

Practical experiences of the Digital Preservation Testbed

By Jacqueline Slats, DLM Forum, Barcelona, 7 May 2002.

The Digital Preservation Testbed is part of the non-profit organisation ICTU. ICTU is the Dutch organisation for ICT and government. ICTU's goal is to contribute to the structural development of e-government. This will result in improving the work processes of government organisations, their service to the community and interaction with the citizens.

The Dutch E-Government house

Government institutions, such as Ministries, design the policies in the area of e-government, and ICTU translates these policies into projects. Together, these projects form what we call the e-government house or ELO-house. In many cases, more than one institution is involved in a single project. They are the principals in the projects and retain control concerning the focus of the project. In case of the Digital Preservation Testbed the principals are the Ministry of the Interior, Jan Lintsen and the Dutch National Archives, Maarten van Boven. Together with Public Key Infrastructure, Digital Longevity is the fundament of the ELO-house.

Digital Longevity

Under the umbrella of Digital Longevity, we have several programs like Record Keeping System, Quality Assurance, Testbed etc. The objective of Digital Longevity is securing the accessibility of reliable government information; the objective of the Digital Preservation Testbed is securing the *sustained* accessibility of reliable government information.

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

The current Dutch Cabinet aims to carry out 25% of its transactions between government and its citizens through digital means by 2002. 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.

Longterm 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 the future with future technologies.

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

Experiments are taking place on text documents, spreadsheets, emails and databases of different size, complexity and nature.

Experiments

The Digital Preservation Testbed is carrying out experiments according to pre-defined problem solving research questions to establish the best preservation approach or combination of approaches. The experiment process started with these basic research questions and each experiment raises new questions.

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 recordtype 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.

This approach requires a multi-disciplinary team. The Testbed team consists of ICT-expertise, record managers, archivists, national and international experts, etc. Very valuable is the evaluation feedback group, consists of archivists from various institutions, e.g. the Dutch National Archives, the Archival Inspection, Graphic Industries, Tax Services, etc. The governmental institutions that provide us with copies of records are participating in the team during the experiments.

Experiments on 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. Migration is currently the most common preservation strategy for digital records, but not always used in a responsible way: when new versions arrive, documents are simply updated into the new versions.

Testbed experimented with migration of text documents:

MS® Word 95, 97, 2000, 2002 step by step, and for example directly from MS® Word 95 into 2002 and Conversion to Adobe Acrobat® PDF 1.2, 1.3, 1.4. Experiments from WordPerfect® into MS® Word are now taking place.

The experiments give good results if the documents are created in a responsible way, (e.g. don't use automated date fields) and captured, and if the migration is well prepared. It is remarkable that migration from MS® Word 95 directly into 2002 gives better results than migration step by step. We don't have a clarification yet, why Adobe Acrobat files sometimes put the last word from one line onto the next line

Still there remain a few disadvantages of migration. Each record must be migrated every few years; this is only feasible if the process is automated. It still requires manual checking of the results, and eventually changes to the format of the record can lead to information loss, thereby compromising the records authenticity.

Experiments on emulation

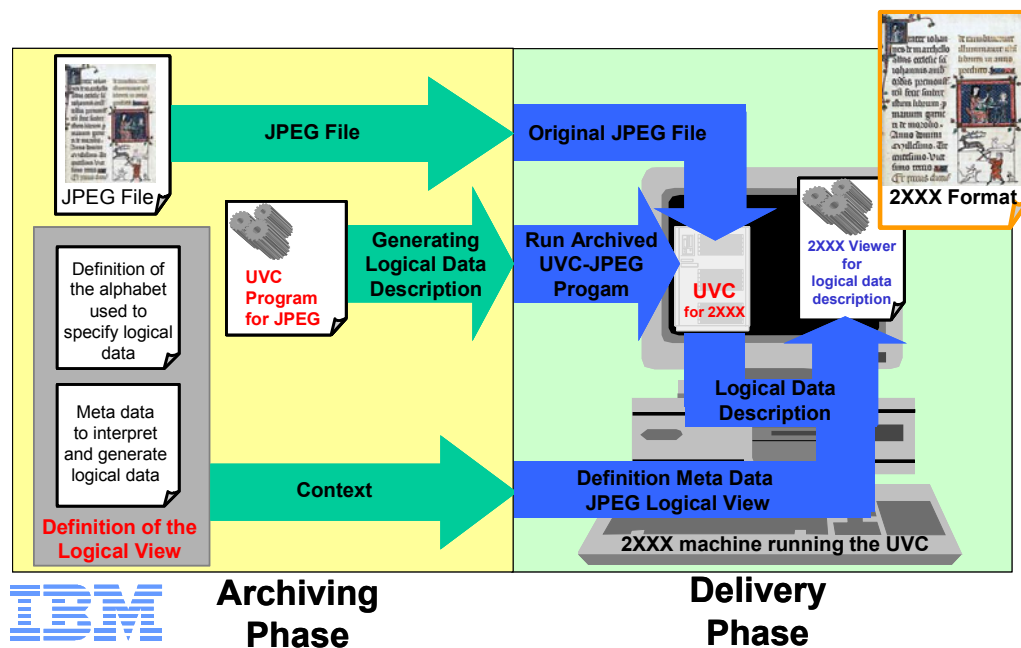
The Universal Virtual Computer (UVC) -based methodology makes a distinction between preserving data and preserving the behaviour of a program. For data, it

implements a conversion program able to decode the original form of the data into a logical format that will be much easier to understand in the future. This conversion program is written in 2000 (for a UVC machine). It can be executed in 2050, on an emulator of the UVC on the 2050 machine. For programs, the UVC-based methodology will rely on an emulator of the 2000 machine (for a UVC machine), written in 2000, and an emulator of the UVC on the 2050 machine. It clearly differs from the emulation method proposed by for instance Jeff Rothenberg, in that it does not require writing in the future an emulator of a real machine of the past.

With the UVC approach we have a solution that can be applied to support multiple preservation strategies both data and program preservation. The UVC architecture relies on concepts that have existed since the beginning of the computer era: memory, registers, and a set of low-level instructions. The fact that the computer is virtual and that performance is of secondary importance allows for a simpler, more logical, maybe less optimised, design. This will guarantee its durability along waves of technology changes.

Because the UVC instruction set is so simple, it is relatively straightforward to write an UVC emulator for any given computer. In the context of long-term preservation of digital data, initiated within IBM Research, they considered an approach that relies only partially on emulation. The approach is applicable to digital object types that do not need to maintain the functionality of the application(s) that were used initially to create or manipulate the objects. IBM refers to this approach as *data preservation*. For data preservation, we propose to save, with the data, a program that can extract the data from the bit stream and return the information to the caller in an easy to understand, technology-independent way, so that it may be exported to a new system.

Data Preservation is the first and simplest mode of operation of the UVC approach



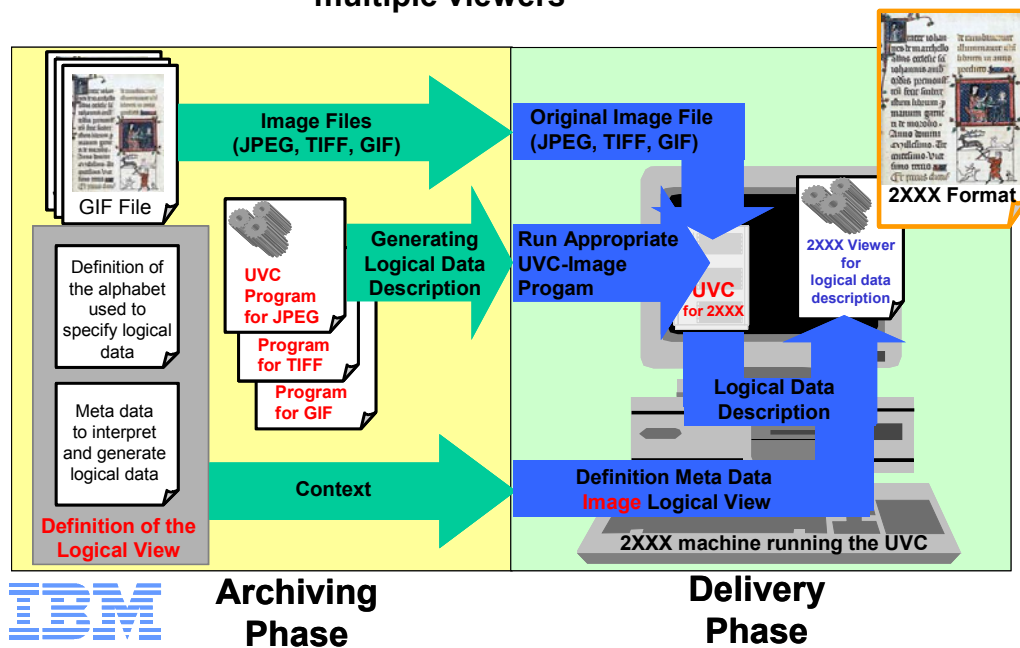
For example the preservation of JPEG files. The UVC JPEG program is written for a Universal Virtual Computer (UVC). This UVC will be stable across technology changes. All that is needed in the future for executing the UVC JPEG program is an interpreter (an emulator) of the UVC architecture.

The execution of UVC JPEG program in the future will return the data with additional information, according to a *logical view* - defined by a *logical view description* or *schema*, which is also archived. This complete data preservation approach enables organisations to always retrieve a technology independent description (logical data description) of any JPEG file in the future with the aid of three components: UVC JPEG program, UVC, and the archived logical data definition for JPEG.

These 3 archived components enable any person in 2XXX to regenerate the information in the current environment, using new data formats. Writing a viewer program that invokes the emulator, runs UVC JPEG program, and processes the returned data as desired. The first proof of concept with this approach was very successful.

In the next step numerous digital object classes, like images, may be translated to the same logical view, eliminating the need for a future client to implement separate viewers for each original format.

Multiple digital object types can produce the same logical data reducing the need for multiple viewers



The data preservation approach is not limited to static information types. Sound and video can also be dealt with by data preservation. Even program dependent applications like relational databases could in essence be described by a logical data description; basically any information that can be described logically in a static way is a candidate for data preservation.

When the approach reaches critical mass probably a small number of logical data descriptions will remain like: general text, images, sound, video etc. However, the beauty of the approach is the fact that an organisation doesn't have to wait for this standardisation process to complete. Everybody at this stage can define their own logical data definitions and support it with the aid of the UVC.

Experiments on XML

We all know XML as a format or mark-up language. But because of its characteristics and because it is an open standard, it is promising to use XML as a preservation strategy.

Testbed experimented with XML and different types of email: MS Outlook 98 and 2000, K-mail, Eudora and Hotmail

The experiments give good results, but development of templates for end-users for internal and external use is needed to create, capture and store the email with required metadata, which makes it possible to interpret the email in the right context. The Testbed will develop these templates for MS Outlook within 6 weeks.

Further Experiments

New experiments expected in 2002 are the migration of spreadsheets, conversion of spreadsheets and databases into XML and a proof of concept with the UVC for text documents and spreadsheets.

Products

Eventually at the end of 2003 the Testbed project will provide:

- advice on how to deal with current digital records
- recommendations for an appropriate preservation strategy or a combination of strategies
- functional requirements for a preservation function
- cost models of the various preservation strategies
- a decision model for preservation strategy
- recommendations concerning guidelines and regulations

For further information about Testbed:

website: www.digitaleduurzaamheid.nl

email: testbed@ictu.nl