CEDARS: Digital Preservation and Metadata

Michael Day
UKOLN: The UK Office for Library and Information Networking,
University of Bath, Bath, BA2 7AY, United Kingdom
http://www.ukoln.ac.uk/
m.day@ukoln.ac.uk

Abstract

CEDARS (CURL Exemplars in Digital ARchiveS) is a UK digital preservation project funded by JISC through eLib. Lead sites in the project are the Universities of Cambridge, Leeds and Oxford. The project aims to promote awareness of the importance of digital preservation, to produce strategic frameworks for digital collection management policies and to promote methods appropriate for long-term preservation. An important strand of CEDARS will concern metadata. Metadata could be used as a means of recording migration and emulation strategies, ensuring the authenticity of digital objects, noting rights management and collection management issues and will also be used for resource description and discovery.

1. CURL Exemplars in Digital ARchiveS (CEDARS)

1.1. Background

University and research libraries have, in recent years, given their users increased access to digital information resources. Some of these form part of their physical collection, e.g. databases on CD-ROM, while others are provided via computer networks and are made available on different commercial terms [1]. At the present time there is no formal mechanism for ensuring that digital resources are preserved for long term use. Indeed in many countries, including the United Kingdom, there is still no formal legal deposit for digital publications [2].

The Joint Information Systems Committee (JISC) of the UK higher education funding councils funds the Electronic Libraries (eLib) Programme. The eLib Programme was set-up in response to a report published in 1993 by a Libraries Review Group appointed by the funding councils [3]. JISC were aware that digital preservation would have an important role in the eventual success (or otherwise) of the eLib Programme. Accordingly, JISC and the British Library co-sponsored a workshop on the "Long Term Preservation of Electronic Materials" which was held at Warwick University in November 1995 [4]. One outcome of this workshop was that JISC, in conjunction with the National Preservation Office (NPO), agreed to fund a programme of studies which would be administered by the British Library Research and Innovation Centre (BLRIC). These JISC/NPO studies covered several distinct areas:

- An analysis of the US Task Force on the Archiving of Digital Information report [5]
- A framework of data types and formats [6]
- Who should be responsible for preservation and access? [7]
- The *post hoc* rescue (data archaeology) of high-value digital material [8]
- The preservation requirements of universities and research funding bodies [9]
- Guidelines for digital preservation [10]
- Comparison of methods of digital preservation [11]

As part of eLib Phase 3, JISC decided to fund a project that would be able to investigate some of the practical issues of digital preservation. The Consortium of University Research Libraries (CURL) is a consortium of research libraries in the British Isles whose mission is "to promote, maintain and improve library resources for research in universities." Digital preservation is a key issue for all CURL members. CURL accordingly submitted a research proposal as part of eLib Phase 3. The result of this is the CURL Exemplars in Digital ARchiveS (CEDARS) project, funded by JISC through the CURL libraries. The project started in April 1998, and will run for three years. The lead sites in the project are the Universities of Cambridge, Leeds and Oxford. UKOLN has some involvement with the parts of the project relating to metadata. Other collaborating institutions include the Arts and Humanities Data Service (AHDS), the British Library, the Data Archive, the NPO and the Research Libraries Group (RLG).

1.2. Objectives

The project aims to investigate strategies which will ensure that the digital information resources typically included in library collections may, with other non-digital objects, be preserved over the longer term. It order to achieve this aim the project plans to:

- Promote awareness about the importance of digital preservation, both amongst research libraries and their users, and amongst the data creating and data supplying communities upon which they depend.

- Identify, document and disseminate strategic frameworks within which individual libraries can develop collection management policies which are appropriate to their needs and which can guide the necessary decision-making to safeguard the long-term viability of any digital resources that are included in their collections.

- Investigate, document and promote methods appropriate to the long-term preservation of different classes of digital resources typically included in library collections, and to develop costed and scaleable models.

1.3. Scope of demonstrators

Several different types of digital resources will be included within the CEDARS project scope.

- Digitised primary resources
- Datasets
- Electronic journals
- Online databases
- Electronic ephemera - pre-prints, Web pages, subject gateways, etc.
- Digital resources where intellectual content is bound to structure, form and behaviour
- Metadata

The CEDARS project is interested in demonstrating the preservation of all material that is the traditional preserve of the research library. It, however, is not concerned with information in the form of sound or video. It is, additionally, only concerned in preserving the intellectual content of resources, not the physical objects upon which they are stored. Each of the lead sites will take responsibility for providing demonstrators for a particular 'flavour' of digital resource. Cambridge will deal with dynamic data, Oxford with primary resources while Leeds will look at digital resources where intellectual content is bound to structure, form and behaviour (e.g. CD-ROMs). In addition, the three lead sites will lead working groups on those related issues that had been identified as most important: Cambridge on rights management, Oxford on metadata, and Leeds on the use of emulation as a preservation strategy.

1.4. Deliverables

Key deliverables of the CEDARS project include the production of:

- Guidelines for developing collection management policies which will ensure the long-term viability of any digital resources included in the collection

- Demonstrator projects to test and promote the technical and organisational feasibility of a chosen strategy for digital preservation

- Methodological guidelines developed by the demonstrator projects providing guidance about how to preserve different classes of digital resources

- Clearly articulated preferences about data formats, content models and compression techniques which are most readily and cost-effectively preserved

- Publications of benefit to the whole higher education community

One of these publications will be a study of digital preservation metadata.

2. Metadata and digital preservation

Discussions of metadata in the library community have largely centred on issues of resource description and discovery [12]. There is, however, a growing awareness that metadata has an important role in digital resource management, including preservation. Accordingly, in May 1997 the Research Libraries Group constituted a Working Group on the Preservation Issues of Metadata. The aim of this working group is to ensure that information essential to the continued use of digital resources is captured and preserved in an accessible form. A preliminary report has been produced which identifies 16 preservation metadata elements and provides a semantic framework for this [13].

The CEDARS project also recognised from an early stage that metadata issues would be important. A working group has been formed to cover metadata. At this preliminary stage of the project it is difficult to predict what particular recommendations this working group will produce but interest is likely to be shown in the following issues.

2.1. Metadata for emulation and migration

The core technical problems of digital preservation relate to inadequate media longevity, rapid hardware obsolescence and dependencies on particular software products. In this context it makes good sense to preserve the data itself, not the physical medium on which it happens to reside. There are several potential technical approaches to this problem. Jeff Rothenberg has suggested, for example, the building of software emulators that would mimic the behaviour of obsolete hardware and software [14]. This would involve encapsulating data together with the application software used to create it and a description of the required hardware environment. To facilitate future use, Rothenberg suggests attaching 'annotation metadata' to the surface of each encapsulation which would both "explain how to decode the obsolete records contained inside the encapsulation and to provide whatever contextual information is desired about these records" [15]. This surface metadata, which could also contain resource discovery information, would be kept in a standard 'bootstrap' format so that it could be converted to new formats as part of the preservation refresh cycle.

Another approach to digital preservation is the periodic migration of digital information from one generation of computer technology to a subsequent one [16]. Using migration, it is important to ensure that preserved documents are what the US National Historical Publications and Records Commission (NHPRC) funded University of Pittsburgh Electronic Records Project describe (in an archives context) as 'inviolate', 'coherent' and 'auditable' [17]. David Bearman defines 'coherent' as follows: "If records are migrated to new software environments, "content, structure and context information must be linked to software functionality that preserves their executable connections or representations of their relations must enable humans to reconstruct the relations that pertained in the original software environment" [18]. Successful migration strategies will, therefore, depend upon metadata being created to record the migration history of a digital object and to record contextual information so that a future user can reconstruct (or understand) the technological environment in which a particular digital object was created.

2.2. Metadata for authentication

In addition to the technical problems of digital preservation, there will be a need to address problems of intellectual preservation [19]. For example, how will users know that the digital object that they retrieve is the one that they want? Again, how can one guard against unauthorised changes being made to the information content of digital objects?

A partial solution to this problem would be the general adoption of unique and persistent digital identifiers. This would mean the assignment of a new identifier each time a particular digital object is updated. Current initiatives include the Uniform Resource Name (URN) which is being developed for the Internet community by working groups of the Internet Engineering Task Force [20] and the Digital Object Identifier (DOI), an initiative of the Association of American Publishers [21]. Legacy identifiers will also continue to be used for some of the digital objects within the CEDARS project scope, so - for example - some publishers will assign International Standard Book Numbers (ISBNs) to CD-ROMs or generate Serial Item and Contribution Numbers (SICIs) for online journal articles. On the other hand, other items in the project scope, electronic ephemera for example, are unlikely to have previously assigned persistent and unique identifiers.

An additional approach to ensuring the authenticity of a given digital object would be to use a simple cryptographic technique like the production of a validation key value or checksum for each resource in a digital archive. An authentication checksum could be computed from each resource in a digital archive and stored with the descriptive metadata. When a user, or the archive, wants to retrieve the resource at a later date this checksum could be computed again and compared with the checksum recorded in the metadata. If the two agree there can be confidence that the document retrieved is the one referred to by the descriptive metadata. This general approach has been adopted for use by the European Telematics for Libraries project BIBLINK [22].

Archivists and records managers have similar concerns with authenticity, integrity and preserving 'evidentiality'. The University of Pittsburgh Electronic Records Project, for example, has defined a metadata model for business-acceptable communications [23]. A University of British Columbia project has also worked on defining the requirements for preserving reliable and authentic electronic records [24].

2.3. Metadata for resource discovery.

Digital resources that have been physically preserved will also need to be retrievable. For this reason, preservation systems will have to interact with resource discovery systems. Recommendations on resource discovery formats (e.g. Dublin Core) or metadata frameworks (e.g., Resource Description Format) will constitute an important part of CEDARS work on metadata.

2.4. Metadata for rights management.

Solving rights management problems in a digital preservation context will be crucial to a practically based project like CEDARS. Within the project, different licensing arrangements will have to be made with relevant stakeholders. This rights management information can be stored as part of the descriptive metadata and this could be used to manage access to digital resources in the demonstrators.

2.5. Metadata for resource evaluation

Not all digital resources will be preserved and, indeed, not all digital resources will be worthy of long-term preservation. CEDARS is interested in helping to develop suitable collection management policies for research libraries. This work could build on work carried out on selection criteria for Internet subject gateways produced by the EU funded DESIRE project [25].

2.6. Metadata management

Another important issue is how this metadata will be generated and where it will be kept. Metadata could be stored either in a centralised or distributed database and linked to the original resource. Alternatively, metadata could also be embedded in or otherwise directly associated with the original resource. Different solutions might be possible for different types of metadata. Resource discovery and rights management metadata could form part of a searchable database, while metadata specifying the technical formats used, the migration strategies operated and a document's use history could be stored with the document itself. Over a long period of time, this metadata will grow in size and will itself have to be subject to migration and authentication strategies.

3. Conclusions

CEDARS is a project that aims to address strategic, methodological and practical issues relating to digital preservation. The project will include the development of demonstrators to check the technical and organisational feasibility of the chosen preservation strategies. One strand of the project will investigate metadata issues. A preliminary report will be made available later this year and a seminar convened.

4. References

[1] Day, M.W., Online serials: preservation issues. In *E-serials: publishers, libraries, users and standards*, ed. W. Jones. The Serials Librarian, 33. Binghamton, N.Y.: Haworth Press, 1998, 199-221.

[2] Hoare, P., *Legal deposit of non-print material: an international overview, September-October 1995*. British Library Research and Development Report, 6245. London: British Library Research and Development Department, 1996.

[3] Joint Funding Councils' Libraries Review Group, *A report for the Higher Education Funding Council for England, the Scottish Higher Education Funding Council, the Higher Education Funding Council for Wales and the Department of Education Northern Ireland* [the Follett Report]. Bristol: HEFCE, December 1993.
<URL:http://www.ukoln.ac.uk/services/papers/follett/report/>

[4] Fresco, M., *Long term preservation of electronic materials: a JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib) organised by UKOLN, 27th and 28th November 1995 at the University of Warwick*. British Library Research and Development Report, 6238. London: British Library Research and Development Department, 1996.
<URL:http://www.ukoln.ac.uk/services/papers/bl/rdr6238/>

[5] Matthews, G., Poulter, A. and Blagg, E., *Preservation of digital materials policy and strategy issues for the UK : report of a meeting held at the British Library Research and Innovation Centre, London, 13 December 1996*. British Library Research and Innovation Report, 41. London: British Library Research and Innovation Centre, 1997.
<URL:http://www.ukoln.ac.uk/services/papers/bl/blri041/digpres.html>

[6] Bennett, J.C., *A framework of data types and formats, and issues affecting the long-term preservation of digital material*. British Library Research and Innovation Report, 50. London: British Library Research and Innovation Centre, 1997.
<URL:http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html>

[7] Haynes, D., Streatfield, D., Jowett, T. and Blake, M., *Responsibility for digital archiving and long term access to digital data*. British Library Research and Innovation Report, 67. London: British Library Research and Innovation Centre, 1997.
<URL:http://www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital-preservation.html>

[8] Ross, S. and Gow, A., *Digital archaeology? Rescuing neglected or damaged digital collections*. British Library Research and Innovation Report, 108.

[9] Data Archive, *An Investigation into the digital preservation needs of universities and research funders*. British Library Research and Innovation Report, 109.

[10] Hendley, T., *Comparison of methods and costs of digital preservation*, British Library Research and Innovation Report, 106.

[11] Beagrie, N. and Greenstein, D., *Strategy for creating and preserving digital collections*. First public consultation and review draft. London: Arts and Humanities Data Service, 24 April 1998.
<URL:http://ahds.ac.uk/manage/framework.htm>

[12] Heery, R., Powell, A. and Day, M., *Metadata*. Library and Information Briefings, 75. London: South Bank University, Library Information Technology Centre, 1997.

[13] RLG Working Group on Preservation Issues of Metadata, *Preliminary report*. Mountain View, Calif.: Research Libraries Group, 7 January 1998.
<URL:http://www.rlg.org/preserv/presmeta.html>

[14] Rothenberg, J., Ensuring the longevity of digital documents. *Scientific American*, 272 (1), 1995, 24-29.

[15] Rothenberg, J., *Metadata to support data quality and longevity*. Proceedings of the 1st IEEE Metadata Conference, NOAA Complex, Silver Spring, Md., 16-18 April 1996.
<URL:http://www.computer.org/conferen/meta96/rothenberg_paper/ieee.data-quality.html>

[16] Task Force on the Archiving of Digital Information, *Preserving digital information: report of the Task Force on Archiving of Digital Information commissioned by the Commission on*

Preservation and Access and the Research Libraries Group. Washington, D.C.: Commission on Preservation and Access, 1996.
<URL:http://www.rlg.org/ArchTF/>

[17] Duff, W., Ensuring the preservation of reliable evidence: a research project funded by the NHPRC. *Archivaria*, 42, 1995, 28-45.

[18] Bearman, D., *Electronic evidence: strategies for managing records in contemporary organizations*. Pittsburgh, Penn.: Archives and Museum Informatics, 1994, p. 302.

[19] Graham, P.S., Long-term intellectual preservation. In *Digital imaging technology for preservation*, ed. N.E. Elkington, Mountain View, Calif.: Research Libraries Group, 1994, 41-57.

[20] Sollins, K. and Masinter, L., *Functional Requirements for Uniform Resource Names*. RFC 1737, 1994.
<URL:http://ds.internic.net/rfc/rfc1737.txt>

[21] Bide, M., *In search of the Unicorn: the Digital Object Identifier from a user perspective*, rev. ed. BNBRF Report 89. London: Book Industry Communication, 1998.
<URL:http://www.bic.org.uk/bic/unicorn2.pdf>

[22] BIBLINK project: <UTL:http://www.hosted.ukoln.ac.uk/biblink/>

[23] Bearman, D. and Sochats, K., Metadata requirements for evidence. Pittsburgh, Penn.: University of Pittsburgh, School of Information Science, 1996.
<URL:http://www.lis.pitt.edu/~nhprc/BACartic.html>

[24] Duranti, L. and MacNeil, H., The protection of the integrity of electronic records: an overview of the UBC-MAS Research Project. *Archivaria*, 42, 1995, 46-67.

[25] Hofman, P., Worsfold, E., Hiom, D., Day, M., and Oehler, A., *Specification for resource description methods: 2, Selection criteria for quality controlled information gateways*. DESIRE: Development of a European Service for Information on Research and Education, Deliverable 3.2 (2), May 1997.
<URL:http://www.ukoln.ac.uk/metadata/desire/quality/>

5. Acknowledgements

UKOLN is funded by the British Library Research and Innovation Centre, the Joint Information Services Committee of the UK Higher Education Funding councils, as well as by project funding from JISC's eLib Programme and the European Union. UKOLN also receives support from the University of Bath, where it is based.

For more information on the CEDARS project, please contact:

CEDARS Project Manager
Edward Boyle Library
University of Leeds
Leeds LS2 9JT

k.l.russell@leeds.ac.uk

<URL:http://www.curl.ac.uk/>