

Practical experiences of the Dutch Digital Preservation Testbed

Jacqueline Slats
Remco Verdegem

This article was originally published in VINE (The journal of information and knowledge management systems); volume 34, number 2, 2004 issue 135, page 56-65)

Abstract

Digital Preservation Testbed is a three-year practical research project with the overall goal of investigating options to secure sustained accessibility to authentic archival records over the long-term, by carrying out experiments in a controlled and secure environment. This allows us to ascertain the effects of undertaken preservation action on archival records.

Testbed is researching three different approaches to long-term digital preservation: migration, XML and emulation. Not only will the effectiveness of each approach be evaluated, but also their limits, costs and application potential.

Experiments take place on four different record types: text documents, spreadsheets, emails and databases of different size, complexity and nature.

At the end of 2003 the Digital Preservation Testbed project will provide:

- Advice on how to deal with current digital records
- Recommendations for an appropriate preservation approach or a combination of approaches per record type
- Functional requirements for a preservation function
- Cost models of the various preservation strategies
- A decision model to select the right preservation strategy
- Recommendations concerning archival guidelines and regulations

Practical experiences of the Dutch Digital Preservation Testbed

The government's digital memory

The digital government: it seemed to be so far away in the previous century. Now in the 21st century, the government is working more and more with digital documents. Email communication is has become part of the daily routine and databases are used everywhere. The government has an obligation to treat information in a responsible manner.

Digital documents must be preserved and remain accessible for coming generations. This principle also applies to paper-based information that is managed and preserved. Building the digital government means that the appropriate digital infrastructure needs to be in place as soon as possible. Records not only have to be found quickly, they also have to be authentic and readable (regardless of the current technology) and remain so in the future.

The current Dutch Cabinet aims to carry out 65% of its transactions between government and its citizens through digital means by 2006. In 2002 the goal was 25% and this was easily achieved. Because of this, there is currently a great deal of work going on to develop strategies, methods, techniques and tools to handle the digital produce of the government in a responsible way.

Digital Longevity

Under the umbrella of Digital Longevity the Netherlands have several programs. Whereas the objective of the Digital Longevity programme is to secure the accessibility of reliable government information, the objective of the Digital Preservation Testbed is securing *sustained* accessibility to reliable government information.

According to Dutch law and regulations the transfer of archival records takes place after 20 years, in a 'good, ordered and accessible state'. For digital records 20 years is more than a lifetime. Therefore the target group of the Digital Preservation Testbed is not only archival organisations, but also the whole Dutch government.

Digital preservation

The most important problem concerning the preservation of authentic digital records is technological obsolescence. Technological change is increasing exponentially. This brings up many questions, such as what to do with files that were made with old hard and software, which cannot be used anymore? Unless action is taken now, there is no guarantee that current files can be read in future with future technologies.

Digital Preservation Testbed

Testbed was established in October 2000 by the Ministry of the Interior and Kingdom Relations and the Ministry of Education, Culture and Sciences (of which the Nationaal Archief of the Netherlands is a linked institution).

Testbed is a three-year research project with the overall goal of investigating options to secure sustained accessibility to authentic archival records over the long-term.

Testbed is a practical research project that carries out experiments in a controlled and secure environment. This allows us to ascertain the effects of undertaken preservation

action on archival records. Our direction is dictated by the Research Questions laid down at the beginning of the project.

Research Questions

The Research Questions have three main areas of interest: General, Metadata, and Attribute-based. General research questions include:

- What are the advantages and disadvantages of implementing the different preservation approaches?
- How can the effectiveness of each approach be measured and or demonstrated?
- What are the factors that affect the effectiveness or appropriateness of each preservation approach? For example, cost? Record type? Authenticity requirements and retention periods?
- What are the basic requirements for preservation functions? For example, what are the requirements for accessing and retrieving records from the preservation function?

Metadata research questions address such issues as:

- What factors affect the metadata required for preservation? For example, record type and preservation approach, and how?
- What are the options for associating metadata with records?

We also consider attribute research questions. The Testbed classifies electronic records according to the five attributes identified by Rothenberg¹. These are: Content, Context, Structure, Appearance and Behaviour. We consider such aspects as

- What are the options for preserving record attributes?
- What is the relationship between the preservation of specific attributes and the cost of preservation?

Experiments

Not only to control the project, but also to run experiments in a controlled environment, we developed a 12-step experiment process. Here we also make explicit, mostly by desk research of available publications, if a record type is excluded from a certain preservation approach. These steps are all fully documented in the experiment database of the Testbed. Records are monitored during experiments to establish whether (and how) a specific method is suitable for long-term preservation.

**Testbed Project
Experiment Process**

Abbreviations:
 PM = Project Manager
 PT = Project Team
 RG = Research Group
 IG = Implementation Group
 EG = Evaluation Group

UPPER CASE = PRIMARY ROLE
 lower case = secondary role

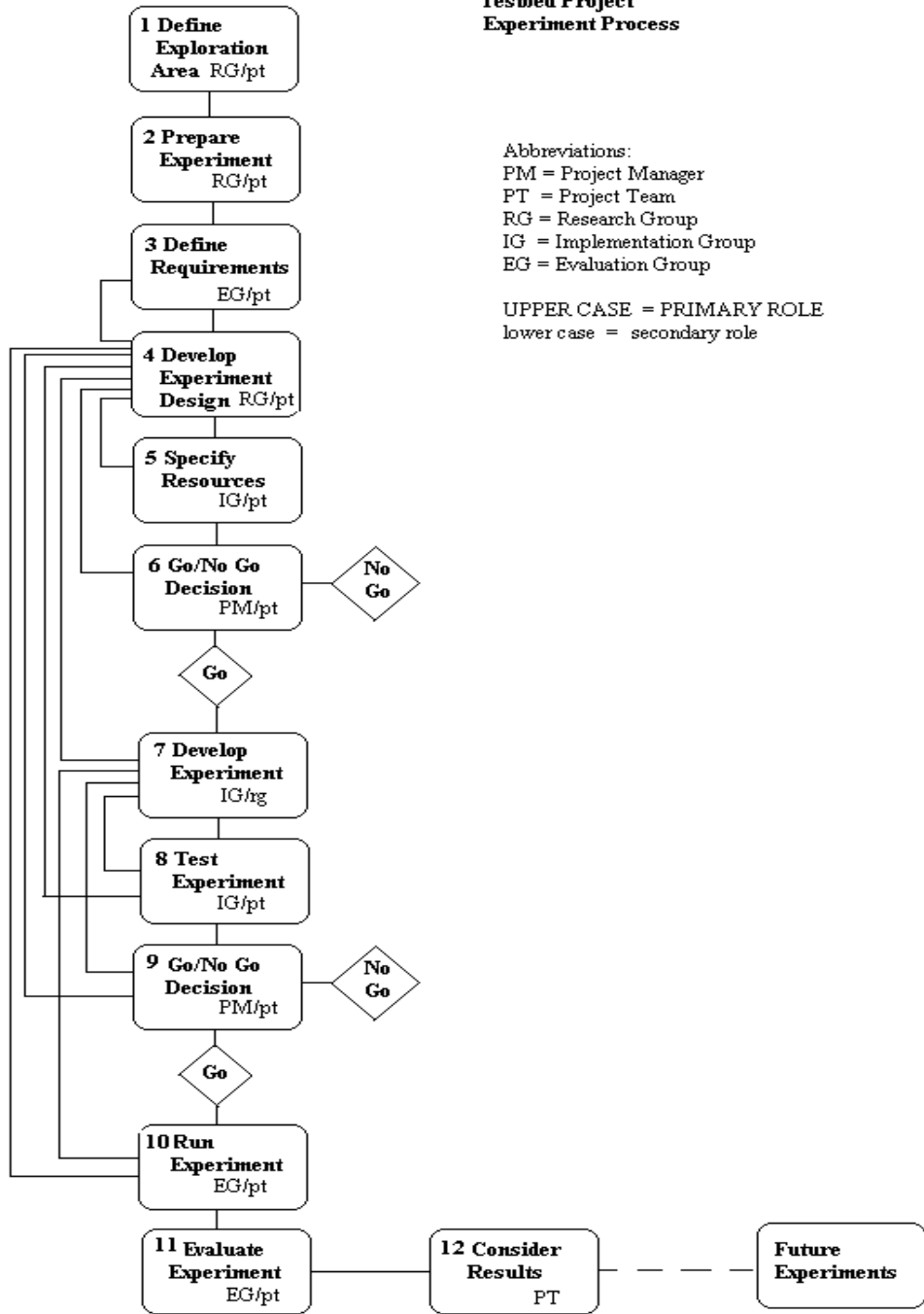


Figure 1: Experiment Process

Testbed team

This approach requires a multi-disciplinary team. The Testbed team consists of ICT-expertise records managers, archivists, national and international experts, etc. Not mentioned in the diagram below, but very valuable is the evaluation feedback group, consisting of archivists from various institutions, e.g. the Nationaal Archief of the Netherlands, the Archival Inspection, Tax Services, etc. The governmental institutions that provide us with copies of records are participating in the team during the experiments.

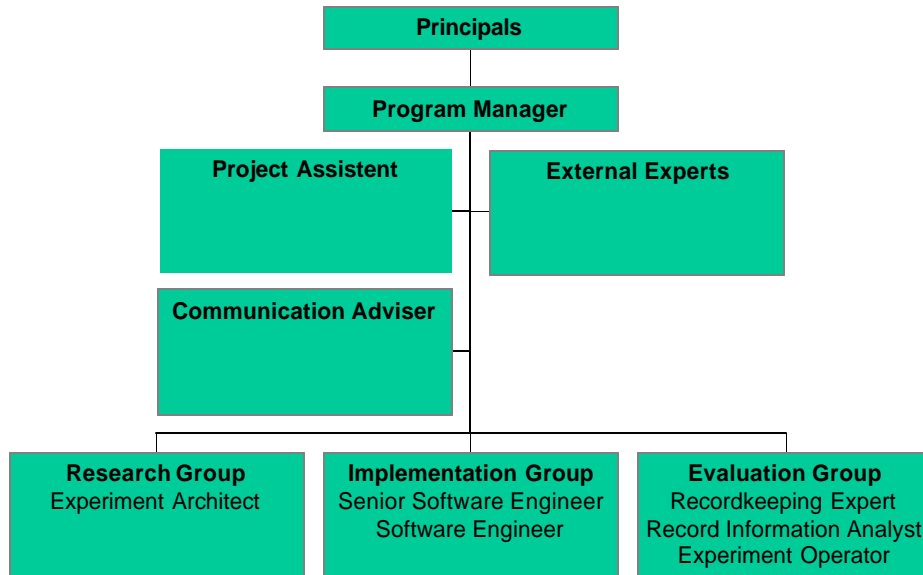


Figure 2: Testbed organisation

Preservation approaches

The Digital Preservation Testbed is researching three different approaches to long-term digital preservation: migration, XML and emulation. Not only will the effectiveness of each approach be evaluated, but also their limits, costs and application potential.

Migration

There are many different definitions of migration. Testbed defines migration as the conversion of records from one hardware and/or software environment to another.

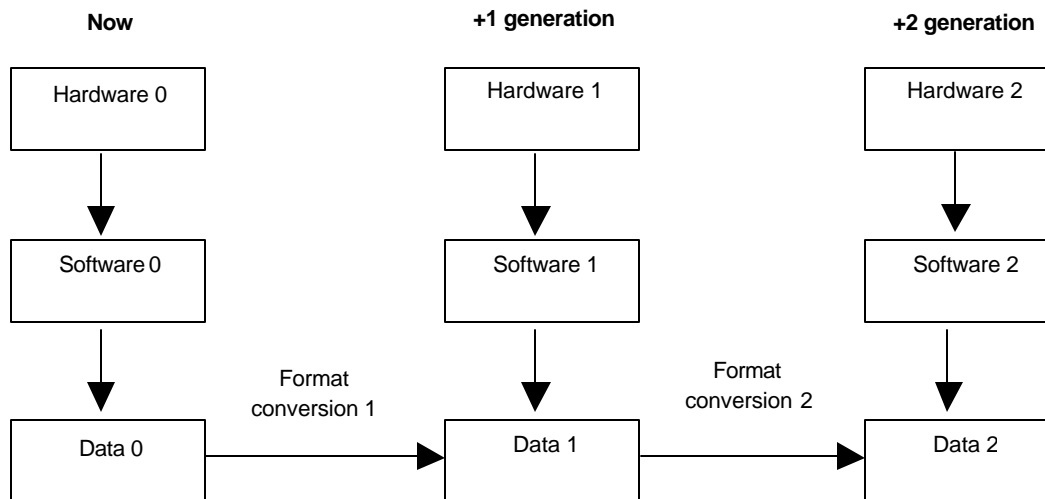


Figure 3: Basic migration diagram

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

- Backward compatibility
- Interoperability
- Conversion to standards

Backward compatibility

Backward compatibility makes it possible to interpret and correctly reproduce a file that has been created in an older version of an application in a new version of the application. New versions of commercial software are often compatible with the previous version(s). One example is that Excel 2002 can read files created with Excel 95 and saved in the Excel 95 file format.

Records maintained using this approach normally need to be re-saved into the new file format, since software only supports a limited number of generations of older file formats. Migration to a higher version usually has to be repeated every few years. Experiments by the Testbed have revealed that every migration carries some risk of a change, however slight, which could adversely affect the authenticity and integrity of the digital record. Thus this strategy can be useful for the short term, but is less suitable for long-term preservation, because of the risk of small errors accumulating.

Backward compatibility is commonly employed to migrate proprietary and unpublished file formats, for example *.xls for spreadsheets in MS Excel. In these cases the digital record continues to be stored in the supplier's own file format thus maintaining dependency on proprietary software.

Interoperability

Interoperability in the technical sense tackles the problem of digital obsolescence by ensuring that files and digital records are no longer, or are less, dependent on a particular combination of hardware and software. Interoperability means that a file can be transferred from one platform or application to a different one and can then still be reproduced in the same, or a similar way.

- A file format which can be read and edited using different versions of the same application made for different operating systems. Software suppliers issue different application versions for different operating systems, such as versions that run under Windows, Linux or Solaris.
- Another example of interoperability is a file format which can be read and edited using a range of different software applications. For example, Excel can read Lotus 1-2-3 files, and Lotus 1-2-3 can read Excel files.
- A last 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.

Conversion to standards

Conversion to standards is really migration from a proprietary (and often closed) file format to a format based on a published (proprietary or public) standard. The advantage is that digital records are no longer dependent on the original hardware and software with which they were created and whose obsolescence may form a threat to their sustainability.

Either 'de jure' or 'de facto' standards are possible. 'De jure' standards come about by a formal process in a formally accredited standardisation body (ISO, NEN, W3C.). A 'de jure' standard also comes into being in an open process, since consensus and participation are the most important motives for formal standardisation organisations. An example of a 'de jure' standard format is the ISO8859 character set. Another example is XML.

'De facto' standards are formats that are widely implemented; there is a critical mass that makes use of the specification. De facto standards can be developed by consortia by means of open processes, but they can also come about by means of closed processes (supplier's own standards)ⁱⁱ. An example of a 'de facto' standard format is PDF.

In general, de jure standards are preferable to a supplier's de facto standards, because maintenance and future development of the standard are controlled by a wider community and not by a single organisation. In some cases, there may also be licensing issues to consider with de facto standards. However, these are not the only

considerations in choosing a standard format for preservation: the technical suitability and popularity of the standard are also important.

XML

The Digital Preservation Testbed also studied XML as an approach towards the long-term preservation of digital records. 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 like specifications, guidelines, software and tools so that the Internet can be used to the fullⁱⁱⁱ.

XML is non-platform specific and can be read by humans as well as machines, using a simple text editor. For these reasons XML can be used for digital preservation. 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 is designed to be easy for computer programs to process, which is one reason why it is a good preservation format; it will be relatively easy to write software in the future to process XML files produced today.

Files can be converted directly to XML or generated directly in XML as a file format. Since XML is not dependent on a particular combination of hardware and software, it is more sustainable than many commercial file formats. The number of conversions will thus be considerably reduced, as will the risk of adversely affecting the authenticity of the digital record.

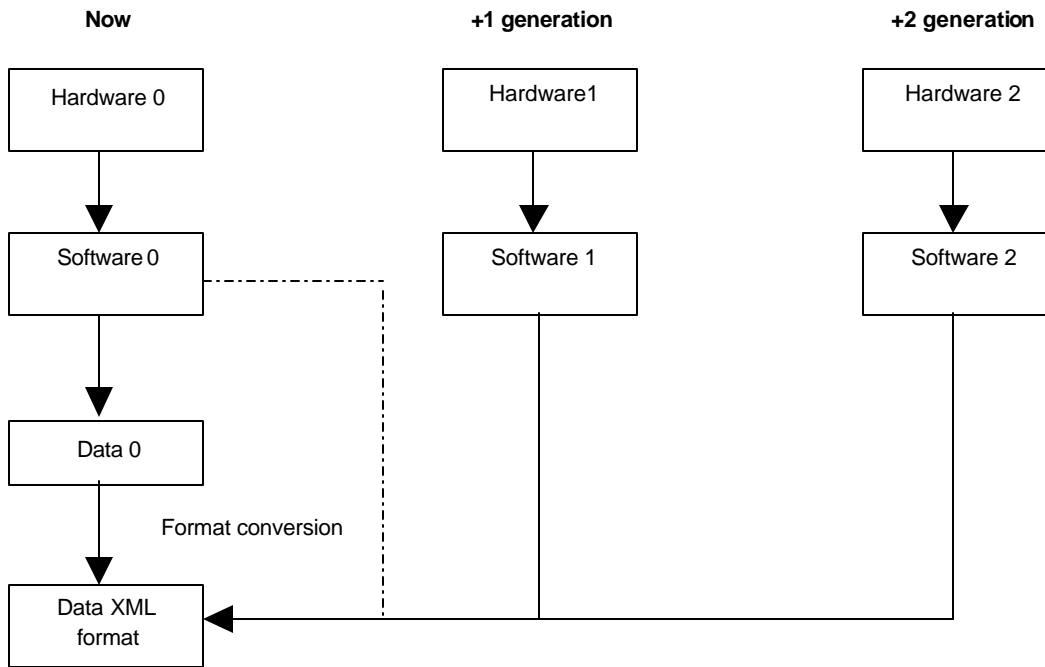


Figure 4. Conversion to XML requires fewer conversions than migration

Encapsulation

This approach is aimed at preservation of the original format. XML is often named as a good language for storing metadata and instructions for the object to be preserved. In this section we review a number of terms used in this context.

Wrappers, containers, encapsulation and framework

The Dutch archival regulation mentions an ‘XML-wrapper’ as a means of adding metadata to PDF and TIFF files. Although one can imagine what this might entail, there is not (as yet) a fixed meaning for this term. For example, the San Diego Supercomputer Centre regards a wrapper as a piece of software that can be used as a ‘mediator’.^{iv} The Roquade project, in contrast, uses the term ‘container’ for the ‘packaging’ of digital archival records.^v A step further than encapsulation is to use XML as a framework to on which a document or parts of a document in e.g. TIFF or PDF format can be hung. In this case, XML forms the backbone of a digital archival record.

Metadata

Metadata, that is to say data about data, is an integral part of XML (in the form of tags). XML also offers excellent facilities for the storage of metadata in the narrower archivist sense. For this reason, XML can be used in combination with other preservation strategies. For example with emulation, XML could be the language used to store technical metadata. Adobe, owner of the PDF standard, has recently launched the eXtensible Metadata,^{vi} which also uses XML for metadata storage.

If a fixed collection of metadata has been agreed (and that is often much more difficult than the technical implementation!), this can be specified in the form of an XML schema, which can be reused by schemas for specific documents. This standardisation is important, because otherwise a digital archive will not know what kind of metadata to expect.

Migration to XML

Structured data such as those found in a database or spreadsheet lend themselves very well to migration to XML. In principle, the contents of database tables can be translated on a one-to-one basis to elements in XML. In addition, a great deal of other information (connected with the technical implementation of the database) should also be transferred to the files to be archived. Migration to XML will have fewer difficulties if in the future the development of a system takes into account the need for a final transformation to XML.

XML (from the beginning)

In the previous strategy, the data was converted to XML after creation in some other format. This approach becomes more complicated if we consider unstructured documents, where the structure must later be made explicit in the form of XML tags. This can not always be done fully automatically and will usually require some human intervention. It is not surprising that there are initiatives to use XML as the underlying format for office productivity software applications. The open source package OpenOffice^{vii} is one example. Given that Microsoft is also making use of XML within Office XP, the trend seems to be towards using XML as the original storage format for office documents.

Emulation

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

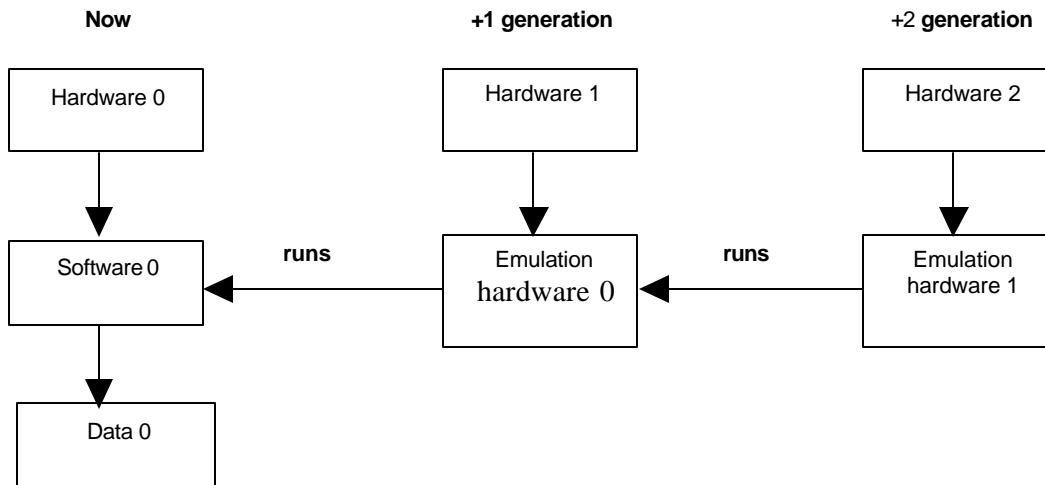


Figure 5: Basic emulation diagram

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

It is important to mention that the approach discussed here involves using software to emulate hardware.

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

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

Universal Virtual Computer

An emulation approach that uses the UVC (Universal Virtual Computer) 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. 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 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.

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.

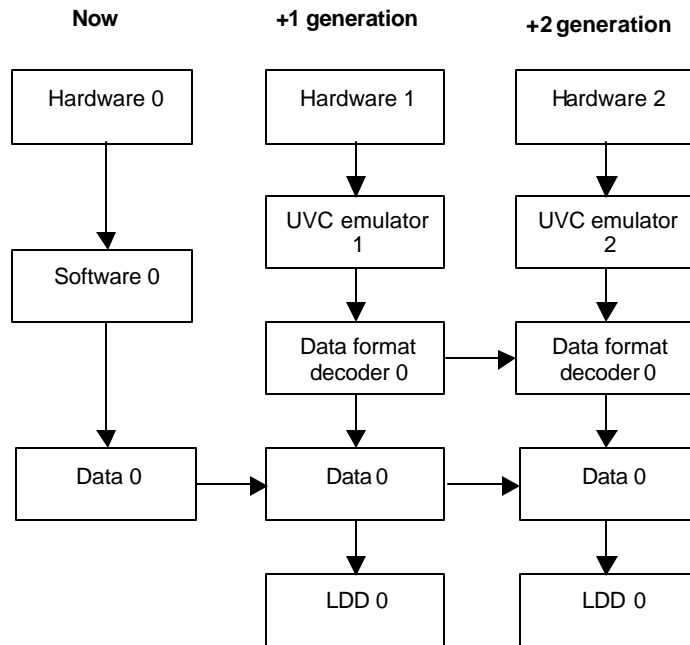


Figure 6: Diagram of the Universal Virtual Computer

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 Dutch archival institutions will be limited by Dutch archival regulation.

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, 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.

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.

In view of the wide variety of file formats and types of digital records, large numbers of decoder programs will have to be developed, if the UVC is to be a feasible and

workable strategy for the long-term preservation of different types of digital records. 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.

Results

Experiments have taken place on four different record types: text documents, spreadsheets, emails and databases of different size, complexity and nature. These are the record types, which are used for more than 90% within the Dutch Government.

Text documents

Starting with text documents we selected migration and XML as preservation approaches to experiment with. For the UVC approach we made use of the reports of the Dutch National Library, which performed a proof of concept preserving electronic publications.

The migration of records from an older version of an application to a newer version of the same application (e.g. Word 97 to Word 2000) is usable for the short-term preservation. We did not encounter significant problems converting the records to a higher version. It was remarkable that the results were even better when we skipped one or more versions. However, after multiple conversions the sum or the minor changes can affect the authenticity of the record. So manual checking is required. Furthermore the migration needs to be repeated every few years and is only feasible if the migration is automated.

For the migration of text records to a standard format we experimented with PDF and RTF. PDF is suitable to represent text documents authentically, especially the content and appearance.

We also migrated old records created in one word processor to another (WP4.2 to Word 2002). This approach only met our authenticity requirements after manual intervention.

Finally the XML approach: XML is able to represent context, content, structure and behaviour of text documents authentically. To represent appearance an additional stylesheet is required.

Spreadsheets

For spreadsheets we selected all three approaches to experiment with. It was an extra challenge to experiment with the UVC data preservation approach using spreadsheets because spreadsheets have more layers (e.g. a data layer and a formulae layer).

Although the concept of the UVC is promising, generating the logical data description appeared to be very difficult. This is not because of the complexity of the UVC, but because of the lack of documentation of the proprietary file formats. From the reports of the Dutch National Library we noticed that they have encountered the same problem.

The migration of records from an older version of an application to a newer version of the same application (e.g. Excel 97 to Excel 2000) is usable for the short-term

preservation. The results of these experiments were comparable with those of migrating text documents to a higher version.

Finally XML is a suitable format to represent spreadsheets authentically, including the different layers.

Email

For email we selected only XML as a preservation approach to experiment with. Based on desk research email has proved to be a particularly suitable record type for XML. There are many similarities between XML and Email formats, and conversion between the two is thus relatively straightforward.

Both are highly specified.

Emails must follow the Internet Message Format to be interoperable on different platforms. This format is well laid out and defines the component parts of a basic email transmission file. (The standard currently in use with emails is RFC 2822, with the MIME extensions specified in RFC 2045 – 2049.) It is controlled by a non-profit organisation, the Internet Engineering Task Force^{viii}, and is well defined, well structured, and text based.

XML is a standardised format, as well as a mark-up language. Again, it is highly specified and controlled by a non-profit organisation – in this case the World Wide Web Consortium^{ix}. The W3C is responsible for organising and maintaining the XML Specification, Schema, Standard and XSLT Recommendation. XML, as the name denotes, is extensible. It can be adapted and extended for any purpose while still remaining true to its spirit. It can operate on any hardware and/or software platform, and can be read on any plain text editor.

The similarities between the two mean that conversion is a relatively straightforward procedure. All individual sections are plainly marked in the email transmission file and can easily be transformed into a similarly well structured XML file.

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**).

Databases

Experimenting with databases we were confronted with the question: ‘What is the archival record’:

- the whole database system [database, DBMS and user application],
- the database itself,
- a row in the database table,
- the record consists of fields spread over different tables,
- database data accessed or presented in a precise manner in the application form

Despite the desk research and a lot of discussion with archival experts we were not able to answer this question unambiguously. Eventually from a pragmatic point of view we decided to experiment with the whole database system and the database itself.

The migration of databases from an older version to a newer version of the same database system (e.g. Access 97 to Access2000) is usable to represent context, content, appearance, structure and behaviour for the short term. The results of these experiments are comparable with those of migrating text documents and spreadsheets to a higher version.

The conversion to XML is suitable to represent the context, content and structure of the database itself. Additionally, in order to preserve the appearance of the application it is necessary to store the technical and functional documentation of the database system, including screen shots.

We were not able to preserve behaviour of database systems for the longer term using migration or XML. Nor is the UVC data preservation approach able to achieve this. Hardware emulation could be a potential approach in this respect, but has not been implemented with an archival focus.

Products

At the end of 2003 the Digital Preservation Testbed project will provide:

- Advice on how to deal with current digital records
- Recommendations for an appropriate preservation approach or a combination of approaches per record type
- Functional requirements for a preservation function
- Cost models of the various preservation strategies
- A decision model to select the right preservation strategy
- Recommendations concerning archival guidelines and regulations

For further information about Testbed:

Website: <http://www.digitaleduurzaamheid.nl>

Email: Testbed@nationaalarchief.nl

ⁱ Rothenberg, J. & Bikson, T. (1999), Digital Preservation - Carrying Authentic, Understandable and Usable Records Through Time.

ⁱⁱ Thomas, W. (2002), XML: de mogelijkheden en valkuilen voor overheid (Dutch document: XML: the possibilities and pitfalls for government).

ⁱⁱⁱ See <http://www.w3.org>

^{iv} “A wrapper is a piece of software that acts as a translator between the native format of an information source and a commonly agreed protocol (XML for us). The end-user or application interacts with a piece of software called mediator that collects information from multiple wrappers “, page 4 of *Methodologies for the Long-Term Preservation of and Access to Software-Dependent Electronic Records*, <http://www.sdsc.edu/NHPRC/Pubs/nhprcf2k.doc>.

^v “It was decided to work out the idea of XML containers. So the Archival Information Packages (AIP), to be stored in the electronic archive, will be wrapped in XML.” *An electronic Archive for academic communities* (Dekker, R. et al, Nov 2001). The AIP concept originates from the Open Archive Information System (OAIS) model.

^{vi} See <http://partners.adobe.com/asn/developer/xmp/download/docs/MetadataFramework.pdf>.

^{vii} see www.openoffice.org

^{viii} See: <http://www.ietf.org/>

^{ix} See: <http://www.w3.org/>