An electronic archive for academic communities

R. Dekker[1], E. Dürr[2], M. Slabbertje[3] en 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
durr@fys.ruu.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

1. Introduction, goals and purposes

Three Dutch scientific communities started the Roquade project: the Library of Utrecht University, the Library of Delft University of Technology and the Netherlands Institute for Scientific Information Services (niwi) (1). The Roquade project aims at enhancing scientific communication in the interest of the academic community. The control of most of its scientific output was outsourced for ages to specialised craft: publishers. But the document management tools that are ubiquitous today, and the fact that the scientific author and the scientific user are members of academic communities, if not colleagues in a virtual project, pressures universities to look after their scientific information. Most likely, every academic community will eventually organise the management of its scientific output via the library. The library is the organisation to control written knowledge in any form: journal articles, preprints, scientific reports, scientific presentations, contributions to scientific discussions and others; and it will have the role of facilitator and stimulator of related developments. The time path for this development in the library is not clear, but it could go fast. As 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.

Electronic archives
The Roquade project involves the setting-up of an infrastructure for organising, co-ordinating, supporting and facilitating the digital publishing process, including an electronic archiving facility. The libraries of Delft, Utrecht and Maastricht Universities participate in the archive project. The experimental university archives will be developed according to one architecture. The collections or parts of them could be virtually merged. Most of the technical problems could be solved. As an example: for virtual merging of collections of items on several servers, there is a standard that is related to http, the Hypertext Transmission Protocol: the collaborative

authoring protocol WebDAV (Web Distributed Authoring and Versioning) (2, 3). The use of WebDAV is relatively simple; the software costs for WebDAV are marginal. The experiments concentrate on the following subjects (4):

- Methods and strategies for preservation of scientific digital sources
- Possibilities to executing different these 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
- Quantities and types of scientific information items that should be kept in an electronic archive of a university

Comparable projects: NEDLIB and Cedars

The responsibility of university libraries differs from national libraries. National libraries have collections of static electronic publications; the electronic publications are in their definitive forms; they must be preserved forever. The authenticity of the electronic documents in the archive is very important. By contrast: a university organises a digital archive of its own scientific information items with the purpose of accessibility and reusability, in a setting of reuse for education and research goals. Keeping these information items over time should be as reliable as possible, but the archives are meant to be working archives, to be used for filtering, for knowledge structuring, more for informational value than for evidential value. The difference in set-up between the two types of archives can also be shown from a description of two well-known and well-documented projects: the NEDLIB project and the Cedars project.

The NEDLIB project, led by the Royal Library of the Netherlands, involved nine European national libraries. It focused on needs of national libraries as they extend their traditional role to take responsibility for the preservation of digital publications and to establish a Deposit System for Electronic Publications (DSEP). A process model was developed to incorporate DSEP functions into library practices for the handling of digital materials (5). The Open Archive Information System (OAIS) model was adopted as a basis for this process model (6). According to ref. 7, 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 Cedars project originally incorporated three universities in the UK: the universities of Oxford, Cambridge, and Leeds (8). Cedars developed a project demonstrator archive. The Cedars Project has included a number of test sites. The purpose is, to test different aspects of digital archiving for these academic communities with various types of digital materials. The 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. Their websites show that the present Cedars test sites have slightly different goals than a national library: the goals of digital resource preservation but also conservation policy issues seen from a university library, and awareness of developments are mentioned.

XML containers, viewers, and conversion or emulation

The first problem is the use of XML and the choice between emulation and conversion. This is a technical problem.
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. 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 it is outside the scope of this project.
For the time being, information items can be described DTD-free, eventually DTD's or XML Schema's will be written.
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 (viewers) become class methods of these groups of objects.
In this way, the preservation issue 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 have no interest in ensuring long-term availability of their programs. They 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 existence of certain viewer software on the client side. An option is to archive an item and reproduce on request later only the bit stream "as is,,, without viewer support. We call that the "null viewer,, approach. In that case the responsibility of maintaining the programs to handle the item content is shifted to the user. But normally, the library will want to 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. Whether or not the library supports a certain view by keeping the viewer running for over time, or refers to external dealers (museums) with old software and hardware, depends on tactic and strategic decisions from the library management. Cost aspects are among the factors that play a role in this decision.
Furthermore, the strategies of conversion and emulation must be compared. 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. That is conversion. For a computer scientist, the results of both strategies are equal. With emulation the conversion is executed at the time of a request ("on the fly,,) and with conversion the intermediate resuilts 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. A more detailed description is given in ref. 9. As for authenticity, 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. Of course, this condition

applies to emulation as well. So, the difference between emulation and conversion is small.
The choice between emulation and conversion may depend upon costs and other 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 strategy based on the estimated future frequency of use of an information item may be decisive. As a consequence, conversion must be enabled. The information item container must be able to contain alternate representations. This means, that the container has the following parts:

- Identification
- Bibliographical metadata
- Preservation metadata
- Viewer info
- Original representation of content (bitstream)
- Alternative representations

Role of costs

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. 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 help to start a 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 the former equivalent. Therefore it is necessary to design it 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, of which in 17 ones the term ‘cost effective’ was used. 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.
This problem has been mentioned before. As a part of the Cedars, a framework for costs has been developed (10). Both the British Library and Birmingham University Library still state that they want to investigate funding for electronic archives (8).
The RLG report of August 2001 (7), 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 then the costs be? Wish that the identified group of potential users, in OAIS terms the designated community, should state their requirements for the electronic archive, but they are hardly aware of the problem, the possibilities and the consequences.

Metadata assignment

It may be surprising that in an electronic archive metadata will be a major cost factor. Nevertheless they are, 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. Moreover, it must be maintained. This cost factor seems sometimes to be overlooked or estimated. The design choices in this field are the length of the metadata list, the quality requirements, the granularity and the amount of work that can be outsourced to the sender of the SIP (Submission Information Package). There is an element on requirements engineering in it. 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.

The authors of the electronic information items are members of the community that builds the archive. That fact makes it feasible to let the users assist in 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'. This principle has been known in electronic record management for several years. For the library it surely is cost-effective to insert user capacity. It is estimated that the personnel costs of assigning metadata is about € 10. That is is line with ref. 11. 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 have to be controlled by the archive and a aprt of the metadata, for instance technical metadata will have to be added by the archive. As a basis for the list of metadata the Dublin Core list is taken (12).

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 (13).

Administration and quality control

NEDLIB report number 6 (5), the process model, gives a high-level overview of the task fields 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 electronic archive.

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 and form part of the bitstream, too. It is not likely that this requirement can easily be automated. Many articles coming from authors are incomplete or overcomplete, 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 have been added twice.
A third matter is the authenticity. The authenticity requires a procedure to assert that only valuable information items are offered to the elctronic archive.
It is estimated that processing SIP's will cost about € 10 per information item.
In the case of conversion of a bitstream to a new format AIP's are processed. The costs of mass-conversion cannot be estimated because of lack of data; moreover, it is not sure whether conversion will be done and how often. This factor is treated P.M.

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 an professional server (about Euro 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 Euro 3 to 5 per GB these days. These costs are marginal.
The total hardware costs including 'everything' is estimated to be k Euro 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, user 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. Maintenance cost of this user support is very limited: about k€ 2 per year (c).
For the technical support for the equipment of the electronic achive 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 DIPs are kept for 20 years and an average DIP is about 500 kB, the costs would be about € 2 per information item, i.e. kEuro 10 per year for all information items (e). It is not clear whether ongoing robotisation will bring these costs down, so we do not reckon on it.

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 kEuro 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.

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 € 10 + Euro 10 + Euro 9 = Euro 29 per information item.
Note 1: if the information items are kept for 50 years instead of 20 years, these costs under (e) are not Euro 2 but Euro 5 per item; if the average DIP would be 1 MB the refreshment costs are Euro 4 for 20 years. But if the average DIP is only 200 kB the costs are less (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 Euro 40 per information item. This figure gives some idea on the elasticity of the costs.
When complete collections of journal articles are stored a complete new situation arises. For that, a new analysis of all elements mentioned has to be made.
Note 2: In this calculation the costs of a project manager or a head of the archive have not been included.

Models of the electronic archive

The analysis of a cost-effective electronic archive leads to the following business models.
First of all, 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 even think of a concise e-learning environment.
Secondly, the format of the SIP's is a ground for selection. Although the electronic archive will accept all formats, this does not mean that all formats are supported. Maintenance of information systems is often performed on a basis of Service Level Agreement (SLA). The SLA describes a level of support in the case of an emergency. The idea of a SLA applies to the situation of an 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. SIP's with other formats will be accepted and saved, but the library does not guarantee maintenance or viewer disposal.
Thirdly, 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, the top 50% of items 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 (14). 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 (15). 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.
Fourth, as announced by the foregoing point: 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. This leads to a discussion on what that means: do all archived materials belong to that collection?
A comparison may be made to the archival materials and 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 (16), our italics). A second selection procedure takes place after a certain time interval from its creation (now 20 years). In the Roquade case, 20 years may be sufficient long 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 the behaviour of publishers with respect to long-lasting electronic publications.
Five: in a way, the collections in the electronic archives could be regarded as a part of the Dutch national scientific output. The possible coupling with other national collections, such as DSEP, will be studied. The simplest solution is based upon the use of a compliant identification number, like the Dutch National Bibliographic Number (NBN). The use of the NBN is described in (17). The practicalities of this have to be investigated further.

Current situation

Maastricht University Library, Utrecht University Library and the Library of 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. Use has been made, too, of the experiences with Utrecht's electronic archive for Ph.D. theses (18).

Conclusion

In the Roquade project, in which three academic communities aim at enhancing the he scientific communication of academic communities, a project has been developed for an electronic archive. In this electronic archive, the AIP's will be stored in XML containers. The electronic archive will have to be economically acceptable. Because of this, the list of metadata will be restricted and the users will compulsory be involved in metadata assignment. Not for all formats the electronic archive will guarantee a high level of service. An updatable list of formats for which a high level of service is given is an attractive alternative. For the archive, it is opted for a first

appraisal selection based on permanent value of scientific information items of the academic communities. Eventually, for instance after 20 years, a second appraisal is foreseen. Under these circumstances, a model is presented in which the operational costs of the electronic archive are € 29 per information item. Due to the economic boundary conditions, the design desiderata and the business model of the electronic archive are subject to requirements engineering. Further experiments with the prototype of the electronic archive will give a basis for improved strategies and cost estimates before the complete electronic archive is rolled out.

References

1. http://www.roquade.nl/ (Last accessed November 2001).
2. http://www.webdav.org (Last accessed October 2001).
3. E.J. Whitehead and M. Wiggins: WebDAV: IETF standard for collaborative authoring on the web. IEEE Internet computing, September - October 1998, 34-40.
4. http://www.library.tudelft.nl/e-archive/Projectbeschrijving/Doel_en_scope/doel_en_scope.html (last accessed November 2001). (In Dutch).
5. T. van der Werf: The deposit system for electronic publications. A process model. NEDLIB report series, report 6. NEDLIB consortium, 2000. Also available from http://www.kb.nl/nedlib/ (last accessed November 2001).
6. 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).
7. 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.
8. http://www.leeds.ac.uk/cedars/testsites.htm (last accessed November 2001).
9. http://www.phys.uu.nl/~durr/EarchiveSite/publications.html (last accessed November 2001).
10. K. Russell and E. Weinberger: Cost elements of digital preservation. http://www.leeds.ac.uk/cedars/documents/CIW01r.html (Last accessed November 2001).
11. S. Puglia: The costs of digital image projects. http://www.rlg.org/preserv/diginews/diginews3-5.html (Last accessed October 2001).
12. http://dublincore.org/ (last accessed October 2001).
13. http://www.schemas-forum.org/registry/schemas/biblink/BC-schema.html (last accessed October 2001).
14. N. Noronha, J.P. Campos, D. Gomes, M.J. Silva and J. Borbinha: A deposit for digital collections. In: P. Constantopoulos and I.T. Sølvberg (Eds.): ECDL 2001. Lecture Notes in Computers Science 2163, (2001), Springer Verlag, Berlin. Pp. 200-212.
15. http://www.leeds.ac.uk/cedars/documents/ABS01.htm (Last accessed November 2001).
16. 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).

17. http://www.kb.nl/coop/donor/rapporten/URI.html (Last accessed October 2001).
18. http://www.library.uu.nl/digiarchief/ (Last accessed October 2001).