR)
ELEMENT 140 [RowNumber] (CHAR)

ELEMENT 600 [Quotient] (CHAR)
ELEMENT 610 [Remainder] (CHAR)
ELEMENT 620 [Exponent] (CHAR)
ELEMENT 630 [Sign] (CHAR)

ELEMENT 640 [Day] (CHAR)
ELEMENT 650 [Month] (CHAR)
ELEMENT 660 [Year] (CHAR)

ELEMENT 670 [VerticalAlignment] (CHAR)
ELEMENT 680 [HorizontalAlignment] (CHAR)

ELEMENT 685 [Wrapped] (CHAR)
ELEMENT 245 [Type] (CHAR)

ELEMENT 690 [ReferenceTimeFormat] (695,700, 705)
ELEMENT 695 [Hour] (CHAR)
ELEMENT 700 [Minute] (CHAR)
ELEMENT 705 [Second] (CHAR)