From digital volatility
to digital permanence

Preserving email

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

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

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

Email testbed@ictu.nl
www.digitalduurzaamheid.nl

Digital Preservation Testbed *From digital volatility to digital permanence. Preserving email.*

The Hague, April 2003.

ISBN 90-807758-1-9

Contents

Foreword

In the Testbed series *From digital volatility to digital permanence* the part entitled *Preserving email* was not originally planned to be the first. The problems and questions surrounding email preservation gave us the incentive to publish this part first.

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 4, *Preserving email.* Parts 3 to 1 will appear in that order during 2003. The titles of these parts are:

Part 3: Preserving text documents
Part 2: Preserving spreadsheets
Part 1: Cost and decision models/Functional specifications
Preserving databases

This publication is written for all those involved in managing and preserving digital information properly for the government. Testbed has tried to avoid the use of jargon as far as possible, or, when it could not be avoided, to explain it. The activities that the various people in an organisation have to undertake to preserve digital information properly, now and in the future, have been divided up per 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 during 2003 and will complete the series, since it contains extra information about all the parts, such as cost- and decision-models and functional specifications.

This part 4, about preserving email, 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 Digital Preservation Testbed to decide on the most appropriate preservation strategy through practical experiments.

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

Chapter 3 discusses the record type that is central to this publication: email. What exactly is email and which authenticity requirements are relevant? In other words, what criteria should emails meet so as 'not to lose authenticity', so that it is clear to everyone that the email message is what it claims to be.

In chapter 4 you will be shown an analysis of the different preservation strategies that are receiving a great deal of worldwide attention. Testbed assesses these strategies in relation to email.

Chapter 5 then looks at the preservation strategy that has emerged from our research as being the most promising for email: XML (Extensible Mark-up Language). There is also a description of an implementation method for the sustainable preservation of email in XML.

In chapter 6 you will find a concrete plan of action for the various target groups in a government or other organisation, i.e. managers, records managers, ICT specialists and end users. Each target group has its own responsibilities in this plan and this chapter gives them the information to enable them to contribute to building a reliable digital government.

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

Appendix A:	Technical description of the email/XML demo
Appendix B:	XML schema to accompany the email/XML demo
Appendix C:	Preservation Transaction Log File
Appendix D:	Various representations of an email message (HTML representation, transmission file, plain text and XML representation).

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 level. Digital preservation however, is not always getting the attention it deserves. Action is needed because a digital government cannot exist without digital memory.

1.1 Developments in digital government

The Dutch government is increasingly working with digital records. The second Kok government formulated its aim of having 25% of the transactions between the government and the public take place digitally by 2002, an aim that was then easily achieved. In the meantime, the government has set new targets: by the end of 2006, 65% of all transactions between the government and the public must be dealt with electronically. Meeting this target fits the image of a government that is operating effectively, where rules have been simplified and bureaucracy has been reduced to a minimum. This policy, summed up by the Minister for the Interior and Kingdom Relations, Mr Remkes, as Better Management for Citizens and Business will stand or fall with the application of ICT in 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 available in a more complete form and can, for example, be used more than once. *Service* to the public can be delivered faster, and *better.* Take, for example, applying for official documents, or identifying hazardous business in a region (as the province of Friesland does on its website www.fryslan.nl), to inform the public and business more adequately. Finally, working digitally not only provides organisational benefits, but also financial ones. Millions of euros can thus be saved [1].

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

[1] See *Winst met ICT in uitvoering* (Dutch publication: Profit with ICT), 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, including email messages.

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).

Moreover, email is a good example of a 'transient' means of digital communication. When the average employee is asked by the ICT department to empty his mailbox because the limit has been reached, he faithfully does so by clicking a button, without realising that certain emails really should be preserved.

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 Archives Act 1995
Every document that has a function in the performance of a task is in principle a record or an archival document. Formally or informally, all forms of electronic mail from the government are potentially related to its operation and are eligible to be preserved. Email can therefore be a record too.

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 Archives Decree 1995. 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 email. These refer to such matters as retaining metadata on the content, form and structure of a document, and technical data on conversion, migration and storage formats.

Open Government Act (WOB) (1998)
When archived records from government organisations are transferred to an archive depository, they are in principle made public by virtue of the Archive Act 1995. Whilst records are still stored in government organisations, their public status is organised differently. In these cases, the WOB comes into effect. The WOB gives everyone the right to request information from a government body. In this, as in the Archives Act, no distinction is made between the type of information carrier for the document, whether it is on paper or digital.

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

Directive on Email Use for Central Government (2001)[2]
This directive on email use is neither an act nor a regulation. A working group from various departments was established to formulate a set of basic rules for dealing with email. This was necessary as email communications were acquiring increasing status and guidelines were required in which email use could responsibly take place.

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 dealings, 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.

---

[2] *Richtsnoer Emailgebruik t.b.v. de Rijksoverheid*) (Dutch document, Directive on Email Use for Central Government) /Ministry of the Interior and Kingdom Relations; working group on email use, The Hague 2001

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 will be freely accessible on the website www.digitaleduurzaamheid.nl

The Testbed project will deliver 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 documents or examine them. In the context of Testbed we looked specifically at text documents, databases, email messages and spreadsheets. Multimedia documents, digital video and sound can also be digital records, but these remained outside the scope of this study.

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

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

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

These aspects will be developed further in the sections below.

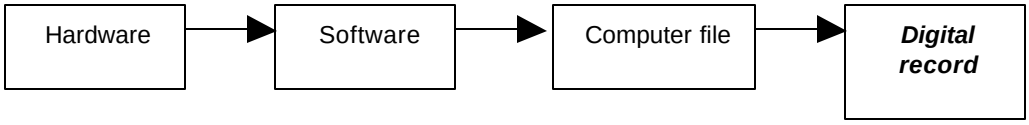
2.2 The digital record as a combination of hardware, software and computer file

In the paper age, the concept of a 'record' was simple. The record as evidence of a transaction was recorded on a physical entity like parchment or paper, possibly in the form of a charter, a receipt, a letter, a memo or a photograph. The evidence was preserved according to fixed and previously determined procedures. Anyone who needed it, contacted the archives and asked for the record.

In the digital age a record is not accessed in the same way. Digital files have to be processed technically before the user can read the records and use them for the

purpose required. It is this dependence on hardware and software that compels us to think differently about the way we make and use digital records.

The diagram below shows the components needed to reproduce and use the digital record[3]:



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

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

2.3 Authenticity as a key concept

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

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

According to the Testbed and Depot 2000 definition, authenticity is 'the representation of a record completely and entirely in accordance with the original recording and function that it was intended to fulfil'. Authenticity has two central concepts:

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

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 person. The authenticity of records can be safeguarded by describing and preserving the original context of the records and by maintaining a chain of unbroken custody;

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.

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

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

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

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

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

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

Context[6]

Context refers to the environment in which the digital record is made. To give the record significance, a certain amount of information about the context is needed. This information only relates to the record, separate from the medium, and does not necessarily include the technical environment in which the record is made and used. The information relates to the business process and the government body in the context of which the digital record is received or made. In addition, the relationship with other records, including those from the same business process, has to be described and preserved. Dossiers are an example of this and also illustrate the concept of external structure

Content

The content is the body of the record, regardless of structure, colour, position or font. In email, content comprises the specific text typed in the message window, known as the body, plus any attachments, and any other inserted or embedded objects. The body of an email usually begins with a salutation and ends with a conclusion (closing remarks, name, etc.).

Structure

The structure of a digital record refers to the structure as it was originally made and reproduced on the screen. This is the logical hierarchy of, and the relationships between, the parts of the record. The structural elements of email are, for example, the headers, the message text and any attachments. The structural elements of a report (a text document) on the other hand can be formed by a cover sheet, a table of contents, chapters (divided into sections and paragraphs) and a bibliography and/or appendices.

It is important that these structural elements are correctly identified and that sections of the email or report are reproduced in the right order. It is also important to know whether there are other essential structural characteristics, for example the presence of footnotes or endnotes in a text document. If this structure is lost as the result of a migration, the record may be reproduced wrongly.

Appearance

The appearance of the digital record refers to the final presentation, what the record looks like when it appears on the screen. Appearance includes aspects such as font and font size. Page- and section-breaks and margin-width can also affect the appearance of digital records. Colour is also a characteristic of appearance; as mentioned above, it can sometimes affect the 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; that which enables us to manipulate and use the record so that new or extra content is displayed. The behaviour can relate to the environment in which the digital record is hosted or placed. If, however, the behaviour is an essential part of the digital record itself and must therefore be preserved, it usually stems from the content of the record. Examples are the inclusion of an email address in a record (e.g. remco.verdegem@ictu.nl) or a reference to a website in the form of a URL like www.ictu.nl

---

6 E*en uitdijend heelal? Context van archiefbescheiden,* (Dutch document: An expanding universe? Context of Records), H. Hofman, Stichting Archiefpublicaties, Jaarboek 2000.

The five characteristics mentioned above, context, content, structure, appearance and behaviour make the digital record special. It is worth noting that the same characteristics are not always equally important for all types of digital records. For example, one can assume that appearance is less important for email messages than for text documents. The characteristics play an important role in the testing of the different preservation strategies that are 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 number of types of metadata can be differentiated, for example:

- Record keeping metadata. This group of metadata focuses mainly on contextual data that give the digital record meaning, but can also contain essential characteristics of the digital record, for example, the presence of a hyperlink to a particular website (a behaviour characteristic).
- Technical metadata. This metadata describes the original technological environment in which the digital record was created and used (hardware and software environment) as well as the new technical environment after a migration for instance.
- Preservation metadata. This group of metadata registers the preservation activities carried out to keep a digital record readable and accessible (for example, a migration that has taken place or a conversion to XML), as well as any changes to the digital record that are the result of these preservation activities.

We can use the 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; it accompanies a digital record throughout its life cycle, and contains information about the creation of the digital record and preservation activities that have been performed. Metadata is therefore vitally important.

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

3. Preserving Email in an Authentic State

Email has quickly secured its place as a fast means of communication, both inside and outside government. Our starting point is that official email must be preserved in an authentic state. To do this, it is necessary to describe the essentials of email and to formulate authenticity requirements.

3.1 The email boom: problems in preservation

The first email was sent in 1971 by the engineer Ray Tomlinson, who worked for the American computer company BNN Technologies. Few people at that time expected email to take off in the way it did. The arrival in the mid-nineteen eighties of the personal computer, the PC, led to an enormous increase in computer usage. In the nineties, email quickly became firmly established as a completely new medium. Email is now frequently used for sending both informal and formal messages throughout the Dutch government. There are now, in 2003, more than half a billion electronic mail boxes worldwide and it is impossible to imagine working life today without email.

People in an organisation create, receive, use and send records to be able to carry out their work; the organisation has to preserve these records and keep them accessible because they might be needed at a later date: for accountability, operations or as a source of knowledge. The introduction of email systems has changed the way in which people communicate inside and outside organisations and therefore has an effect on decision-making processes. Digital communication takes place faster and is more widely accessible.

The benefits of the digital exchange of information are enormous. But the problems involved in preserving sustainable digital records also apply to email. Email is a transient medium that many people think has no official status. There are various factors involved in preserving email that affect one another. They have a bearing on the organisation (or its culture), legal, and technical aspects.

Organisation and organisational culture
Email is a relatively new application for the government and its development is ongoing. As a result, little experience has been gained as regards the long-term preservation of email messages. Email is an outstanding example of a distributed means of communication that is therefore difficult to control. As a rule, it falls outside the normal, registered paper post. Employees often decide for themselves what should and should not be kept and save or delete their email messages at their own discretion. They wrongly view electronic mail as part of their own personal working domain. There are guidelines for using and organising email[7], but it is doubtful whether they are applied properly. If email messages are actually preserved, they are often not preserved correctly: people simply print them out. Part of the context or

[7] For more information on email policy, see: *Archivering van elektronische post. Methoden, meningen en alternatieven,* (Dutch document: Archiving Electronic Mail. Methods, opinions and alternatives), P. Horsman, Amsterdam 1999 and *Richtsnoer Emailgebruik t.b.v. de Rijksoverheid* (Dutch document: Directive on Email Use for Central Government)/Ministry for the Interior and Kingdom Relations; working group on email use, The Hague 2001.

other information is thus lost and the accessibility lessens. A further point is that it is impossible to print multimedia attachments.

Legal aspects
Existing legislation provides a framework for preserving email and other digital records, such as the *Regulation on the Arrangement and Accessibility of Records*. These and other acts were mentioned in chapter 1. The problem is that it is very difficult to translate these regulations into practicalities. There are no or very few legal precedents about the use of email as evidence. Proof that this is a vague area can be found in the fact that many organisations include disclaimers in their email messages.

Technical issues
Hardware and software applications quickly become obsolete, which means that digital files are no longer accessible. Little practical research has been done, either nationally or internationally into technical ways of preserving email. Here and there projects are underway, such as the DAVID project for the city archive in Antwerp where research is being done, but not on a mass-scale[8]. The problem is not only that of preserving email, but also of preserving the attachments. Should they be preserved in their original format or not? Should they be stored with the email message or separately?

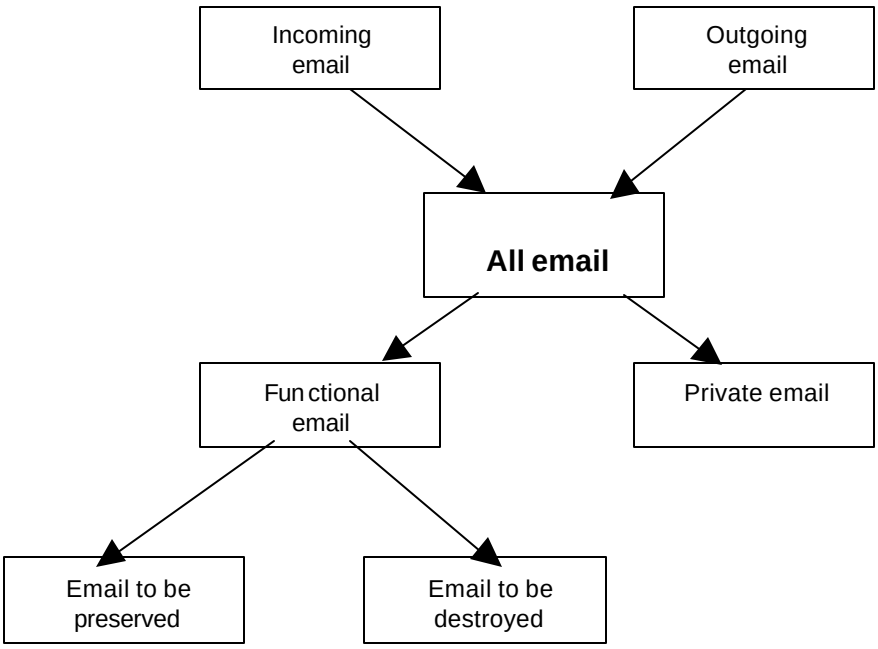
As a consequence of the complexity of these problems and developments, email messages are not, or seldom, preserved properly. The solution lies in an approach in which all the aspects mentioned are dealt with. A technical or legal solution on its own is not enough: people must become aware of the importance of good preservation. To arrive at a practical approach, it is necessary first to describe the characteristic features of email and to formulate authenticity requirements: which criteria should a well-preserved email meet?

3.2 The status of email

At the risk of stating the obvious: not all email messages have to be preserved. The choice of which email messages need to be considered for preservation partly depends on the tasks of the organisation where the email was created or received. This is described in an Institutional Research Report (RIO). The Basic Selection Document (BSD) that is based on this RIO provides the basis for either destroying email messages or transferring government outpourings to an archival location for long-term preservation.

[8] DAVID – Digitale Archivering in Vlaamse Instellingen en Diensten

Without going into too much detail on this issue, the following arrangement[9] may be helpful in answering the question of which email should be preserved.



Incoming or outgoing email
This distinction has a different character to the classifications below, but is nonetheless relevant to the regulations for dealing with email. A difference between internal and external email can also be made in this category, distinguishing between electronic messages exchanged within an organisation and messages exchanged with outside parties.

Functional email versus private email
Email that an employee sends or receives as part of his/her job is official email. Email that an employee sends or receives as a private individual, which is not related to the fact that the employee holds office in the organisation, is classified as private email.

Email to be preserved versus email to be destroyed
If an email message is functional, a decision has to be taken on whether it is eligible for preservation. In principle, the same criteria as for 'normal' paper post apply here too.

Currently, the default file format offered by many email applications is, in the first instance, the suppliers proprietary format for example, *.msg for Outlook. This is the way in which most people save their email, even if only temporarily. As an intermediate solution, or for the lack of anything better, some people also print out their email, which, as demonstrated earlier, results in the loss of information and does not take advantage of the benefits of email.

[9] Directive on Email Use, p. 4.

3.3 Characteristics of email

The concept of electronic mail (email) has two aspects: on the one hand the email system that transports messages along the electronic highway and, on the other, the email messages themselves. That is in fact analogous to ordinary mail, which refers to both the service and the letters themselves. An email system consists of applications, a transport medium (like network facilities), and computers. It enables people to exchange messages asynchronously to and from electronic mailboxes. Long-term preservation of email relates in the first place to the email messages and their essential attributes. When we talk about email in this document, we mean the email message itself.

Email thus consists of messages sent or received by means of an email system. Just as with paper post, email messages can take on a variety of forms. A message may be constructed as a simple announcement, but it may also consist of a complex digital document, with moving images or sound linked to the email message. Such an attachment can be made with a variety of applications, for example a text processing application, an arithmetic application or a graphics application.[10]

Users and end users usually see an email message as a single entity, an electronic message in which information is transferred by computers in a digital form. Strictly speaking, an email message consists of two components: a message header and the message body. A message header contains information about the message, like sender, addressee, subject, date and many other things. The message itself, including any attachments, can contain data in any conceivable form. This could be simple ASCII text, but nowadays data consists increasingly of images, text processing files, multimedia files, executable program files or HTML.

In the perception of the user, the email message could also consist of three components: the header, the body and the attachment. This perception results from the way in which email messages are represented in email applications, in other words, what you see on screen.

The representation of an email message produced when an application reads the bit stream can differ widely per email application. The transmission file (carried in the bit stream) is expressly developed to safeguard interoperability between different hardware and software platforms. Whichever email application you use, the basic email body should always be readable. This principle serves as the basis for the success of email as a communication medium, but it does mean that the same email message may look different in Hotmail than in Microsoft's Outlook2000, for example.

3.4 Email and the transmission file

The email message as it is visible on the screen, is only one representation of the transmission file, (the file that was actually sent). The email application being used reads this transmission file and processes it into the record, which is displayed in a human-digestible form on the screen.

[10] Ibid, p. 7.

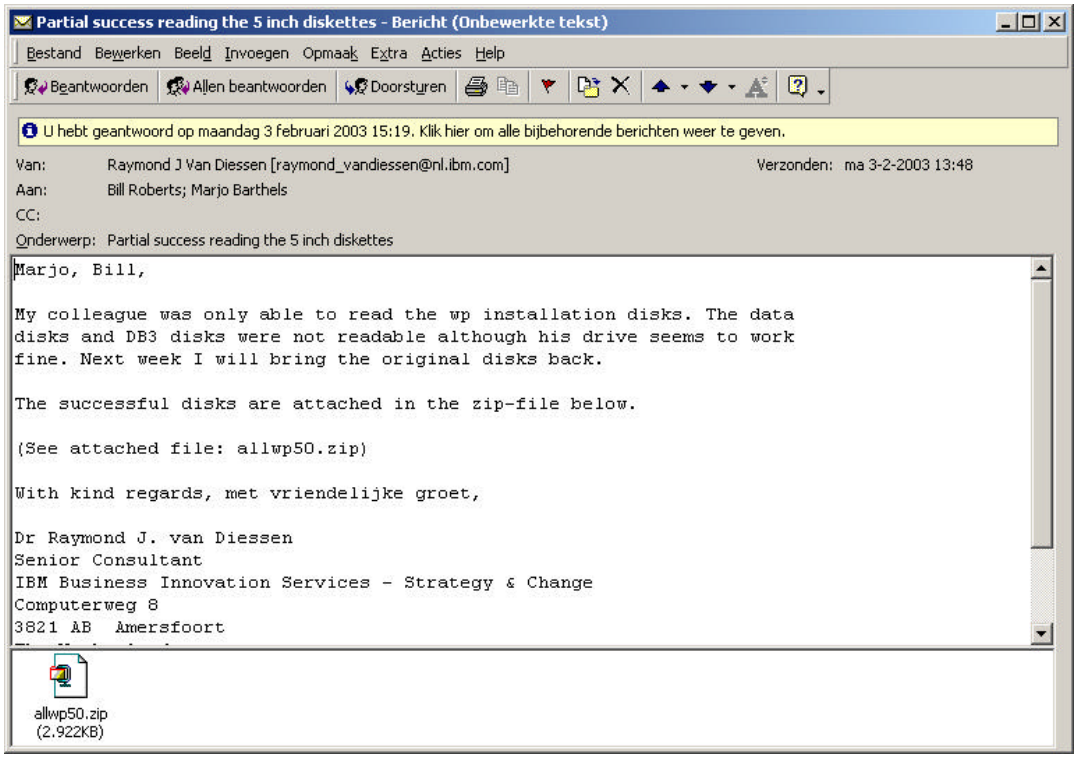


Figure 1. Example of an email message seen on the average user's screen.

The email message on the screen is not representative of the way in which it is sent. The file that is sent consists of plain text only (see appendix D for an example). The email application generates this file behind the scenes when the user creates an email message. This transmission file is the primary source of content and meta-data for all email messages. In some email applications, for example in Outlook, 'Message options' can be used to show a limited part of this transmission file (part of the header information, see figure 2).

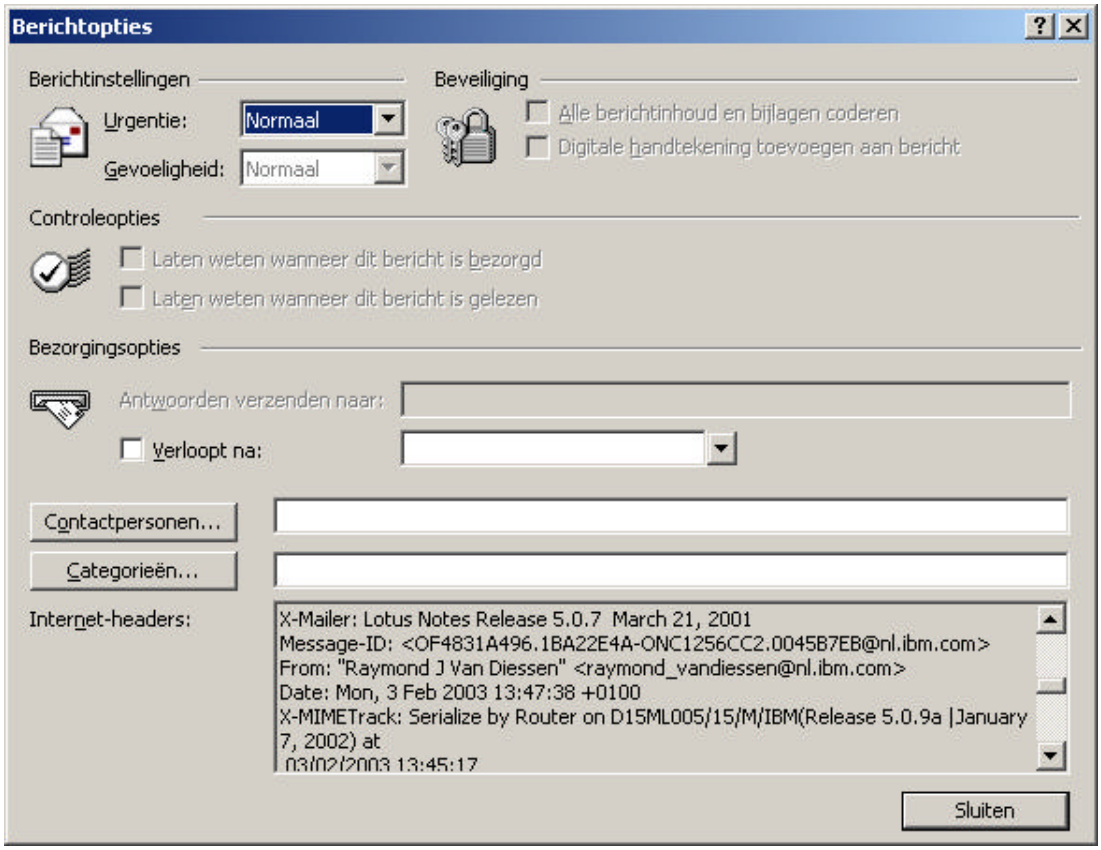


Figure 2. *Example of part of the transmission file (part of the header information) as shown in Outlook.*

The structure of transmission files is in fact the same for both incoming and outgoing email messages, regardless of how they are made and with which applicaton. The only difference between incoming and outgoing transmission files is that incoming files contain extra headers on receipt. These headers contain the date and the time when the messages passed through each server whilst being sent.
Outgoing email messages do not have these headers, either because they have not yet been sent, or because the sender has selected copies of those records from the 'Sent items' folder. In any case, outgoing messages have not gone through servers from the perspective of the sender, (but have of course for the recipient). See appendix D for a complete overview of a Transmission File.

Just as for the email message, the transmission file also consists of a header, body and attachments. These components determine the entire layout of the email, and every email message will look a little different. The basic components of the email are however always the same. The three components of the transmission file are discussed in detail below.

The header
A transmission file starts with header information. This contains far more than most email applications usually display. The basic information in the header consists of:

- Date and time on which the servers received the message (only for incoming email);

- MIME version
- Message-ID (unique number)
- Date
- Subject
- From
- To
- cc

Extra header information may consist of message categorisation, the name of the application with which the message was made, the priority and the urgency of the messages, etc. The main header section is an important source of metadata about the message, the content and its use (who sent and/or received this message and when, what was the subject). The components of the message, like the message text and any attachments, follow the main header. Each component also has its own header. These headers are important for management and preservation because they enable the parts of the transmission file to be put together again accurately. The second, smaller header contains the following information:

- Content type (for example, text/plain; multipart/mixed)
- Character set (for example ISO 8859-15)
- Content-Transfer-Encoding (for example Quoted Printable or Base64)

The body

The body is the message that the user types or inserts into the message window. Normally the text starts with a salutation and ends with a closing phrase. The message text is usually the first part of the content that is defined in the transmission file. The header field 'Content type' defines the composition of the content and shows whether the email message has been sent as plain text or styled.

Messages in plain text layout have a single text component with the content type 'text/plain'. Messages with a more complex layout may consist of several components. For example, the transmission file for messages composed in HTML often contains two representations of the message: the first is in plain text, the second in HTML. This means that users whose email applications do not support HTML can still see the content in plain text, although not with the intended layout. Messages of this type are defined as 'content type = multipart/alternative'.

Rich Text messages, as made in Outlook, can be sent in a similar way, but the layout and any attachments are often linked together in a unique file with a *.dat extension. This coded version of the message text can only be decoded in Outlook. This can cause problems for users who receive Outlook RTF messages with a different application, such as Eudora or even Outlook Express.

Messages with more complex layouts may consist of several text components or inserted images. The various file components are assembled by the email application so that the body can be displayed in its entirety. These types of messages are usually defined as 'content type' = 'multipart/related'. In these messages, the first header 'Content type' identifies the entire sent file as content type 'multipart'. This tells the application receiving the sent file that certain components of the message have to be assembled for a complete reproduction of the body. The components, each with their own individual header, are below the first header 'Content type'. These headers indicate the position the component occupies in the digital document that is to be reassembled, whether it is an inserted image, an inserted file, an inserted object or a background image. This information is extremely important for being able to properly reproduce the email message. If it is missing or gets lost, the authenticity of the digital record is endangered.

Attachments
The third component of the email message consists of attachments and metadata about these attachments. The attachments can be almost any type of file, for example, a Word document, a PowerPoint presentation, or an image in bitmap format. The only limitation to the sending of attachments is the size of the recipient's mailbox. Attachments that are too large will be rejected. The header 'content type' shows which application was used to make the inserted file and which coding was used. Binary attachments (not plain [ASCII] text files) are coded before being sent either in Quoted Printable or in Base64, in line with the MIME (Multipurpose Internet Mail Extensions) standard. This is the standard for sending non-text files. The email application first has to decode incoming attachments such as these before the digital record can be fully displayed.

3.5 Authenticity requirements for email

As discussed in chapter 2, the concept of authenticity is extremely important when preserving information, whether paper or digital. However, different types of digital records, such as email, databases or text documents, can have different authenticity requirements. These requirements play a crucial role when choosing a preservation strategy.

The Testbed has carried out extensive experiments with email and with methods for safeguarding its authenticity. Based on these, Testbed has drawn up a number of authenticity requirements for email in which the attributes of an email record are specified, and, the minimum that should be preserved to enable accurate reproduction. The requirements formulated below correspond to the characteristics or attributes of a digital record: context, content, structure, appearance and behaviour. In addition to these, extra authenticity requirements may be formulated from the work process. For example, some work processes may require the preservation of a particular colour on a digital map, because the colour indicates a specific meaning which must not be lost.

Context
Almost all the information about the immediate context of email messages is in the message headers. Some applications allow you to print the header data, other do not. The header information contains not only meta-data (sender, recipient, subject) but also the technical aspects of the file.

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

Essential header elements are:
- The email address, the organisation, and the full name of the sender
- The email address, the organisation, and the full names of all recipients
- For outgoing messages, the date and time the message was sent
- For incoming messages, the date and time the message was received, as well as the date and time this message was sent
- The subject of the message
- The security and/or confidentiality settings
- The file name and the file format of any attachments (all messages)

Almost all the data above is already in the transmission file, although not all email applications display this information in its entirety.

Maintain a 'preservation logbook' containing at least the following information:
- Data about the original and current file formats

- Information needed for interpreting the current file format
- Information about the preservation strategy being used and any changes to the digital record, like the date the time and the person or persons responsible

Determine the organisational context, such as:

- The dossier that the digital record belongs to (classification or dossier code)
- The name of the organisation that made the digital record
- The business process that it belongs to

It is worth commenting here that the organisational context is mainly recorded in a RMA or DMS. This is outside the scope of the Testbed research.

Content

The content of an email is often accompanied by an attachment created on an arbitrary application with a matching file format. The content of an attachment must likewise be preserved.

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

The actual content must always be readable

The displayed content of a reproduced email (this is the email after migration or emulation) must be as clear and readable as in the original digital record. The appearance of the record need not be identical, but the actual content must of course remain the same.

Attachments in the content must also be preserved

Any extra files such as photos or images that were inserted in the original message must likewise be preserved, as must the links and hyperlinks.

Structure

The structure of an email message relates to the arrangement of the message, the interpretation or the order of the components of the email. Every arbitrary change to the structure of the content can influence how the record is interpreted. The structure of email messages is usually limited to header information, content data, information about the layout, attachments and inserted items, (for example, a digital business card).

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

The structure of the content of the original message must be faithfully preserved

The parts of the message must be reassembled in the correct logical order. Attachments and inserted items must be clearly distinguishable and sub-headers must refer to the right part.

Appearance

The appearance of an email message is, as a rule, not the most important of the attributes. After all, the 'original' appearance of a message depends on the email application with which it was displayed. Different applications display messages differently and a user's personal settings can also influence some features of appearance. This means that the appearance of the message is variable right from the start.

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

The appearance of the preserved email may differ from the original

The appearance of the original message and that of the preserved version do not have to be identical, but the new appearance may not in any way alter the original meaning of the digital record

The position of an attachment in an email message is not important; the mention of it is
Any indication of an attachment in the content of the original message must be displayed in a similar manner in the preserved version. The position of the attachment (relative to the message body) is not important and may therefore be different.

Behaviour
Email messages do not usually have complex behaviour characteristics. The behaviour often associated with email is firstly the possibility of opening attachments. Secondly, there is the opportunity to reply to the message or to forward it. This second behaviour characteristic does not, however, belong with the message itself, but to the application with which the message was made. Usually the behaviour that occurs the most often in email messages is in the form of hyperlinks to websites, for example, or other documents.

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

The ability to open and access attachments, using suitable software, must be preserved.
The facility for opening inserted attachments in a suitable application must remain, but not the facility for sending, forwarding or changing the message.

Hyperlinks to websites should be preserved.
Hyperlinks to websites should be preserved, including the URL. It is not generally necessary for the authenticity of the email to preserve the content of the website to which the email is referring. Whether or not this is preserved depends on the importance of the linked website for the organisation.

Links to other documents must also be preserved
References to other documents must also be preserved. The title of the linked document must be recorded in the metadata of the message. Just as for linked websites, preserving the content of the linked document is not required to for authentic preservation of the email record. Whether linked documents will be preserved depends on the content of the linked document and on the importance the organisation assigns to it.

3.6 The digital signature

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

This development will be strengthened by legislative actions (Electronic Signatures Act, Electronic Administrative Transactions Act) and also by development of the required technical infrastructure, of which the government PKI is an important representative within the 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) and historical data about cancelled certificates. This data may be highly significant in the event of a dispute about the authenticity and applicability of a digital signature.

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

Digital Preservation Testbed recommends that, until further research has taken place, all data about the digital signature and the identity of the signatory, data relating to authentication of the signature, and the accompanying certificate, should be preserved as metadata at the time of signature authentication within the work process.

3.7 Summary

Integrity and authentication are crucial in establishing the authenticity of digital records:

In the first place, the characteristics of the digital record, as mentioned in the authenticity requirements, must be preserved so that the integrity of the digital record is safeguarded. This can be accomplished largely by developing a strategy in which the important aspects of the content, structure, appearance and the behaviour of the digital record can be preserved. The requirements in this book relate to the characteristics of email as a unique type of digital record.

In the second place, authentication is the important point. The context in which the digital record is made and used, and any changes that have been made as a result of management and preservation activities, are described in the metadata. This makes it possible to demonstrate or verify the extent to which the digital record is authentic in creation and contemporary use.

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 email.

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 email, as described earlier in chapter 3. Based on these considerations, we will decide which is the most suitable strategy for the long-term preservation of email.

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 file format are transferred to a new format that will run on modern computer platforms. A text document can, for example, be transferred from Word 95 to Word 2002 or from WordPerfect 5.1 to Adobe's PDF 1.4 (Portable Document Format).

Every migration requires advance research. After all, the target file format must be compatible with the source file 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:

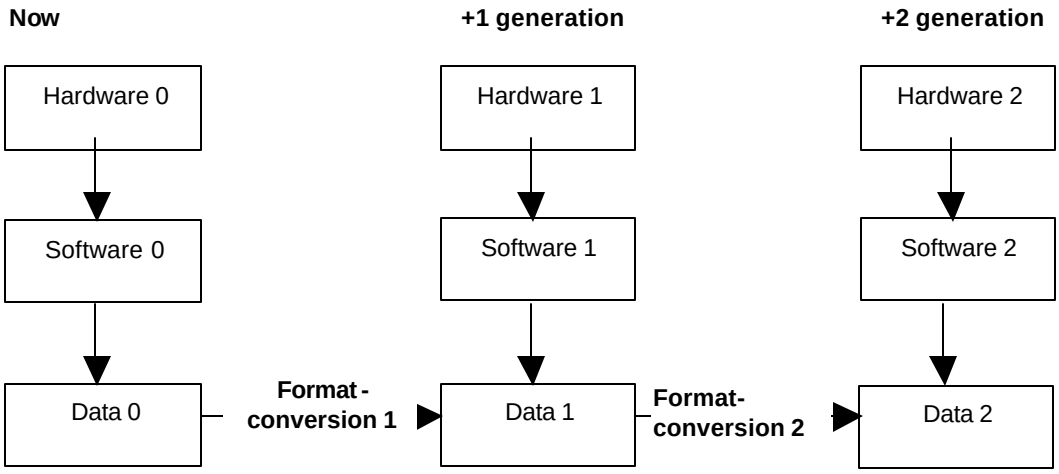


Figure 3. Basic migration diagram

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 or twenty years?

4.2.1 Backward compatibility

Backward compatibility makes it possible to interpret and correctly reproduce a record that has been made in an older version of an application in a new version of the same application. Software suppliers often guarantee that new versions of their software are compatible with the previous version. One example is the use of Word 2002 to read files made with Word 95 and originally saved in the Word 95 file format.

Records maintained using this approach must be re-saved into the new file format, in view of the fact that, as a rule, software only supports a limited number of generations of older file formats. This strategy can be used for files that are not too old and still in use, but is less suitable for long-term preservation as problems in authenticity will occur after a few generations. Although suppliers have an interest in providing backward compatibility - in view of the fact that the user can then easily upgrade to a new version - it may be technically difficult to continue fully supporting old, partly obsolete properties of the old version due to the properties of the new version. A disadvantage is therefore that only a limited number of generations are compatible with each other and even then some properties of the digital record can be interpreted differently by the new version of the software. This can have adverse effects on the authenticity and integrity of the digital record. No application can live forever.

Another disadvantage of backward compatibility as a preservation strategy is that the digital record continues to be stored in the supplier's own file format (for example *.msg for email messages in Outlook), thus maintaining an undesirable dependency on proprietary software.

A final disadvantage is that migration to a higher version has to be repeated every few years. Testbed experiments have revealed that with every migration, the risk of damage to the record is increased. This damage, however slight at the outset, could adversely affect the authenticity and integrity of the digital document and become worse with each subsequent migration.

Is backward compatibility suitable for preserving email?
In view of the disadvantages of using backward compatibility as a preservation strategy (storage in the supplier's own format, the need to repeat it every year, and the risk of adverse affects to authenticity and integrity of the digital document) the conclusion is that backward compatibility an approach that may be functional for the short or medium term (while the file is still being used) but that it is not a realistic option for tackling the digital obsolescence of email messages in the long term.

4.2.2 Interoperability

Interoperability in the technical sense tackles the problem of digital obsolescence by reducing 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. In its simplest implementation two forms are possible:

- Dependency on the application, but not on the operating system
- Independency of both the operating system and the application

In the first option a scenario can be envisaged in which applications can run under various operating systems. Software suppliers are aware that not everyone has the same computer platform and so they issue different application versions for different operating systems, such as versions that run under Windows, Linux or Solaris. Often, however, manufacturers have to bring out different application versions for each operating system, which can limit their interoperability. Measures are still needed to prevent the loss of information, authenticity and functionality. The strategy often involves migration to a file format that is still proprietary, and thus further migration of the files is required over time.

In the second option a scenario can be envisaged in which files are independent of both a specific application and operating system. An example of this is the situation in which the chosen file format is a widely accepted standard, such as XML. There are other ways to apply extensive interoperability, but their complexity makes it more difficult to guarantee longevity. What is more, our focus is more on the longevity aspect of the process rather than on interoperability. Nevertheless, a strategy in which both interoperability and longevity are possible would be preferable to a strategy based solely on interoperability. A strategy that does combine interoperability and longevity (the application of XML) is discussed in detail in section 4.3. This type of interoperability is the same as the method of conversion to standards that is discussed below.

A more complex form of interoperability requires the use of an interim conversion program. In this approach, files are converted from a proprietary format, such as MS Word, into an interchange format, such as ASCII (American Standard Code for Information Interchange) or RTF (Rich Text Format), which can then be interpreted by another application, like WordPerfect. Such an approach carries a substantial risk that essential characteristics of the digital record are lost, particularly when it has a complex layout or multimedia content.

Is interoperability suitable for preserving email?
Total, problem-free interoperability for complex digital records between and within platforms without the use of a conversion program is rare. Proprietary applications produce files that should be rendered on a specific version of an application for the best results. It is unlikely that files made with an application under a particular operating system can be read accurately and authentically by a version of that application that was designed for another operating system.

It is important to emphasise that although the email transmission file was developed expressly for interoperability, the email message is usually stored in proprietary format such as *.msg for Outlook messages or *.vew for Novell GroupWise messages after it has been processed by the email application. If such a message is simply stored as a *.txt or *.html file, it is unlikely that all of the essential transmission data needed for long-term preservation will be saved.

As described earlier interoperability can involve converting the file to an interchangeable format that can be read by other applications, often also on other platforms. The biggest problem with this option is that the files first have to be converted to an interchange format such as ASCII. This will probably not capture all aspects of the digital record. When the file is read into a new application, certain properties and aspects of the digital record may be rendered in such a way that the authenticity and integrity are adversely affected. Moreover, when using this strategy the problems of using proprietary file formats are not necessarily tackled.

In view of the disadvantages mentioned above (loss of essential properties of the email message; storage in the supplier's own format), a strategy of interoperability alone is not suitable for preserving email over the long-term. Nevertheless, the strategy can be used in combination with the method of conversion to standards, in which interoperability is often an implicit result of the conversion. This option is discussed below.

4.2.3 Conversion to standards

Conversion to standards relates to migration from a proprietary (and often closed) file format to a format with a broad support base that is widely used across computer platforms. The advantage is that digital records are no longer dependent on the original hardware and software with which they were made that threatens their longevity. Two types of standards are possible here: 'de jure' and 'de facto'. 'De jure' standards are constructed through a formal process by a formally accredited standardisation body (ISO, etc.). A de jure standard is always developed through an open process, since consensus and availability are the most important motives for formal standardisation organisations. 'De facto' standards are not normally accredited by such an official body. Such standards are widely implemented and there is a critical mass that makes use of the specification. De facto standards are often

developed through open process by consortia, but they can also be developed by means of closed processes ('proprietary standards').[11]

To preserve digital records over the long term, two forms of conversion to standards can be used:

- Conversion to a published, proprietary standard like PDF (a de facto standard)
- Conversion to an open standard developed by the community, such as XML

Conversion to a proprietary standard is not preferable because although the specification is freely available and tools can be designed for it, the basic software and licence costs still have to be paid and the supplier controls the latest edition of every new version. The manufacturer can end development of the software at any time and is not obliged to ensure that a third party continues with the development. This means that if the manufacturer goes bankrupt, the standard may disappear. That is the reason why conversion to open standards is preferred over a standard that is owned by a single company.

Although software for open standards still has to be developed and/or purchased for processing and conversion, licence costs and suchlike are not needed. A large number of people are involved in developing open standards and the specifications are free.

Is conversion to standards suitable for preserving email?
Most email applications do not (yet) offer possibilities for storing email messages directly off-line in a non-proprietary format that provides access to the entire digital record. For that reason, other formats have to be considered. Conversion to a standard format such as XML, which is not dependent on a particular supplier, results in backward compatibility and interoperability of the file and record. In that case, backward compatibility and interoperability are advantages of this strategy and not the strategy itself.

In the ministerial 'Regulation on the Arrangement and Accessibility of Records' referred to earlier, several standards are mentioned for long term preservation of text documents and images [12]. There is however no specific format mentioned for preserving email messages. For the purpose of these standards, email messages can be considered as a type of text document containing alphanumerical data and possibly images. Naturally the standard has to be carefully chosen and it must be suitable for reproducing the digital document in question. The Testbed has analysed PDF, RTF and XML in this context.

Converting email messages to PDF is not advisable. The majority of email header information (including context information) will be lost, and the authenticity of the email message can no longer be guaranteed. The only information captured by this approach will be the static image of the message as displayed by a particular email application. Attachments are not opened and the clickable icons that indicated the presence of an attachment in the original message will appear in the PDF file simply in the form of an image. PDF is a format that is primarily intended for publishing fixed records, where an identical version of the digital records has to be preserved. This is not a

---

11 See: *XML*: *de mogelijkheden en valkuilen voor overheid* (Dutch document: the possibilities and pitfalls for government); W. Thomas, 19 September 2002.

12 Regulation on the Arrangement and Accessibility of Records, February 2002

requirement for email messages since email messages do not have a fixed appearance. After all, it is dependent on the email application being used.

Apart from PDF, Testbed also examined RTF (Rich Text Format). RTF is no more suitable than PDF for preserving email over the long term. Essential header information is lost here too. Added to that, neither PDF nor RTF are open standards, meaning that the dependence on hardware and/or software suppliers is maintained, which is not desirable for digital longevity.

Finally Testbed experimented with using XML. The results are positive: in contrast to PDF and RTF, XML is capable of holding on to all the email message's header information. Furthermore, XML is an example of an open standard that is independent of a specific platform and application. More detailed advantages and disadvantages of the use of XML as a preservation strategy for digital records are discussed in the next section.

4.3 XML as a preservation strategy

The Digital Preservation Testbed also studied XML as an approach towards the long-term preservation of digital records in its own right. XML stands for Extensible Mark-up Language. It is a mark-up language for enriching data with information about structure and meaning that can also be used as a file format. It is an open standard defined by the World Wide Web Consortium, a non-profit organisation that develops interoperable technology such as specifications, guidelines, software and tools so that the Internet can be used to its full potential[13].
XML is non-platform specific and can be read by humans as well as machines. For these reasons XML can be used for digital preservation and interoperability.

Depending on the way the XML approach is implemented, it may overlap with the other strategies described above. For example, the conversion of files to XML can be seen as a specific type of migration (see Conversion to standards, above). XML can also be used in combination with the emulation strategy and the UVC variant, or as a file format and linking tool within an object oriented solution. In the following sections we will describe first how the problem of technological obsolescence in digital records can be tackled with XML. After that, we will examine alternative and additional applications of XML in digital preservation.

4.3.1 Wrapper and framework

Wrapper and framework are terms used to describe a single basic way of implementing XML for digital preservation. When this approach is used, the file (or part of it) usually remains in the original format. A wrapper, which describes the content, is placed around the outside. The content of the wrapper, its structure and the way in which the strategy is implemented depend on the requirements of the parties involved.

This means that a single wrapper can preserve or enclose a large number of related objects. A wrapper can for instance contain a copy of the file in its original format (the 'master file'), a converted version of the file in a chosen file format, such as PDF (the ‘access’ file) and metadata to describe these objects and their history. All objects can be stored together with an XML wrapper that not only describes the content, but also the access and preservation requirements. This gives authorities access at a single safe location to:

[13] See http://www.w3.org

- The master file, the authentic digital original for future preservation activities;
- A converted version to which researchers have access by means of modern software and;
- Metadata that describes the digital record. Metadata relating to access and preservation can be updated by adding extra data to the wrapper. If necessary, extra digital objects can also be added.

The wrapper strategy has been tested and implemented in a number of ways. The VERS project in Australia[14] wraps the records, context and authenticity information in a single object. XML is used to mark-up the wrapped digital records in such a way that the history of the digital document can easily be updated by placing a new layer of meta-data above the previous one (known as the onion model).

The e-Archiving project[15] have a similar approach but use the term containers instead of wrappers. Their containers hold scientific data for archiving, and the choice between a conversion or an emulation approach depends on cost considerations and obsolescence factors.

14 http://www.prov.vic.gov.au/vers/welcome.htm

15 http://www.library.tudelft.nl/e-archive/

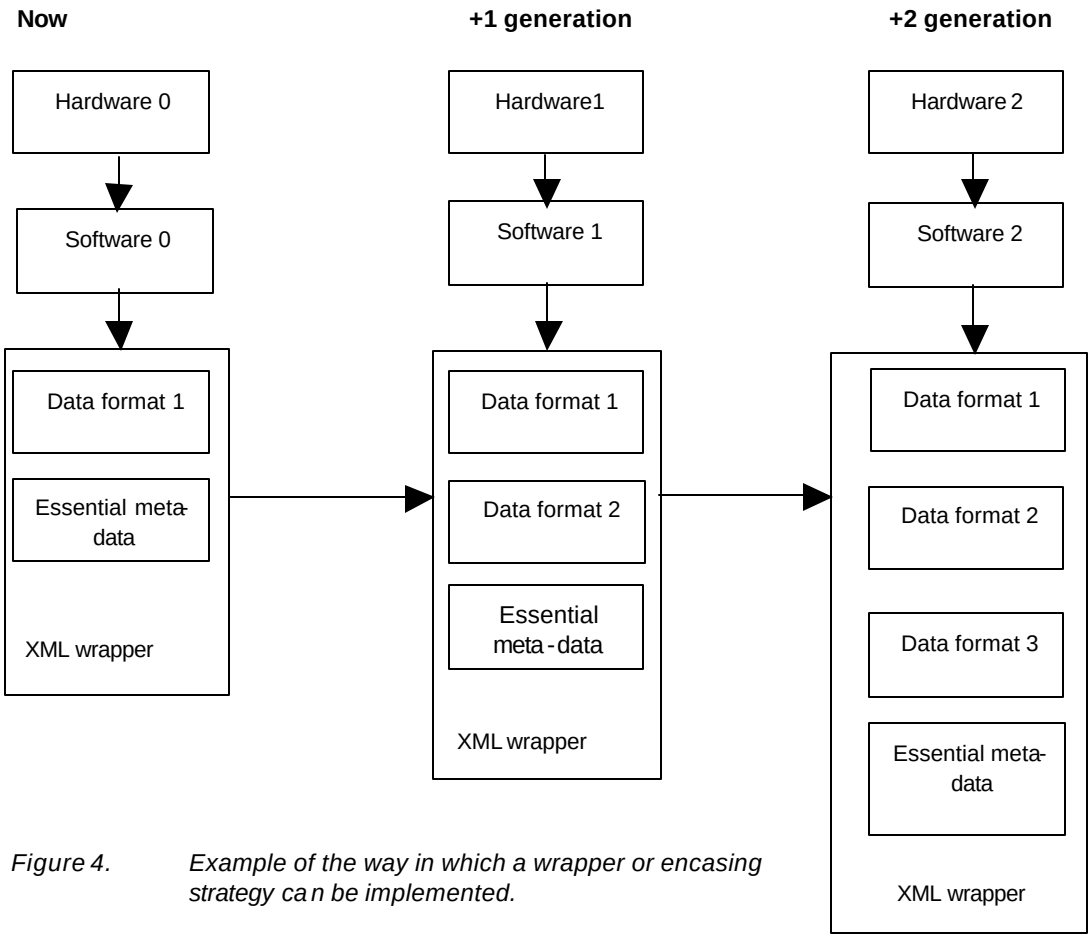


Figure 4. *Example of the way in which a wrapper or encasing strategy can be implemented.*

Comment on this figure:
Obsolescence of hardware/software is avoided because XML is non-platform specific. If the files in the wrapper are at risk of becoming obsolete, it will be indicated in the XML wrapper and authorities can make a new version of the digital record using modern applications. This is just an example. Such a strategy can also be implemented in other ways.

Organisations may also prefer only to pack a single file in XML and to add extra preservation metadata to the wrapper when it is created. This can be done if the source file is relatively simple and straightforward. Complex or proprietary files are probably more suited to the more complex wrapping strategy pictured above.

The so-called framework approach uses XML slightly differently, namely as a tree structure in which the digital record and its various components can be incorporated. The complexity of the digital records and the requirements or preferences of the institution determine the way in which the approach is implemented. Naturally, suitable software is still needed to read the files. Although XML is not platform specific, it must still be processed by software. This software may or may not be platform-independent and must therefore be checked for signs of approaching obsolescence.

XML is suitable for a wrapper strategy because it is a descriptive meta-language, independent of specific hardware and software combinations, and readable by both man and machine. All of this makes it possible to keep several files together and associate them with the necessary metadata, through which digital obsolescence can be resisted. The original file, or at least the major part of it, remains intact and can be stored together with other representations of the digital record.

This description of wrapper and framework is not an exhaustive or definitive explanation of the ways in which XML can be implemented.

Are XML wrappers and frameworks suitable for preserving email?
By using wrappers or framework in an XML approach, related files can be linked to each other smoothly via a single digital record object. Not only that, but extra files can also be added to the digital record object during subsequent preservation activities (see figure 11 in the next chapter). The related files are linked together with XML, thus making explicit the structural relationships between them.

Email messages generally consist of several files or components, like header information (the address), the body of the message and any attachments. This structure of email makes the application of the wrapper or framework approach an especially suitable strategy with which Testbed has had good results:

Wrapper approach
A major advantage of the wrapper approach is the straightforward management of the XML file. After all, there is only one file that contains all the components needed to reassemble the record. The disadvantage of this approach is that unless separate action is taken, any attachments remain encoded (e.g. in base64) in the XML file. Just like the email transmission file, XML files can only contain textual data, not binary files. A binary file consists of a collection of bits that contain more than simply plain text. Extra codes are in there that can only be used by a specific system or application, like Word, Excel or PowerPoint. The consequence of this is that decoding algorithms must also be preserved to decode the attachments at a later stage (and to reassemble them into their original binary format).

Framework approach
The framework approach does not have this disadvantage: attachments are instead stored in their original file format. This means an alternative preservation strategy can be devised for different types of attachments with

different preservation requirements. This is the Testbeds preferred approach. A detailed description of the implementation of this approach can be found in chapter 5 'Recommended approach to email'.

4.3.2 XML as a file format

An alternative to wrapping is to convert the files directly to XML or to generate them directly in XML, using XML as a file format. This technique is related to the method of converting to standards, described under Migration, but to a specific standard: XML. Since XML is not specific to a particular combination of hardware and software, it is more sustainable than many proprietary file formats. The number of conversions needed will thus be considerably reduced, as will the risk of adversely affecting the authenticity of the digital record.

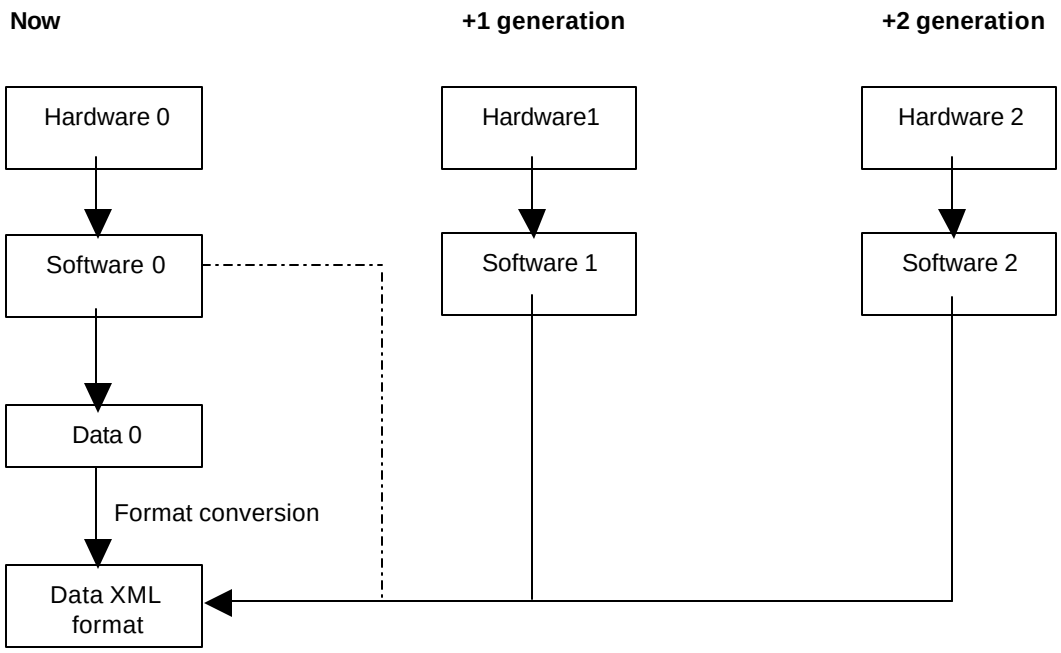


Figure 5. Conversion to XML requires fewer conversions than migration

XML is a suitable format for registering metadata and reproducing the five characteristics of a digital document described earlier: 'content', 'context', 'structure', 'appearance' and 'behaviour':
Content and context can be marked up in XML so that it is clearly described and readable without the use of a machine. XML is a well-structured language enabling the *structure* of the digital record to be easily reproduced. The structure of the digital record can also be made explicit by way of an XML schema or DTD, a separate file.

The XML file itself contains no instructions relating to the *appearance* of the digital record. The way in which the digital record should appear when it is reproduced is represented by an XSL style sheet (XSL, Extensible Style sheet Language). Finally, the *behaviour* of the digital document can also be reproduced with XML. Simple behaviour like hyperlinks and email addresses are identified in the content through suitable tagging, and initiated using standard PC software when the record is rendered onscreen. More complex behaviour is more difficult to reproduce using XML and has to be looked at separately.

These attributes make XML a suitable choice as the target format for the sustainable preservation of digital records in many cases.

The conversion to XML can be implemented in several ways and the specific features of the implementation are largely dependent on the type of digital record. Conversion tools are needed to convert the files from their original format to XML. There are a number of types of commercial conversion tools available that may be suitable but the quality of the output varies tremendously. This means that the tool must be chosen with care to ensure that the output meets the long term technical and preservation requirements. Such requirements may, for example, be a suitable DTD for simple digital documents and, for more complex digital documents, suitable linking between the digital components and the digital record. This link enables them to be reassembled in such a way that they correspond sufficiently with the original.

Is XML suitable as a file format for preserving email?
XML can be used as a file format and is a good choice for the long-term preservation of email: XML is an open standard, interoperable and capable of reproducing practically all five characteristics of a digital document. Because email and XML have a number of common characteristics, XML pre-eminently lends itself to the long-term preservation of email messages. After all, email messages must meet a technical standard, the MIME format, to be interoperable on different platforms. This message format clearly defines the components of an email transmission file[16]. It is managed by a non-profit organisation, the Internet Engineering Task Force[17], and is well defined, well structured and text based. The similarities between XML and MIME mean that conversion of email into XML is relatively straightforward

XML is not only a good target format for converting email messages at a later date, but is also suitable for generating new email messages now. The use of XML in the creation of new email messages does have a limitation. XML is a non-binary file format and that means that this approach does not lend itself to the sustainable preservation of binary attachments or other attachments (like Word documents, Excel spreadsheets and PowerPoint presentations). These can only be stored in XML in a converted form, as text, and thus encoded (see also the section 'Are wrappers and frameworks suitable for preserving email?').

16 The standard currently used for email is RFC 2822, with the MIME extensions being specified in RFC 2045 – 2049. The RFC2822 specification Itself can be found at http://www.ietf.org/rfc/rfc2822.txt?number=2822. See also http://www.nacs.uci.edu/indiv/ehood/MIME/2045/rfc2045.html for more information on standard (and non-standard) MIME headers and their continuing development.

17 See http://www.ietf.org/

4.4 Emulation as a preservation s trategy

Emulation is a more technically complex strategy than migration for preserving digital records. Just as with migration, there are various ways in which an emulation strategy can be implemented. The complexity varies according to the type and level of emulation required. In emulation the old, original computer environment is recreated on a new, modern computer.

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

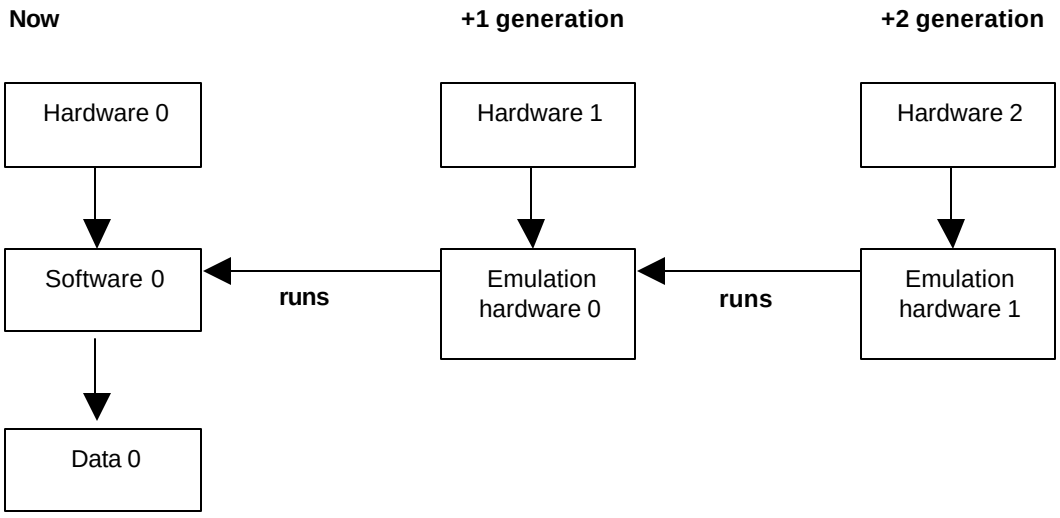


Figure 6. Basic emulation diagram

The theory behind emulation as a preservation strategy for the long-term preservation of digital records lies in the preservation requirements of digital records. Many people think that the authenticity and integrity of a digital record can only finally be guaranteed by keeping that record accessible in the original environment, in other words, with the original operating system and application. The file is then not damaged through conversion or change. The original behaviour and appearance (the 'look and feel') of the environment and the record remain intact. To do this, however, an emulator has to be built that emulates the old software or hardware environment (including drivers for peripherals) on a new, modern computer.

Unlike migration, there is no conversion of the file format and no change to the file bit stream when employing an emulation approach. This means that the digital record will in principle always be reproduced in the same way, irrespective of the time that has elapsed and the computer platform on which the emulation software will run in the future.

Over the years, various forms of emulation have been developed:

1. Hardware emulation
2. Software emulation

3. UVC approach (Universal Virtual Com puter)[18]:
    - Data preservation
    - Application preservation

Each of these options is described in more detail below.

4.4.1. Hardware emulation

Hardware emulation is one of the oldest and best-known concepts in the area of computer emulation. In hardware emulation, as many enthusiastic fans of the Commodore 64 or BBC Micro will know, an old and possibly obsolete hardware environment is recreated on a modern computer. To do this a program is written that imitates the structure of an old computer on a new compute r platform. An example of this is a Commodore 64 emulator running on a modern Pentium 4 computer. The user can do everything in the new environment that was possible in the old one. The Pentium 4, which behaves like a Commodore, executes applications, understands commands and reads files in the same way that the Commodore 64 used to. The files are thus reproduced in the same way and can be used in the same way as when they were first created and used in the original environment.

Through hardware emulation, the entire computer environment can be recreated so that the behaviour and the appearance (look and feel) of a digital record remains intact. This may be necessary for extremely complex types of digital records in which functionality and behaviour were important aspects of their use. These might be complex multimedia digital records or CAD objects. For records of this type hardware emulation is probably the only way of guaranteeing that their authenticity and integrity can remain intact. The major advantage of hardware emulation is that the original file does not need to undergo any migration or conversion. The drawback is that hardware emulation is not an easy exercise. It is labour-intensive and therefore expensive to emulate the entire, original computer environment.

4.4.2. Software emulation

In software emulation the problem is tackled in a similar way, but at a higher level of abstraction. Only the software that runs on the old hardware needs to be recreated, not the hardware environment itself. The digital records are thus accessed on a modern computer, but with the help of special software that imitates the original software environment.

In this way, an operating system – for example Linux - can run a programme designed to emulate, or mimic, another operating system (for example, Windows). Applications that run under Windows can then be used in the normal way via the Windows emulator.
As with hardware emulation, the advantage is that the original digital record undergoes no migration or conversion but remains intact. The disadvantage is still that emulation is labour intensive and therefore an expensive preservation strategy.

An emulator specification has to be made for the required aspect of the original computer environment for both har dware and software emulation.

[18] The concept of a Virtual Machine/Computer has been around for many years; the concept of a Universal Virtual Machine/Computer is more specific and originated with Raymond Lorie of the IBM Almaden Research Centre.

This specification must then be stored together with the digital objects, so that they can be reproduced. In theory, a single emulation specification is sufficient for a number of digital records, as long as the same type of hardware or software was used for these records. The emulator specification must, however, be extremely accurate and, for this reason, many people think that this strategy is too complex and too susceptible to errors. It is possible that the specification will only be fully tested when it is needed for the actual emulation. By then the original application used to make the digital records may no longer exist. In such a case, if there is a problem with the emulator specification, the environment cannot be emulated with certainty and the authenticity and integrity of the digital records may be at issue.

Is hardware and/or software emulation suitable for preserving email?
Hardware and software emulation are too powerful a tool for email messages because the appearance (look and feel) of an email message is of minor importance, as shown in the authenticity requirements for email in chapter 3. Although email is interoperable, this characteristic is only applicable to the transmission file, not to the ultimate version of the record that is processed and reproduced by the email application. The transmission file must be readable by any number of email applications, because the sender can never be sure which email application the recipient is using. It is partly because of this interoperability that email has become such a fantastic success: it makes no difference which email application is being used, since all applications can interpret the transmission file. However, that is where the interoperability stops. Email applications convert the transmission file to a different format (often into the supplier's own proprietary format), to reproduce the digital record onscreen. Depending on the settings, possibilities and interface, email applications will translate an email message in different ways. Specific properties of one email application (such as urgency flags) may not be a standard component of another application, which might result in the recipient not being able to see certain information in the record when it is translated onscreen. The behaviour and appearance of a sent email message will therefore not always be the same as the received email message. There is thus no reason to deploy a labour intensive and consequently expensive preservation strategy like emulation, which is pre-eminently suitable for retaining the original behaviour and appearance of a digital record, for the long term preservation of email messages.

4.4.3. The Universal Virtual Computer strategy (UVC)

An emulation approach that uses the UVC differs somewhat from the original emulation concept. An emulator must still be written, but in this case it is for a non-existent, virtual computer, called the UVC (Universal Virtual Computer).

The UVC is a computer with such a simple architecture and basic set of instructions that any software developer in the future will be capable of writing an emulator for the UVC. The UVC is then used to run an application (UVC data format decoder) that takes the original record as input and delivers a Logical Data Description (LDD) as output for the data. This logical data description is built up of tags that provide additional information about the content of the digital record. The additional semantic information is set up in such a way that, in the future, people will be able to interpret the logical data description without additional resources. After that, a viewer built in the future processes the logical data description, which displays the authentic digital record on the screen.

The Universal Virtual Computer preservation strategy is a variation on the emulation strategy and needs to be applicable in every future computer environment for complete preservation efficiency. The strategy only partly relies on emulation and contains some aspects of the migration strategy. Using the UVC, original data files are converted into a Logical Data Description (LDD) via a program written in the UVC programming language. This LDD is an independent, self-descriptive and clearly structured data format that contains all the information needed for re-assembling the digital record in the future.

An advantage of the UVC approach as opposed to the classic emulation approach is that the UVC data format decoder can be tested immediately to determine whether the decoder meets requirements. If it does, this decoder can be preserved until it is needed in the future to reproduce a similarly preserved data file again on a current computer.

IBM is currently developing the UVC as an alternative strategy for long-term preservation. A successful Proof of Concept was carried out in the Royal Library in the Netherlands in 2002 using publications, particularly PDF files.

As far as the application of the UVC is concerned, there are two forms of implementation: data preservation and program preservation.

UVC data preservation
'Data preservation' is the first and simplest implementation form of the UVC strategy. In it, the data – the original file in its original format – is stored with a program that extracts the data out of the bit stream and describe this data simply and independently, so that a viewer can process the data.

The original file – for instance a JPEG file – is stored together with the specific UVC data format decoder program for JPEG. In the future this UVC JPEG program will be run on the UVC emulator. The UVC JPEG program reads the bit stream of the original file and produces an LDD as output (Logical Data Description). The LDD is reproduced on a future computer platform using a viewer that can be developed in the future based on the LDD Schema.

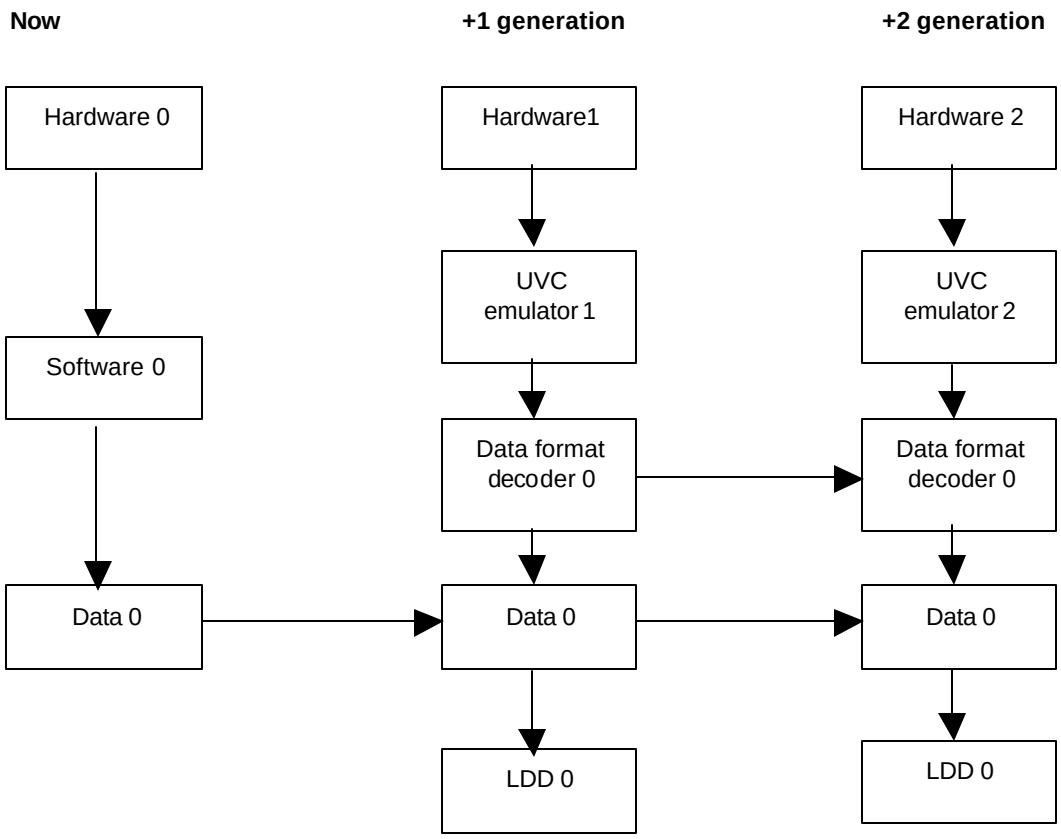


Figure 7. Diagram of the Universal Virtual Computer

The original bit stream is not changed in this strategy and the new file (the LDD), made when running the UVC data format decoder program, is not saved. The LDD is displayed by way of a viewer. The format and the structure of the Logical Data Description are so clearly defined that designing and writing a new viewer should be straightforward. If necessary, new viewers can be developed for future computer platforms.

At present, a separate viewer is needed for each type of LDD. This means that, in theory, possibly hundreds of viewers could be used. In practice, the number of different formats accepted by the Archives will be limited by of the Regulation on the Arrangement and Accessibility of Records. In the next phase of the UVC development, classes of objects will be formed that behave according to the same logic. A class of objects like this (for example, comprised of files in different image formats) will produce one LDD, for which only one viewer will have to be developed. It will, however, still be necessary to develop an individual UVC data format-decoding program for each of these file formats.

This strategy, which preserves the original digital record, is suitable for records in which the functionality of the application is separate from the record and is not needed to manipulate and use the record. Nonetheless, depending on the

original record, a few important aspects of the digital object may have been dependent on properties of the original application. In that case, these must also be included or described in the Logical Data Description.

A disadvantage of the UVC emulation approach is that UVC data format decoder programs have to be written for each file type (to generate the logical data description). In addition, a new UVC emulator must be written for each new generation of hardware that differs so much from previous generations that the old UVC emulator can no longer reliably run on it. Responsibility for this lies with the market players, at present IBM. IBM, however, intends making the UVC an open standard.

At present there are only UVC data format decoder programs for PDF documents and Excel spreadsheets. In view of the wide variety of file formats and types of digital records, large numbers of decoder programs will have to be developed as quickly as possible, if the UVC is to be a feasible and workable strategy for the long term preservation of different types of digital records[19]. The ultimate success of the UVC strategy is partly dependent on the extent to which this strategy is accepted by the software and computer sector. Software suppliers would have to develop a UVC data format decoder programme for their software that can make a logical data description based on the original file. When that happens the UVC strategy could expand enormously.

Is UVC data preservation suitable for preserving email?
In UVC emulation the environment is partially emulated and the file partially converted. Data preservation with the UVC shows great promise for long-term preservation, but is still in the developmental phase and cannot therefore be used in practice in the coming years.
Deploying the UVC data preservation strategy for preserving email is attractive from a conceptual point of view. Translating the transmission file into a logical data description, which resembles XML, and which can then be reproduced on the screen at any time one wants using a viewer built when required, could be a workable strategy. The output file can be read by man and machine and is no longer dependent on the original email software. There is, of course, a small risk of the UVC's dependence on IBM. IBM has, however, expressed its intention of wanting to deploy the UVC as Open Source Software. The major practical problem is that no UVC program has yet been developed and tested in practice for email.

UVC: program preservation
'Program preservation' works in the same way as data preservation but has more in common with the full emulation strategy. When deploying the UVC strategy for program preservation, the UVC emulator again runs on a future computer, but instead of using a UVC data format decoder program to read and convert the data files to a LDD, a software emulator is run that imitates the required application for reading and opening these files. The files are then opened in their 'original' software environment (in other words, an emulated

19 Another approach, Migration On Request, was recently advanced by the CAMiLEON project in the UK. This proposal has much in common with the UVC strategy, but the emphasis lies on the conversion program, which is written in C and is expected to stand the test of time. This means that the Universal Virtual Computer, on which the UVC conversion program was dependent, is not needed. See http://www.si.umich.edu/CAMILEON/reports/migreq.pdf for more information.

version of this original environment) with the advantage that any behaviour characteristics that are a part of the application remain available.

Is UVC program preservation suitable for preserving email?
This approach is still in a conceptual phase and will have to prove itself in practice. However, in view of the interoperability of the transmission file and the 'independence' of email from the email application used, this implementation of the UVC, just as the UVC data preservation, is not immediately the most obvious one for long-term preservation of email messages.

4.5 Conclusion

Based on the advantages and disadvantages of the various preservation strategies for the long-term preservation of email messages, combined with the results of the Digital Preservation Testbed experimental research, the conclusion at present is that the use of XML is the most suitable strategy for the sustainable preservation of email messages.

Migration of email to a higher version of the application format (backward compatibility) has the drawback that the email message remains saved in the supplier's own format, which is not desirable from the point of view of sustainable preservation. It is, after all, best to be independent of the specific hardware and software with which the digital record was created, because of the risk of it becoming obsolete. Migration must also be repeated regularly, and every migration carries the risk of the file undergoing a change that affects the authenticity and integrity of the digital record.

When employing interoperability, the second form of migration, the email remains stored in the supplier's own format: not a desirable situation. Conversion to, for example, ASCII or RTF carries the risk that important characteristics of the digital record are lost.

Finally, the conversion to standards (the third form of migration) offers different perspectives depending on the standard chosen, with an open, not a proprietary, standard being preferred. Conversion to PDF, an example of a proprietary standard, is not recommended. On the one hand, this is to avoid an undesirable dependence on the owner of the standard (in this case Adobe) and on the other because PDF does not capture the entire header and other essential data of an email message. The authenticity of the email message is thus at issue.

XML is interoperable and can be used independently of a specific combination of hardware and software, just as the email transmission file. XML is text based and therefore able to be read by man and computer. It is also an open standard, which means that a market player cannot unilaterally change the standard. At the same time, XML makes a distinction between content/structure and appearance, something that helps the longevity because it is precisely the record appearance that currently provides much of its dependence on software.

Emulation as a preservation strategy certainly offers good potential, but is too powerful and expensive for email messages. Emulation is particularly desirable in those cases where the functionality of the software with which the digital record was originally created has to be retained: the original appearance (the look and feel) has to remain intact. This is in no way a requirement for email since every email application processes the transmission file and reproduces it on the screen in its own way.

In the next chapter we describe a specific strategy and implementation for the sustainable preservation of email messages by making use of XML.

5 Recommended Email Approach

Chapter 4 described various preservation strategies with regard to their suitability for the record-type email, and XML emerged as the most promising. In this chapter we describe the way in which XML can be implemented.

5.1 Introduction

Based on the Testbed experiments, XML appears to be the most suitable preservation strategy for email at present. In chapter 4, we explained in great detail the advantages of using XML for the long-term preservation of email messages. In this chapter we will examine the question of the way in which XML can be deployed as a preservation strategy for email.

To better understand the approach Testbed is proposing and the way in which the conversion to XML takes place, it is necessary to know something about the properties of email as a computer file: the transmission file. This was discussed in detail in chapter 3 and is repeated briefly here.

5.2 Email as a transmission file

The transmission file is the most complete and definitive source of the content and metadata of the original email message. It contains all the information that was communicated about the time and space in which the transaction was carried out. The transmission file is the best starting point for converting sent and received messages to XML. After all, the transmission file, intended to guarantee interoperability, is independent of the application with which it was made. A conversion from the transmission file to XML will be less complex than a conversion from a proprietary format (e.g. *.msg in Outlook), because the file does not first have to be converted back into its original transmitted form. Email messages in proprietary formats interact with the environment in which they are being reproduced (and take certain information from it). For example, some email applications extract the sender's full email address from the system and the network hosting them. If the email message in a proprietary format is physically transferred to another system, the sender's email address could undergo changes. It is possible that the address may then contain the wrong domain, which will affect the context of the email message.

The transmission file consists of a number of components. The components determine the overall layout of the email, and every email will be different, depending on the content of the components of the transmission file. The basic principles of email are, however, always the same. All email messages consist of headers and body and may also contain attachments, images and other items. Every original file, its relationship with other files, and every component of the message is clearly indicated.

A significant difference between incoming and outgoing transmission files is that incoming files contain 'receipt headers' and outgoing files do not. The 'receipt header' contains the date and the time when the message passed through the mail server(s) during the process of being sent.

5.3 Conversion procedures

The conversion to XML can be done with several tools. Commercially available tools should be assessed with regard to their suitability, since the XML created by these tools will differ. Organisations may prefer to develop their own conversion tool and share the responsibility for this with allied organisations. Testbed can, if required, give further advice on suitable conversion tools.

There are two different possible scenarios for converting to XML:

- Post-use (converting to XML later on) and
- Pre-use (generating directly in XML).

The post-use scenario is intended for existing email messages (both already sent and incoming messages) that have to be preserved for an unspecified length of time (these messages are thus converted to XML **later on**).
The pre-use scenario can be used for new outgoing email messages and is the first step in the direction of making and sustainably storing official email messages (the messages are generated in XML **directly, at source**).

The chosen tool should enable extra information, like 'dossier' or 'work process', which is not usually in the transmission file, to be stored in the XML file. This extra information, which is necessary for both scenarios, is metadata for preserving context in digital records. The pre-use scenario requires a more integrated tool, which will be described later in a separate section.

There are three categories of email eligible for conversion to XML

- Incoming email messages. Email messages that are received can only be converted to XML later on, using the transmission file as the source for the conversion.

- Outgoing email messages (existing). Email messages that were sent in the past and that are eligible for sustainable preservation can later be converted to XML. In this case, too, the transmission file serves as source for the conversion.

- Outgoing email messages (new). New outgoing email messages belong to the only category that is eligible for the pre-use approach. New outgoing email can be generated in XML directly at source. A conversion later on is then not necessary.

The actual XML files for these three categories of email messages scarcely differ. The starting point for the conversion to XML for existing email messages, received and already sent, is the transmission file. The only significant difference between these two types of messages is the receipt header, mentioned in the previous section. Both types of email messages can be converted to XML in the same way. Further details of this can be found later in this section[20].

A different approach is taken for newly created email messages. To make things as easy as possible for end users, we recommend that organisations extend their current email application with a facility for generating new outgoing email messages directly in XML. Testbed took Microsoft Outlook as an example to show how this application can be made suitable for the long-term preservation of email messages with a reasonably simple adjustment. Testbed has developed an add-in[21], in which, users are first

[20] See Appendix A for information about the Testbed conversion processes.
[21] See Appendix A for information about the Testbed 'Email to XML demo', that was designed as an add-in for an existing email application and which can be used to make email messages correctly in XML.

presented with a specially adapted window where they must fill in certain metadata if using email for official purpos es

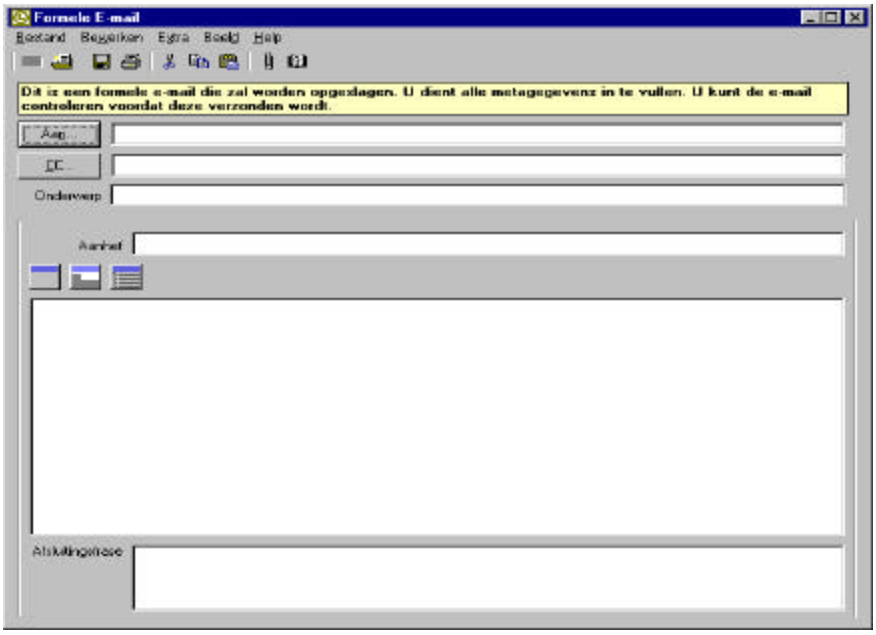


Figure 8. Example of a specially adapted window in Outlook.

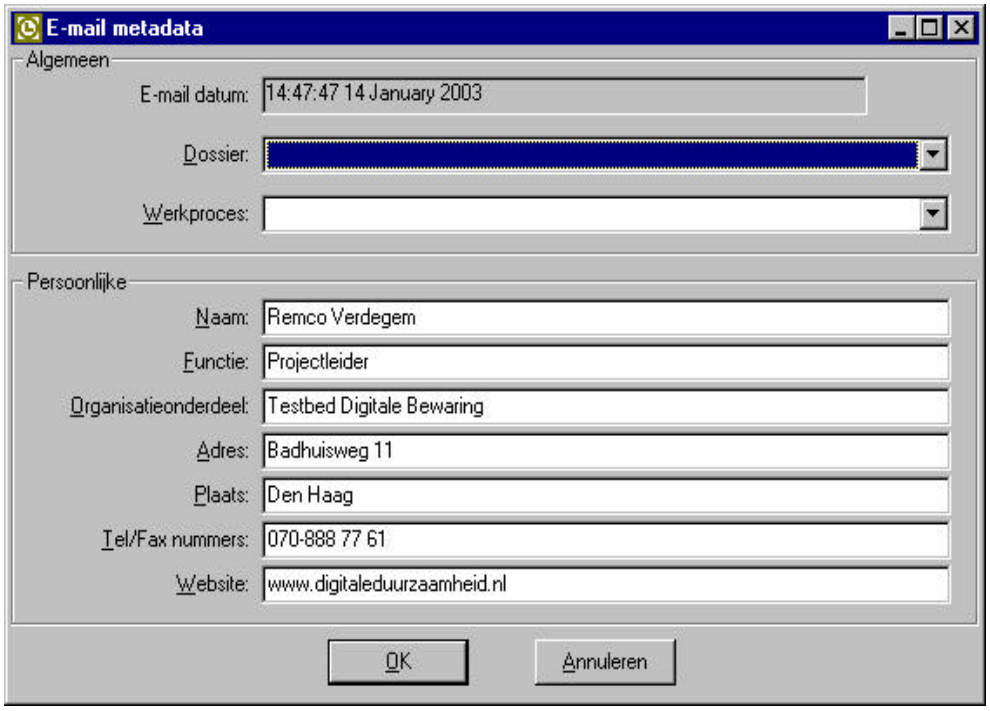


Figure 9. Example of compulsory fields in an email message.

A conscious effort was made to limit the number of compulsory fields to a minimum. In the Testbed demo only two fields were compulsory, dossier and working process. The personal data only has to be filled in once and is retained in Outlook for subsequent messages.

The specially adapted version of Outlook combines the metadata and the em ail message together in an XML file. This XML file is saved on a separate server that checks whether the XML meets the defined XML schema. An XSL style sheet is then used to convert the XML to HTML. This involves collecting the body of the email message and the metadata, and reproducing it in a formatted form, to which the overall layout of the email message, the choice of font, colours and a logo or image are added. This record, now in the form of an HTML document, is returned to Outlook and can then be sent to its desired recipient.

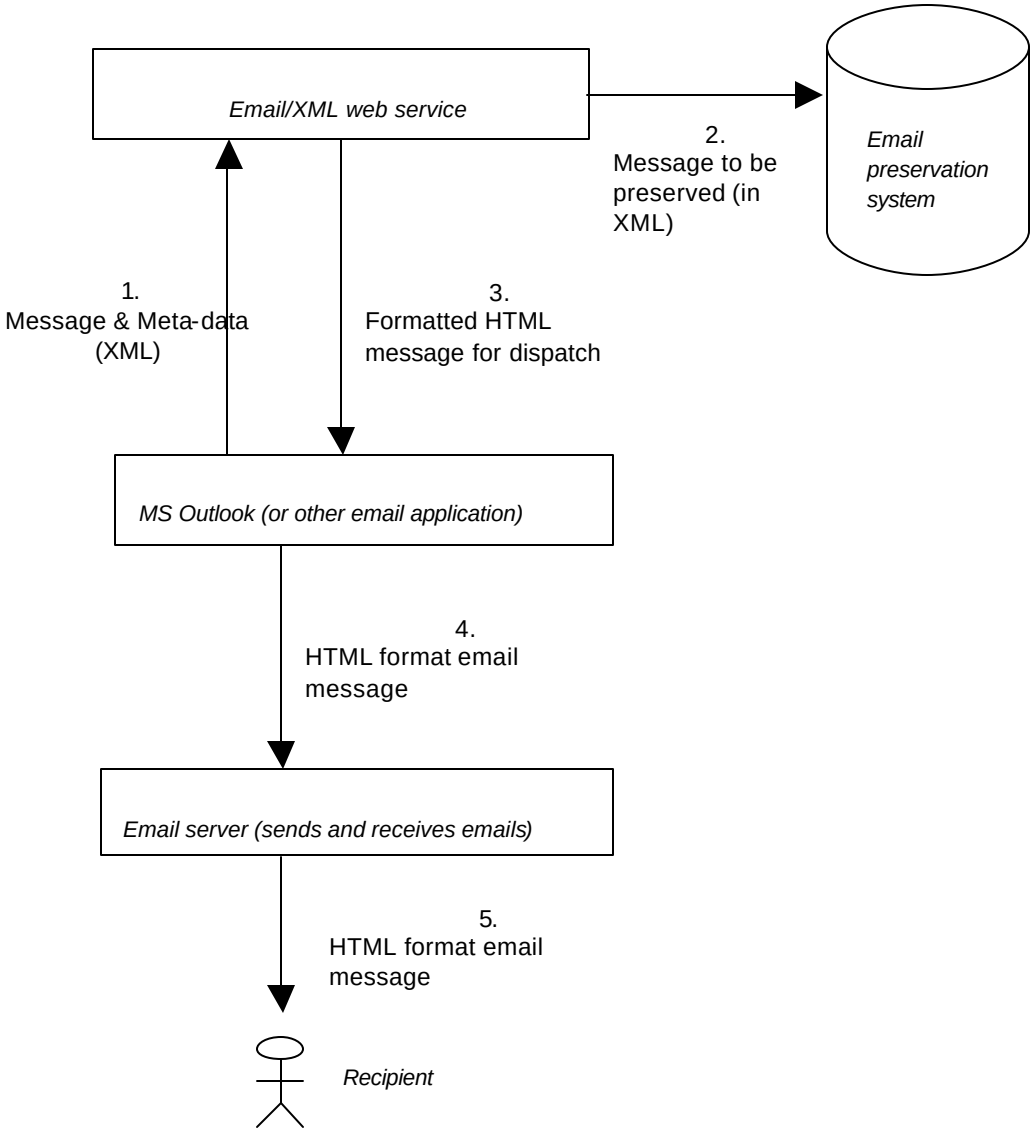


Figure 10. Conversion to XML using web service.

To increase acceptance by users, it is important to integrate this activity into their current procedures as much as possible.

Extra metadata must be collected before the conversion and attached to the message. The XML file must contain a basic set of metadata (described in the XML schema). Where necessary, links and references should be inserted (for example to the body or the attachments) and the XML file must be linked directly to the preservation object (see section 5.4). This strategy uses XML as a file format in combination with a wrapper and framework approach.

5.4 Long term preservation of email

As stated previously, XML is good as a file format in combination with the application of a wrapper or framework approach. The message converted to XML has a more complex structure than the transmission file, but it also contains considerably more information. In this strategy, the XML file is linked to a larger preservation object. This preservation object ensures that the links between the various components of the email message remain intact.

The preservation object consists of at least four components linked to each other via a central framework:

- XML file
- Preservation log file
- Transmission file
- (Extra) metadata

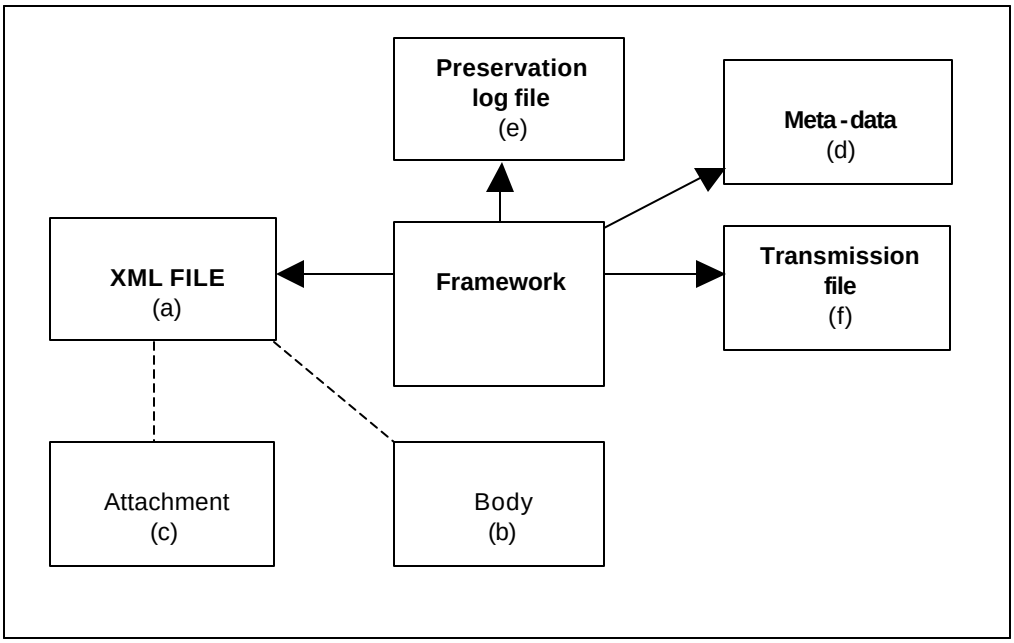


Figure 11. Diagrammatic representation of the email preservation object

The solid lines in figure 11 show those components that Testbed considers must always be present. The dotted lines indicate the optional components, like the body, that are written in HTML or RTF (so not in XML), or any attachments.

The four components mentioned above are needed to preserve the digital record over the long-term. In the next sections we discuss all the components (a to f) of the preservation object.

The XML file (a)

The XML file is the most important component of the email preservation object. This file is made by either converting the contents of the transmission file, or by making use of the pre-use approach (in which email messages are created directly in XML). The XML file is incidentally more than just a version of the transmission file marked up in XML: the information it contains comes from various sources, for example metadata added by a user.

The XML file contains a subset of headers from the transmission file, including all known addressees, the sender, the subject, the date, and the information in the header of each individual component. A limited set of extra metadata is added and tagged, as are references to the transmission file and any attachments. The body of message text can be reproduced by either a reference to a separate file or by direct tagging.

Extra context data is included in the XML file in connection with authenticity requirements. The Testbed email demo includes the following items:

- Conversion date (in the standard MIME date format)
- Dossier
- Business process

All organisations are free to decide which metadata are necessary to safeguard the continuing accessibility of the email messages.

When the pre-use approached is used, the author of the email can make use of a specially adapted application so that this information is included in every new email message that is made. When employing the post-use approach, this information has to be added to the XML file later, during the conversion process.

The figure below is a diagrammatic representation of the XML schema used in the pre-use approach of developing the add-in for Outlook, mentioned previously. Incidentally, when implementing the Testbed XML schema, an empty BCC field does not mean that no one has been included in that field. The content of the BCC field can only be preserved for outgoing messages. The BCC field never contains data in incoming messages. This is because content in the BCC field is removed when the message goes through a server for the first time. This explains the name: Blind Carbon Copy. The nature of the BCC field prevents this information from being recorded for incoming messages.

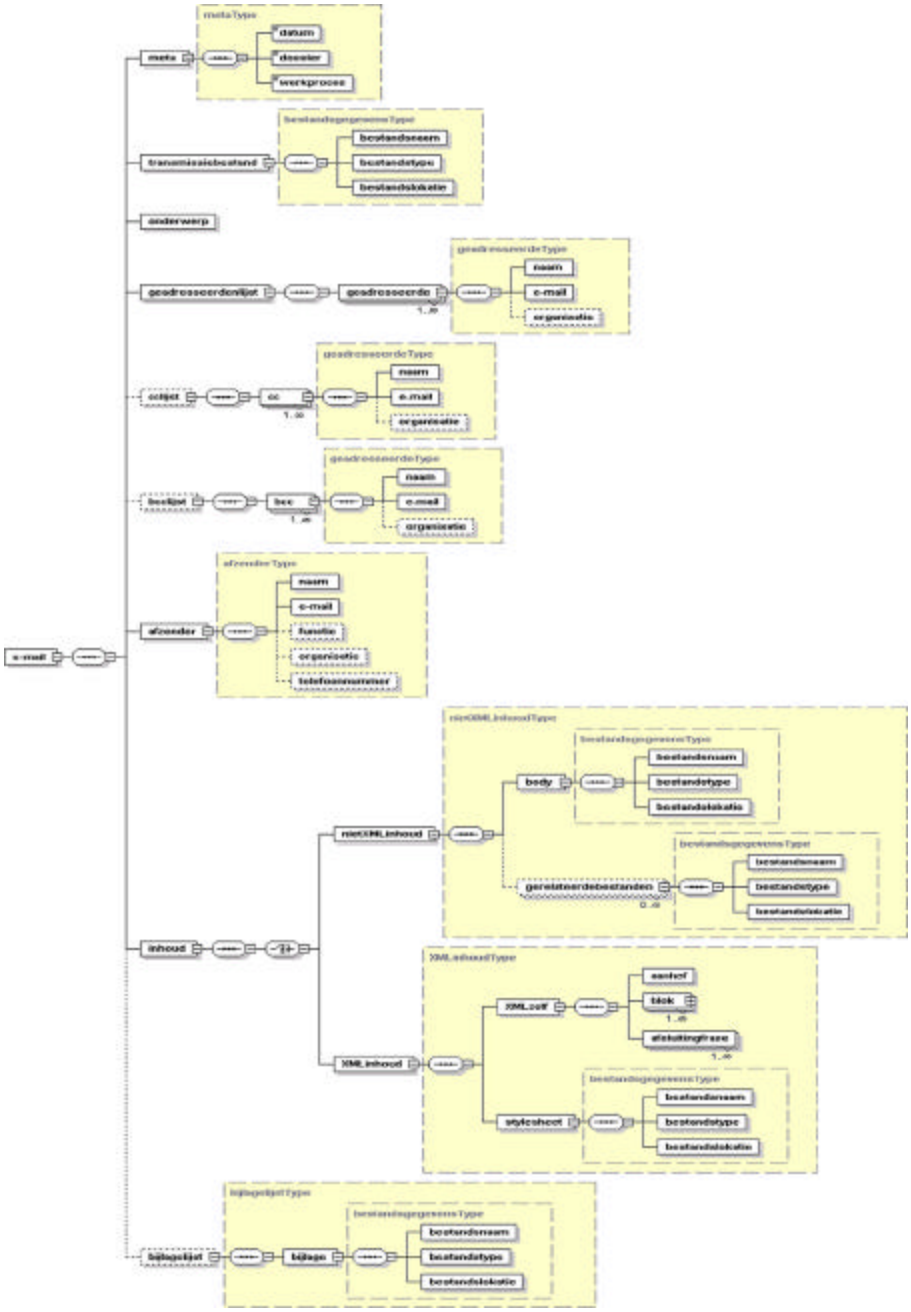


Figure 12 *Overview of an XML file for email. See Appendix B for the XML schema for this overview.*

Body (b)

The body, or message text, has to be stored in such a way that the content remains intact. If the 'pre-use approach' is used, in which an XML version of the email message is made and an HTML version is sent to the recipients, the content of the message can immediately be tagged in an XML file. In that case, a reference must be included to the style sheet used for making the HTML version of the message when it was sent.

If the message is converted to XML from the transmission file (post-use), a separate file may be made to store the text part. If it is a message in an unprocessed text arrangement, it can simply be stored as a plain text file. If it is a more complex HTML document, the HTML representation of the message text must be stored. Any attached multipart/related files can be stored with the message text, with clear instructions for reassembling the digital record.

The XML file must contain the following references to each of the above files:

- File name: the name of the file as it appears in the transmission file, for example, 'body.txt'
- File type: this describes the file format in which the content is stored, for example, HTML. A MIME designation can be used for this, for example 'text/html' or 'app/msword'
- File location: this contains a unique ID indicating where the file can be found. This location can be the Windows' system for directory names and file names, a database ID or a URL. The exact ID depends on the chosen method of implementation

Using the method described above, the software can find the components of the body reasonably easily and reassemble them.

Attachments (c)

Attachments have to be stored separately. Organisations that use the pre-use approach must design this specially adapted email application in such a way that a copy of the attachment can be added to the stored XML file. Attachments to messages converted from the transmission file (post-use) have to be decoded from their coded transmission form and stored in their 'native' format (Excel, MS Word, WordPerfect, etc.). Attachments are reproduced in the XML file by separate metadata and a reference. Each attachment is defined by the tag <*Attachment*> and has the following characteristics:

- File name, for example 'pic23004.JPG'
- File type, for example 'Image/JPEG'
- File location: this indicates the new storage location of the linked file, for example 'C:\xmail\files\loc'

References to multiple attachments are preceded by the parent tag <*List of attachments*>.

As long as attachments are stored in their 'native' format, the records manager can take appropriate action to preserve those files, according to their file format. After all, each type of digital record has different preservation requirements. One might decide to transfer certain attachments from proprietary, closed formats, suh as Microsoft Word, to Adobe's PDF, which is one of the formats recognised by the Regulation on the Arrangement and Accessibility of Records for long-term preservation of digital records. The preservation object approach enables attachments to be treated from their native format. Original files can be supplemented with newer representations.

When adding a new file to the preservation object, a new reference must also be created. The records manager has to update the preservation log file (e) with new metadata, but the other parts of the object may be left as they were.

Metadata (d)

The long-term preservation of authentic digital records requires an extra metadata file. The exact content of this metadata file is decided by the organisations themselves. Many institutions already register this metadata, whether in a Records Management Application (RMA) or a Document Management System (DMS).

In relation to this, it is important that organisations take all the header information out of the transmission file that has not been tagged in the XML file and place it in the metadata file as extra contextual metadata. It should be possible to carry out searches on these items within the boundaries of an RMA or DMS. All transmission data is preserved and searchable through this method, including references to message discussions, 'flag' categorisation data (which the recipient may not have seen) and information about the application. Incidentally, the XML file contains a considerable quantity of metadata about the original context of the digital record. This metadata can easily be processed and included in an existing metadata infrastructure, for example, EAD (Encoded Archival Description).

Preservation log file (e)

The preservation log file contains all the information about the preservation activities carried out on the record. The log file can also include information about preservation and access requirements.

The log file serves two purposes. Firstly, it can be used to verify whether the digital document is still authentic and accessible as time goes by. Secondly, it makes it possible for a component of the email message (this will often be an attachment) to be managed easily and continuously, without endangering the authenticity of the digital record.

The preservation log file is made when the record is first converted to XML. We will not prescribe a format for the log file, but only point out that it has to be a format that enables the preservation log file to be updated easily and continuously without data that is already present being overwritten. A database may be suitable for this and the use of XML might also be considered. The first content of the log file must consist of information about the original form of the digital record, as it was when it was received for conversion to XML. This content must be followed by information about the conversion, including the conversion tool used, the date and the time the conversion took place and the new file format.

The preservation log file must be updated each time preservation activities are carried out on the record. The preservation log file must also contain information about any changes in the digital record caused by preservation activities. In Appendix B we discuss the possible content of the log file.

Transmission file (f)

The transmission file is the most complete and reliable source for converting existing emails and should also be preserved. If the conversion process fails or if, for one reason or another, there is doubt at a later stage about the authenticity of the

converted record, it can be compared against this file for verification. Although this may not always be possible for all aspects of the digital record (the attachments for example) it should be possible for the headers and possibly also for the message text.

In the future, we may be able to use the transmission file in a new preservation strategy. It is to be expected that other strategies will emerge for preserving email, making better use of the new opportunities that future technology will have on offer. Preserving an original copy of the file in its original format may therefore hold advantages that cannot at present be examined or predicted.

Preservation strategies will not last forever and will have to be assessed and possibly altered as time goes by. After all, technology does not stand still. At some point, the preservation strategy recommended by Testbed for email messages may no longer be the most suitable. When that moment will arise, no one yet knows. Until then, the XML strategy for email messages that we have outlined is the best strategy for preserving them in an authentic manner for the long term.

6 Concrete Actions

The previous chapters dealt with the problem of digital obsolescence and proposed the best strategy for preserving email. 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 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 persistence: Preserving email* you will have discovered the advantages of working digitally, but also the specific problems that arise in the sustainable preservation of digital records in general and email messages in particular. The Digital Preservation Testbed has tested preservation strategies for the record type 'email'. The best way of preserving email at present is to use XML. The publication also discussed in detail how the proposed application of XML might be implemented.

But that's not the end of the story. In an organisation, different people are involved in the sustainable preservation of email messages: from the line managers in an organisation through to the end users who have the email facilities at their disposal. The concrete actions listed below are specifically oriented towards:

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

These four players have a specific responsibility in this matter. This final chapter sets out the concrete steps each target group has to take to make the sustainable preservation of email 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 a communication tool, email, that improves 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 email, just like other means of communication, 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 email is an official document, with all the consequences this implies, is a condition for working successfully with this medium.

It is also important to take *action quickly*. Examples of cases in which good preservation of electronic mail was the cause of major problem s are increasing in number, because email usage has multiplied in the last few years.

Concrete actions for managers

Formulate email policy*:* how does your organisation deal with email? How does it fit into the information and archive policy in your organisation? The purpose of an email policy is to aim at achieving the same quality procedures for email as for paper douments (see also the NEN-ISO standard 15489).
A digital government has many advantages when it comes to using email in the work process. Although the medium has the same status as a letter on paper, many people are not aware of this. Many organisations use a disclaimer because they are uncertain of the legal status of en email, but in current discussions people are

increasingly taking the view that email messages are also legally binding. Testbed therefore advises against the use of disclaimers: a digital government must communicate reliably and email is part of that as long as it is used properly.

Design procedures for and with your employees: these must state clearly who is responsible for what, who can be called to account and which people (positions) should keep each other informed. The following is the minimum that should be done:

- Agreements about the use of email (which communications it is to be used for)
- Agreements about the management and preservation of email

Consultative partners are: the records managers in your organisation, ICT managers and office managers.

Introduce email templates for your organisation*:* Email templates in combination with a web service ensure that all the relevant data is recorded correctly and that it can then be centrally stored in XML. In addition to this, email templates can give your email an official character by, for example, adding your organisation's logo. In this way digital information can be preserved easily and cheaply, even over the long-term.

Email templates are easy to integrate into existing applications like Outlook. A working group could be formed to implement them, made up of a records manager, an ICT specialist and a communications advisor. This publication even offers general instructions for making email templates and for storing them centrally in XML. Alternatively, you may consider hiring a company to do this for you.

Inform all employees about email policy and procedures. Train all employees to use the email templates correctly.

Evaluate policy and procedures from time to time.

6.2 Action plan for records managers

Introduction

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

But that's not the end of the story. In an organisation, different people are involved in the sustainable preservation of email messages: from the line managers in an organisation through to the end users who have the email facilities at their disposal. The concrete actions listed below are specifically oriented towards:

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

These four players have a specific responsibility in this matter. This final chapter sets out the concrete steps each target group has to take to make the sustainable preservation of email 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 variety of problems that have to be solved before email messages can be managed in the same way as paper records. Most people treat their email chiefly as a fast and transient medium; they view it more as a telephone conversation than an official record. What can you do to convince your colleagues that email messages are also digital records? How can you convince them to arrange their email messages and to classify important email messages? And, last but not least: how can you convince management to make money and means available to enable email messages to be sustainably preserved? This is not something you can achieve on your own in the organisation. As records manager it is important to look for co-operation with line management, the ICT department and with end users.

Concrete actions

The best way of preserving email for the long term depends on a number of factors, such as the culture of the organisation, the demands the environment places on the organisation, the political context, the state of the technology and the way in which the records function is designed. The concrete steps to be taken are:

a. Analyse the current situation
b. Formulate the desired policy and
c. Draw up procedures

The description below is based on the publication A*rchivering van Elektronische Post (Dutch document: Archiving Electronic Mail).*[22]

[22] *Archivering van elektronische post. Methoden, meningen en alternatieven*

a. Analysis of the current situation

Make sure that preserving email is a priority
Procedures only have a chance of succeeding if they are based on explicitly disseminated policy. It must be clear what the organisation wants to achieve with its email, the importance it attaches to it and the vision it has of developments. This is mainly a line management matter, but you, as records manager, must play the role of catalyst and source of inspiration. It is particularly important for you to get the preservation of email onto the agenda.

Bring in the knowledge and skills you need
How clear is the current archiving policy on the preservation of email messages? Your department is the important factor in determining and implementing this. Remember that the sustainable preservation of digital records requires different knowledge and skills than the preservation of paper records. Make sure that you have that knowledge in-house!

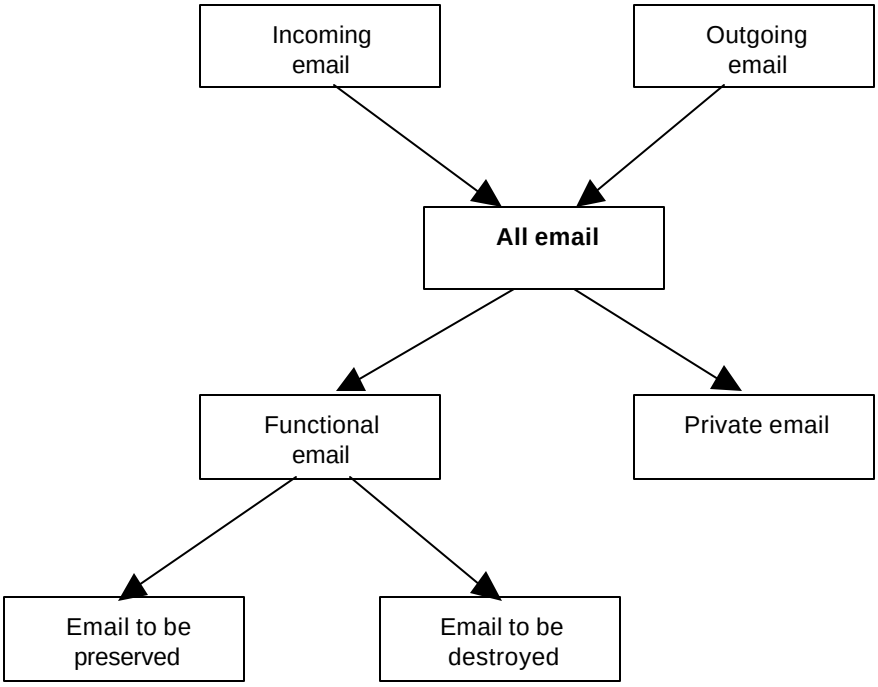
Look for partners and interested parties
Policy-making is not your responsibility in the first instance, but you can play an important role in the process to get the matter put on the agenda. To do this, it is important to trace other interested parties, such as heads of department who need particular information to conduct their business, the ICT department and the interests of all the email users.

b. Formulating the desired policy

Determine the selection criteria
Even if your organisation is fully aware of the need to preserve email messages, the question of which email messages should be preserved remains. Without going into too many details on this question, the following classification[23] may be helpful in answering the question of which email should be preserved.

---

[23] (Dutch document: Archiving electronic mail. Methods, opinions and alternatives), P. Horsman, Amsterdam 1999.
Richtsnoer Emailgebruik t.b.v. de Rijksoverheid (Dutch document: Directive on E-mail Use for Central Government) /Ministry for the Interior and Kingdom Affairs; working group on email use, The Hague 2001, p. 4.



It is important to select a suitable strategy for preserving records such as email messages at a very early stage. Early selection of the strategy assumes more significance with digital records than with paper records. This is because it is then possible to convert email that is to be preserved to a more sustainable format, **directly at source,** whilst the creation or original format is still accessible. This is more suitable than a proprietary format, which is the default save option in many email applications

Formulate the selection criteria for preserving email
These will have generally already been laid out in a 'document structuurplan' (a structured plan of records), or 'Basis Selectie Documenten' (Basic Selection of Documents) (BSDs). The point now is to apply these criteria to email. Insist that this selection be made at source.

The decision whether or not to preserve email lies with the person receiving or sending the email message. If you want to rely on users identifying and preserving the right email messages, it is your job to make them realise that official email messages are also records and that they have to be handled with the requisite care.

Retain the authenticity of email messages
Choosing the right method for saving email messages is essential since it affects their authenticity. It is clear that printing them out onto paper damages the authenticity of email messages, as much contextual information is lost in this way. From your own background you will realise that such an option must therefore be resolutely rejected.

In chapters 4 and 5 of this publication you will see that Testbed recommends the use of XML as a storage format. Link up with other disciplines in your organisation and use this information to make out a case for this solution.

Decide which metadata are needed
A subset of information from every email message is important for determining its origins, destination, receipt date and dispatch date. This metadata is needed to determine the authenticity and the function of the document. You now have to decide which metadata has to be registered[24].
Make sure that an accurate record is made of which metadata is important for use or re-use and interpret the information, and which metadata your organisation needs for reasons of accountability.

Determine the method of arranging and classifying
The aim of arranging and then classifying is to make the structure visible, the connections between records, and between records and processes in which they played a part. It is conducive to their accessibility and can support structured searches. This means that a classification diagram will have to be developed, based on tasks or business processes (see also NEN-ISO 15489). Involve the ICT department in this to determine search entries and connections between records. Email messages should be classified in the same way as paper records.

Formulate policy
The passage through the previous steps and the choices made in each must be recorded in a policy document. When making each choice, establish what is feasible and what is ideal. This document is the basis for the follow-up process, which focuses chiefly on implementation and is where the actual procedure has to be written.

c. Drawing up procedures

Indicate who is to take the actual decision to preserve an email message
It is usually the DIV employee who decides which paper records are to be preserved. In the case of email, it is usually the end user who decides whether to save an email message. In this case it is not only the email message that has to be stored, but also the accompanying contextual metadata.

Describe the classification and dossier-making procedure
By applying a classification scheme (as described above) an email message is assigned to a dossier. If the classification scheme is based on tasks or activities, the relationship with the work process can also be established and classification take place. In the email templates Testbed has developed, the end user is forced to indicate the dossier and work process to which an email message belongs when sending and receiving email. There are of course other methods of establishing this link. In view of the nature of email, we recommend that the individual end user establish the link at the desktop when email messages are being created or received. Remember that if end users have to add too much metadata manually to email messages, this will undermine their acceptance of the procedure.

Regulate access to the stored email messages
Access opportunities are closely related to your choice of storage format and the quality of the metadata. If the email messages are stored on a central server, as in the Testbed email demo, the email messages can in theory be made generally accessible. The question is whether the organisation wants that; management will

[24] For determining meta-data, see the Regulation mentioned previously, under section 12, or *Een uitdijend heelal? Context van archiefbescheiden,* (Dutch document: An expanding universe? Context of records), H. Hofman, Stichting Archiefpublicaties, Jaarboek 2000.

have to decide. Taking the policy of the organisation as the starting point, management should assign authorisation for access or access control, or possibly delegate this to your department. The ICT department does the actual implementation, of course.

Make sure that email is preserved in XML
Email messages eligible for long-term preservation must be converted from a proprietary format to XML. This is an open standard that is very suitable for the long-term preservation of email messages. A complicating factor is the attachments, which, for the time being, have to be preserved in their original format. The preservation strategy for these attachments will depend on the file format and record type.

Make sure that policy is tested regularly
Information technology changes quickly, and that applies to organisations too. The requirements placed on digital archiving are likewise under development. This means that policy should be regularly tested and/or adapted.
It is to be expected that better software for managing digital records will be available in the future. Testbed therefore also advocates preserving the original data file, in this case the transmission file.

The coming years will represent a transition period in which your department can prepare to give new shape to its task. Implementing the preservation of email offers you plenty to go on.

6.3 Action plan for ICT specialists

Introduction

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

But that's not the end of the story. In an organisation, different people are involved in the sustainable preservation of email messages: from the line managers in an organisation through to the end users who have the email facilities at their disposal. The concrete actions listed below are oriented specifically towards:

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

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

Prior conditions

The use of email has increased exponentially over the last few years. Many organisations are developing policy for dealing with email. At the same time, people are increasingly realising that email messages have to be preserved too, for example for operational management or for accountability. To steer this in the right direction, various people in the organisation have to take action. The starting points are that an email policy has been formulated and that the records manager has formulated procedures for selecting the email messages eligible for preservation. In addition to this, end users must have been trained to use the email application provided in your organisation.

The ICT department is indispensable in enabling the successful preservation of email messages. The technical ICT issues under discussion when implementing an email preservation strategy are described below. It is, however, not possible to indicate exactly how the proposed preservation strategy should be implemented. This is because it depends on the existing computer environment and the specific requirements of the organisation in question, and these are different in every case. We will discuss the most important requirements, however, and will summarise possible system architecture.

Concrete actions

The concrete actions you have to undertake relate to:

a. General principles
b. Recommendations about format and implementation possibilities and
c. Practical matters

a. General principles

Save the email messages that are to be preserved on a centrally managed system and not on the computers or in the personal files of the individual users. This helps to prevent email messages from being deleted, either intentionally or accidentally (for example if the user gets the message that his or her personal mail box is full); access to the centrally stored email messages can be controlled, both to keep the information available for those who need it and to prevent illegal access. A central system also enables the storage media- usually a combination of discs and tapes - to be controlled and managed. This also covers the making of copies and backups. Remember that, in the context of digital longevity, there is a world of difference between preserving backups and the sustainable preservation of digital records, including email messages.

Record metadata automatically as far as possible
Much of the metadata of an email message is found in the email headers: for example, who the message is from, who it is addressed to, the subject and the date. Make sure that this data is picked up automatically by the email preservation system so that the user has less work to do and the risk of errors is reduced. You of all people know that it is important that the preservation system, just as any other system, must be user-friendly if it is to be broadly accepted by those working with it.

There is, however, metadata that must be filled in by the users at the time they select the email message for preservation. Make things as easy as possible for them by developing templates with default values and drop-down menus from which the user can choose the right value. This increases uniformity in the data entered and reduces the risk of errors. An important example of this is the metadata about the recipients of an email message. A person's email address may be insufficient to clearly identify that person. It is important to store not only the email address but also the full name and job and data about the organisation. Bearing in mind maximum user-friendliness, the email application can be linked to a suitable database (like the Contacts directory or the Outlook address book). The information then only has to be entered once.

Metadata about the classification and the context of an email message, for example the dossier the message belongs to, is required by the central preservation system for sorting the stored messages, especially to support a search function.

Make sure that the preservation system adds a preservation log file (audit trail-information) to every stored email message
A log file of this type must contain metadata about the computer environment, such as the version of the email application in use, the version of the preservation system in use and a list of any preservation activities performed on the email, such as the date and the time the email was received into the preservation system. See Appendix C for more information on the recommended content of the Preservation Log File.

Apply the organisation's house style to the email's appearance
How easily email can be preserved for the long term is generally highly dependent on the way in which it is made. Users can be helped to make official email messages by a combination of guidelines and software tools. Tools can also be used to apply a general house style to official email messages, in other words, a fixed structure and appearance as regards font, colours and logos. This can be done, for example, by creating the email message directly in XML (a strategy investigated in the Testbed email demo) via a user-friendly interface, so that the user does not need to know anything about XML, and then having a standard XSL layout model applied by a central system to be able to send the message as HTML.

Do not save email in a proprietary format but in XML
Chapter 3 of this publication describes the authenticity requirements for the long-term preservation of authentic email messages. The preservation system must enable the messages to meet these requirements. The storage format must of course be able to reproduce all the important characteristics of the message. As described in chapter 3, these can be classified as follows: content, context, structure, appearance and behaviour. These characteristics must be stored in such a way that they can be reproduced as easily and accurately as possible both now and in the future.

We strongly advise against storing email messages in proprietary formats, since they require specific software for data interpretation. Although other record types, such as text documents, are more difficult to manage without the software that goes with them, email messages can be preserved relatively simply in a 'neutral' storage format. This is because the Internet Engineering Task Force (IETF RFC2822 and related extensions, MIME in particular) has developed a well-defined standard for the format, in which email messages must be sent. Storing messages in this RFC2822+MIME format could be one option for the long-term preservation of email messages. We therefore recommend this format, but only as one component of a preservation strategy (see chapter 5). This format alone is not enough and that is why we recommend a strategy based on XML, which we describe below.

Dealing with attachments
Attachments in email messages make the problem of long-term digital preservation much more difficult. Since, in theory, any type of file can be attached to an email message and sent with that email message, a preservation system for email must therefore be able to preserve many different types of attached files. In other parts of this series we will discuss the preservation strategy for these file types. Here we will restrict ourselves to describing only the way in which the email preservation system has to be arranged so that attachments can be dealt with as simply as possible.

b. Recommended format and implementation possibilities

The strategy recommended by Testbed for preserving email messages is described in detail in chapter 5. A brief summary is given below, followed by a few comments on the possibilities for implementing this strategy.

The most important components for a sustainably preserved email message are:
- The message text (body), if possible in more than one form (for example, in both a plain text layout and in HTML)
- Attachments: images or other objects (often present in email messages marked up in HTML) that belong with the email message
- Metadata about context

To be able to preserve email messages authentically over the long-term, the components must be present in an email preservation object, together with the following extra components:
- A preservation log file (in fact an audit trail of the actions carried out on the email message, and the technical data needed for preservation)
- The original transmission file and
- Extra metadata

Testbed recommends the use of XML as a framework to create a preservation object for preserving email messages. This preservation object contains the most important contextual metadata for the digital record (for example, information about the subject, the sender and recipients and data about date and time) and contains the structure of the email message. It also contains links to separate files like the message text

(body), attachments and any related objects. At the same time, the framework indicates the relationship between these files. Please refer to chapter 4 for an extensive description of the 'XML as framework' strategy. The structure of the preservation object discussed in chapter 5 is illustrated in the diagram below.

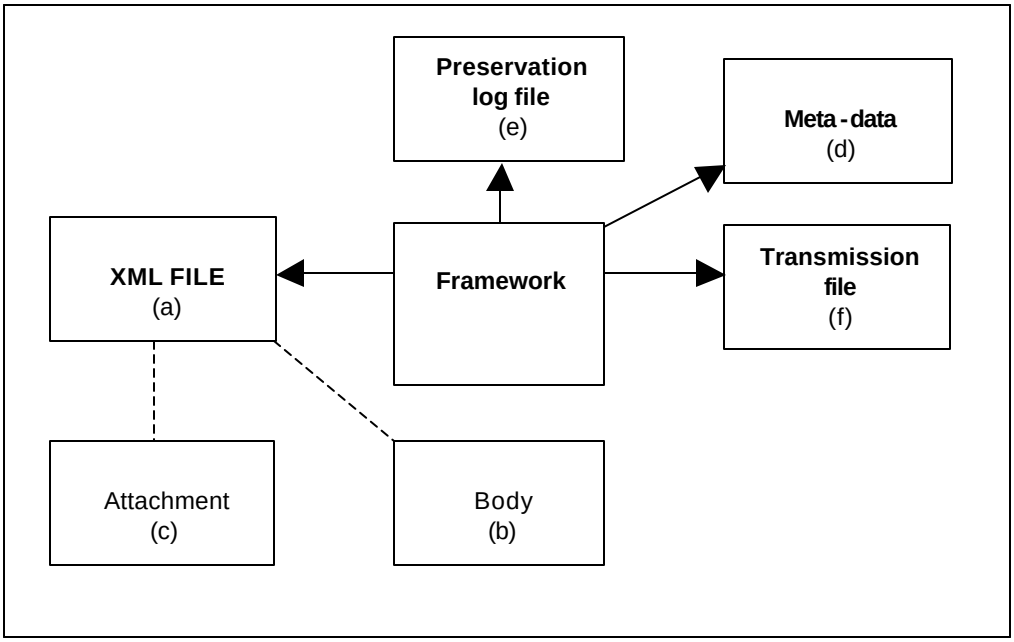


Figure 14. Diagram of the email preservation object

The mechanism for linking the components of the email preservation object is determined by the chosen implementation strategy, although it is essential that every component has a unique ID. This could be, for example, a database ID, a URL or the path name of a file system. Care must be taken with this last ID that no name conflicts occur. The XML file that represents the email message can be linked to the separate files with the message text and attachments via the components' IDs.

The structure of the XML file must be described in a DTD or XML schema so that individual XML files can be validated. The advantage of separating the components of the message (see figure) is that the structure and content of the message can be preserved more easily. The different types of files can also be kept up to date more easily in the system (for example, plain text, HTML, MS Word, PDF or JPEG), so that the most suitable preservation strategy can be applied to each of these types.

This email preservation strategy can be implemented in various ways. The most suitable manner depends on the existing computer environment and the requirements of other related systems. We have set out three points below that deserve special attention.

- The email messages can be stored via an RMA (Record Management Application) or DMS (Document Management System). Although such systems offer many useful facilities and a well-designed environment for controlled information storage, consideration must be given to the long-term preservation of digital documents, which the system does not provide for. It is

crucial that the preservation object is stored in the RMA with all its aspects intact.

- The standard email application (for example, MS Outlook) has to be adapted in such a way that metadata can be collected and that communication with a central service is possible with a view to preservation. Implementing the central system's message storage interface as a web service, in which communication between client and server takes place by means of SOAP messages sent via HTTP, is a relatively simple solution that is well-supported by modern software tools such as .NET and J2EE.
- The central storage system can consist of a file system, a database or a combination of the two. The most suitable method depends on the size of the system, on the question of whether it is a long-term or short-term solution and on the requirements for access to the stored email messages. A method that is often used is to store metadata about the email message is a relational database (so that the data can be easily looked up) and making links to files stored on a server. A 'native' XML database can be used for this method instead of a relational database.

Searching for, finding and opening a stored email message

A storage system for email is only practical if the information stored in it is accessible. It is important to be able to search for specific email messages. The central system must offer browse and search functions to allow users to search for the information they need. Searches should be possible by metadata (for example, by sender, recipient, date, subject, dossier, work process, etc.), but also full text in the content of an email message. Discuss with the records manager what the classification codes and names should be used as search entries. These are important for maintaining the relationships between different records.

Once the email has been found, the user must also be able to read it and interpret it. To make this possible, the email message should either be converted into the format used by the organisation's standard email application or, for example, with a HTML Internet-based viewer.

Installation, maintenance, training and support are needed

Assuming that most government employees work with email, many will also make use of the email preservation system. Installation of the specially adapted email application, as developed by Testbed, should be accomplished within existing systems with a view to centralised management and a central rollout of applications. The use of a 'thin client' strategy based on web browsers is one way of keeping problems to a minimum during installation and configuration. The disadvantage of this, however, is that the functionality of the system is limited and that the users have to transfer to a new, less well-known email application.

Since this strategy requires new working methods and changes to the existing email application, the users will need training. By making the software as user-friendly as possible and ensuring that it resembles existing standard applications as far as possible, training can be kept to a minimum. Nonetheless, training is still important for teaching users how to use the software effectively. At the same time, the ICT department, together with the records manager, will have to offer them support in what will be a considerable change in their working methods.

c. Practical issues

When designing and configuring the preservation system, the following practical issues have to be considered:

- Security: access to the central storage system will have to be carefully controlled to prevent unintentional or deliberate damage to the stored information (set up an access classification system, see also NEN_ISO 15489);
- Backup: just as for every important IT system, an appropriate backup strategy is required so that the system can be reinstated if a system crash occurs, if the system is unintentionally or deliberately damaged or in the event of a calamity like fire or flood;
- Flexibility: different groups in an organisation may need other metadata and these needs may change as time goes by, so it is an advantage to keep this aspect of the system design as flexible as possible. The records manager will indicate this after consultation with the user;
- Distributed email storage systems: in a large organisation with many departments it may well be more practical to have a number of smaller email storage systems than one very large system. In this case it is important to ensure that the cohesion is monitored and regulated, because searches throughout the system are a requirement;
- Response times and reliability: because users will often need to use the storage system as part of their day-to-day work, short response times and the best possible reliability is called for. Two things are important here. Firstly the user must be able to save an email message in the storage system quickly and easily. Secondly, the information already stored in the system must be easy to access and easy to find. There may be widely divergent patterns of use throughout the business processes in this regard.

6.4 Action plan for end users

Introduction

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

But that's not the end of the story. In an organisation, different people are involved in the sustainable preservation of email messages: from the line managers in an organisation through to the end users who have the email facilities at their disposal. The concrete actions listed below are oriented specifically towards:

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

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

Prior conditions

Ultimately, you as the user of email are the one who determines whether it will be possible to actually implement the recommendations in this document. After all, you are at the start of the chain, at the source, by which we mean that that you create, send, receive and manage the email messages. In so doing, you determine to a great extent whether your organisation is capable of the sustainable preservation of email messages.

The fundamental prierequisite is that your organisation has formulated an email policy and that there are agreements and procedures in your organisation about how to deal with this issue. Different parties have a role in all this, like management, the records management department, the ICT department and you as end user. The following section describes the things you should think about, especially when creating email messages. Because if anything has emerged from our research, it is that digital longevity begins at the source, with you.

Concrete actions

Three areas of focus can be distinguished with regard to your concrete actions:

a. Addressing email messages (header information)
b. Drafting an email message (the message and any attachments or inserted items) and
c. Managing email messages (incoming and outgoing)

a. Addressing email messages

The information typed in the header of an email (the fields To; CC, BCC, Subject) is very important because it provides mainly contextual information such as who is the sender and the recipient, what is the subject of the message, etc. It is therefore important that you fill in these fields correctly. The creator must record this information at source, because you, above all, are the one who knows the context in which the email was made.

Always use the address book in your email application. The address book contains extra information about the people to whom you are sending messages. This information is stored together with the email message so that this context information is always present. If you insert addresses from the address book into the email message header, not only is the email address automatically added to the message, but also, for example, the recipient's name as it appears in the address book. This is particularly important in those cases where people use 'exotic' email addresses, like ens-p45ht@planet.nl. Of course, the address book must be filled in correctly and maintained for this approach to be successful. An organisation's general address book is usually managed centrally. We also recommend that you fill in and maintain the address information for your personal contacts as completely as possible.

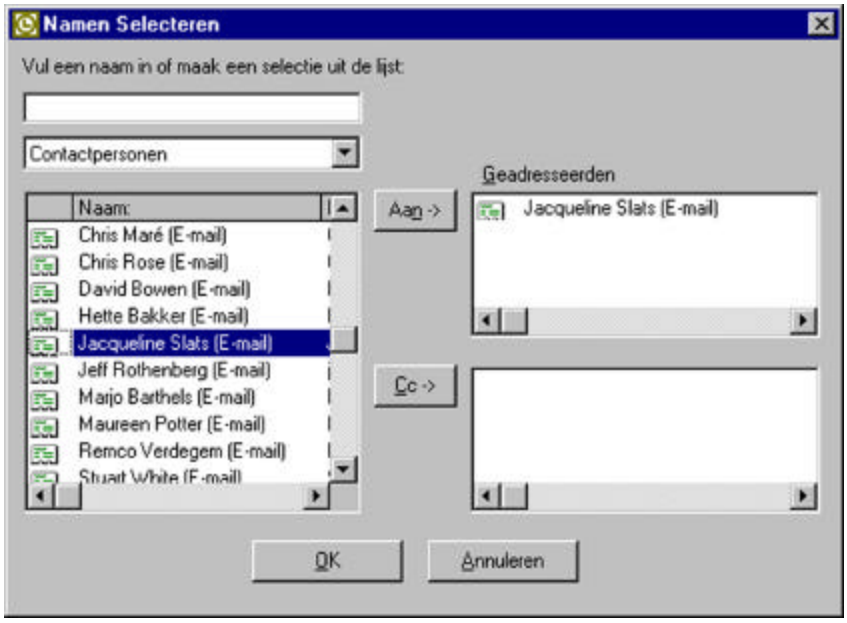


Figure 15. Example of the use of an address book

Be circumspect when using distribution lists
We have already said that it is important to know the names and addresses (and preferably the job) of the recipients. This information is not, however, always registered (or correctly registered) on distribution lists. Depending on the email application and the type of email (internal or external), the name of the list and the names of the people on the list do not always appear in the message. It is possible for a distribution list to be sent with an internal email message, without the names of the members of that distribution list being visible. You can, however, find out who is a member of the list in question by using the address book. But a distribution list is a dynamic list. New names are added to the list and old ones are removed as time goes by. The status of the distribution list is often not registered during its lifetime, so that it becomes nigh on impossible to find out who was on the list at a particular time.

If you want to keep a list of the names of the people to whom you have sent a message and for which you used a distribution list, add this information clearly to the content of the email message. Whether this is necessary depends on the type of message you are creating. Not all recipients names need to be saved in the content of a message sent by way of listserv facilities, for example. After all, that could signify an invasion of those people's privacy. On the other hand, an email message sent to members of a distribution list to consult them about draft documents should probably contain the names of the recipients, since it should be possible to prove that certain people have been able to see the document and comment on it, before it was released. In this situation you are better to include the email addresses of the recipients in the 'To' field rather than to use a distribution list.

In the case of distribution lists used in external email messages, the distribution list is usually 'translated' by the email application into a list of separate members. The name of the distribution list is then lost. If you attach importance to this name (for example as context information), you will have to include the name of the distribution list in the content of the email message.

Always give your email messages a subject
It sounds logical, but it is amazing how many people forget to do so. The subject line should be unique and informative. It is important information that you can use to sort and evaluate your messages. The subject line of an email is often also the title of the email message, so make sure that the subject is relevant and useful.

Only use message options like urgency when absolutely necessary
Many modern email applications enable the user to specify Urgency and Sensitivity settings for the email message and to add flags to the message that provide extra information about the message straight from the header (priority, for instance).
This advice has not so much to do with preservation- email applications send flags reliably and they can be preserved - than with the differences between email applications. Although flags and urgency settings can be useful, not all email applications are equipped to reproduce them correctly. In some cases, the recipient does not get to see them at all and it is possible that any extra information or significance sent with the message by means of flags or settings become lost. If you are certain that the recipient uses the same email application as you do, you can use flags and settings. If you do not know this for certain, try to add this extra information to your email message in another way, for example, by including this information directly in the subject line or in the body of the email message.

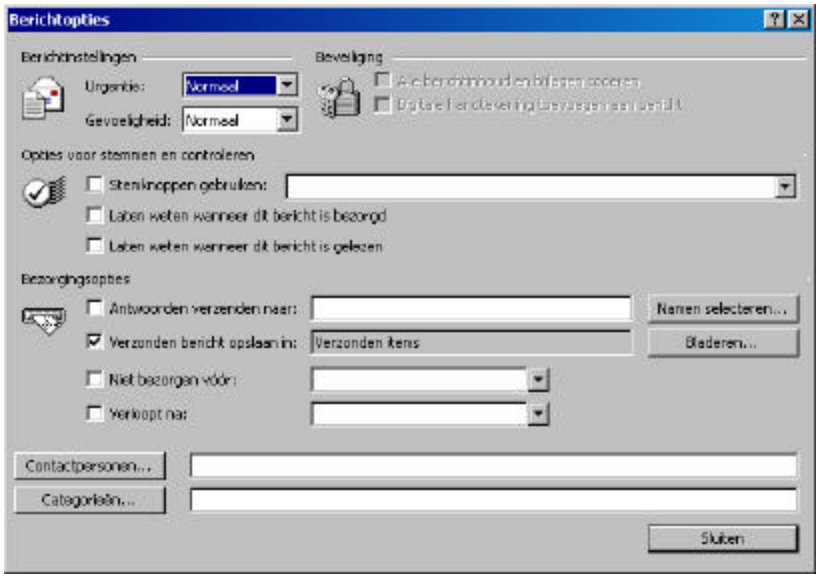


Figure 16. The message options for urgency in Outlook2000.

b. Drafting email messages

Our study has shown that email messages in plain text can usually be sent and received without any problems. The simpler the email message is, the greater the chance that the addressee (recipient) sees exactly what you intended to send. The more complex the email becomes, the less likely this is. The moral of this story? Keep your email as simple as possible.

Where possible, make and send your messages in plain text or in HTML format[25]
Messages in plain text are usually supremely suitable for simple communication. More formal and official email messages, possibly containing images and logos, can be made in HTML. Remember, however, that the more complex email messages are, the more difficult it is to sustainably preserve the email messages. What is more, old email applications do not support HTML.

Beware of using MS Outlook's Rich Text Format (RTF), because this format is specific to Outlook and, when sent, is coded in a file that may also contain attachments and layout information. This data has to be translated so that it can be read by other email applications. MS Exchange, which is often used together with MS Outlook, does this translation. What is then sent depends on the MS Exchange settings.

[25] You can decide on the format your messages are sent in by way of your email application's settings. Outlook 2002 offers the options Plain Text, Rich Text and HTML. Many government organisations work with Outlook, which is why we have used this application as an example. Other systems, like Novell Groupwise and Eudora, offer different format options. When using these systems, users will have to investigate their options and act according to the general advice given above.

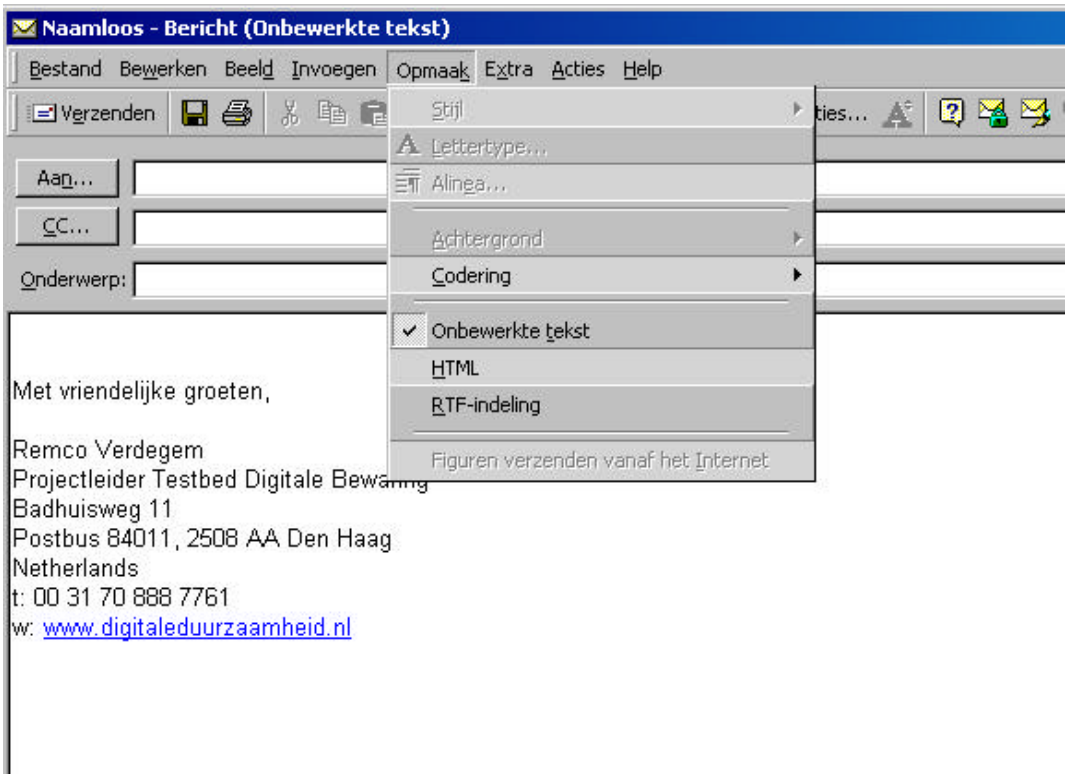


Figure 17. The different format options in Outlook.

Do not use automatically updating fields in your email messages, such as automatic date and time fields. Always enter this data as 'hard' text. Automatic fields are not stable and may update every time you open the email, thus causing the content and context of the email message to be lost immediately.

Use attachments sensibly
Make sure that you send your files in the correct file format. For example, send an image as a bitmap or JPEG file. Images are all too often pasted into an application like MS PowerPoint or Word. These applications produce files in proprietary formats, while almost every viewer can render bitmap and JPEG files.

Do not 'insert' when replying to email
Normally you reply to an email message you have received by clicking on <Reply>. There is nothing wrong in that, as long as it is done properly. If you want to respond to the content of a message, type your comments above the original message and leave some space between your signature and the headers of the original message. In that way, your comments are kept separate from those of the other person. Do not forget that your message may be read again in twenty, thirty or forty years' time and you will not be around to give a detailed explanation.
This procedure is even more important if you are corresponding with several people in one and the same email message. If five people respond to this email message one after the other and each one inserts his or her comments in your email message, it will soon no longer be clear who has written what.

Use a signature block
By adding a 'signature block' to the end of your email message, you provide the recipient with important contextual information. Remember that a signature block is not a digital signature. The term 'digital signature' is often used wrongly, but nowadays refers to the technology with which a certain type of encoding can be applied to files to send them via the Internet.

Signature blocks are blocks of information added to the end of an email, which provide information about the identity of the sender and some contact information. It is an addition that will give future users of the digital record extra contextual information.

It is worth using two signature blocks: one for internal use among members of your team, department, project or organisation and one for external use that is added to all official, outgoing correspondence.
An internal signature block must contain at least the information needed to identify an employee. This is:

- The sender's name (for example, Jacqueline Slats)
- The official job of the sender (for example, Programme Manager)
- The project or the department (for example, Digital Preservation Testbed)
- The organisation (for example Stichting ICTU)

External signature blocks must contain sufficient information that recipients can find out who the sender is without clicking on the Reply button:

- The sender's name (for example, Jacqueline Slats)
- The official job of the sender (for example, Programme Manager)
- The project or the department (for example, Digital Preservation Testbed)
- The organisation (for example Stichting ICTU)
- The address and the location (for example, Nieuwe Duinweg 24 – 26, The Hague)
- The telephone number (for example, +31(0) 70 888 77 69)
- The sender's email address (for example, Jacqueline.Slats@ictu.nl)
- The project or department's website (for example, http://www.digitalduurzaamheid.nl)

c. Managing incoming and outgoing email messages

Initially, you are responsible for managing your incoming and outgoing email messages. This means that if you are the recipient of an official email message from an external contact that is eligible for preservation, you must store that message in the appropriate directory. If you are the sender of an email message that is to be preserved, you are also responsible for this.
Internal (administrative) messages within the organisation only have to be managed and stored by the sender. They do not therefore have to be stored for everyone who has received a copy. Non-administrative email messages between different departments in the same organisation may relate to different work processes and must therefore be stored by both sender and recipient.

Ensure that the Inbox is well managed
Delete transient incoming messages immediately after you have read them. In that way you keep your Inbox workable. Delete transient messages from the Sent Items window too and empty the Deleted Items directory regularly.

Good temporary storage
If a specially adapted system has not yet been implemented for preserving your email messages, you will have to store the email in another way until such a system has been installed. A simple way of doing this is to create directories for those email

messages that have to be preserved, so that not all email messages are placed in one huge directory, which would make tracing them extremely difficult. Create the directories in consultation with the records manager in your organisation. Your email messages will then be stored in their 'original' format until it is possible to convert them to XML. The same applies to email messages that you share with others, such as in a department mailbox. Make sure that incoming and outgoing email messages that belong together are put in the same directory so as to maintain their relationship.

As has been said already, this is no more than a temporary solution until a specially adapted email system has been installed that enables you to generate your email messages directly in XML and store them in a central location. Attachments have to be stored in their original format together with the message. Since attachments are often in a proprietary format like Microsoft Word, Excel, PowerPoint or Adobe's PDF, they may require a different approach for long-term preservation than the email message itself. By storing them separately, their transmission encoding is immediately decoded and the relevant record-type can then be checked for approaching obsolescence so that appropriate preservation action can be taken, for example, migration to a higher version of that format.

Never paste the content of an email message into a different application like Microsoft Word and then delete the original message. This would seriously damage both the authenticity and the integrity of the message. As soon as you carry out this action, a large quantity of important meta-data is lost, which cannot be retrieved from the new format.
Ultimately, email messages in XML can be transferred to a Document Management System (DMS) or a Record Management Application.

Glossary

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

ASCII
American Standard Code for Information Interchange
Generally accepted standard containing the meaning of symbols (bytes). After all, computers only understand numbers. ASCII code 65, for example, stands for capital A. ASCII is also a protocol for transferring files from one computer to another. It is really only suitable for text files.

Attachments
Additional binary or text files sent with email messages.

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 read in files made with earlier versions of the same software. Incidentally, most software is only backward compatible to a limited degree.

BCC
Blind Carbon Copy
The option in email applications to send a copy of an email message to another person, without the recipient being informed of this. See also *CC*

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

CC
Carbon Copy
In this case the recipient can see in the header who has received a copy of the same email message.

Computer file
A grouping of data in a particular storage format

Context
All the administrative and organisational, managerial and 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
Converting or transforming data to a different file format.

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

DIV
Documentaire Informatie Voorziening (Documentary Information Provision). The process of communicating by way of documents; this concept applies to paper and digital documents, both text and numeric, process control data and images.

DMS
Document Management System
A system that offers functionality for acquiring, storing, and retrieving documents, including their management while implementing, administering, passing on and authorising users. Document Management Systems monitor access to files and maintain a content audit trail

Email
Electronic mail

Emulation
Imitation of the old hardware and/or software environment by means of software. Emulation software runs on current and future hardware platforms, thus avoiding the problem of hardware and software becoming obsolete.

Form
The outward appearance in which the structure and layout are visible

HTML
Hyper Text Mark-up Language, a descriptive language for form atting hypertext documents. Is used to write pages on the World Wide Web

Integrity
Property of a digital document such that, when consulted, its form, content and structure are the same as the form, content and structure at the moment it was laid out.

J2EE
Stands for Java 2 Enterprise Edition and is a development platform that has grown over the last few years into an industry standard for developing large-scale Java applications.

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.

MIME
An Internet protocol that enables the binary content of email messages to be coded. MIME can, for example, be used to code a graphics file or a Word document and include them as attachments to a text-based (ASCII) message. The recipient must also be working with MIME to be able to decode the attachment.

PDF
Portable Document Format. A file format developed by the Adobe Company for exchanging documents while retaining their quality and design.

PKI
Public Key Infrastructure. A system of digital certificates, certificate authorities and other registration authorities that can verify the validity of every party for an electronic transaction.

Platform
All the equipment and operating software on which the application software runs.

Preservation
Processes and activities relating to ensuring the technical and intellectual maintenance of authentic records through time

RMA
Records Management Application. Application software for recording and managing records and making them available.

RTF
Rich Text Format. A format for a text document with layout. A Microsoft protocol for arranging files that contain bold, markings, underlining and many other layout characteristics.

SOAP
Simple Object Access Protocol. A protocol for access to applications via HTTP and XML. It is an XML based specification for the execution of Remote Procedure Calls (RPC). A RPC is a set of protocols for giving instructions that can be sent and executed over the network.

Storage
The structured preservation of digital information, like files and records, on magnetic or optical media.

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

Transmission file
The email file as it is received or sent by a mail server.

URL
Uniform Resource Locator (address). 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
Software that enables certain files to be looked at, but not necessarily manipulated.

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

Wrapper
A term that here refers to an approach in which XML is used as a sort of envelope, a casing.

XML
Stands for eXtended 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 cominations.

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

Bibliography

Boudrez, Filip	*<XML/> en Digital Archiveren (Dutch document: <XML/> and Digital Archiving (2002))* http://www.antwerpen.be/david/teksten/xml_digitalarchiveren.pdf
Crocker, David H	Standard for the Format of ARPA Internet Text messages – RFC # 822 http://www.ietf.org/rfc/rfc0822.txt
Giesbers, Saskia (RMC)	Records Management Terminology (6 March 2002)
Horsman, Peter	Archivering van Elektronische Post; methoden, meningen en alternatieven (Dutch document: Archiving Electronic Mail; methods, opinions and alternatives (Digital Longevity Programme, The Hague 1999))
Feeney, Mary (Ed)	Digital Culture: Maximising the Nation's Investment (National Preservation Office UK, 1999)
Hovy, Lodewijk	Sporen nalaten of uitwissen?Het bewaren van persoonsgegevens / (Dutch document: To leave tracks or cover them up? Preserving personal data / (Archieven blad, December 2001))
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
Lourens, Wim, et al	*Emulation and Conversion: Organisational and Architectural Overview of an electronic Archive* http://www.library.tudelft.nl/e-archive/Documents/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
Natural Resources Canada	Guidelines on Managing Electronic Mail Messages (January 2000)
Ploeg, F. van der	Regeling geordende en toegankelijke staat archiefbescheiden (Dutch document: Regulation on the Arrangement and Accessibility of Records)
Prins, Prof. J.E.J. Matthijssen, L.J.	De digitale overheid en de wet; juridische kaders voor gebruik van digitale documenten bij overheden (Dutch document: The Digital

	Government and the Law; legal framework for the use of digital documents by governments (Digital Longevity Programme, The Hague, 2000)
Digital Longevity Programme	*De Digitale Overheid en de wet* (Dutch document: The Digital Government and the Law) http://www.digitalduurzaamheid.nl/bibliotheek/docs/dig-overh-wet.pdf
Digital Longevity Programme	*Archivering van Elektronische Post* (Dutch document: Archiving Electronic Mail) http://www.digitalduurzaamheid.nl/bibliotheek/docs/archelp.pdf
Digital Longevity Programme	*Naar een verantwoorde archivering van email (Dutch document: Towards Responsible Email Archiving (Digital Longevity Programme, The Hague, 1998))*
Editors	Email dertig jaar oud (Dutch document: Thirty Years of Email) (Archievenblad, December 2001)
Rothenberg, Jeff & Bikson, Tora	Carrying Authentic, Understandable and Usable Records Through Time (1999) http://www.digitalduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf
National Archives Inspectorate	Wet- en regelgeving (Dutch document: Legislation and Rules) www.rijksarchiefinspectie.nl/wetgeving/
Digital Preservation Testbed	XML en digitale bewaring (Dutch document: XML and Digital Preservation (2002)) http://www.digitalduurzaamheid.nl/biblioth eek/docs/white-paper_xml-nl.pdf
Thomas, Wimpe	XML: de mogelijkheden en valkuilen voor de overheid (Dutch document: XML: the Opportunities and Pitfalls for Government; 19 September 2002)
VERS	Victorian Electronic Records Strategy Final Report http://www.prov.vic.gov.au/vers/published/final.htm
Working group on the use of Email	Richtsnoer Emailgebruik t.b.v. de Rijksoverheid (Ministry for Internal Affairs and Kingdom Relations, The Hague, 2001) (Dutch document: Guidelines for Email Use for Central Government)
Zuurmond, A, Mies, K.	Winst met ICT in uitvoering. (Dutch document: Profit with ICT; Zenc, The Hague, June 2002)

Appendix A

Email/XML Demo: A Technical Description

Email/XML demo: a technical description

Introduction

In the Testbed Digitale Bewaring project we have recently developed prototype software to allow us to investigate in detail the issues around the long-term preservation of email messages and to illustrate possible solutions to the problems that many government organisations are faced with.

Objectives

There were three aspects of email use that we considered:

- Email is now becoming more and more widely used for official communications, but in some cases people still use quite an informal approach to composing their emails. We wanted the content and style of official messages to resemble more closely a letter written on official headed paper.
- If an email needs to be filed, many organisations either print the message onto paper and store it in a paper file, or messages are stored in an ad hoc way by individuals in their personal email folders. We wanted to set up a central filing system for email messages, without involving the complexity of a full document management system.
- The email messages must be stored in a way suitable for easy and reliable long-term preservation, so that they continue to be accessible and understandable for as long as required: that could be 10 years or 100 years or more.

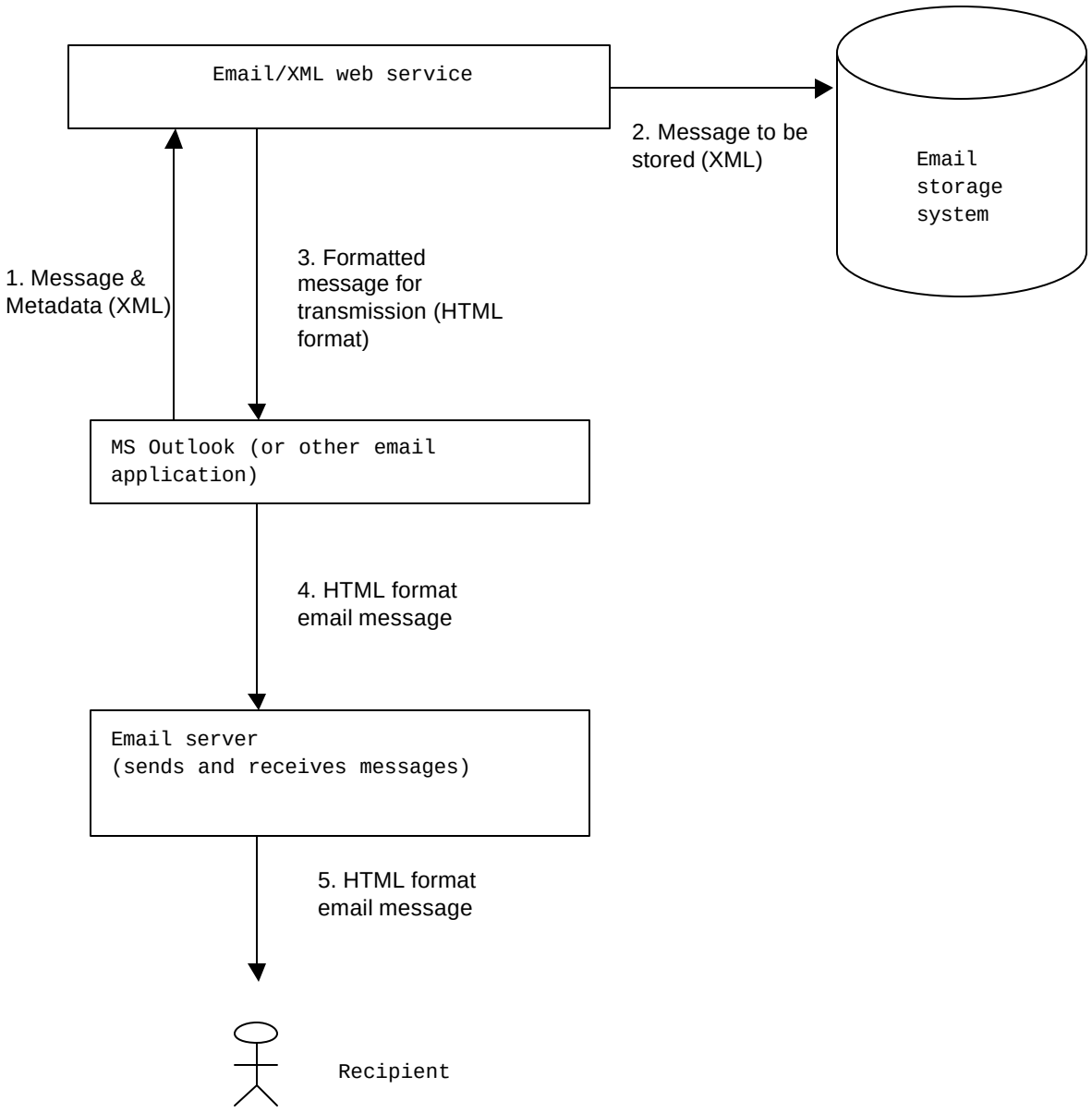
We wanted a solution that would be simple to use and relatively simple to implement.

Our solution

We based our approach around Microsoft Outlook: this is because this is the most widely used email application in the government and because we wanted our solution to be integrated into the familiar working environment of the users and also to be able to make use of many of the facilities provided by Outlook, without having to re-create them for ourselves. Although we chose Outlook for our demonstration, a very similar approach could be taken with other email applications and we do not intend to imply that Outlook is better or worse than any other choice.

Our approach involves a customisation of Outlook, using an ActiveX DLL written in Visual Basic. This adds a number of additional features to the standard Outlook and also prevents access to some of Outlook's usual functions. The overall system follows a client-server architecture. The clients, that is the instances of the customised Outlook running on users' desktops, communicate with a central server that takes responsibility for archiving the messages and for applying the house style to outgoing emails. We refer to this as the "archiving server", to distinguish it from the usual email server, which is still required in the normal way.

The figure below shows the system in use.



Metadata

The standard format for email messages (as defined by the Internet Engineering Task Force) contains a lot of useful information, but for meaningful long-term preservation, it is beneficial to provide additional information. In our demonstration, we collect the following additional information:

- Context information such as dossier, work process;
- Information about the sender of the message, such as full name, job title, postal address, telephone number;
- Information about the recipients, such as full name and organisation.

An additional metadata entry form must be completed before a formal email can be sent. However, this requires very little effort from the user: the personal information only needs to be entered once and can then be filled in automatically by the software and the information about dossier, handeling and so on is given by choosing from a list of predefined options.
The email address alone is often not sufficient to identify a person from the point of view of preserving the context of a message. Therefore our demonstration makes use of the Contacts folder in Outlook, where additional information about the recipients of a message can be defined. Before the user can send a formal email to someone, they must enter the recipient's full name and organisation details into the Contacts folder. This is then extracted by the software and stored in the message metadata.

Storage in XML format

In our demonstration, both the message contents and the metadata are stored in an XML document. There are two main reasons for this: one is because XML is a well-defined open standard that is widely believed to be a good format for long-term preservation; the other is that use of XML for the message content allows the use of XML Style sheet Language (XSL) to define a transformation from XML to HTML. This takes the content of the message and the metadata and presents it in a formatted way, specifying the overall layout of the message, the choice of fonts and colours and the inclusion of a logo or other images. By doing this centrally, a common house style can be applied to all formal messages. Any change in the style only needs to be made in a single central place.

The Outlook user interface has been modified to guide the user through creating the different elements making up the content of the message. This is then converted to XML behind the scenes. The Outlook extension combines the message content and metadata into a single XML document. This is transmitted to the archiving server, encapsulated in a SOAP[26] message. The archiving server stores a copy of this XML file in the archive. It applies the XSL style sheet to create the formatted HTML message and sends that back to Outlook, also as a SOAP message.

Another function of the archiving server is to verify that the XML produced by Outlook satisfies the XML schema defined for the message. By checking that the XML document obeys the schema, then we can be sure that the XSL transformation will work correctly.

Attachments can be added to messages in the normal way. These are transmitted from Outlook to the archiving server. Each attachment file is stored separately in the archive and a link to the attachment and key metadata items are stored in the archived XML file. Because essentially any type of file can be attached to an email message, the long-term preservation of attachments is a difficult problem, not directly addressed in our demonstration. However, our approach of storing email messages as XML with metadata about the attachments, including information on the type of the file, allows easier control and monitoring of the type of attachment files in the archive and so is a first step to allowing a preservation solution to be applied.

Received email messages

The software also provides a function to save email messages from the Outlook Inbox (or other personal folders). The user is prompted to provide the necessary metadata items and the message is then converted to XML, following the same XML schema as that for outgoing messages. The only difference is in the way that the message content is handled. Because the user has no control over the format and style of incoming messages, the XML schema must be

[26] Simple Object Access Protocol. See www.w3.org/TR/SOAP for more information.

able to handle any kind of message that is allowable by the general standards for email. For this reason, the body of the message is saved in a separate file (typically an HTML or plain text file) and a link to this file is stored in the XML representation of the message.

Appendix B

XML Schema

XML Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<!--Dit is een schema dat als afdwingbaar sjabloon de structuur van een emailbericht vastlegt.-->
<!--Gemaakt door Hette Bakker voor Testbed, ICTU, nov 2002. -->

<xs:schema targetNamespace="xmail.ictu.nl" xmlns="xmail.ictu.nl" xmlns:xs="http://www.3.org/2001/XMLSchema"
elementFormDefault="qualified">
	<!--Het element email is de 'wortel' van het schema, dit element bestaat weer uit de elementen meta, transmissiebestand,
onderwerp, etc. -->
	<xs:element name="email">
		<xs:complexType>
			<xs:sequence>
				<xs:element name="meta" type="metaType"/>
				<xs:element name="transmissiebestand" type="bestandsgegevensType"/>
				<xs:element name="onderwerp"/>
				<xs:element name="geadresseerdenlijst">
					<xs:complexType>
						<xs:sequence>
							<xs:element name="geadresseerde" type="geadresseerdeType"
maxOccurs="unbounded"/>
						</xs:sequence>
					</xs:complexType>
				</xs:element>
				<xs:element name="cclijst" minOccurs="0">
					<xs:complexType>
						<xs:sequence>
							<xs:element name="cc" type="geadresseerdeType"
maxOccurs="unbounded"/>
						</xs:sequence>
					</xs:complexType>
```

```
                                        </xs:element>
                                        <xs:element name="bcclijst" minOccurs="0">
                                                <xs:complexType>
                                                        <xs:sequence>
                                                                <xs:element name="bcc" type="geadresseerdeType"
maxOccurs="unbounded"/>
                                                        </xs:sequence>
                                                </xs:complexType>
                                        </xs:element>
                                        <xs:element name="afzender" type="afzenderType"/>
                                        <xs:element name="inhoud">
                                                <xs:complexType>
                                                        <xs:sequence>
                                                                <xs:choice>
                                                                        <xs:element name="nietXMLinhoud" type="nietXMLinhoudType"/>
                                                                        <xs:element name="XMLinhoud" type="XMLinhoudType"/>
                                                                </xs:choice>
                                                        </xs:sequence>
                                                </xs:complexType>
                                        </xs:element>
                                        <xs:element name="bijlagelijst" type="bijlagelijstType" minOccurs="0"/>
                                </xs:sequence>
                        </xs:complexType>
                </xs:element>
                <!-- Einde van het element email. -->
                <!-- Hieronder volgen de verschillende 'types' waarnaar hierboven wordt verwezen. -->
                <xs:complexType name="metaType">
                        <xs:sequence>
                                <xs:element name="datum" type="xs:dateTime"/>
                                <xs:element name="dossier">
                                        <xs:simpleType>
                                                <xs:restriction base="xs:string">
                                                        <xs:enumeration value="Experiment 49"/>
```

```
                                        <xs:enumeration value="Experiment 53"/>
                                        <xs:enumeration value="Jansen 12"/>
                                        <xs:enumeration value="Pietersen 19"/>
                                        <xs:enumeration value="Overzichten adviesorganen van de centrale overheid"/>
                                        <xs:enumeration value="EmailXML Demo"/>
                                </xs:restriction>
                        </xs:simpleType>
                </xs:element>
                <xs:element name="werkproces">
                        <xs:simpleType>
                                <xs:restriction base="xs:string">
                                        <xs:enumeration value="Het uitvoeren van migratie-experimenten met spreadsheets"/>
                                        <xs:enumeration value="Het verstrekken van bijstandsuitkeringen"/>
                                        <xs:enumeration value="Het houden van enquête t.b.v. het verzamelen van informatie
over adviesorganen"/>
                                        <xs:enumeration value="Het verstrekken van advies over het langdurig bewaren van
emails"/>
                                </xs:restriction>
                        </xs:simpleType>
                </xs:element>
        </xs:sequence>
</xs:complexType>
<xs:complexType name="geadresseerdeType">
        <xs:sequence>
                <xs:element name="naam"/>
                <xs:element name="email"/>
                <xs:element name="organisatie" minOccurs="0"/>
        </xs:sequence>
</xs:complexType>
<!-- Hieronder volgen de twee types inhoud -->
<xs:complexType name="nietXMLinhoudType">
        <xs:sequence>
                <xs:element name="body" type="bestandsgegevens Type"/>
```

```
                        <xs:element name="gerelateerdebestanden" type="bestandsgegevensType" minOccurs="0"
maxOccurs="unbounded"/>
                </xs:sequence>
        </xs:complexType>
        <xs:complexType name="XMLinhoudType">
                <xs:sequence>
                        <xs:element name="XMLzelf">
                                <xs:complexType>
                                        <xs:sequence>
                                                <xs:element name="aanhef"/>
                                                <xs:element name="blok" maxOccurs="unbounded">
                                                        <xs:complexType>
                                                                <xs:sequence>
                                                                        <xs:element name="kop" minOccurs="0"/>
                                                                        <xs:choice maxOccurs="unbounded">
                                                                                <xs:element name="regel"/>
                                                                                <xs:element name="lijst">
                                                                                        <xs:complexType>
                                                                                                <xs:sequence>
                                                                                                        <xs:element
name="lijstelement" maxOccurs="unbounded"/>
                                                                                                </xs:sequence>
                                                                                        </xs:complexType>
                                                                                </xs:element>
                                                                        </xs:choice>
                                                                </xs:sequence>
                                                        </xs:complexType>
                                                </xs:element>
                                                <xs:element name="afsluitingfrase" maxOccurs="unbounded"/>
                                        </xs:sequence>
                                </xs:complexType>
                        </xs:element>
                        <xs:element name="stylesheet" type="bestandsgegevensType"/>
```

```
            </xs:sequence>
        </xs:complexType>
        <xs:complexType name="bijlagelijstType">
            <xs:sequence>
                <xs:element name="bijlage" type="bestandsgegevensType"/>
            </xs:sequence>
        </xs:complexType>
        <xs:complexType name="bestandsgegevensType">
            <xs:sequence>
                <xs:element name="bestandsnaam"/>
                <xs:element name="bestandstype"/>
                <xs:element name="bestandslokatie"/>
            </xs:sequence>
        </xs:complexType>
        <xs:complexType name="afzenderType">
            <xs:sequence>
                <xs:element name="naam"/>
                <xs:element name="email"/>
                <xs:element name="functie" minOccurs="0"/>
                <xs:element name="organisatie" minOccurs="0"/>
                <xs:element name="telefoonnummer" minOccurs="0"/>
            </xs:sequence>
        </xs:complexType>
</xs:schema>
```

Optional and repeatable elements are clearly recognizable to an ‘unbounded’ value. Government organisations must formulate their own categories for Dossier and Working process. The values used in this figure are just an example

Appendix C

Preservation Transaction Log

Preservation Log File

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

Technical Metadata

- Details of the original computing environment: client software = application (e.g. Outlook) + server (e.g. Exchange) + operating system.
- Details of interim formats (e.g. ASCII, RTF).
- Details of new computing environment (sufficient details must be recorded access to the records in their current format).

Preservation action metadata

- Date and time of any and all preservation action
- Person in charge for preservation action
- Details of the transformation software and
- Transformation results

Metadata which refer to the access of the records

- Privileges
- History

Appendix D

Various representations of an email message

Various representations of an email message

This appendix illustrates the various representations of an email message. Figure D1 is an example of an email message represented by the email application, which is in this case Outlook 2000'. Figure D2 shows the complete transmission file of the same email message, including all transmission data and different representations of the message body. Figure D3 represents the message body in plain text. Figure D4 illustrates the email message as an XML file as generated by the email/XML demo, developed by the Testbed team.

Figure D1: HTML representation of an email message

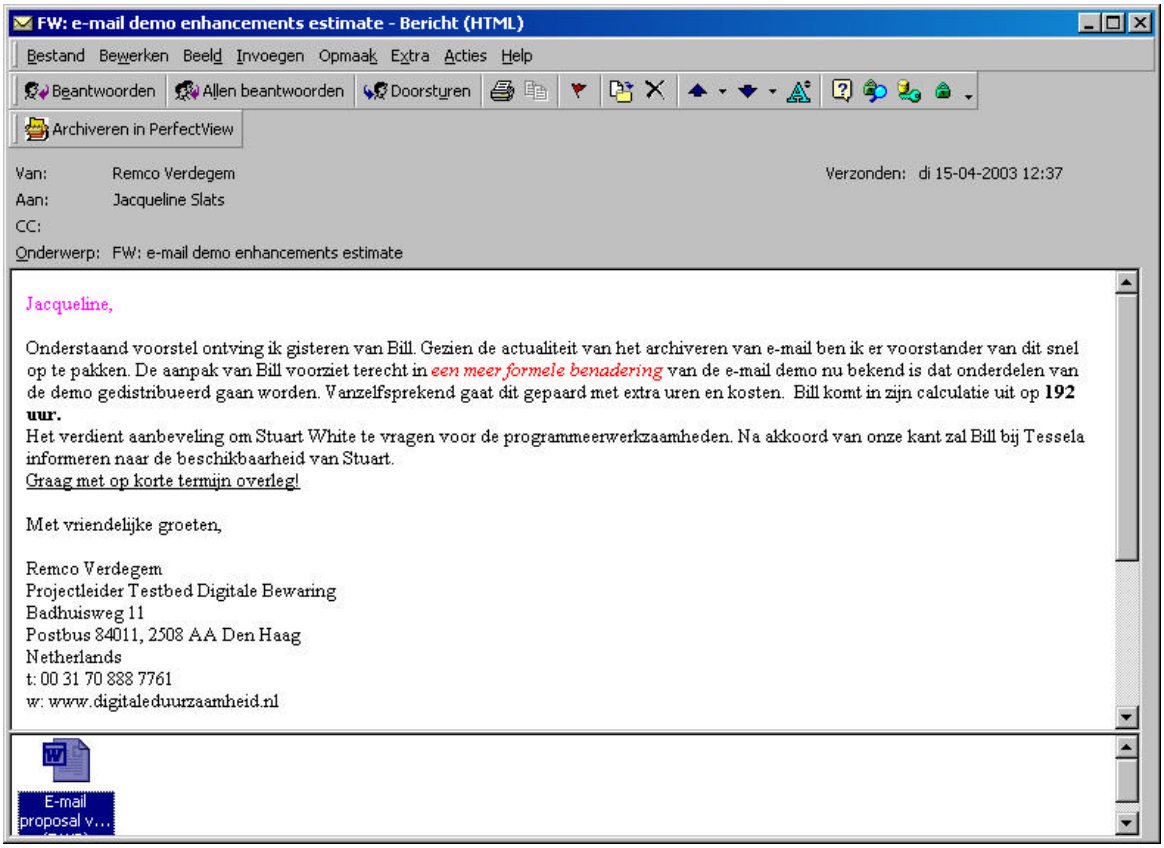


Figure D2: Complete Transmission File

```
X-MimeOLE: Produced By Microsoft Exchange V6.0.6249.0
content-class: urn:content-classes:message
MIME-Version: 1.0
Content-Type: multipart/mixed;
    boundary="----_=_NextPart_003_01C3033A.2C220780"
Subject: FW: email demo enhancements estimate
Date: Tue, 15 Apr 2003 12:31:49 +0200
Message-ID: <F5714832E4D128479178CA6877D2280805F542@gw02.dh01.ictu.nl>
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
Thread-Topic: FW: email demo enhancements estimate
To: "Jacqueline Slats" <Jacqueline.Slats@ictu.nl>

This is a multi-part message in MIME format.

------_=_NextPart_003_01C3033A.2C220780
Content-Type: multipart/alternative;
    boundary="----_=_NextPart_004_01C3033A.2C220780"


------_=_NextPart_004_01C3033A.2C220780
Content-Type: text/plain;
    charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

Jacqueline,

Onderstaand voorstel ontving ik gisteren van Bill. Gezien de actualiteit
=
van het archiveren van email ben ik er voorstander van dit snel op te =
pakken. De aanpak van Bill voorziet terecht in een meer formele =
benadering van de email demo nu bekend is dat onderdelen van de demo =
gedistribueerd gaan worden. Vanzelfsprekend gaat dit gepaard met extra =
uren en kosten.  Bill komt in zijn calculatie uit op 192 uur.
Het verdient aanbeveling om Stuart White te vragen voor de =
programmeerwerkzaamheden. Na akkoord van onze kant zal Bill bij Tessela =
informeren naar de beschikbaarheid van Stuart.
Graag met op korte termijn overleg!

Met vriendelijke groeten,

Remco Verdegem
Projectleider Testbed Digitale Bewaring
Badhuisweg 11
Postbus 84011, 2508 AA Den Haag
Netherlands
t: 00 31 70 888 7761
w: www.digitaleduurzaamheid.nl


> -----Oorspronkelijk bericht-----
> Van: Bill Roberts
> Verzonden: maandag 16 september 2002 14:48
```

```
> Aan: Remco Verdegem
> Onderwerp: email demo enhancements estimate
>
>




------_=_NextPart_004_01C3033A.2C220780
Content-Type: text/html;
    charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML><HEAD>
<META HTTP-EQUIV=3D"Content-Type" CONTENT=3D"text/html; =
charset=3Diso-8859-1">
<TITLE></TITLE>

<META content=3D"MSHTML 5.00.3103.1000" name=3DGENERATOR></HEAD>
<BODY>
<P><FONT size=3D2><FONT =
color=3D#ff00ff>Jacqueline,<BR><BR></FONT>Onderstaand=20
voorstel ontving ik gisteren van Bill. Gezien de actualiteit van het =
archiveren=20
van email ben ik er voorstander van dit snel op te pakken. De aanpak =
van Bill=20
voorziet terecht in <EM><FONT color=3D#ff0000>een meer formele=20
benadering</FONT></EM> van de email demo nu bekend is dat onderdelen =
van de=20
demo gedistribueerd gaan worden. Vanzelfsprekend gaat dit gepaard met =
extra uren=20
en kosten.  Bill komt in zijn calculatie uit op <STRONG>192=20
uur.<BR></STRONG>Het verdient aanbeveling om Stuart White te vragen voor
=
de=20
programmeerwerkzaamheden. Na akkoord van onze kant zal Bill bij
Tessela=20
informeren naar de beschikbaarheid van Stuart.<BR><U></FONT><FONT =
size=3D2>Graag=20
met op korte termijn overleg!<BR><BR></FONT></U><FONT size=3D2>Met =
vriendelijke=20
groeten,<BR><BR>Remco Verdegem<BR>Projectleider Testbed Digitale=20
Bewaring<BR>Badhuisweg 11<BR>Postbus 84011, 2508 AA Den=20
Haag<BR>Netherlands<BR>t: 00 31 70 888 7761<BR>w:=20
www.digitaleduurzaamheid.nl<BR><BR><BR>&gt; -----Oorspronkelijk=20
bericht-----<BR>&gt; Van: Bill Roberts<BR>&gt; Verzonden: maandag 16 =
september=20
2002 14:48<BR>&gt; Aan: Remco Verdegem<BR>&gt; Onderwerp: email demo=20
enhancements =
estimate<BR>&gt;<BR>&gt;<BR><BR><BR></FONT></P></BODY></HTML>

------_=_NextPart_004_01C3033A.2C220780--
```

------_=_NextPart_003_01C3033A.2C220780
Content-Type: application/msword;
    name="Email proposal version 3.doc"
Content-Transfer-Encoding: base64
Content-Description: Email proposal version 3.doc
Content-Disposition: attachment;
    filename="Email proposal version 3.doc"

d

0M8R4KGxGuEAAAAAAAAAAAAAAAAAAAAAPgADAP7/CQAGAAAAAAAAAAAAAAABAAAARgAAAAAAAAAA
EAAASAAAAAEAAAD+////AAAAAEUAAAD///
//
//
//
//
//
//
///s
pcEATSATBAAA8BK/AAAAAAAAEAAAAAAABAAAFRsAAA4AYmpiauI94j0AAAAAAAAAAAAAAAAAAAAA
AAATBBYAIjYAAIBXAACAVwAAMBUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAD//w8AAAAA
AAAAAAD//w8AAAAAAAAAAAAD//w8AAAAAAAAAAAAAAAAAAAAAAGwAAAAAAOAAAAAAAAAAA4AAAAOAA
AAAAAAAA4AAAAAAAAADgAAAAAAAAAOAAAAAAAAAA4AAAABQAAAAAAAAAAAAAAPQAAAAAAAAAAxAoA
AAAAAADECgAAAAAAAMQKAAAAAAAAxAoAAAAwAAADQCgAAVAAAAPQAAAAAAAAAGywAADIBAAAwCwAA
AAAAADALAAAAAAAAMAsAAAAAAAAwCwAAAAAAADALAAAAAAAAMAsAAAAAAAAwCwAAAAAAADALAAAA
AAAAgisAAAIAAACEKwAAAAAAAIQrAAAAAAAAhCsAAAAAAACEKwAAAAAAAIQrAAAAAAAAhCsAACQA
AABNLQAAIAIAAG0vAACsAAAAqCsAAC0AAAAAAAAAAAAAAAAAAAAAAAAA4AAAAAAAAAAAwCwAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAwCwAAAAAAADALAAAAAAAAMAsAAAAAAAAwCwAAAAAAAKgrAAAAAAAA
MAsAAAAAAADgAAAAAAAAAOAAAAAAAAAAMAsAAAAAAAAAAAAAAAAAAADALAAAAAAAA1SsAABYAAAAw
CwAAAAAAADALAAAAAAAAMAsAAAAAAAAwCwAAAAAAAOAAAAAAAAAAMAsAAAAAAADgAAAAAAAAAADAL
AAAAAAAAgisAAAAAAAAAAAAAAAAAADALAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAMAsAAAAAAACCKwAAAAAAAADALAACeCAAAMAsAAAAAAADOEwAA
4gAAAE4nAACkAAAA4AAAAAAAAADgAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAPikAAAAAAAAwCwAAAAAAACQLAAAMAAAAcEGR/n5d
wgH0AAAA0AkAAMQKAAAAAAAAMAsAAAAAAADyJwAAFgAAAAAAAAAAAAAAPikAAEQCAADrKwAAMAAA
ABssAAAAAAAACCgAADYBAAAZMAAAAAAAAADALAAAAAAAAGTAAAAAAAAA+KQAAAAAAAADALAAAAAAA
9AAAAAAAAAD0AAAAAAAAAOAAAAAAAAAA4AAAAAAAAADgAAAAAAAAAOAAAAAAAAAAgDZAAAAUHJv
cG9zYWwgZm9yIGUtbWFpbCBkZW1vIGVuaGFuY2VtZW50cw0NSW50cm9kdWN0aW9uDQ1CaWxsIGFu
ZCBDaHJpcyB3aWxsIHNldCB1cCB0aGUgY3VycmVudCB2ZXJzaW9uIG9mIHRoZSBlLW1haWwgZGVt
byBvbiB0d28gbGFwdG9wcyBpbiB0aW1lIGZvciB0aGUgT3ZlcmhlaWQgJiBJQ1QgY29uZmVyZW5j
ZSBvbiAyNyBTZXB0ZW1iZXIuICBUaGF0IHdvcmsgZG9lcyBub3QgZm9ybSBwYXJ0IG9mIHRoaXMg
cHJvcG9zYWwuDQ1UaGUgZW5oYW5jZW1lbnRzIGluY2x1ZGVkIGhlcmUgYXJlIGludGVuZGVkIHRv
IGltcHJvdmUgdGhlIHNvZnR3YXJlIGZvciBkZW1vbnN0cmF0aW9uIHB1cnBvc2VzLiAgQSBwb3Nz
aWJsZSBzZWNvbmQgcGhhc2UsIG5vdCBpbmNsdWRlZCBpbiB0aGlzIHByb3Bvc2FsLCBjb3VsZCBp
bnZvbHZlIG1ha2luZyB0aGUgc29mdHdhcmUsIG9yIGF0IGxlYXN0IHBhcnRzIG9mIGl0LCBhdmFp
bGFibGUgZm9yIGRpc3RyaWJ1dGlvbiB0byBpbnRlcmVzdGVkIHBhcnRpZXMgdG9nZXRoZXIgd2l0
aCBndWlkYW5jZSBvbiBob3cgdGhlc2Ugb3JnYW5pc2F0aW9ucyBjYW4gaW1wbGVtZW50IGFuIGUt
bWFpbCBhcmNoaXZlIGZvciB0aGVtc2VsdmVzLiAgQXMgcGFydCBvZiB0aGlzIHNlY29uZCBwaGFz
ZSwgdGhlIHNvZnR3YXJlIHdvdWxkIGJlIHVzZWQgb24gYW4gZXZlcnlkYXkgYmFzaXMgYnkgdGhl
IFRlc3RiZWQgYW5kIFRhc2tmb3JjZSBwcm9qZWN0IHRlYW1zLiAgVGhpcyB3b3VsZCBhY3QgYXMg
YW4gZXh0ZW5kZWQgdGVzdCBvZiB0aGUgcHJhY3RpY2FsIGlzc3VlcyBpbnZvbHZlZCBpbiB0aGUg
dXNlIG9mIHN1Y2ggc29mdHdhcmUuIEFsdGhvdWdoIHRoZXNlIGFzcGVjdHMgd2lsbCBiZSBsZWZ0

/ /

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABAAAAAgAAAAMAAAAE
AAAABQAAAAYAAAAHAAAACAAAAAkAAAAKAAAACwAAAAwAAAANAAAADgAAAA8AAAAQAAAAEQAAABIA
AAATAAAAFAAAABUAAAAWAAAAFwAAABgAAAAZAAAAGgAAABsAAAD+////HQAAAB4AAAAfAAAAIAAA
ACEAAAAiAAAAIwAAACQAAAAlAAAAJgAAACcAAAAoAAAAKQAAACoAAAArAAAALAAAAC0AAAAuAAAA
LwAAADAAAAAxAAAAMgAAADMAAAA0AAAA/v///zYAAAA3AAAAOAAAADkAAAA6AAAAOwAAADwAAAD+
////PgAAAD8AAABAAAAAQQAAAEIAAABDAAAARAAAAP7////9////RwAAAP7////+/////v//////
//

//
//
///1IAbwBvAHQAIABFAG4A
dAByAHkAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAWAAUB////
//////8DAAAABgkCAAAAAADAAAAAAAAARgAAAAAAAAAAAAAAAEByn/5+XcIBSQAAAIAAAAAAAAAA
MQBUAGEAYgBsAGUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAA4AAgD///////////////8AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAcAAAAGTAAAAAAAAABXAG8AcgBkAEQAbwBjAHUAbQBlAG4AdAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAGgACAQUAAAD//////////wAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAiNgAAAAAAAAUAUwB1AG0AbQBhAHIAeQBJAG4AZgBvAHIAbQBh
AHQAaQBvAG4AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAoAAIBAgAAAAQAAAD/////AAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAANQAAAAAQAAAAAAAABQBEAG8AYwB1AG0AZQBu
AHQAUwB1AG0AbQBhAHIAeQBJAG4AZgBvAHIAbQBhAHQAaQBvAG4AAAAAAAAAAAAAAAAADgAAgH/////
//////////8AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA9AAAAABAAAAAAAAAAB
AEMAbwBtAHAATwBiAGoAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAEgACAQEAAAAGAAAA/////wAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAABqAAAAAAAAAAE8AYgBqAGUAYwB0AFAAbwBvAGwAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAWAAEA////////////////AAAAAAAAAAAAAAAAAAAAAAAAAAAABA
cp/+fl3CAUByn/5+XcIBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAD///////////////8AAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABAAAA/v///////////////
//
//
//
//
//
//
//
//
///wEA/v8DCgAA/////wYJAgAA
AAAAwAAAAAAAAAEYYAAAATWljcm9zb2Z0IFdvcmQtZG9jdW1lbnQACgAAAE1TV29yZERvYwAQAAAA
V29yZC5Eb2N1bWVudC44APQ5snEAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

------_=_NextPart_003_01C2E6F3.34C731CA--

Figure D3: The body of the email as plain text

```
Beste Jacqueline,
=20
Onderstaand voorstel ontving ik gisteren van Bill. Gezien de actualiteit
=
van het archiveren van email ben ik er voorstander van dit snel op te =
pakken. De aanpak van Bill voorziet terecht in een meer formele =
benadering van de email demo nu bekend is dat onderdelen van de demo =
gedistribueerd gaan worden. Vanzelfsprekend gaat dit gepaard met extra =
uren en kosten. Bill komt in zijn calculatie uit op 192 uur.
Het verdient aanbeveling om Stuart White te vragen voor de =
programmeerwerkzaamheden; hij heeft immers de juiste ervaring. Na =
akkoord van onze kant zal Bill bij Tessela informeren naar de =
beschikbaarheid van Stuart.
Graag op korte termijn overleg!!!

Met vriendelijke groeten,

Remco Verdegem
Projectleider Testbed Digitale Bewaring
Badhuisweg 11
Postbus 84011, 2508 AA Den Haag
Netherlands
t: 00 31 70 888 7761
w: www.digitaleduurzaamheid.nl


>  -----Oorspronkelijk bericht-----
> Van:  Bill Roberts=20
> Verzonden:    maandag 16 september 2002 14:48
> Aan:  Remco Verdegem
> Onderwerp:    email demo enhancements estimate
>
>=20
```

Figure D4: XML representation of the email message

```
<?xml version="1.0" encoding="UTF-8" ?>
- <ictu:email xmlns:ictu="xmail.ictu.nl"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="xmail.ictu.nl Email.xsd ">
- <ictu:meta>
  <ictu:datum>2002-09-16T12:16:08+01:00</ictu:datum>
  <ictu:dossier>EmailXML Demo</ictu:dossier>
  <ictu:werkproces>Het verstrekken van advies over het langdurig
    bewaren van emails</ictu:werkproces>
    </ictu:meta>
- <ictu:transmissiebestand>
  <ictu:bestandsnaam>TransmissieBestand.txt</ictu:bestandsnaam>
  <ictu:bestandstype>text/plain</ictu:bestandstype>
  <ictu:bestandslokatie>c:\xmail_files\archive\EmailXML
    Demo\mail9\TransmissieBestand.txt</ictu:bestandslokatie>
    </ictu:transmissiebestand>
  <ictu:onderwerp>FW: email demo enhancements estimate
    </ictu:onderwerp>
- <ictu:geadresseerdenlijst>
- <ictu:geadresseerde>
  <ictu:naam>Jacqueline Slats</ictu:naam>
  <ictu:email>jacqueline.slats@ictu.nl</ictu:email>
    </ictu:geadresseerde>
    </ictu:geadresseerdenlijst>
- <ictu:afzender>
  <ictu:email>remco.verdegem@ictu.nl</ictu:email>
  <ictu:naam>Remcol Verdegem</ictu:naam>
  <ictu:functie>Projectleider</ictu:functie>
  <ictu:organisatieonderdeel>Testbed Digitale
    Bewaring</ictu:organisatieonderdeel>
    </ictu:afzender>
- <ictu:inhoud>
- <ictu:nietXMLinhoud>
- <ictu:body>
  <ictu:bestandsnaam>Body.html</ictu:bestandsnaam>
  <ictu:bestandstype>text/html</ictu:bestandstype>
  <ictu:bestandslokatie>c:\xmail_files\archive\EmailXML
    Demo\mail9\Body.html</ictu:bestandslokatie>
    </ictu:body>
    </ictu:nietXMLinhoud>
    </ictu:inhoud>
    </ictu:email>
```