World Library and Information Congress:
69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 209-E

Meeting: 165. Information Technology and Preservation and Conservation
Workshop

Simultaneous Interpretation: -

Challenges to Digital Preservation and Building Digital Libraries

Seamus Ross
Director HATII and ERPANET[1], U.K

June of this year (2003) was the tenth anniversary of a workshop sponsored by the visionary, but now defunct, British Library Research and Development Department (UK), The British Academy, and The International Association for History and Computing. This event, which I designed and ran with Edward Higgs (then of the Public Record Office), was among, if not, the first international workshop on digital preservation hosted in the United Kingdom.[2] Awareness of the problem was in 1993, even among many of the participants, sketchy. It had been Stephanie Kenna and Brian Perry of the British Library and the historian Sir Keith Thomas, at the time President of the British Academy, who had taken the risk to fund the workshop on a then very little discussed, and little understood, problem.[3] Participants included Lynne Brindley, Peter Doorn, Daniel Greenstein, David Ryan, Kevin Schurer, Doran Swade, and Ron Zweig, all of whom have since played (and most of whom still are playing) a leading role in the area of digital preservation. About eighteen months earlier Charles Dollar, who was then at NARA (National Archives Records Administration) in Washington DC, in response to a request for guidance as to the available literature on digital preservation sent me package enclosing all the publications worth reading on the topic—there were not many.

---

[1] Seamus Ross, Director of Humanities Computing and Information Management at the University of Glasgow, runs HATII (Humanities Advanced Technology and Information Institute) [http://www.hatii.arts.gla.ac.uk], which he founded in 1997. He is Principal Director of ERPANET (Electronic Resource Preservation and Network) (IST-2001-32706) a European Commission activity to enhance the preservation of cultural heritage and scientific digital objects [http://www.erpanet.org] He is a lead partner in The Digital Culture Forum (DigiCULT Forum, IST-2001-34898), which works to improve the take-up of cutting edge research and technology by the cultural heritage sector in Europe [http://www.digicult.info]. email: s.ross@hatii.arts.gla.ac.uk

[2] The issues raised by the workshop were picked up by David Millward of *The Daily Telegraph* who later that summer published 'History is going down computer black hole' in *The Daily Telegraph*, on 2 August 1993.

[3] The British Library's Research and Development Department offered a substantial grant (RDD/C/160) to make it possible for the seminar to take place. With the additional assistance of a British Conference Grant from the British Academy the workshop met in London on the 25th and 26th of June 1993. Its results were published as Seamus Ross and Edward Higgs (eds.), *Electronic Information Resources and Historians: European Perspectives*, St Katharinen: Scripta Mercaturae, 1993 (and as British Library Report 6122).

A decade on the digital preservation bibliography now includes thousands of articles, items of grey literature, and project websites often rich with resources. The number of specialists talking about the problem runs into the hundreds. It was also fair to say that up to 1993 a 'certain narrow-mindedness ha[d]s pervaded studies of electronic information as the focus ha[d]s been predominately by national archives on the preservation of records about the national governments themselves'.[4] Librarians and other cultural specialists now recognise that access to and preservation of digital resources, whether the product of digitisation initiatives or born digital, are crucial activities whether they aim to preserve contemporary culture for posterity or to develop information resources to underpin digital library services.[5]

There appear to be as many definitions of digital libraries as there are institutions and individuals developing digital libraries and digital library services. For our purposes today, and borrowing from a definition prepared for the National Library of New Zealand (NLNZ) as part of a review of their digital preservation initiatives, we will describe a digital library as 'the infrastructure, policies and procedures, and organisational, political and economic mechanisms necessary to enable access to and preservation of digital content.'[6] As the report goes on to argue, in some instances a digital library may be a new entity, but in other cases it will be the electronic or digital face of a traditional library or information holding entity such as archives. Digital library activities will be embedded within current and evolving library service structures, although in some emerging models digital libraries may provide the raw materials that a wide range of other digital library service providers could package as part of a variety of aggregate information services meeting the needs of different user communities. Worldwide there are numerous digital library experiments both within commercial organisations, public and private information providers, and national, regional and university libraries and archives.[7] Some are services provided through many libraries, others subscription services[8], and still others are the digital resource face of traditional libraries. For some the variation may lie in the types of content which they manage and deliver. For example, some handle documents, others audio, others moving image resources and still others engineering, scientific, or social science data sets. For some the resources may be homogeneous but for others the content they hold and supply may be of a heterogeneous in nature. In either instance types of institutions need to handle content held in a variety of representations, such as websites, databases, or packaged digital objects, and reflecting different modes of creation with some coming in as digitised representations of analogue materials and others as borne digital resources.

What is lacking though is general agreement as to what a digital library is? But our understanding of the possibilities and the types of user communities and their needs is evolving. It may, therefore, be no bad thing that currently the term digital library is a flexible concept that is moulded in a variety of ways by content users, providers, and

---

[4] Seamus Ross, 'Historians, Machine-Readable Information, and the Past's Future, in Seamus Ross and Edward Higgs (eds.), 1-20.
[5] Spanish Presidency Resolution on Digital Preservation, Council Resolution of 25 June 2002 (2002/C162/02) *on preserving tomorrow's memory - preserving digital content for future generations*, http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/c_162/c_16220020706en00040005.pdf
[6] Seamus Ross, *Digital Library Development Review*, National Library of New Zealand, (Wellington, 2003), (ISBN Number: 0-477-02797-0), http://www.natlib.govt.nz/files/ross_report.pdf, 5.
[7] Projects run by these types of organisations and the National Science Digital Library are defining the expectations for digital library services, see for example Carl Lagoze, et.al., (2002), 'Core Services in the Architecture of the National Science Digital Library NSDL', *JCDL'02*, (Portland Oregon) July 13-17 2002, 201-209.
[8] For example, IEEE Computer Society Digital Library, http://www.computer.org/publications/dlib/

owners. It is evident that our expectations of and the ways we may use digital libraries will continue to evolve in terms of types of content, services, and even the kinds of organisations that will act as digital library providers.[9] They are also the subject of substantial research efforts.[10] Increasingly as institutions invest in developing digital libraries they come to recognise that the digital assets on which their library depends—their capital assets, so to speak—are fragile and require substantial curation effort if they are to remain accessible over the longer term. Even the viability of the digital library that holds them tends to be at risk. Digital repositories which lie at the heart of digital library developments have become an increasingly significant area of research; current design and implementation guidelines remain in their infancy.

The European Commission and the Swiss Federal Government recognising the risks faced by digital materials supported beginning in 2001 ERPANET (Electronic Resource Preservation and Access Network[11]). ERPANET works to enhance the preservation of cultural and scientific digital objects through awareness raising, improving practices, providing access to experience, research, and sharing policies and strategies. ERPANET's work is made possible not merely by the funding of the Commission and the Swiss Government, but also by the commitment of professionals from across Europe, Australia and New Zealand, and Canada and the USA who have given time, thought and effort to make its activities possible. Between our first seminar in June 2002 and August 2003 more than seventy colleagues from the public and commercial sectors have contributed to make the ERPANET seminars, which have been attended by nearly 500 participants, a success. Alongside its workshops, seminars, and content building activities ERPANET has been examining how data holding and creating organisations manage risk of information loss. During our initial twenty-one months the contributors to ERPANET have enabled us to identify standards and best practices that can improve the handling and long term curation of digital materials. Details of this work can be found at our website.[12] Research is needed, though, in many areas. These have been identified in a report prepared by members of the digital library community under the auspices of the European Commission
funded digital library network DELOS and the National Science Foundation (NSF) in the United States.[13] Digital curation, which encompasses the description, management, preservation, conservation, and delivery of digital objects, lies at the heart of all sustainable digital library service provision.

Despite recognition in the library, archive, and records management communities that the survival of digital information requires action[14], casual discussion with professionals

---

[9] DELOS, (2001), *Digital Libraries: Future Directions for a European Research Programme (Brainstorming Report)*, San Cassiano (Alta Badia), Italy, June 13-15, 2001, http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf . See also the details of the workshop at: http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/1st-ws.html
Knowledge Lost in Information, Report of the NSF Workshop on Research Directions for Digital Libraries, Chatham, MA, June 15-17, 2003, http://www.sis.pitt.edu/%7Edlwkshop/JISC/NSFreport.pdf

[10] Work under the European Commission's Sixth Framework Programme funded DELOS2 Network of Excellence (http://www.delos.info), which will start in January 2004, will help to advance thinking in this area, as will ongoing work at the National Archives of Australia, the Netherlands, the United Kingdom, and the USA. A look at the research agenda identified by DELOS will indicate the breadth of activity that is required if we are to address the digital preservation challenge.

[11] ERPANET, a European Commission funded activity (IST-2001-32706), is led by HATII (University of Glasgow), Schweizerisches Bundesarchiv, ISTBAL (Università di Urbino), and Nationaal Archief van Nederland. Details of the project can be found at http://www.erpanet.org. Current funding for ERPANET will run to November 2004.

[12] http://www.erpanet.org

[13] *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation* (2003), (http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf)

[14] One line of argument holds that without action to promote preservation all digital materials will be lost.

from these communities indicates that calls for action have so far not resulted in effective and commonly adopted digital curation and preservation strategies. In an effort to understand what organisations are doing to promote preservation of their digital materials, the team at ERPANET have been conducting case studies. These studies are helping us to:

- build a picture of digital preservation methods within the context of different institutional structures. These results will inform our thinking on good practice;
- accumulate and make accessible details of the experiences of different digital resource creating, managing, and using communities;
- identify issues which could benefit from new research;
- enable comparisons of the strategies and practices used by institutions from different sectors;
- provide sources of experience and methods to underpin our creation of guidance of preservation; and,
- create material for training seminars and workshops.

Organisational and sectoral requirements, awareness of digital preservation, availability of resources, and the nature of the digital object created, place unique and specific demands on organisations. In designing these case studies we have selected sectors to represent a wide scope of information production and digital preservation activity. ERPANET's first studies examined the pharmaceutical, broadcasting, publishing, and telecommunications sectors. We are attempting to balance our case studies to ensure a range of institutional types, sizes, and locations as well as selecting sectors and organisations that will be representative of different kinds of business activity, include organisations from a diversity of regulatory frameworks organisational cultures. The ERPANET interview instrument[15], which takes account of the strengths and weaknesses of instruments used by earlier projects to support their study of digital preservation practice, facilitates the exploration of three main areas:

- awareness of the issues surrounding digital preservation,
- the planning and implementation of digital preservation strategies, and
- the anticipated needs or opportunities.

To build as comprehensive and representative a picture as possible we are interviewing not merely archivists/records managers in our target organisations, but we are also interviewing information systems or technology managers, and business managers. This broader assessment of awareness and activity in organisations is providing us with detailed information about the extent of knowledge and practice in organisations, giving us an indication of where ownership for the problem lies, and offering us material to determine where digital preservation activity is likely to be promoted within organisations. In conducting our interviews we are examining:

- perception and awareness of risk associated with information loss;

[15] Available at the ERPANET Website., http:///www.erpanet.org

- how digital preservation affects the organisation;
- the actions organisations are taking to prevent data loss;
- how organisations monitor these activities; and,
- what mechanisms organisations have put in place to enable them to define their digital preservation needs.

Interviewees are asked to describe what they think the main difficulties associated with digital preservation are and what value information has in their organisation and the sector to which their organisations belong more generally. The risks associated with not preserving information and the justifications for preserving are becoming evident through our interviews. This instrument enables us to explore existing policies, strategies, and standards employed to tackle digital preservation concerns. We are, also, accumulating information about selection, preservation techniques, storage, access, and costs.

The results of the first set of case studies, which examined broadcasting, pharmaceuticals, publishing, and telecommunications involved twenty organizations and roughly fifty staff. The results will appear in a forthcoming paper. [16] Here I shall only summarise the findings that appear to be of direct value to the library community. Some conclusions seem very obvious but are not (widely) documented and others are surprising. Analysis of the interviews has allowed us to draw the following general conclusions against which the subsequent and more detailed initiatives need to be planned:

- Organisations appear already to have substantial quantities of digital information to handle;
- the categories of digital objects in use within organisations themselves varies more between sectors (e.g. broadcasting, engineering) than it does across organisations within a particular sector;
- organisations retain information for different reasons; and
- there is no approach to preservation that has been broadly adopted. This may be explained by a general lack of agreement as to which are the most effective approaches to preservation and a limited level of understanding of the risks and preservation challenges.

In Europe pharmaceuticals and broadcasting organisations are among the most highly preservation aware and broadcasting professionals displayed the broadest knowledge of the issues. External regulation (e.g. FDA), compliance requirements, and perceived market advantage and exploitation opportunities have created an environment which has prompted pharmaceuticals to develop an awareness of digital preservation challenges. Of these factors, though, the need to comply with statutory requirements appears to have been the main reasons why pharmaceuticals, such as Pfizer, developed preservation technologies. Competition within the sector has led to solutions being developed independently. Our interviews in 2002 and 2003 indicate that publishers had only just

[16] S Ross, M Greenan, and P McKinney, in press 2004, 'Digital Preservation Strategies: The Initial Outcomes of the ERPANET Case Studies' in the *Preservation of Electronic Records: New Knowledge and Decision-making*, (Ottawa, Canadian Conservation Institute).

begun giving serious consideration to how they should tackle the preservation of digital information and what shape their business cases should take.

We found, though, collaborative effort to tackle the problems of digital preservation rare in all sectors except broadcasting.[17] While our survey found collaborative initiatives in the publishing sector unusual, publishers are aware of what their competitors are doing. They anticipate that collaborative work with libraries will provide a way forward. The collaboration between Koninklijke Bibliotheek and Reed Elsevier may be indicative of similar future initiatives[18].

All of the organisations questioned were aware of the need to identify and implement mechanisms to support long-term access to digital entities. Even where organisations had created strategies and policies for managing and maintaining digital objects, these were often not implemented across the organisation and interviewees noted that they were applied with different degrees of rigour across different parts of the organisation. [19] As a result of collaborative initiatives broadcasting organisations exchange information on defining costing policies, approaches to constructing technical solutions, standards, and implementation guidelines. The broadcasting groups we interviewed had internal directives, standard procedures, and programmes focusing on preservation requirements, recovery, formats, and metadata systems.

Interviewees frequently reported that, in their view, a good strategy was to keep everything; this, they claimed, at least ensured that the material would be there in the future. The approach, however, begs questions about documentation, formats, curation, and risk. Few organisations included in the studies completed so far showed awareness of the critical role that selection and appraisal played in making certain that appropriate content was available and suitably documented for future use. And even fewer institutions reported that they had established selection policies in consultation with internal departments or other units with a stake in long-term access to digital objects.

So far we have identified few organisations actively developing solutions to enable digital longevity. There was a widely held opinion that software and system developers would eventually provide the necessary tools. Organisations tended towards preservation models that were reactive, pragmatic, and *ad hoc*. For instance, several organisations reported that migration to new data formats would be undertaken when the need arose, although few organisations seemed aware of the complexities associated with migration. Some organisations report that, to be safe, they had concluded that it was essential to retain the digital object in its original format alongside the migrated version. In regulated sectors, such as pharmaceuticals the need to guarantee the authenticity, integrity, and confidentiality of the records was acknowledged. Validating these features for each digital object would be prohibitively expensive. Therefore the optimum validation point was

[17] International organisations in the sector such as EBU (European Broadcasting Union) and European Commission funded research projects including PRESTO (Preservation Technology for European Broadcast Archives) have fostered the development of strategies for the preservation of film, video, and audio material and are contributing to the development of standards and best practices.
[18] http://www.kb.nl/kb/resources/frameset_kb.html?/kb/pr/pers/pers2002/elsevier-en.html
[19] The ERPANET Seminar on Policies and Procedures held in Fontainebleau in January 2003 added further weight to the findings of these case studies. There was a demand for guidance in the development of policies and even those organisations which had them in place acknowledged that they were not always implemented. Even where policies are in place they are often unrealistic and, as a result, unimplementable. See the seminar report, http://www.erpanet.org/events/2003/paris/ERPAtraining-Paris_Report.pdf

seen as system level.[20]

For digital documents PDF (Portable Data Format) emerged as the most widely used format for documents to which long-term access was required. A number of interviewees reported that their organisation had taken a policy decision to limit how PDFs were created, and even several reported that they had decided to restrict the use of its special features to enhance its suitability as a preservation vehicle. Of course it was recognised that there was an urgent need for agreement on preservation standard formats for a wide variety of data types from images, to audio, to database file types. Most organisations noted that they were waiting for industry agreed preservation enabled formats to emerge.

Their inability to predict the costs of digital preservation concerned all organisations interviewed. For example, the broadcasting sector expends substantial resources on digital preservation, and is now trying to streamline activities in order to rationalise that spending. This may seem surprising because, as Peter Lingaard Holm of Danish TV2 has noted this is the one sector that has proven its consumer base does not suffer from significant price sensitivity.[21] Several interviewees from the publishing sector stressed that in their opinion greater investment in digital preservation was necessary, but recognised that a better understanding of the costs involved and how, if at all, return on investment (ROI) would be achieved or measured.

Compliance and risk management have provided the major impetus to efforts to secure long-term access. It is not surprising, therefore, that less regulated sectors have not been as quick to address preservation challenges. Few of the companies included in our first surveys had succeeded in transforming digital holdings into assets. Moreover, only a couple of the publishers and broadcasters recognised the cultural or historical value of digital information.

It was certainly widely recognised that the need to preserve an increasingly large quantity of records and information had to be linked to a business case to improve and expand access to the material itself. With the exception of the broadcasting sector, institutions are waiting for external developments that they can adopt, or off-the-shelf solutions they can implement. Few sectors are aware of the enormity of the preservation problem or of the techniques that have been developed in other sectors that might be of value within their own sector. While we have many more case studies to conduct over the coming fifteen months we believe that six areas require immediate action:

- standardized preservation policy statements which can be easily adapted should be made available[22];
- the development of business cases and strategies that records managers, archivists, librarians, and other information professionals could use to convince business managers to fund digital preservation longevity initiatives;

[20] A conclusion, often attributed to David Bearman, but which was common practice at NARA by the mid-1970s. NARA first announced a strategy for scheduling and appraisal of records from a systems level in its first version of General Records Schedule 20, Automated Data Processing Records. See *Data Automation Program Records - General Records Schedule No. 20*, Federal Property Management Regulation 101-11.4, April 28, 1972 and especially 'Part V. Procedural Analysis of Data Processing Systems -- Guidelines for Appraising Files and Data Sets for Permanent Retention" 18-33.
[21] Personal communication.
[22] 'Policies for Digital Preservation: Seminar Report', ERPANET Seminar, Paris, January 29-30, 2003 http://www.erpanet.org/events/2003/paris/ERPAtraining-Paris_Report.pdf, 18-19.

- clear guidance on how different technologies impact on preservation pathways and options needs to be made available;
- improved models (e.g., reference, costs, standards, functional requirements) are required;
- preservation workflow modeling tools need to developed; and,
- production of guidance on creating, managing, and auditing digital repositories.

It is with this in mind that ERPANET is now beginning to undertake: the development of a suite of tools that will offer guidelines to organisations to measure their 'preservation effectiveness' and to improve digital preservation practices, as well as enable communication with suppliers and developers. For many organisations as well as developing a recognition of the critical role that appraisal plays in identifying materials for preservation, ensuring that suitable repository infrastructures and workflow practices are in place pose significant challenges. The lack of easily implementable repository models exposes organisations to unnecessary design and development risks. It makes the curation of digital materials challenging. In my recent review for the National Library of New Zealand of their preservation activities the lack of off-the-shelf repository models was identified as an obstacle to the widespread development of digital libraries and digital library services.[23]

As we focus more on providing access to and curation of digital information the distinction between the different types of information holding institutions begins to blur. Several presentations at our 1993 seminar anticipated this.[24] The continued growth of digital information increases the demand for adequate repositories and for those that recognised interconnection of information. Creating repositories is challenging and collaboration between groups of public sector organisations may be essential if high quality repositories are to be available and adequately maintained. Large scale repositories can achieve significant economies of scale as data repositories have demonstrated over the last couple of decades.[25] As more digital repositories emerge and we increasingly recognise the interconnection and interdependence of information resources we also recognise that we have more in common across the archives, libraries, and museum sectors than we tend to acknowledge. What is evident is that collaboration is essential if we are to establish mechanisms to address preservation challenges and to ensure that those approaches are widely adopted and implemented. ERPANET is one vehicle helping to do this, there are others, and we hope others will emerge.

[23] Ross, 2003, 24-29 and 51-52.
[24] Most explicitly in the paper by W Boyd Rayward, 'Electronic Information and the Functional Consolidation of Libraries, Archives, and Museums', in Ross and Higgs (eds.), 1993, 227-243.
[25] G Hunolt and A Booth, (9/2001), *ESDIS Data Center Best Practices And Benchmark Report*, (Science Operations Office, Earth Science Data and Information Systems Project, Goddard Space Flight Center, NASA Contract NAS5-00154).