nationaal archief

Costs of Digital Preservation

The Digital Preservation Testbed was founded bij the Ministry of Education, Culture and Science and het Ministry of the Interior and Kingdom Relations. It was established in October 2000 to research different strategies to preserve digtal documents over the long-term. Since July 2003 Testbed Digital Preservation is adopted by and continued within the Nationaal Archief of the Netherlands.

Nationaal Archief
Prins Willem-Alexanderhof 20
2595 BE Den Haag

Tel. +31 70 331 54 00
Fax: +31 70 331 55 40

Email testbed@nationaalarchief.nl
www.digitaleduurzaamheid.nl

Digital Preservation Testbed *Costs of Digital Preserving* (version 1.0)

The Hague, May 2005

Table of content

Costs of Digital Preservation

Cost indicators and the cost model

Testbed has studied the costs involved in the long term preservation of digital records, drawn up a list of indicators which exert an influence on the total cost of preservation, designed a computational model for the calculation of these costs, and compared the costs involved in the various methods for the creation of digital records and in the various preservation strategies.

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

This discussion is, in the first instance, focused on the larger archives; however, it is also applicable to the local storage needs of ministries and other (government) agencies. The costs 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 the storage system or repository and the relative sizes of the different components of the installation may vary, the following cost factors will in any case need to be taken into consideration.

Although the following list might initially appear to be extremely detailed, it is nevertheless important not to overlook any of these factors. 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.

Summary of the cost indicators

The costs involved in the long term preservation of digital records are influenced by a number of factors. These are summarised below. In the following discussion, a digital archive is included as an element of the costs, as is the storage of the digital records, for example, in an RMA (Records Management Application) or a DMS (a Document Management System). It is often difficult to specify the demarcations between the actual use of the records, their local storage, local preservation, archiving, and long term preservation. The concept of the 'records continuum', which is ideally suited to use in this context, can be defined as:
'a consistent and coherent regime of management processes from the time of the creation of records (and before creation, in the design of record keeping systems), through to the preservation and use of records as archives.'

Consequently digital preservation is not just a necessity for archival repositories, but also for every organisation. Each organisation will need to specify its own requirements, determine its own demarcations, and tailor the cost model discussed in this Chapter to its specific situation.

Testbed makes a distinction between the following cost indicators:

1) The cost of the digital archival system (a digital depot or repository) and functionality for the long term preservation of digital records[1]
2) Personnel costs
3) The cost of the development (or procurement) of software and methods for the preservation of digital records
4) The cost of the actual storage of digital records
5) Other factors that exert an influence on the total cost

1. The cost of a digital repository and functionality for the long term preservation of digital records

The cost of a digital archival repository and functionality for long term preservation is comprised of various components. The cost model (see page 75) indicates the major factors and 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.

1.1 The physical space
- 1.1.1 Server room, with air-conditioning
- 1.1.2 Sufficient office space
- 1.1.3 Conference room
- 1.1.4 Toilets and kitchen
- 1.1.5 Security

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. This can reduce the risks and increase productivity. There will also be a need for offices and conference rooms, for both the staff and visitors.

1.2 Hardware for the digital archival repository
- 1.2.1 Servers for the storage of digital records
- 1.2.2 Disks, tapes, or other storage media
- 1.2.3 Backup equipment
- 1.2.4 Network communications

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.

Appreciable costs can be incurred in the purchase of storage media (tapes, CDs). Make sure an adequate amount is budgeted for effective storage and backup media.

Network facilities may also be important. Archival repositories that receive large numbers of digital records from diverse locations may require a high-speed connection or a flexible connection capable of accommodating varying loads.

[1] See: Functional specifications for a preservation system

1.3 Software for the digital archival repository

- 1.3.1 Operating system
- 1.3.2 Security
- 1.3.3 Specific software for archives management
- 1.3.4 Old software applications
- 1.3.5 New (current) software applications
- 1.3.6 Display programs
- 1.3.7 Communications software
- 1.3.8 Database licences

This Section covers issues such as the purchase of operating systems and the standard software for databases. There will also be a need for protection software (against viruses, unauthorised access, and tampering with the archives by unauthorised persons). There may also be a need for specific software for the receipt and storage of authentic digital archival records, such as Depot 2000 or the Digital Archive System of the United Kingdom National Archives [2]. Every organisation which works with digital records will, irrespective of its size, have a need for a DMS (Document Management System) or an RMA (Records Management Application).

The following discussion assumes that the archival repository possesses functionality to provide access to the stored records. Consequently in addition to the customary storage software, there will also be a need for specific applications or display software which enable users to display (or use) the stored records.

Communications software and network and database licences are two other issues that are often overlooked when preparing the budget.

1.4 Hardware for the preservation system

- 1.4.1 Servers for the preparation of software
- 1.4.1 Servers for the testing of software
- 1.4.3 Servers for the storage of records before preservation action
- 1.4.3 Servers for the storage of records subsequent to preservation action
- 1.4.5 Work stations for programming work
- 1.4.6 Disks and tapes
- 1.2.3 Backup equipment
- 1.2.4 Network communications
- 1.4.9 Reading equipment for tapes and disks

The preservation system may require computer systems (servers and storage) of the same type and size as the archival repository. 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, for example, performing preservation operations (such as migration or emulation), assessing the results of preservation actions, and returning the preserved records to the archival repository (digital depot).

In addition more servers and storage equipment may be required for the development and testing of:

- preservation methods,
- software to evaluate the results of preservation operations,
- other tools.

When automated tools are to be developed and tested, t he test system may need to test large datasets so as to collect sufficient statistical proof of the tools' success.

The preservation system may need a variety of different types of reading equipment to read the various formats of tapes and/or disks.

[2] http://www.pro.gov.uk/about/preservation/digital/archive/default.htm

1.5 Software preservation system
    1.5.1 Operating systems
    1.5.2 Program environments
    1.5.3 Security
    1.5.4 Old software applications
    1.5.5 New (current) software applications
    1.5.6 Software for the preservation of documents
    1.5.7 Test and evaluation software
    1.5.8 Communications software
    1.5.9 Database licences

Extremely comprehensive software may be required for the sustainable preservation of records. 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.

The preservation system will also need to cater for a comprehensive package of software applications. This will allow for research into the different options for preservation and experimentation with a range of records and record-batches.

Automation is a significant factor in controlling the costs of large-scale digital preservation. Manual processing is one of the largest cost items for digital preservation. Consequently, automated preservation actions and automated evaluation (tests) are significant factors in controlling the costs of digital preservation.

2. Personnel costs

This Section reviews the staff duties involved in the operation of 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 have 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). In yet other situations, preference might be given to the use of temporary employees, or to the use of the services of a specialised company.

2.1 Duties of digital archives personnel
    2.1.1 Compile requirements
    2.1.2 Obtain funds and support
    2.1.3 Design and construct the digital archives

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.

Even designs obtained from other organisations or purchased from a commercial supplier will still require modification to meet the needs of the specific organisation.

- 2.1.4 Stocking the digital archive
- 2.1.5 Process management
- 2.1.6 Managing the digital repository
- 2.1.7 Security management
- 2.1.8 Quality control system and documents
- 2.1.9 Standard Operating Procedures (SOPs)
- 2.1.10 User manuals

Once the digital archive has been constructed the next step will be to develop the procedures and commence the management of the digital archive. The internal management encompasses the security and access procedures, a comprehensive quality system (to ensure the authenticity of the records stored in the digital repository), everyday SOPs (Standard Operating Procedures), and user manuals.

Management can also incorporate external activities, such as the identification of records, the arrangement of records, acquisitions, and cataloguing.

2.2 Duties of preservation system personnel

- 2.2.1 Compile requirements
- 2.2.2 Obtain funds and support
- 2.2.3 Design the preservation system
- 2.2.4 Construct the preservation system
- 2.2.5 Process management
- 2.2.6 Management of the preservation system
- 2.2.7 Ongoing security management
- 2.2.8 Quality control and documents
- 2.2.9 SOPs
- 2.2.10 User manuals

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.

2.3 Duties of Public-services staff

- 2.3.1 Access management
- 2.3.2 Training and schooling

3. The cost of the development (or procurement) of software and methods for the preservation of records

3.1 Determine authenticity requirements
3.2 Analyse authenticity requirements

One important Testbed conclusion 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 aforementioned tasks can be carried out solely by the organisation that created the records. 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 these requirements will be specified by the records managers. 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 cost.

3.3 Design preservation approach
3.4 Develop preservation approach
3.5 Preservation software (parser, etc.)
3.6 Viewing software

Once the authenticity specifications 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.

3.7 Test the approach
    3.8 If approved, continue
    3.9 If not approved, return to 3.1, 3.2, or 3.3
3.10 Document 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.

4. Cost 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 performance of a migration on request
- The use of emulation to retain the accessibility of records

These can be specified with OAIS terminology [3]. A migration or other form of transformation of the records results in changes to the Archival Information Package (AIP) stored in the archives. AIP1 is changed into AIP2. AIP2 serves as the basis for the DIP (Dissemination Information Package) issued to applicants. Migration is one of the preservation strategies examined by Testbed.

The performance of a 'migration on request'[4] has no influence on the AIP. It is possible to produce a DIP which differs from the last DIP produced by the same AIP. It will in any case be necessary to produce a DIP that is both authentic and accessible. Migration on request will require the preparation, testing and issue of an appropriate software tool before the migration (or another form of transformation) is carried out. If a user requests a record, the correct tool is retrieved and the record subsequently transformed. The user receives the transformed copy in the form of a DIP. The transformed copy can be stored, or deleted once the user has finished with it.

Another form of transformation which Testbed examined as a possible approach to the sustainable preservation of records is conversion to XML. The costs for this are included in the migration. 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 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 in part on emulation, and in part on migration.

In fact, and as will be revealed by our cost model, the cost of digital preservation activity is only a small fraction of the total cost of the digital archive. The cost of digital preservation also depends on the size of the batch: the cost model will reveal that grouping records in larger batches is a particularly cost-effective approach.

4.1 Determine which digital records will need to be preserved
4.2 Construct the interface with the archives-management system
4.3 Incorporate systems for electronic records management, DMS, RMA
4.4 Receive digital records
4.5 Select the preservation strategy and approach

[3] http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[4] '*Cedars Guide to Digital Preservation Strategies*', Clare Jenkins, April 2002.

Records which will need transformation, or which require the development and testing of an emulator, can be identified by an automated process (a technology watch, or a process within an RMA or DMS) or by means of manual identification. Records requiring transformation will first need to be selected and transferred to the preservation system, after which a preservation strategy can be assigned to them.

4.6 Preparing records for transformation
    4.7 Supply metadata
    4.8 Repair or modify records

In specific circumstances it will be necessary to prepare records for transformation. Records will be in need of 'repair' if metadata are missing, or when the records possess properties that could pose risks with the selected preservation strategy. This can be a slow and labour-intensive process that accounts for the majority of the costs.

Automated methods for the assignment of metadata or the repair of records can greatly reduce the associated costs.

4.9 The transformation of records using the selected method
4.10 Evaluation of the transformation
    4.11 The records are accessible
    4.12 The records are authentic
    4.13 If NO, return to 4.9 or to an earlier step
4.14 Storage of preserved records in the digital archives

The final step is to transform the records. Every transformation will need to be evaluated to demonstrate that the records are still accessible and to ensure their authenticity and integrity after the transformation. Within the context of this discussion 'transformation' refers both to migration and conversion to XML.

Transformed records shall need to be transmitted to the digital archive, where they will be stored until further preservation actions are required and where access to the records can be managed.

Note: it is assumed that hardware emulation is not employed. When hardware emulation is used there will be (virtually) no transformation costs, and no recurrent costs for the preparation of records. The cost model reviews the long term cost benefits offered by emulation.

5. 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.

5.1 Public services
- 5.1.1 Number of users
- 5.1.2 Required training and support tools
- 5.1.3 Required maintenance and support

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. Kevin Ashley has drawn up a summary of the costs incurred in providing public services[5].

5.2 The time between preservation actions
The time between preservation actions is a critical cost factor. The more preservation actions, the higher the costs. In addition, more preservation actions increases the risk of affecting the authenticity and integrity of the records, and there may also be a need for additional tests.

The costs can be reduced with longer periods of time between the preservation actions. However, preservation actions carried out at excessively great intervals of time can increase the risk of problems with digital preservation and the cost of preservation.

5.3 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.

5.4 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.

5.5 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.

5.6 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).

---

[5] *Digital Archive Costs: Facts and Fallacies* Kevin Ashley, DLM Forum 1999 http://www.europa.eu.int/ISPO/dlm/fulltext/full_ashl_en.htm

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 it will a large number of small but diverse batches that take up the same amount of storage space.

5.7 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.

5.8 Preservation of the systems themselves

5.8.1 Preservation of the archival system

5.8.2 Preservation of the preservation system

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 object[6], as a result of which these will need to be preserved separately.

One example of such an element is the preservation log file, the logbook of earlier preservation operations, which will also need to be preserved. Another example is the emulator, which will need to be preserved to ensure for the continued accessibility of the records and for their possible reuse.

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

Cost model: Results

A cost model has been prepared in the form of an Excel spreadsheet. This computational model (version 1.0; 20 April 2005) is available from the Testbed website: (http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&categorie=6).
The following sections review a number of the most important conclusions, and relate these to other general information about the costs of preservation and of archiving records. The sections are arranged in the lifecycle of a record, i.e. its creation, acquisition (by the archive), transformation and, where applicable, emulation.

1. 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:

- Administrative support (1), hourly wage EUR 14
- Archivist/Records Manager (2), hourly wage EUR 25
- Supervisor (3), hourly wage EUR 32
- Data-management assistant (4), hourly wage EUR 18
- Programmer (5), hourly wage EUR 25
- Senior IT assistant (6), hourly wage EUR 32.

The above salaries have been assumed for the purpose of calculations with the computational model; they are, of course, open to discussion and/or amendment. Amendments can be made to all calculations in the computational model that make use of the rele vant salary.

It is assumed that the archive and preservation system will need four types of space. An estimate has been made of the cost of furnishing each type. A proposal has also been made for the number of staff, based on full-time equivalents (FTEs), 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.

- Space for the digital archive system: 1 FTE, categories 2 and 4; costs EUR 1,897,400 in the first year and EUR 632,403 in every successive year; storage capacity 100 Tb.
- Office space: 1.4 FTE, categories 1 and 3; costs EUR 7,400 in the first year and EUR 2,466 in every successive year. Standard office equipment and furnishings.
- Development and test area: 3 FTEs, categories 4, 5 and 6; costs EUR 50,200 in the first year and EUR 16,731 in every successive year. This room will be equipped with additional software (program environments) and hardware (inclusive of older systems and newer systems). These facilities will be used for the development of digital preservation tools.
- Space for the digital preservation system: 3 FTEs, categories 2, 4 and 5; costs EUR 279,600 in the first year and EUR 93,190 in every successive year. This room is intended for the acceptance, transformation and testing of digital records. The equipment will be capable of storing 10 Terabytes of records each time. If so required, this room can be combined with the space for the digital archive.

The facility specified here should be able to manage a total of 100 Tb of records, and could readily be expanded to 1000 Terabyte (1 Petabyte) or more.
This facility would be able to cater for the annual acquisition and transformation of 40 batches of e-mails (2000 e-mails in each batch), 20 batches of text documents (200 documents in each batch), 20 batches of spreadsheets (20 in each batch), and 20 databases. A batch of 4000 e-mails costs about the same as a batch of 2000 e-mails, provided that they have been created in a durable manner.
Every year this facility would be able to develop 55 ‘preservation approaches’, each for a different batch of records. It would be possible to carry out more work at the same cost in the event that different batches resemble each other. For example, it would be possible to use the same preservation approach for two batches of text documents which differ only with respect to the process in which they were used, but that were created by the same application, possess the same properties, and are governed by the same authenticity requirements. The use of the same preservation approach would result in

substantial cost savings.

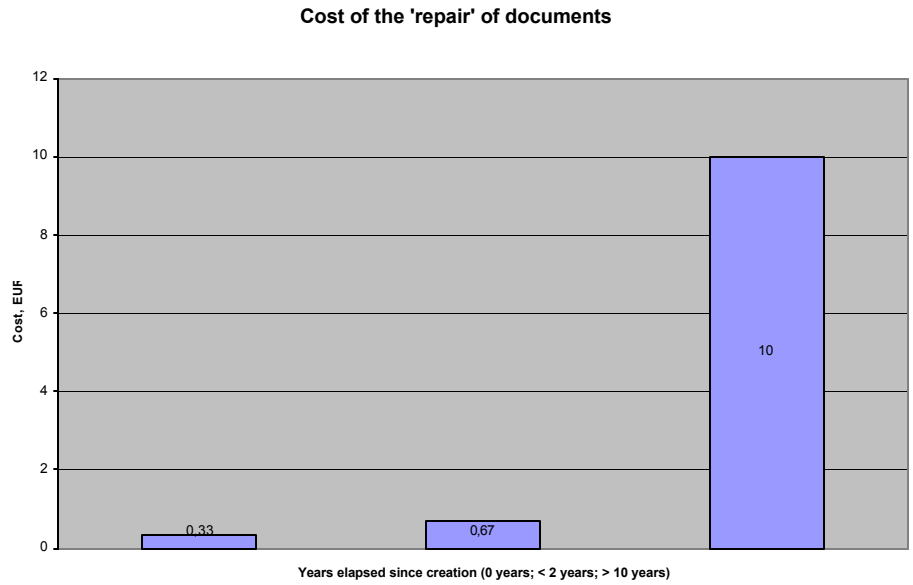
The aforementioned batches would require a total storage capacity of 5 Gb (on the basis of 50 kb per e-mail, 100 kb per text document, 250 kb per spreadsheet, and 2 Mb per database). The discrepancy between these figures and the total storage capacity of the model facility compared on the one hand with the NARA figures for electronic records or on the other hand with the Amsterdam Municipal Archives, indicates two things:

- the estimated number of digital records that can be processed by a team is probably a conservative estimate;
- the automation of the procedures used for the acquisition, inspection, and transformation of digital records will result in tangible benefits.

2. The creation of documents

From the recommendations that Testbed has made for four types of records, it will be clear that digital preservation begins at source, i.e. 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 suitable durable records, compared to the 'repair' of those records at a later date.

Testbed has estimated the costs incurred in the creation of durable documents with the appropriate metadata, the addition of metadata and performance of repairs after 2 years or less, and the addition of metadata and performance of repairs after 10 years or less.



Graph 1: Cost of the 'repair' of records

The differences in cost per record become substantial when they are applied to batches of 1000 records.

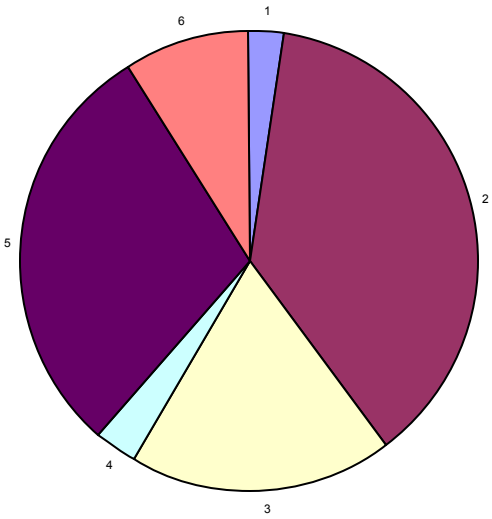
Approximately EUR 333 must be paid for the creation of a batch of 1000 records in an appropriate manner (the first bar in the graph). This calculation is based on a cost of EUR 0.33 for the creation of a well-constructed record.

Conversely, it will cost EUR 10,000 (the third bar in the graph) to 'repair' a batch of 1000 badly created records.

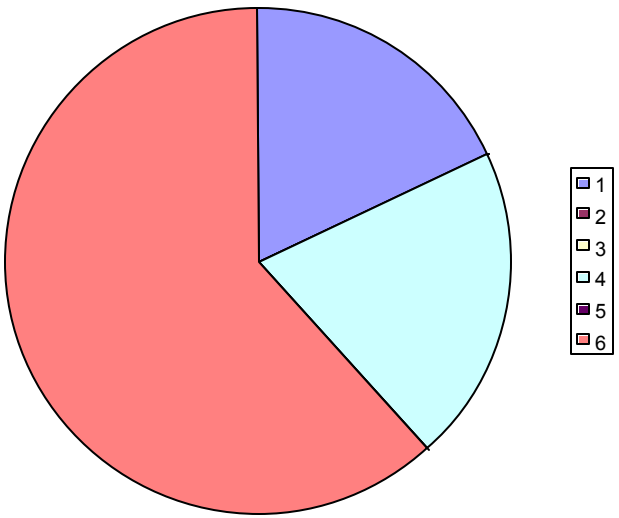
Another good example of this effect is the cost difference of emails that have been preserved from a standard email application (such as Outlook), compared to e-mails preserved using a system focused on durability, such as Testbed's XML/e-mail application. The usual cost incurred in the acquisition and

input of metadata amounts to EUR 1.41 per normal e-mail, whilst the cost is no more than EUR 0.06 per XML e-mail. The difference is that the XML emails are already equipped with the appropriate metadata and structure.

Graph 2: Cost of acquistion and preservation of existing e-mail messages



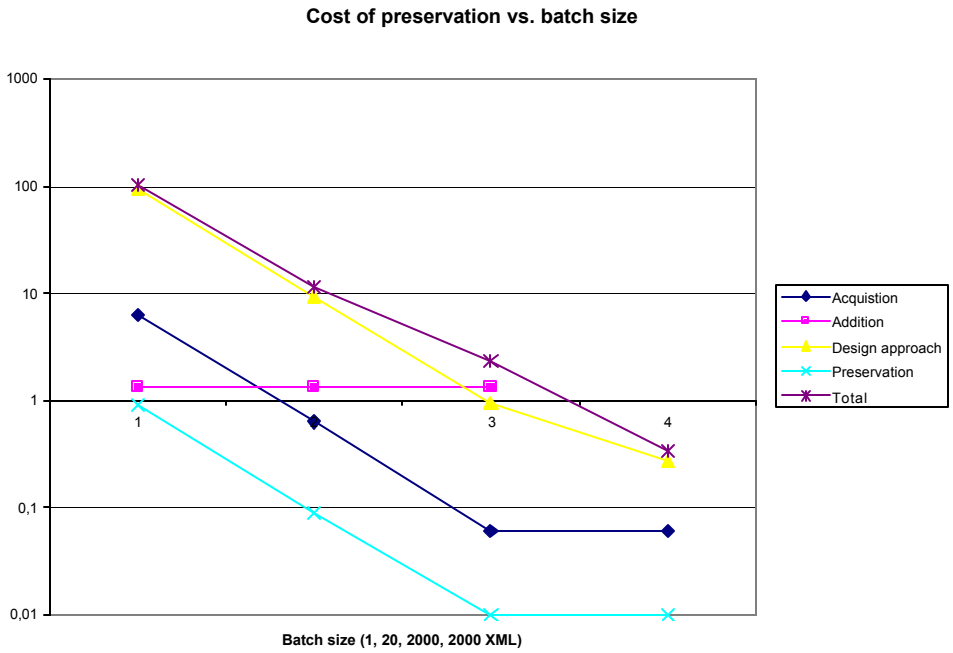
Graph 3: Cost of acquisition and preservation of existing XML e-mail messages



Legend:
1 = Acquire and appraise batch of e-mail,
2 = Necessary metadata,
3 = Repair digital records,
4 = Determine authenticity requirements for batch,
5 = Develop and test the preservation approach,
6 = Perform preservation and assess of e-mails.
The above legend relates to both graphs. In the right-hand graph items 2, 3 and 5 are all 0.

Accession to an Archive

The accession phase shows the effect of batch size on costs. This effect continues into the sustainable preservation phase. At every stage, larger batches of records cost less to manage. The reason for this is that metadata and processes can be added to 10,000 similar records as quickly as they can be added to one record. The time required to process a batch of records does not depend greatly on the size of the batch. For this reason the cost per record is much lower for large batches.
How can records be grouped into batches? A group of records from the same application (such as an e-mail or a word-processing program) that all contain explicit and correct metadata and which can all be processed in the same manner, can be processed together in a batch.



Graph 4: *Cost of preservation vs. batch size*

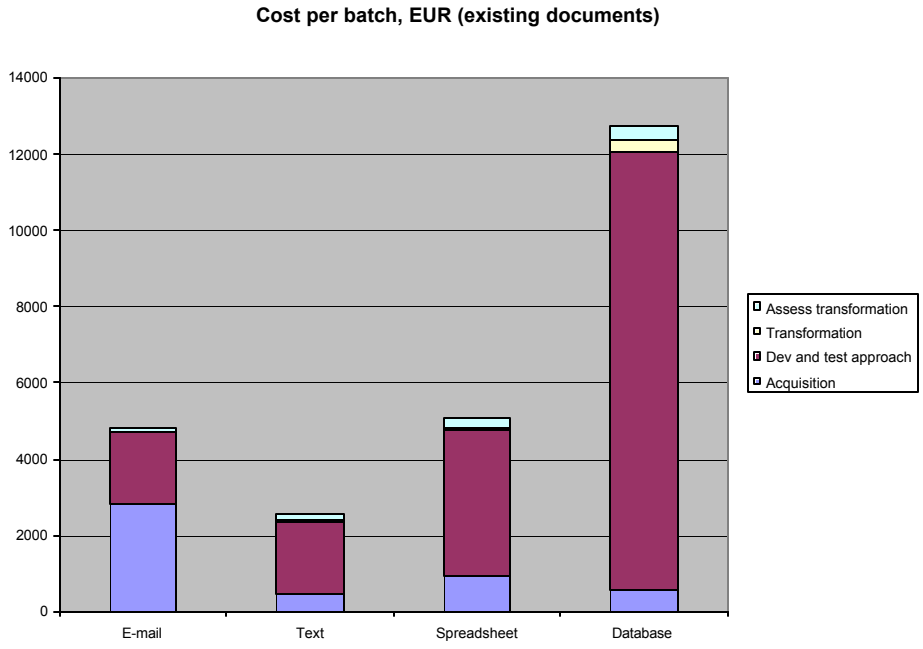
The graph reveals that the total cost for acquiring and preserving the e-mails decreases from EUR 102 per e-mail to EUR 2.35 per e-mail when the batch size is increased from 20 to 2000 e-mails. Acquiring and preserving e-mails already created in XML (for example, using the application developed by Testbed) costs even less (EUR 0.34 per e-mail for a batch of 2000 e-mails). The development of the approach is the most costly element of the process. If the approach can be used by several batches of records, through careful process design, then a significant saving can be made.

3. Transformation of different types of records

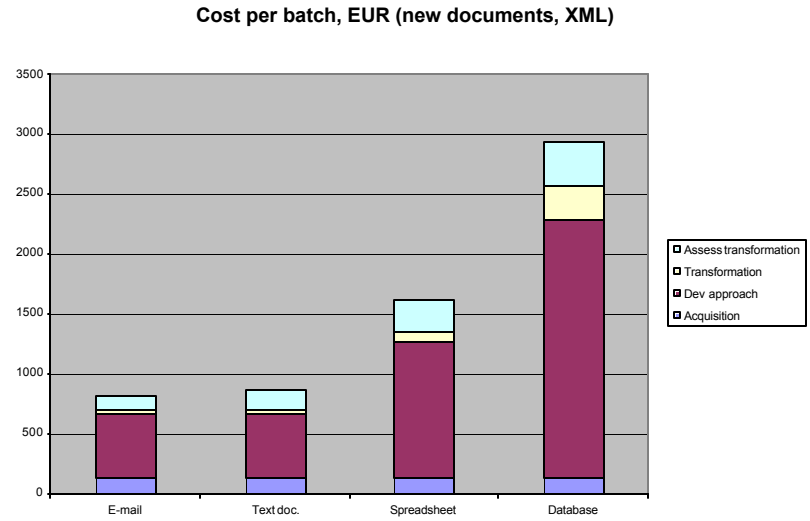
It is difficult to compare the costs of different preservation methods because they depend on so many different factors, the specific values of which will vary between different organisations. The migration costs have been calculated for each of the four record types investigated by Testbed. The costs for transformation do not differ substantially. The majority of the costs relate to the development and testing of the approach, and testing the transformed records. These costs remain constant for the migration from one application (version) to another, or for transformation to a standard format.

As discussed above, the size of the batch is one of the most significant variables for the transform ation. For the purposes of this comparison it is assumed that e-mails are supplied in batches of 2000, text documents in batches of 200 and spreadsheets in batches of 20, and that databases are received and managed on an individual basis.

Testbed's experience has also revealed that the amount of work involved in adding metadata to each type of record, and in developing and testing the requisite transformation, varies with the type of record. The following graphs show the costs for existing and new records (such as records created in XML, whereby the correct metadata can be added at the point of creation.).



Graph 5: *Cost per batch, EUR (existing documents)*



Graph 6: *Cost per batch, EUR (new documents)*

The total acquisition and preservation costs for batches of well-formed records are much lower than the costs for batches of badly created records. Acquisition (which here encompasses the repair of records and the metadata) is a significant cost item for badly formed records. Substantial costs will also be incurred when developing the approach. The development costs per batch can be substantially reduced if the approach can be shared by several batches.

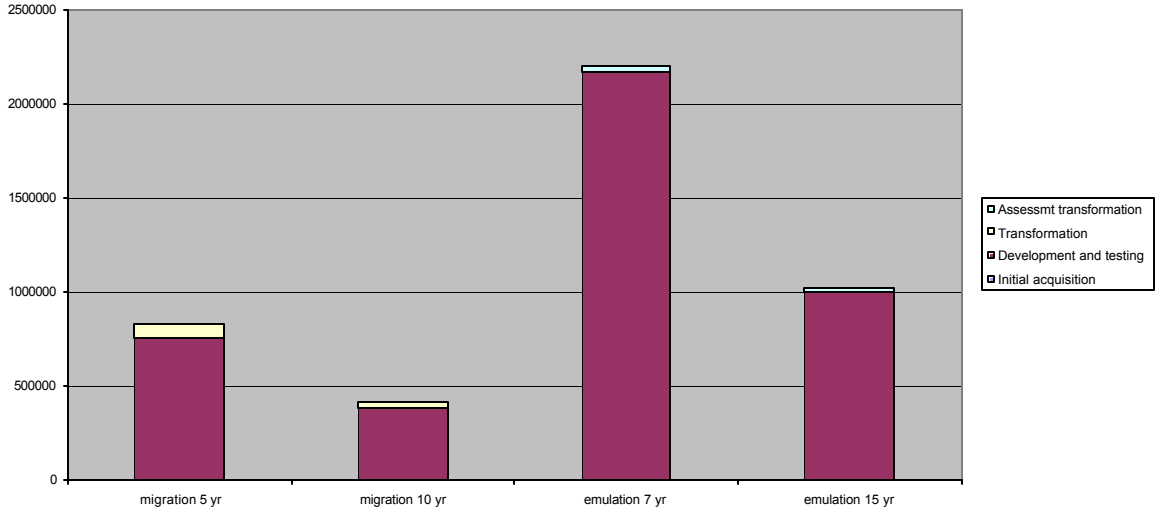
4. Emulation (inclusive of the UVC approach)

Hardware emulation, the only form of emulation included in this review of the costs, offers two benefits when compared to migration as a preservation strategy. The first advantage is that the records themselves do not need to be modified. The second is that one emulator can be used for a variety of record types. Consequently one Pentium-PC emulator would cover text documents (whether prepared in Word, Open Office or Lotus SmartSuite), spreadsheets, and desktop databases.

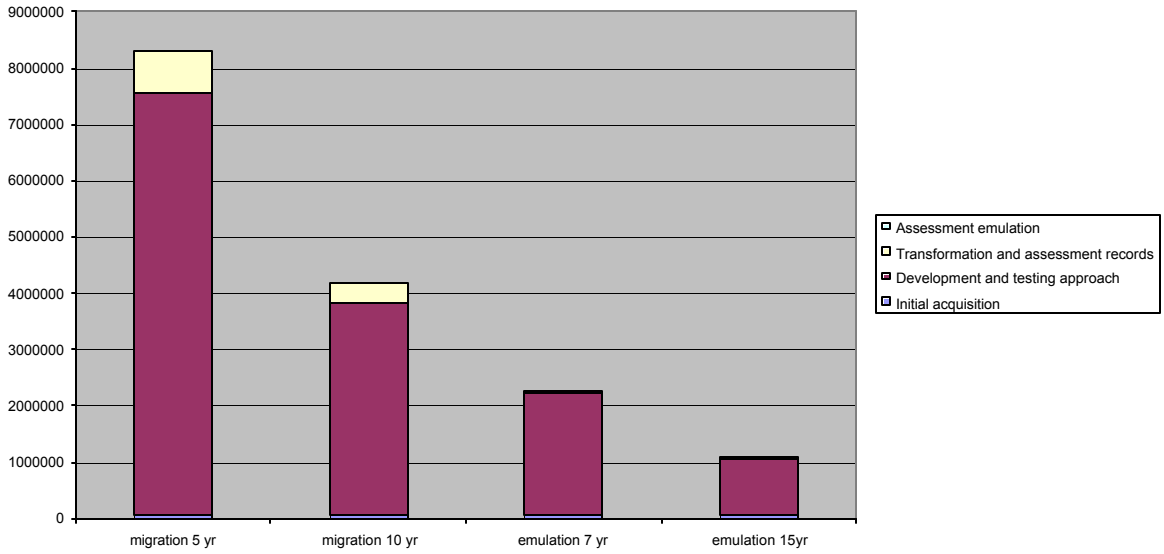
The following graph compares the costs for the maintenance of a collection of records, including e-mails, text documents, spreadsheets and databases, over a 100-year period. This collection resembles a shopping basket: arbitrary, but nevertheless illustrative. It is assumed that:

- the records need to be migrated every 5 or 10 years.
- new emulators are needed every 7 or 15 years.
- once the authenticity requirements for a batch of records have been identified, they do not have to be re-assessed.
- once added, metadata do not need further modification (other than the automatic addition of new preservation metadata about each preservation operation).
- this cost calculation uses a 'basket' of records containing one batch of each type of records, together with a mixture of existing and newly-created (durably-created) records.
- it is estimated that the development of an emulator will require 1.75 years' of work by a team of three persons.

Cost of digital preservation for a 100-year period, EUR



100-year digital preservation: 10x batches



In this second graph it is also assumed that ten batches of each type of records must be preserved and that each type can make use of the same emulator, but that different migration operations are required for each batch, for example as a result of different authenticity requirements.

The advantage of emulation in this example is that only one emulator need be developed for all records created by application software that used to run on the old emulated platform. In contrast, records from different software applications will require different migration approaches, and differences in authenticity requirements between batches from the same application can require different approaches.

5. Cost information from other sources

Testbed has obtained information about the cost issue of digital preservation from a number of other sources. This information can be used for an independent check of the calculations and the cost model.
For the purposes of these cost comparisons it may prove useful to consider the following equivalences:

- 1 metre of records (linear, on a shelf) is approximately equal to
- 0,09 cubic metre (based on A4 pages), which is about
- 3 cubic feet, with about
- 6,500 pages that, in digital form, would occupy about
- 65 Mb of storage space (as flat text).

Delft University of Technology, Utrecht University

The team in Delft and Utrecht has published a cost estimate for medium-term digital preservation that is one of the most detailed and carefully-considered estimates published to date. Dekker, Durr *et. al.*[7] have analysed the costs for the maintenance of records over twenty years.

National Archives of Canada

The National Archives of Canada employs about 661 FTEs and has an annual expenditure of about CAD 49,000,000. This excludes certain services supplied by other agencies, such as accommodation. The total costs of the National Archives amount to about CAD 90,000,000 per annum.

Of the main expenditure, EUR 30,460,000 (at CAD 1 = EUR 0.6217), about 30%, was destined for the acquisition and management of records, 18% for the management of government information, and 26% for services, awareness, and assistance.

The Annual Report of the National Archives of Canada (see www.archives.ca) notes that during his 15-year period of office, the then Prime Minister Mr Pierre Trudeau created some 1.3 Mb of electronic information. The offices of the current Prime Minister, Mr Jean Chretien, create about 1.3 Mb of digital records every day.

The current records stored by the National Archives comprise 111 km of paper government text records, 45 km of paper text documents from private individuals, 3.2 Gb of digital records, 2,500,000 maps and architectural drawings, 345,000 hours of audio, video or film, and 22,734,000 other items.
The cost of managing government records amounts to about CAD 80 per meter shelf, equivalent to about EUR 50 per shelf meter.

The UK National Archives (formerly the Public Record Office)

The UK National Archives (formerly the Public Record Office looks after about 176 km of records at a cost of about GBP 97 per metre (EUR 135) for the selection and preservation of the records.

---

[7] An electronic archive for academic communities, November 2001.

Providing access to the records results in additional costs for each visit, namely GBP 5.59 (EUR 7.80) for on-site access and GBP 0.13 (EUR 0.18) for online access.

The UK National Archives has a workforce of 451 FTEs and an annual budget of about GBP 35,600,000 (= EUR 49,662,000). The annual reports are published on www.nationalarchives.gov.uk and www.pro.gov.uk.

National Archives of Australia

The National Archives of Australia (www.naa.gov.au) has a workforce of about 435, of whom 322 work full-time. The annual budget is about AUD 148,000,000 (EUR 83,428,000 at an exchange rate of AUD 1 = EUR 0.5637) of which AUD 38,274,000 (EUR 21,575,000) is destined for personnel and operational costs. (The remaining costs relate to depreciation, which in Australia also includes changes in the value of the collection itself.)

US National Archives and Records Administration (NARA)

The NARA Performance Report 2002[8] indicates that the budget for 'space and preservation' amounted to USD 128,000,000 (EUR 109,000,000 at USD 1 = EUR 0.8524), with 338 FTEs. This also includes digital preservation, which is not stated separately in the accounts. NARA reported that in the preceding year the volume of 'logical records' (digital records) increased by 60% and that the volume of digital records during the presidency of President Clinton increased by 1500% (a factor of 15) in comparison with the preceding presidency.

Of these significant electronic holdings, 98% records are accessible, irrespective of their original format. NARA now manages about 4×10^9 'logical records'.

It should be noted that the total NARA budget amounted to USD 286,000,000 (EUR 244,000,000), and that in 2002 it had a workforce of 2829 FTEs.

NARA's digital preservation programme has 48.5 FTEs. Of these, 42.5% are active in ingest, i.e. maintaining relations with records producers, receiving records, generating Archival Information Packages (AIPs). A further 20 work on digital archival storage: receiving the AIPs, managing the storage hierarchy, monitoring quality, and generating access copies. 8% are engaged in data management, 7.5% in accounting, and 14% in user services and access. Preservation planning takes up 8% of staff time.

Amsterdam Municipal Archives

In 2001 the Amsterdam Municipal Archives acquired 350 metres (and 16,435 objects) of new archives. The digital archives were also expanded by 185 Gb. The Archives have a workforce of about 160. The reports are available from http://gemeentearchief.amsterdam.nl.

[8] http://www.archives.gov/about_us/strategic_planning_and_reporting/2002_performance_report.html#goal3