An electronic archive for academic communities

R. Dekker[1], E.H. Dürr[2], M. Slabbertje[3] and K. van der Meer[4]

[1]Library of Delft University of Technology
P.O. Box 98, 2600 MG Delft, The Netherlands
R.Dekker@library.tudelft.nl
[2]Utrecht University, Dept. Computational Physics
P.O. Box 80195, 3508 TD Utrecht, The Netherlands
E.H.Durr@phys.uu.nl
[3]Library of Utrecht University
P.O. Box 16007, 3500 DA Utrecht, The Netherlands
M.Slabbertje@library.uu.nl
[4]Delft University of Technology, Dept. ITS
P.O. Box 5031, 2600 GA Delft, The Netherlands
winfvdm@is.twi.tudelft.nl

Abstract. Related to the Roquade programme to enhance scientific communication for the academic community, an electronic archive of scientific publications is being developed. This article describes the design of this archive. For technical reasons the scientific information items will be wrapped in XML containers. The choice between the emulation or conversion strategy can be postponed until the moment that hardware threatens to become obsolete. Electronic archives are expensive. Because of cost considerations the submitters of documents must be involved in the assignment of metadata, a SLA-like construction must be designed for the acceptance of formats of SIP's, and explicit appraisal procedures will be applied. Still the costs are considerable; for metadata assignment, administration and quality control, and technical infrastructure of an electronic archive that accepts 5,000 items per year the costs are estimated to be 29 per information item.

1 Introduction, Goals and Purposes

The Roquade programme was started by the Library of Utrecht University, the Library of Delft University of Technology and the Netherlands Institute for Scientific Information Services (niwi) (1). The Roquade programme aims at enhancing scientific communication in the interest of the academic community. For ages, management of scientific output was outsourced to specialised craft: publishers. But now scientific authors and scientific readers are often members of one virtual academic community, or even colleagues in a virtual project. And as document management tools are ubiquitous, universities look after their scientific information. The libraries will have a central role in scientific information management. They will control written knowledge in any form: journal articles, preprints, scientific reports, scientific presentations, contributions to scientific discussions and others; and moreover facilitate and stimulate related developments. The time path for this development in the library is not clear, but it could go fast. Stevan Harnad once remarked: if a significant number of universities worldwide would mount and use archive software, the freeing of the research literature in a global public archive could take place within months.

The digital library needs an electronic archive. A project to develop this was started by the Roquade partners together with Maastricht University. In this project, the digital library archives are developed according to the same architecture. Eventually, the collections or parts of them could be virtually merged to form common collections. Even a common authoring environment to present such collections to a designated community could be created. There are technical tools for services from collections of items on different servers, for instance the collaborative authoring protocol WebDAV (Web Distributed Authoring and Versioning (2, 3); it is a standard related to http).

2 Other Electronic Archive Projects

The NEDLIB project is well known. It was led by the Royal Library of the Netherlands and involved nine European national libraries. It focused on needs of national libraries as they extend their traditional role to the preservation of digital publications and to establish a Deposit System for Electronic Publications (DSEP). As a part of the project, a process model was developed to incorporate DSEP functions into library practices for the handling of digital materials (4). The Open Archive Information System (OAIS) model was adopted as a basis for this process model (5). According to ref. 6, the main concern that NEDLIB raised about using OAIS for this goal was its lack of explicit functions and strategies for continuing access and how these might be chosen and effectively implemented in an OAIS repository; and 'NEDLIB reworked the OAIS model itself and initiated changes that take this preservation perspective into account'.

The Nedlib project refers to static electronic publications in their definitive forms, to be preserved forever. Authenticity of the electronic documents in the archive is very important. But in our case, a digital archive is needed for our own scientific information items with the purpose of accessibility and reusability, in a setting of reuse for education and research goals. Keeping information items over time should be as reliable as possible, but our archives are meant to be working archives, the scientists need the informational value far more than the evidential value.

The Cedars project, well documented as well, originally incorporated the universities of Oxford, Cambridge, and Leeds (7). A project demonstrator archive was developed. The purpose is, to test different aspects of digital archiving for these academic communities with various types of digital materials. A number of test sites acted as both content providers and end users of the demonstrator archive. A Web-based front end was developed to allow access from additional sites.

The Cedars project differs slightly from Nedlib: digital resource preservation but also conservation policy issues seen from a university library, and awareness of developments are mentioned. Just like the NEDLIB project the OAIS model has been adopted by the Cedars project. Special attention was given to the preservation metadata of the OAIS model. An important deliverable of the project was the translation of the abstract preservation metadata of the OAIS model into a practical preservation metadata set (8). The British Library and Birmingham University Library want to investigate funding for electronic archives (8). A framework for costs has been developed (9).

Many initiatives on electronic archives have recently been developed. There is much experience on E-print archives (10), there are many initiatives on digital repositories for historical (often visual) sources and other examples exist. But on top, Nedlib and Cedars are important references for the development of electronic archives for academic communities.

3 Electronic Archives for Scientific Communities

For the present case, experimental results will be needed for high-level decisions on methods and strategies for electronic archives and preservation of scientific digital sources; including possibilities to execute these different methods and strategies; experience with tools to preserve scientific digital sources, standards for maintenance of an electronic archive and standards for retrieval and access of the information items in the electronic archives; ways to keep, search and access heterogeneous durable sets of scientific digital information items; and quantities and types of scientific information items that should be kept in an electronic archive of a university (11). The way of working to come to a working archive, however, starts with research and experiments on the following topics:

- The structure of information objects
- The viewers
- The digital longevity strategy: emulation or conversion
- The role of costs for the way of operation of the electronic archive
- The size of costs, the amounts of costs of the electronic archive

4 Technical issues

4.1 Structure of Information Objects: XML containers

It was decided to work out the idea of XML containers. The Archival Information Packages (AIP), to be stored in the electronic archive, will be wrapped in XML. XML, the extensible mark-up language, has the enormous advantage that the language is self-descriptive. So, the contents of the packages can be deciphered as long as the characters are recognised. That recognition is the subject of a list of standards that is too long to be convincing (ISO 646 – ASCII; ISO 6937; ISO 8859; ISO 10646 - Unicode), but that problem is outside the scope of this project.

XML Schema's will be used for the logical structure (or 'content model') of the information objects.

Why was PDF not chosen? An academic library cannot prescribe the format to the submitters of SIP's. And if the library could do so (and the submitters would even accept that prescription) it would be of little value for information items containing a spreadsheet, Dynamic HTML or other software elements.

4.2 Viewers

For generic maintainable long-term preservation of the information items in the containers, the object-oriented (OO) approach is used. In this approach, information items are perceived as objects. The objects must be made visible. One or more 'views' on its content can exist. As in any OO design there are many objects with similar characteristics. Such a collection of similar objects forms a class. The view methods, i.e. the viewers, become class methods of these groups of objects.

In this way, the preservation of the viewer for an electronic information item is separated from the preservation of the bitstream. There is a wide variety of electronic formats, even for common documents, such as the MS-Word versions, PDF and bitmapped picture formats. They represent different views realised by different viewers (= programs). Guaranteed availability of programs to 'view' these documents on the long term is questionable. Commercial organisations rarely guarantee their products beyond their commercial lifetime and cannot guarantee them beyond the lifetime of their organisation. So, on the long run one cannot rely on the availability of certain viewer software on the client side. An option is to archive an item and reproduce on request later the bitstream only 'as is', without viewer support, the 'null viewer' approach. In this case the responsibility of maintaining the programs to handle the item content is shifted to the reader. But normally, the library should ensure that there is always a possibility to 'run' the viewer at the library server and view the results in some browser, without plug-ins, running at the client side.

So the library has the task to support viewers by keeping the viewer running over time, supported by external dealers (museums) with old software and hardware.

4.3 Digital Longevity: Emulation or Conversion

The emulation strategy means that an original bitstream is converted by a sequence of emulators (probably integrated). The conversion strategy means that after emulation the resulting bitstream is stored. On a conceptual level the results of both strategies are the same. On emulation, conversion is executed at the time of request ('on the fly') and on conversion the intermediate results are stored. In both cases the same conversion programs will be used. The difference boils down to a trade-off between computing power and storage capacity ((12) for a more detailed description). If conversion has been performed neatly, with respect to all characteristics of the information item, the level of authenticity of an information item might be acceptable. But the same condition applies to emulation. So, for the aspect of authenticity the difference between emulation and conversion is so small that it can nearly be neglected.

The choice between emulation and conversion will be more dependent on costs and such contingency factors. One can calculate (on the one hand) the cost of complete (sequenced) emulation from the original bitstream and (on the other hand) storage of intermediate results after any emulation for future use. The cost of each and estimated future use of an information item may be decisive. Only if and when emulation becomes doubtful, as no hardware may longer be available, conversion becomes the last solution.

4.4 The structure of an information item

As a result of this section, the information item will be contained in an XML container with the following parts:

- Identification
- Bibliographical metadata
- Preservation metadata
- Viewer info
- Original representation of content (the bitstream)
- Alternative representations from conversion (zero or more bitstreams)

5 Costs and the Way of Operation of the Electronic Archive

A university is a semi-public organisation that has to control its costs. In the Netherlands, there is no legal prescription that an academic community should build an archive of its electronic scientific output. By the way, there is not even a legal deposit of Dutch publications; the deposit of Dutch publications has a voluntary basis.

Formally, there is no budget for the electronic archive. Subsidies helped to start the project, but the exploitation of the operational electronic archive should ideally be cost-neutral or, if the electronic archive replaces traditional functions, cost no more than that traditional function. Therefore it is necessary to design the electronic archive in such a way that the operational cost will be acceptable and to try and estimate the cost factors of an operational electronic archive.

Not much is known of the cost of operation of electronic archives. We found in September 2001 nearly a hundred recent scientific studies on this subject, of which 17 ones had the keyword 'cost effective'. But there are hardly any comparative studies or hard criteria for cost effectiveness, and so it is difficult to find a foundation for an economically acceptable electronic archive. The RLG report of August 2001 (6), at the moment of writing this article the most recent broad overview, states that 'not a great deal is known about the costs of preserving complex digital objects over time, there is an accepted wisdom in the library community that digital preservation will require ongoing resource commitments - potentially more than for traditional materials, but certainly different. Traditional and digital preservation should be compared with some caution, because the complex dependencies between long-term maintenance and continuing access make comparison problematic.' The report points out that preserving digital materials will require resource commitments over time, and that digital preservation is also likely to draw on resources longer than traditional preservation does, and it may be the case that different technical strategies (e.g., different types of migration or emulation) will prescribe quite different costing timeframes and schedules.

So next to the technical problem there is an economic problem: how to design a cost effective, economically acceptable electronic archive? And what will those costs be?

Ideally, the identified group of potential users, in OAIS terms the designated community, should state their requirements for the electronic archive, but they are still hardly aware of the problem, let alone the possibilities and the consequences.

5.1 Metadata assignment

In an electronic archive metadata are a major cost factor for two reasons. Firstly, the cataloguing process of a library document takes a lot of time from expensive employees. Cataloguing electronic documents costs more than paper documents. Secondly, updating metadata files is an important cost factor.

From the viewpoint of cost control, the list of metadata should be kept limited. Each metadata element in the list is expensive because it may or must be assigned to every document, and it must be maintained. Maintenance seems sometimes to be overlooked or underestimated. The design choices are the length of the metadata list, the quality requirements, the granularity and the amount of work that can be outsourced to the submitter of the SIP (Submission Information Package). Here is an element of requirements engineering. The more the information items in the archive will be used, the more the expenses on the assignment of metadata are justified, that saves on search efforts for wanted information items; it is a well-known library trade-off. Anyway, agreement on a restricted list of metadata has to be reached.

The authors of the electronic information items are members of the community that builds the archive. So submitters could share the responsibility for the assignment of metadata. It is an application of the closed loop principle: 'if a user wants that his/her scientific output can ever be found and reused, (s)he shares the responsibility to enable that'. For the library it surely is cost-effective to insert user capacity. It is estimated that the personnel costs for the library to assign metadata is about 10 per information item. That is in line with ref. 13. It is easy to spend far more on it. The metadata item is probably the most cost-sensible item in the design.

Of course, the user-assigned metadata will be controlled by the archive; technical metadata and some others will be added by the archive. As a basis for the list of metadata the Dublin Core list is taken (14). If the electronic archive collection is regarded as a part of the national bibliography (extended use), elements of the Biblink DC list come into consideration (15).

5.2 Administration and quality control

NEDLIB report number 6 (5), the process model based on OAIS, gives a high-level overview of the tasks of processing SIP's and AIP's. The corresponding steps have been described in the sections on registration, verification, storage handling and preservation. Although no figures on costs for these activities have been given, the description of archival storage and data management tasks is a starting point for our design.

Moreover, for electronic archives of scientific communities, the SIP's must be controlled on correctness, completeness and authenticity. For an information item that has been delivered by e-mail (as for most items will be the case) a checksum calculation must be performed. A checksum is a count of the number of bits in a transmission unit that is included with the unit so that the receiver can check to see whether the same number of bits arrived. If the counts match, it is assumed that the complete information item was received.

It must be controlled whether all figures, references, footnotes and so on have been received. This cannot easily be automated. Many articles coming from authors are incomplete or over-complete, or even both: an example being an article with six figures, accompanied by five figures as separate electronic files and ten captions for the figures: the captions occurring twice.

A third matter is authenticity. The authenticity requires a procedure to assert that only valuable information items are offered to the electronic archive; harvesting URL's from web pages of scientists is less attractive by far.

Processing SIP's will cost about 10 per information item.

5.3 Technical infrastructure

An electronic archive needs equipment, such as servers, workstations for library employees, network capacity, storage media and printers. For a number of 5,000 items per year we recommend to assign 6 PC's with a network card and AV facilities to the archive (costs about 1500 each), as well as a professional server (about 5000) and a back-up storage facility.

The electronic archive will bring costs for storage media, too. These costs have been decreasing for years, due to technical improvements and the mass use of these media. Optical disks are needed with sufficient capacity to store the 5,000 documents (including original file, converted files, metadata files) per year. Generally it is expected, that the price reduction of media and other hardware will continue to decrease. Optical disk storage costs about 3 per GB these days. These costs are marginal.

The total hardware costs including 'everything' is estimated to be k 32; as the equipment will be written off and renewed in four years, the costs are k 8 per year (a).

Software is more expensive. Elementary document management functionality demands installation and configuring. Licenses for Operating systems etc. and viewers for the e-archive equipment are needed. For packages for which the number of workstations is a main factor in the licensing strategy, one must try and reduce the number of workstations where the system is accessible. We estimate the costs to be k 15 per year (b). If public domain software will be used, the costs will be slightly less.

As indicated in the section on metadata, reader support should be provided, a help screen and an elementary e-learning environment. The development costs of it are outside the scope of this aspect: it may be subsidised via a project structure. This maintenance does not cost much: about k 2 per year (c).

For the technical support for the equipment of the electronic archive only we reserve 0.2 full-time equivalent, one employee gets assigned the task to service it for one day per week. The costs are k 9 per year (d).

Preservation requires data refreshment. It is suggested to refresh the data every 5 years. The cost of this is 1 per MB, and if it is assumed that conversion is executed once every 5 years, the DIP's are kept for 20 years and an average DIP is about 500 kB, the costs would be about 2 per information item, i.e. k 10 per year for all information items (e). It is not clear whether future mechanisation will bring these costs down, so we stay with this figure.

Adding up (a) – (e): for the infrastructural costs including technical support for the infrastructure of an electronic archive in which 5,000 information items per year are stored of 500 kB each for 20 years costs about k 44 per year, and as every year 5,000 information items are added to the electronic archive it costs 9 per information item. Mark that 'per year' is not mentioned in the last figure. These are the total costs for 20 years. This figure of course may further vary with quality requirements, configuration details, the wanted level of support, the amount of licences in-house and the way of cost accounting.

6 Costs Estimates

The addition of the last three paragraphs yields the result, that for metadata assignment by the library plus administration and quality control plus the infrastructure of the operational electronic archive the estimated costs are now estimated to be 10 + 10 + 9 = 29 plus extras per information item.

Note 1: if the information items are kept for 50 years instead of 20 years, these costs under (e) are not 2 but 5 per item; if the average DIP would be 1 MB the refreshment costs are 4 for 20 years. But if the average DIP is only 200 kB the costs are less (to give an indication: this article without metadata needs less than 100 kB; a PhD thesis may need 20 MB). Moreover, in these cases the size of the infrastructure varies too, as do the costs under (a) – (d). For the worst case that each information item needs 1 MB and is stored 50 years we add 10% for all other costs. In that case, the total costs according to this calculation are 40 per information item. This figure gives some idea on the elasticity of the costs.

Note 2: what happens if a viewer becomes obsolete? In that case mass-conversion of a bitstreams to a new format is foreseen. These costs are extra. They might be in the order of magnitude of 5 to 10 per information item but they cannot be estimated reliably because of lack of data and lack of evidence how often this would happen.

Note 3: With this model, one can calculate start-up costs of a controlled growth path, starting with 500 items in year 1, 1500 in year 2, and 5000 in year 3. When complete collections of journal articles are stored a new situation arises, but again this analysis can be used as a basis.

Note 4: These are the costs to control the electronic document flow. The costs of, for instance, a project manager or research activities are not included. This type of costs depends on priorities of the library.

7 Way of Operation of the Electronic Archive

The analysis of a cost-effective electronic archive leads to the following way of operation or 'business models' for the electronic archive.

1. The user will compulsory have a role in assigning metadata to the electronic information items (s)he delivers. Due to that compulsory character, facilitation is needed, too. The facilitation requires far more than a few examples in a help screen; one could on the contrary think of a concise e-learning environment.

2. The format of the SIP's is a ground for selection. Although the electronic archive will accept all formats, but not all formats are supported. Maintenance of information systems is often performed on a basis of Service Level Agreement (SLA). The idea of a SLA applies to the electronic archive. Full service is guaranteed for SIP's that have been submitted in one of the formats that have been defined by the electronic archive beforehand. Among others, the experiences with keeping viewers running over time will decide what the library will offer in its SLA. SIP's with other formats will be accepted and saved, but the library does not guarantee maintenance or viewer disposal.

3. Selection of items could be performed on a scientific ground. Not all the publications of an academic community have the same historical value. This helps to bring the flood of items down. The annual scientific report of Delft University of Technology lists over 10,000 items. Will the top 20% contain 80% of the historical value; will the top 50% of items contain 99% of the historical value? The study from a Portuguese university and the national library mentions 'historic interest' as a reason for selection, and begs the question (16). In the Cedars project, selection for preservation will also need to be closely tied to the long-term research or cultural interests of the organisation (17). In several projects it has been stated that the future use of electronic information items is uncommonly difficult to predict. Perhaps the archive should not be too defensive and currently aim at a large part of the scientific items. But anyway, norms for acquisition and acceptance for a first selection must be drawn up.

4. A second appraisal is foreseen. Old documents are used far less than recent documents and as old electronic archives will brings even higher costs than a paper archive. In the Netherlands, university libraries must save their collection. What does that mean? Do all archived materials belong to that collection?

A comparison may be made to the record-keeping policy of the State of the Netherlands. The State keeps 'all' records. But in the National Archive, only a small percentage of all records that could be carried through time will be kept. The State *cannot and does not want to* keep all records infinitely (congress of (18), our italics). A second selection procedure takes place after a certain time interval (now 20 years). In our case, 20 years may be long enough to predict what scientific output is worth keeping forever, or at least at that time is easier than immediately after publication. The process of removal and keeping must be executed according to archival rules. So, further study may lead to rules for a second appraisal for the information items in the electronic archive.
As a corollary, it will be interesting to see what publishers do with respect to longevity of electronic publications.
5. The collections in the electronic archives can be cosidered as a part of the Dutch national scientific output. The possible coupling with other national collections, such as DSEP, will be studied. This starts from a compliant identification number, like the Dutch National Bibliographic Number (NBN). The use of the NBN is described in (19). The practicalities of this will be investigated further.

8 Current Situation

The Libraries of Maastricht University, Utrecht University and Delft University started analysing scientific information items for a requirements analysis of the electronic archive. Based upon current electronic information items, lists for preservation metadata and the content metadata have been drawn up. The global organisation and architecture of data management, access and storage have been drawn up. They will be published elsewhere. Use has been made of the experiences with Utrecht's electronic archive for Ph.D. theses (20), too.

9 Conclusion

For an electronic archive the technical problems are not the most difficult ones. In the case of an electronic archive for the scientific community the AIP's will be stored in XML containers and conversion will take place if necessary. A bigger problem is its economical viability. Because of their size, the costs prove to influence the way of working (the business model) of the electronic archive. Thus, they give rise to design criteria. In this article, a restricted metadata list and involvement of the users metadata assignment, a SLA for the formats the electronic archive with guaranteed high level of service, and a primary selection based on scientific and historical value of scientific information items of the academic communities are discussed. Eventually, for instance after 20 years, a second appraisal is necessary. Still, in this model the costs of control of the electronic document flow (assignment of metadata, administration and quality control, and equipment of the electronic archive) are estimated to be as much as 29 per information item. Further experiments with the prototype of the electronic archive, as well as requirements engineering with the designated community, will give a basis for high-level decisions on methods and strategies for electronic archives and preservation of digital sources.

References

1. http://www.roquade.nl/ (Last accessed November 2001)
2. http://www.webdav.org (Last accessed October 2001)
3. Whitehead, E.J., Wiggins, M.: WebDAV: IETF standard for collaborative authoring on the web. IEEE Internet computing, September - October (1998) 34-40
4. van der Werf, T.: The deposit system for electronic publications. A process model. NEDLIB report series, report 6. NEDLIB consortium (2000). http://www.kb.nl/nedlib/ (Last accessed November 2001)
5. Consultative Committee on Space Data Systems, Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-R-2: Red Book. Issue 2. (June 2001). www.ccsds.org/RP9905/RP9905.html (Last accessed November 2001)
6. Attributes of a trusted digital repository. Meeting the needs of research resources. An RLG-OCLC report. Draft for public comment. Research Libraries Group, Mountain View (August 2001)
7. http://www.leeds.ac.uk/cedars/testsites.htm (Last accessed November 2001)
8. The Cedars Project Team and UKOLN metadata for digital preservation. The Cedars Project outline specification. Draft for public consultation. The Cedars Project Team and UKOLN (March 2000) http://www.leeds.ac.uk/cedars/cedars.pdf (Last accessed November 2001)
9. Russell, K., Weinberger, E.: Cost elements of digital preservation http://www.leeds.ac.uk/cedars/documents/CIW01r.html (Last accessed November 2001)
10. http://xxx.lanl/gov (Last accessed January 2002)
11. http://www.library.tudelft.nl/e-archive/Projectbeschrijving/Doel_en_scope/doel_en_scope.html (Last accessed November 2001). (In Dutch)
12. http://www.phys.uu.nl/~durr/EarchiveSite/publications.html (Last accessed November 2001)
13. Puglia, S.: The costs of digital image projects http://www.rlg.org/preserv/diginews/diginews3-5.html (Last accessed October 2001).
14. http://dublincore.org/ (Last accessed October 2001)
15. http://www.schemas-forum.org/registry/schemas/biblink/BC-schema.html (Last accessed October 2001).
16. Noronha, N., Campos, J.P., Gomes, D., Silva, M.J., Borbinha, J.: A deposit for digital collections. In: Constantopoulos, P., Sølvberg, I.T. (eds.): ECDL 2001. Lecture Notes in Computers Science, Vol. 2163. Springer Verlag, Berlin Heidelberg New York (2001) 200-212
17. http://www.leeds.ac.uk/cedars/documents/ABS01.htm (Last accessed November 2001).
18. Documenten uit de tijd. Behoud en beheer van digitale informatie. (Documents out of time. Retention and maintenance of digital information). Eindrapport MLG project fase 2A. 's-Gravenhage (1993). (In Dutch)
19. http://www.kb.nl/coop/donor/rapporten/URI.html (Last accessed October 2001)
20. http://www.library.uu.nl/digiarchief/ (Last accessed October 2001)