FROM DIGITAL VOLATILITY TO DIGITAL PERMANENCE

Jacqueline Slats [*]

UDC: 004.3:930.253

***Jacqueline Slats: From digital volatility to digital permanence**. Technical and Field Related Problems of Traditional and Electronic Archiving. Conference Proceedings, Maribor 4/2005, No. 1, pp. 135-144.*

Original in English, abstract in English, summary in German.

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 National Archives of the Netherlands initiated in October 2000 the Digital Preservation Testbed project together with the Ministry of the Interior and Kingdom Relations to provide recommendations for an appropriate preservation approach or a combination of approaches for digital records.

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 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.

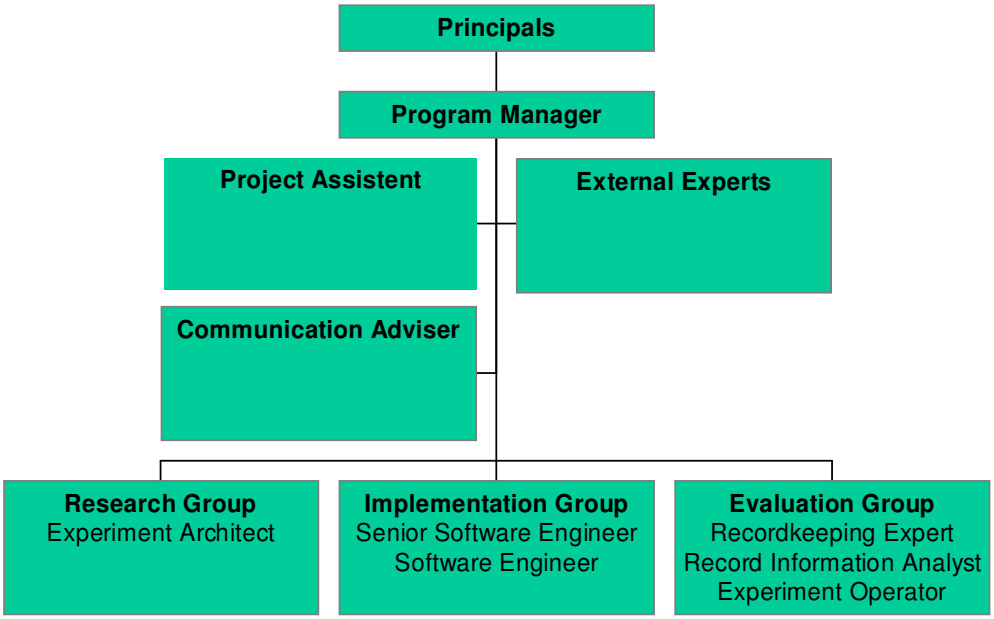
[*] ***Jacqueline Slats**, Head Digital Longevity, Nationaal Archief of the Netherlands, The Hague, The Netherlands.*

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 National Archives of the Netherlands is a linked institution). Testbed was 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 was a practical research project that carried out experiments in a controlled and secure environment. This allowed us to ascertain the effects of undertaken preservation action on archival records. Our direction was dictated by the Research Questions laid down at the beginning of the project.

TESTBED TEAM

The approach required a multi-disciplinary team. The Testbed team consisted of ICT-expertise records managers, archivists, national and international experts, etc. Not mentioned in the diagram below, but very valuable was the evaluation feedback group, consisting of archivists from various institutions, e.g. the National Archives 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.



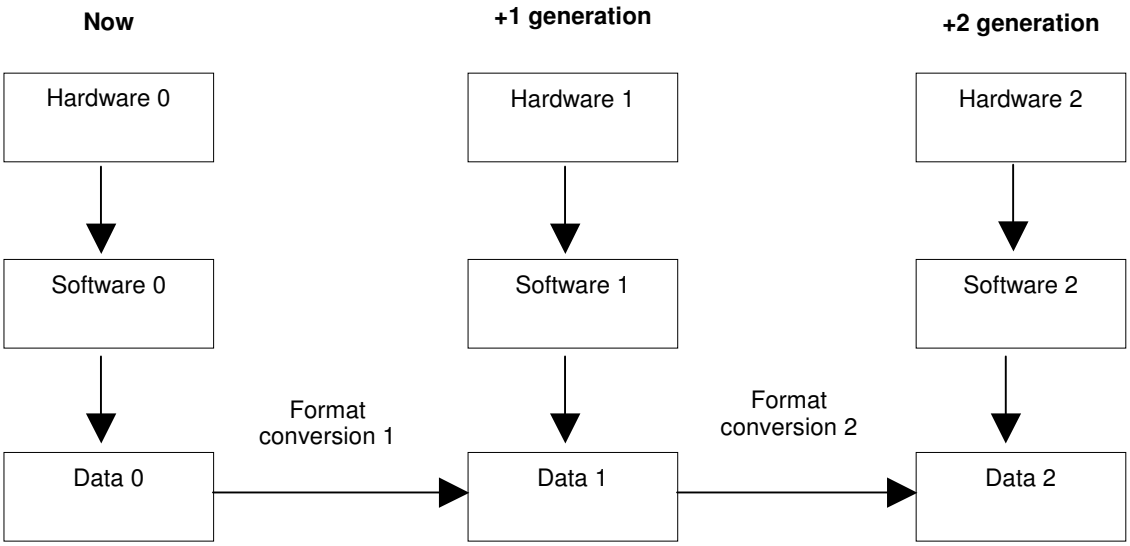
Testbed organisation

PRESERVATION APPROACHES

The Digital Preservation Testbed was researching three different approaches to long-term digital preservation: migration, XML and emulation. Not only the effectiveness of each approach was 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.



Basic migration diagram

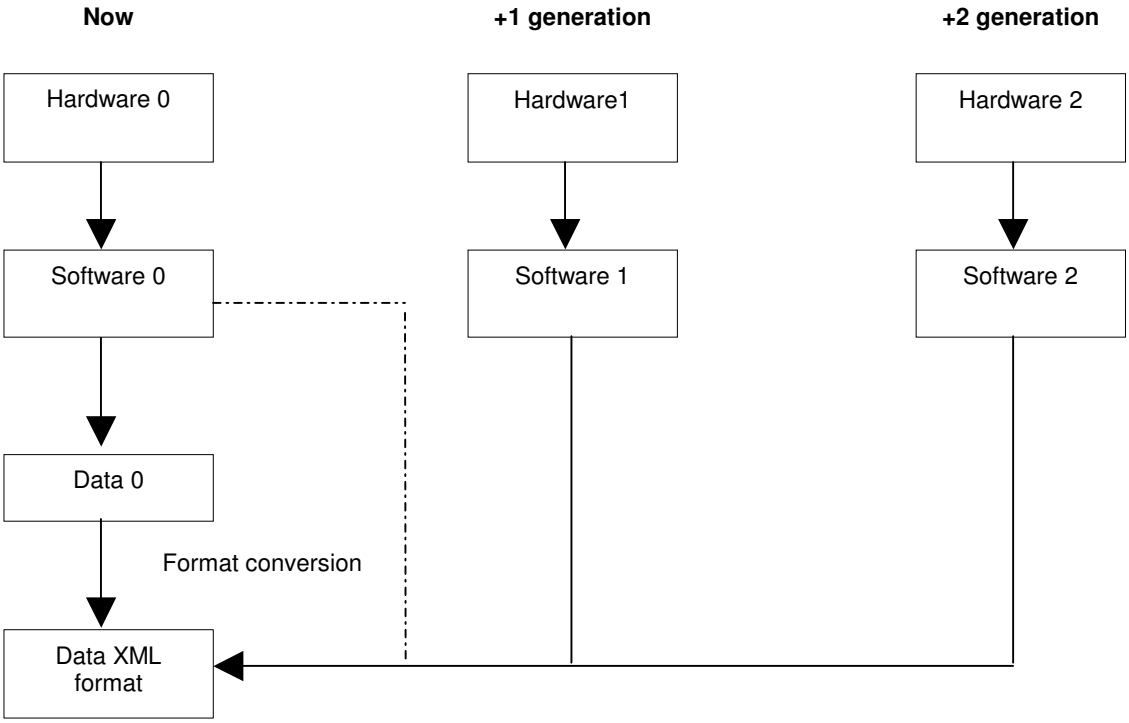
Testbed has studied and experimented with the following forms of migration:

- Backward compatibility
- Interoperability
- Conversion to standards

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 full. 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.



Conversion to XML requires fewer conversions than migration

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 a simple architecture and basic set of instructions that any software developer in the future should 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 should 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.

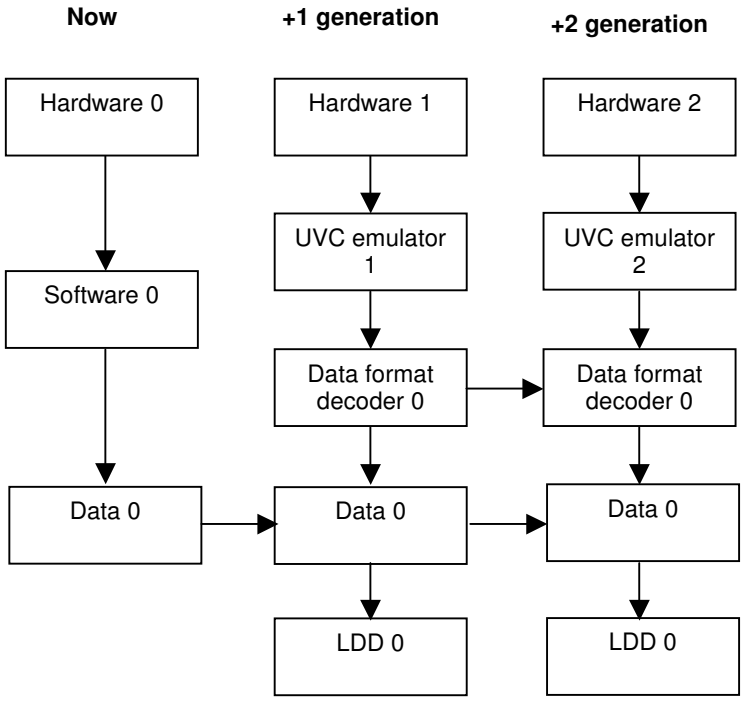


Diagram of the Universal Virtual Computer

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 possibly hundreds of viewers must be used. 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.

The 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.

EXPERIMENTS

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. We classified electronic records according to the five attributes identified by Rothenberg. These are: Content, Context, Structure, Appearance and Behaviour.

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?

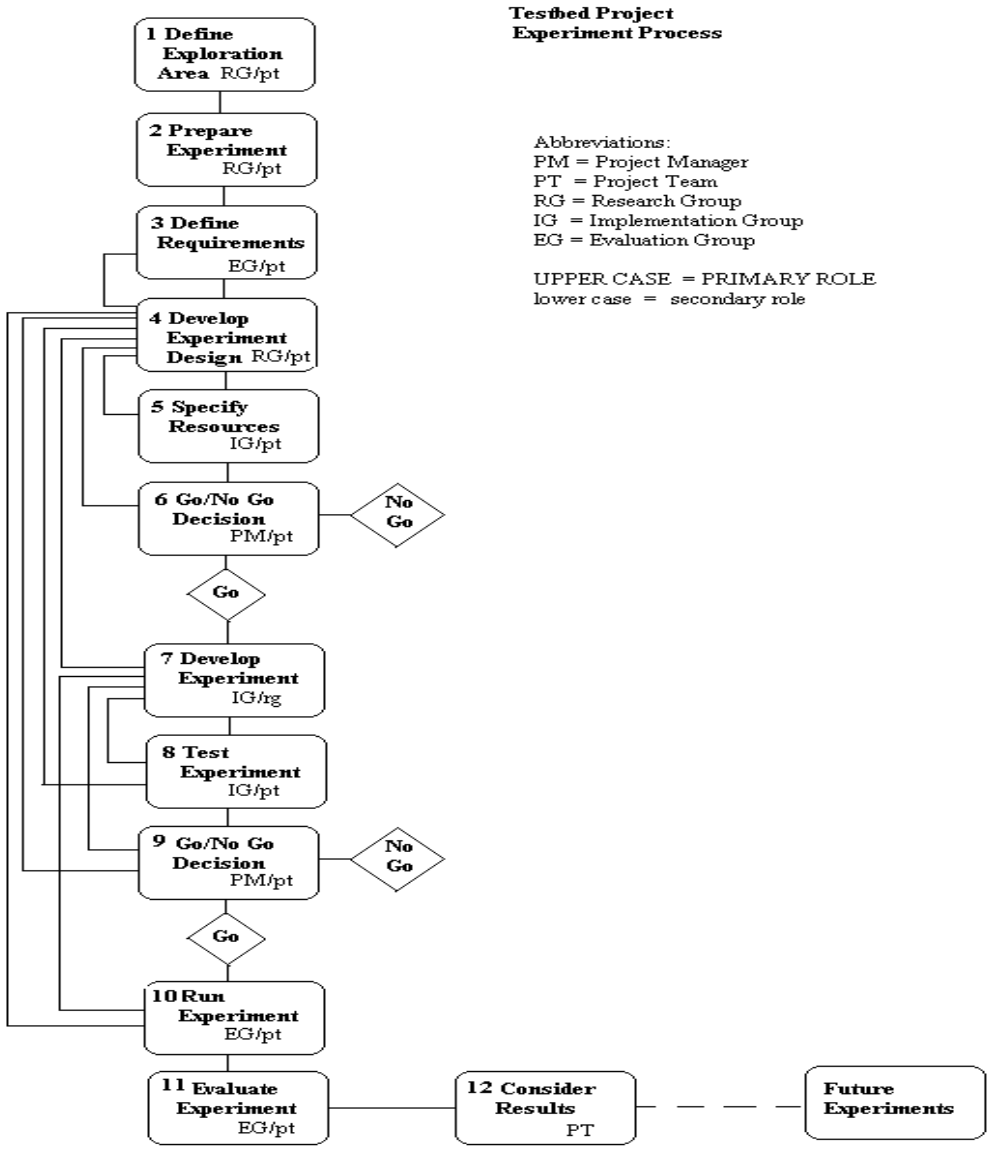
Metadata research questions address such issues as:

- 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?

EXPERIMENT PROCESS

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.



Experiment Process

RESULTS

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, 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. 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**).

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.

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.

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.

FOLLOW UP

The National Archives of the Netherlands defined three new projects for the coming years:

- Recommendations concerning archival guidelines and regulations
- Recommendations of the Testbed integrated in the Electronic Document and Records Management Systems of ministries
- Hardware emulation project

ZUSAMMENFASSUNG

VON DIGITALER UNBESTÄNDIGKEIT ZU DIGITALER BESTÄNDIGKEIT

Das Versuchsverfahren zur Erhaltung digitaler Aufzeichnungen war ein dreijähriges praktisches Forschungsprojekt mit dem Hauptziel, die Möglichkeiten zu untersuchen, die bisher gewahrte Zugriffsmöglichkeit auf authentischen Archivbestände längerfristig zu gewährleisten, indem Versuche in einer kontrollierbaren und sicheren Umgebung durchgeführt wurden. Dies brachte Gewissheit über die Auswirkungen bislang gepflogener Maßnahmen, um Archivbestände zu erhalten.

Das Versuchsverfahren bezog sich auf drei Punkte der Langzeitaufbewahrung digitaler Bestände: Migration, XML und Emulation. Dabei wurden nicht nur die Wirkungsweise der einzelnen Punktes bewertet, sondern auch deren Grenzen, Kosten und Anwendungsmöglichkeiten.

Die Versuche wurden an vier verschiedenen Arten von Archivalien durchgeführt: an Schriftstücken, Statistiken, E-mails und Datenbanken verschiedener Größe, Komplexität und Art. Am Ende des Jahres 2003 lieferte das Versuchsverfahren zur Erhaltung digitaler Aufzeichnungen:

- Ratschläge zum Umgang mit aktuellen digitalen Akten
- Empfehlungen für eine entsprechende Erhaltungsweise oder die Kombination verschiedener Methoden je nach Art der Akten
- Grundvoraussetzungen für die Erhaltung
- Kostenaufstellungen für verschiedene Erhaltungsmaßnahmen
- Entscheidungshilfen für die Wahl der richtigen Erhaltungsmethode
- Empfehlungen hinsichtlich der Archivrichtlinien und –bestimmungen

Jacqueline Slats, *After her study in Information Management Jacqueline worked for 7 years at the computer centre of the Ministry of Transport, Public Works and Water management. In 1994, she joined the Dutch State Archive Service, where she was responsible for different Information Technology projects. Last few years she was the program manager of the Digital Preservation Testbed and the Taskforce Digital Longevity, which were sponsored by the Dutch State Archive Service and the Ministry of the Interior and Kingdom Relations. Now she is the head of the Digital Longevity department of the Nationaal Archief of the Netherlands.*