

Paper : **Cost Model for Digital Preservation**
Authors : Jacqueline Slats & Remco Verdegem
Organisation : Nationaal Archief of the Netherlands

1. Introduction

Digital Preservation Testbed was a 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 allowed the project to ascertain the effects of undertaken preservation action on different archival records.

Testbed researched three different approaches to long-term digital preservation: migration, XML and emulation (more specific: the UVC [Universal Virtual Computer] approach). Not only the effectiveness of each approach has been evaluated, but also their limitations, application potential and costs.

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

In the spring of 2005 the Digital Preservation Testbed project has completed its deliverables:

- 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 system;
- Decision model to select the right preservation strategy;
- Recommendations concerning archival guidelines and regulations;
- Cost indicators and Cost model for digital preservation.

This paper describes the different cost indicators of digital preservation and focuses on comparing the costs of the different preservation approaches that Testbed has investigated. A list of indicators, which exert an influence on the total cost of preservation, has been drawn up. Furthermore a computational model in Excel has been developed for calculating the total cost of preservation, and for comparing the costs involved in applying the different preservation approaches.

These costs are estimates based on Testbed's studies and experience, published information, information others have supplied to Testbed and on common sense. These detailed estimates are intended to encourage others to submit their comments on these figures, and to report the costs incurred in practice.

The costs indicators identified will always be incurred, irrespective of whether the relevant records need to be stored for no more than 10 or 20 years or come into consideration for permanent preservation. Although the scale of a digital archive system and a digital preservation system and the relative sizes of the different components of the installation may vary, the cost factors described below will in any case need to be taken into consideration.

Although the following list might initially appear to be detailed, it is nevertheless important not to overlook any of these cost indicators. It will, in particular, be necessary to calculate capital and personnel costs. Digital preservation will continue to develop and change. Consequently the functionality for sustainable preservation of digital records will also need to change. The costs incurred in making future changes need to be incorporated in the computational model right from the very beginning.

2. Cost Indicators

Testbed has made a distinction between the following cost indicators:

- (a) The costs of a digital archive system (= a digital depot or repository) and a functionality for the long term preservation of digital records (= preservation system)
- (b) Personnel costs
- (c) The costs of the development (or procurement) of software and methods/strategies for the preservation of digital records
- (d) The costs of the actual performance of certain preservation actions
- (e) Other factors that influence the total costs of preservation

- a) Costs of a digital archive system and a functionality for the long term preservation of digital records

The cost of a digital repository and a preservation system is comprised of various components. The cost model indicates the major factors as well as the minor factors. It also explains which indicators are influenced by the different ways in which records can be created and by the different preservation strategies and which indicators are not sensitive to (strategic) choices of this nature.

- The physical space
Physical space is required for systems for storage and long term preservation. Servers will be required for the storage of digital records and for the management of long term preservation. It may be advisable to set up separate development, test and production facilities for long term preservation.
- Hardware for the digital archive system (e.g. servers, storage media, back up equipment)
Hardware is required for the storage of records (in a file system, archival repository, or RMA). It will also be necessary to configure the storage equipment once an impression has been gained of the number of records that will need to be stored, although it will be difficult to predict transfer volumes of the coming years.
- Software for the digital archive system (e.g. OS, application software, security software, rendering software, communications software)
This covers issues such as the purchase of operating systems and standard software for databases. There will also be a need for protection software (against viruses, unauthorised access, and tampering with the records by unauthorised persons). There may also be a need for specific software for the acquiring and storing of authentic digital archival records
- Hardware for the preservation system (e.g. servers, workstations, storage media, back up facilities)
The preservation system may require computer systems (servers and storage) of the same type and size as the digital archive system. This is necessary so that the preservation system can receive groups of records with a total size in excess of several terabytes. The system will need to store these records in a safe manner ready for performing preservation operations (such as migration or conversion to XML), assessing the results of preservation actions, and returning the preserved records to the digital archive system.
- Software for preservation system (e.g. OS, application software, security software, preservation tools, test and evaluation software, communications software)
The preservation system may require more than one operating system, since it may be necessary to transfer records from their original operating system to another operating system capable of an improved preservation performance. In addition, there may also be a need for more than one programming environment if the organisation plans to develop in-house software tools or modify third-party tools.

b) Personnel costs

This Section reviews the staff duties involved in the operation of a digital archive system and a preservation system. The discussion reviews the numbers and types of staff that will be required. The cost model discussed later in this chapter is based on the time that will be needed from such staff, who has various qualifications and skills.

Personnel costs always form a major factor. The necessary staff could be selected or recruited specifically to work upon the digital archive and preservation system. However, in some instances it may be preferable to second staff from other disciplines (such as the records-management department or the ICT department).

- Digital archive system personnel
The staff will need to begin by designing and constructing the digital archive. This will require a budget for between one and two man-years. Cost calculations should not underestimate this.
- Preservation system personnel
The staff responsible for the preservation system will also first need to design and construct the system. They will then need to establish the quality control system, the SOPs, and the procedures. Finally, they will need to begin the development of the preservation methods and evaluation tests, and start work on the sustainable preservation of the records. Once again, the costs incurred in the development and construction phase are easily underestimated.
- Public-services staff
The staff responsible for helping the customers accessing the 'open' records

c) The costs of the development (or procurement) of software and methods for the preservation of records

One of the Testbed conclusions was that digital preservation is not a question of all or nothing. In many instances the characteristics of records that are essential to the records' integrity and authenticity can be separated from other less important characteristics. Digital preservation activities can then focus on those aspects of essential importance to the integrity and authenticity of the record.

The initial costs incurred in digital preservation relate to issues such as determining the authenticity requirements for each batch of records. In an ideal situation the records managers will specify these requirements. However, in some situations it may be necessary for the (authenticity) requirements to be determined by a multidisciplinary team comprised of specialists such as archives-management and IT specialists, whereby every member of the team has some experience in the other specialists' fields.

The cost model assumes that (authenticity) requirements will need to be determined for each batch of records. A batch contains records all made with the same application, the acquisition or preservation of which all takes place at the same time. It will later be shown that the size of the batch is a critical factor in the costs.

Once the authenticity requirements have been determined, the next step is to design and develop a suitable preservation approach. Since this is a lengthy process that requires a large number of skills, it is assumed that an international collection of shared preservation strategies will gradually be developed. However, even then it will still be necessary to evaluate these strategies in terms of the specific requirements of the batch of records in question. In some instances it will ultimately be necessary to modify the approach.

Finally, each strategy or approach will need to be tested and documented. All of the IT operations involved in each preservation system will need to comply with the most stringent quality standards. A

high level of quality is of essential importance to the authenticity, since the quality systems and the documentation are needed to prove that the preservation actions have achieved the intended results, and that they have had no influence on other records. A high level of quality also increases the probability that the approach will be re-used in this or other preservation systems.

d) Costs of the performance of preservation actions

This Section reviews the costs incurred in the performance of preservation actions on digital records. Within this context the 'performance of preservation actions' can relate to diverse activities:

- The migration of records (transformation)
- The conversion of records to XML
- The use of the UVC to retain the accessibility of records

These can be specified with OAIS terminology¹. A migration or other form of transformation of the records results in changes to the Archival Information Package (AIP) stored in the digital archive system. AIP1 is changed into AIP2. AIP2 serves as the basis for the DIP (Dissemination Information Package) issued to applicants.

Another form of transformation, which Testbed examined as a possible approach to the sustainable preservation of records, is conversion to XML. The possible cost benefits of XML (because it is an open standard, is expected to have a long and useful life, and can be interpreted by a variety of applications) are explained below. Conversion to XML changes the AIP from AIP1 to AIPX, whereby AIPX is in XML.

Retaining access to digital records through the use of hardware emulation has no influence on the record contained in the AIP. In principle the DIP will also remain unchanged in the future, although in practice it may be necessary to implement a number of small modifications to the DIP to accommodate future technology. In this respect, Testbed has examined the UVC approach formulated by IBM. This approach is based partly on emulation, and partly on migration.

In fact, and as will be revealed by our cost model, the costs of digital preservation activity (operational costs) are only a small fraction of the total costs of preservation. The costs of performing digital preservation also depend on the size of the batch: the cost model will reveal that grouping records in larger batches is cost-effective.

e) Other factors that exert an influence on the total costs

Other factors not mentioned in the above summary are also of relevance. These indirect factors can, however, account for a substantial proportion of the total costs. In addition, they can also have an influence on the impact of a number of the aforementioned factors.

- Public services
The degree to which users draw upon the services of the archive and preservation systems will have a great influence on the costs; however, the provision of services also offers an opportunity for the generation of income.
- The time between preservation actions
This is a critical cost factor. The more preservation actions, the higher the costs. In addition, as the number of preservation actions increases the risk of affecting the authenticity and integrity of the records, additional tests may be needed.
The costs can be reduced with longer periods of time between the preservation actions. However,

¹ http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

preservation actions carried out at excessively great intervals of time can increase the risk of problems with digital preservation and the cost of preservation.

- **Technology watch – assessing when the hazards increase**
A technology watch requires the monitoring of the hardware, software and systems used for the current records. The threatened obsolescence of components on which the digital records are dependent will give cause to the need for an evaluation and implementation of the necessary measures.
- **Supplementary storage requirements**
Testbed recommends that the original files of the preserved records also be stored. We advise that text document records are stored in both PDF and XML. These recommendations increase the storage space required for each record. In some instances the space can be increased by a factor of between three and five. Although storage is relatively cheap, this will result in additional costs.
- **Links to the management systems for electronic records**
Testbed has not examined links in DMSs or RMAs. However, it is to be expected that these links will be desirable at some point in the future. Extra costs will be incurred in the construction and maintenance of these links.
- **Volume of records**
The expected volume of the records to be stored and managed will have substantial consequences for the costs. The storage costs increase linearly with the volume. Moreover the required space will increase even more rapidly when the records need to be stored in a variety of formats (for example, the original file format and two migrated formats).
More expensive servers and storage systems may be required for large volumes of records (more than 500 Terabytes), in particular when there is a need for rapid access to the records.
It should be noted, however, that the cost of digital preservation is influenced more by the variety (diversity) of the records than by the volume of the records. Records that make use of various functions of an application or different application software will generally require different preservation strategies, or at least a variety of tests for the preservation strategies. For this reason it will cost less to preserve a few large batches of records which all use the same application (maybe also the same template) and have the same authenticity requirements, than a large number of small but diverse batches that take up the same amount of storage space.
- **Requirements for authenticity and reliability**
The authenticity requirements for a specific type of records constitute a significant cost factor. Consider, for example, a text document. Preservation of this will be a relatively simple task when only the plain text (the content) needs to be preserved. Highlights can also be preserved, at a slightly increased cost. However the costs will increase if the exact position of each character on the page and the exact colour must to be preserved. This will also complicate the preservation tests for the approach.
For this reason it is important that the authenticity requirements are determined in as comprehensive and realistic manner as possible.
- **Preservation of the systems themselves**
Finally, it will also be necessary to preserve the systems themselves. These costs will in part be covered by depreciation, as a result of which funds are made available for a three to five-year replacement cycle. However, it is also possible that specific elements of the preservation system form part of the digital record or preservation object², as a result of which these will need to be preserved separately. In any case, exporting digital records from system A and importing them into system B will not be a trivial undertaking.

² See Chapter 5 of the Database recommendations for a further explanation of what is referred to as the 'preservation object'.

3. Cost model

In this part of the paper the cost model will be explained. This computational model (version 1.0, 20 April 2005), developed in Excel 97, is available at our website (<http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=6>) and consists of three spreadsheets. The first one (Cost Basis) is the basis for the other two spreadsheets and describes the costs for staff and a digital archive system and a preservation system. The second spreadsheet (Time Calculation) describes the time it takes to perform several record keeping activities, to develop preservation approaches and to actually execute preservation activities (per person type, per year and over a period of N years). The third spreadsheet (Cost Calculation) focuses on the comparison between the preservation approaches Testbed project has looked into (migration [including the conversion to PDF]; the use of XML; and emulation [Universal Virtual Computer] in terms of costs (per batch of records, per year and over a period of N years).

Again, it must be emphasised that the cost model is not intended to be fixed and final; it is to provide a starting point for improvements and enhancements based on community experience and feedback. These costs are estimates, based on experiments and experience of the Testbed project, on published data and data shared with Testbed by other workers and on common sense. These detailed estimates are published to encourage other groups to comment on these costs, and to encourage groups to report real cost measurements.

The following sections review a number of the most important conclusions, based on the outcome of the cost model. The sections are arranged in Record Keeping Activities, Development Activities and Performing Digital Preservation. These costs are calculated over a period of one year till N years. We start by making explicit some of the assumptions the computation model is based on.

Assumptions made for the computational model

It is assumed that there are six categories of staff, who are paid four different hourly salaries. It is also assumed that each category of staff works 1620 hours per annum. A cost category is assigned to each category of staff and will be used in later parts of the spreadsheet.

It is assumed that the digital archive and the preservation system will need four types of space. An estimate has been made of the costs of furnishing each type. A proposal has also been made for the number of staff, based on full-time equivalents, to be based in each category of space. Once again, amendments can be made to the computational model that will influence the results from the calculations.

The facility specified should be able to manage a total of 100 Terabyte of records, and could readily be expanded to 1000 Terabyte (1 Petabyte) or more.

It is assumed that 10% of the records being ingested into the digital archive system will need repair (remove password, deal with unusual fonts, automated date fields, macro's, etc.).

It is assumed that records created directly in XML do not need repair, or metadata adding.

It is assumed that migration (based on backward compatibility) will be repeated every three years while the conversion to XML (and/or its successor) will be repeated every ten years. There is no repeating factor for the UVC (and the viewer), because it is assumed that adjustments to implement the UVC and the viewer on future hardware are marginal.

For the UVC approach it is assumed that each year a data format decoder program needs to be developed.

Record Keeping Activities

From the recommendations that Testbed has published for the four types of records it has investigated (email, text documents, spreadsheets and databases), it will be clear that digital preservation begins at the time of the creation of the records. The creation of records in an appropriate manner is a quicker, cheaper and less risky manner of obtaining sustainable records, compared to the ‘repair’ of those records at a later date.

A good example of this effect is the cost difference of emails that have been preserved from a standard email application (such as Outlook), compared to e-mails preserved using a system focused on durability, such as the TestbedXMaiL application. The calculated costs incurred in the acquisition and input of metadata amounts to EUR 2.30 per email, whilst the cost is no more than EUR 0.06 per XML email. It is important to note that these figures are based on the variables used in the cost model (e.g. a batch of 2000 emails and 4 hours to acquire and appraise this batch).

The main difference is that the XML emails are already equipped with the appropriate metadata and structure (as described in the XML schema).

Develop preservation approach

To develop a preservation approach is a major part of the costs involved in digital preservation. The example underneath is based on preserving just a hundred spreadsheets a year (5 batches of 20 spreadsheets). There is a significant difference between developing a migration strategy (33,50 Euro per spreadsheet) and developing the UVC approach, more specifically developing a data format decoder program (769,10 Euro per item). Applying the XML approach is free of costs (assuming that there is an application available storing spreadsheets in XML with the appropriate metadata).

Develop preservation approach	Personel costs/hr in Euros		Existing spreadsheet (migration)	Existing spreadsheet to XML	Existing spreadsheet (UVC)	New spreadsheet in XML
Gather requirements	35,63	costs per record type	285,00	285,00	285,00	0,00
Develop approach	35,63	costs per record type	2850,00	4275,00	71250,00	0,00
Test approach	26,88	costs per record type	215,00	322,50	5375,00	0,00
Costs in Euros per record type			3.350,00	4.882,50	76.910,00	0,00
Costs per item			33,50	48,83	769,10	0,00

Table 1: Costs of developing a preservation approach

Digital Preservation Activities

The operational costs of performing preservation actions are marginal compared to the costs of developing the actual preservation approach. For evaluation purposes automatic tools are needed to evaluate the results of the specific preservation activity. Manual evaluation can be used for random testing of a set of ‘preserved’ records.

Conclusions

Migration (including the conversion of text documents to PDF) is on the long run an expensive preservation approach, caused by the fact that development costs will recur over time. In our model we assume that migration has to be repeated every three years.

The use of the UVC approach is the most expensive preservation approach according to our cost model. This is caused by the huge development costs (R&D). In the Testbed project we have experienced that it was very difficult to develop a ‘data format decoder’ program for Excel. Furthermore, for each (new) file format a (new) ‘data format decoder’ program has to be developed. On the other hand, these costs will not recur over time: a once developed ‘data format decoder’ program will last “forever”. International collaboration on this area can strengthen the application of the UVC approach.

On the basis of our cost model XML proves to be a cost effective preservation approach, compared to migration (backward compatibility and conversion to PDF) and applying the UVC approach, especially when records are created directly in XML. Nevertheless, even records created in XML (or converted to XML) need some sort of preservation action, because XML will not last forever either. In our cost model it is assumed that records in XML (created directly in XML or converted to XML) need to be converted to a new open standard after 10 years.

Preservation approach	Existing spreadsheet (migration)	Existing spreadsheet to XML	Existing spreadsheet (UVC)	New spreadsheet in XML
Batch size	200	200	200	200
Number of batches per year	5	5	5	5
Total costs in Euros per year	9.213,79	11.252,29	107.061,79	855,00
Total costs in Euros after 20 years	709.461,96	337.568,75	2.141.253,83	129.622,92

Table 2: Four preservation approaches compared

Preservation approach	Existing text (migration)		Existing text to XML		New text in XML	
	200	2000	200	2000	200	2000
Batch size	200	2000	200	2000	200	2000
Number of batches per year	10	10	10	10	10	10
Total costs in Euros per year	20.479,33	112.579,33	24.574,33	116.674,33	1.710,33	1.710,00
Total costs in Euros after 20 years	1.578.294,67	8.669.994,67	737.230,00	3.500.230,00	279.943,33	1.200.943,33

Table 3: Three preservation approaches and the effect of batch size