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

1-9 August 2003, Berlin

Digital Preservation Activities in the U.K – building the infrastructure

Maggie Jones
Digital Preservation Coalition Co-ordinator,
U.K

Abstract:

Digital Preservation has been regarded as a matter of increasingly urgent priority in the U.K for a number of years. The Joint Information Systems Committee (JISC) and the British Library hosted a Workshop at the University of Warwick in 1995 to determine what needed to be done. Since then, there has been a great deal of effort and progress, coupled with a growing imperative to move beyond theoretical research and into practical archiving efforts. This paper will provide an overview of the major digital preservation activities being undertaken in the U.K and the highly collaborative nature of digital preservation, recognising that the issues cross both sectoral and geographic boundaries. The Digital Preservation Coalition, launched in February 2002, aims to leverage maximum benefit from the activities being undertaken individually and co-operatively by member institutions, and to provide a catalyst for further action.

In 1995 the Final Report of the Taskforce on Digital Archiving, a U.S report commissioned by the Research Libraries Group and the Commission on Preservation and Access, was released. For the first time, the disparate elements and range of stakeholders involved in the lifecycle of digital preservation were articulated and potential solutions were offered. The report sought to encourage an approach which simultaneously acknowledged the enormous complexity of the challenges, without being paralysed by them.

> *…we expect that the best use of the work of the Task Force will ultimately be to heighten awareness of the seriousness of the digital*

preservation problem, its scope and complexity – and its manageability. There are numerous challenges before us, but also enormous opportunities to contribute to the development of a national infrastructure that positively supports the long-term preservation of digital information. (Waters and Garrett, 1996)

In the U.K, the report inspired much of the discussion at a Workshop held at the University of Warwick and hosted by JISC and the National Preservation Office. A number of action points emerged from this Workshop, including a series of seven commissioned research papers, each one focussing on an identified aspect of digital preservation. The papers were subsequently synthesised into a single publication, *Digital Culture.* (Feeney, 1999)

It has been assumed that digital preservation will need to be a distributed responsibility. This is partly because of the scale of digital material being produced, and partly because of the nature of digital technology. Decisions which will affect the long-term viability of a digital object need to be taken so early in the lifecycle of the digital object, that those creating them are more logically the ones to undertake that initial activity.   It is also a factor that solutions are not going to be of the nature of "one size fits all". Different approaches will be perfectly valid for different types of digital resources and, while duplication of effort is to be avoided, a certain amount of judicious overlap is beneficial, particularly in these early stages of developing digital repositories.

The role of creators of digital materials is both crucial and difficult to integrate into a coherent infrastructure for preserving digital materials. Some creators of digital materials may be best placed to undertake preservation responsibility because of their in-depth knowledge of the subject matter, but they may lack the necessary archiving skills. The optimum solution in these cases might be an alliance between an organisation skilled in managing digital data, and the creators, so that those with the greatest knowledge of the material maintain control over decisions on what content needs to be preserved and at what intervals.

However there may be reluctance on the part of creators to hand over responsibility for their materials elsewhere. Libraries and archives have established their credentials for preserving print materials over a very long timeframe. In these very early stages of developing digital repositories, it may be difficult for creators to assign the same level of trust to them for preserving digital materials.

Some creators may simply be unaware that their material could be managed elsewhere.   By the same token, digital repositories may be unaware of valuable material being created. In both of these cases, digital materials are particularly vulnerable to loss. It is unlikely they will have been created with longer term management in mind, increasing the prospects of loss of the digital resource before a digital repository potentially able to manage it is even

aware of it. Alternatively, by the time such a repository is aware of it, it may be too late for cost-effective preservation to be undertaken.

It is possible to envisage three basic types of digital resource creator. Those who don't trust the ability of digital repositories to take care of their material; those who are unaware that such a possibility exists; and those who would love to be able to hand over their materials but no obvious repository yet exists for them to do so. Even assuming a trusted digital repository does exist, bringing them together with those creating digital materials in a mutually beneficial relationship is far from simple.

It is these complexities that have led some commentators on digital preservation to observe that organisational and legal issues are more intractable than technical challenges, formidable though the latter may be, particularly for complex digital materials.

So what has happened in the almost eight years since the first Warwick Workshop? How far have we come in building the infrastructure which will support the long-term preservation of digital information? There has certainly been a great deal of activity, and for the purposes of this paper, I will focus primarily on the work of the UK deposit libraries, in particular the British Library, the National Archives, and on the Joint Information Systems Committee (JISC).

JISC has made digital preservation a strategic priority for some time and has sponsored a number of projects and initiatives aimed at advancing knowledge and understanding of the many challenges still to be faced before long-term preservation of digital materials can be assured.

 The Cedars Project was funded by JISC with the initial objective of *addressing strategic, methodological and practical issues and provide guidance for libraries in best practice for digital preservation*. Collaboration is always a theme running through any discussion of digital preservation and the Cedars project began as a collaborative project between three CURL institutions, Leeds (the main site), Oxford, and Cambridge. Outcomes and deliverables form the Cedars project, including the five guidance documents which focussed on specific aspects of the work of Cedars (Collection Management, Intellectual Property Rights, Metadata, Digital Preservation Strategies, Digital Archiving Prototype) are all available from the Cedars website.

There was also constructive collaboration between the Cedars and the CAMiLEON project. The latter was funded jointly by JISC and the National Science Foundation in the U.S. The U.K part of the project was also based in Leeds, allowing a mutually beneficial working relationship between Cedars and CAMiLEON. CAMiLEON worked primarily on testing the viability of emulation as a preservation strategy and successfully re-created the 1986 BBC Domesday Project, which had been copied onto video disk, a long obsolete technology. It also developed tools to assist in future rendering of

digital materials when the inevitable technological obsolescence renders them incapable of being accessed on current technology.

The JISC Continuing Access and Digital Preservation Strategy 2002-2005 signalled a different direction.

> *Much has been achieved by JISC and others with relatively modest investment in recent years. However the escalating scale and complexity of digital resources to be curated and the subsequent urgency of developing a critical mass of expertise, shared services and tools, for long-term digital preservation, will require a step change in investment and approaches. Over the next three years a greater emphasis on development of production services and tools will be needed to build on previous research studies and projects.* (A continuing Access and Digital Preservation Strategy for the Joint Information Systems Committee (JISC) 2002-2005)

The 2002-2003 Digital Preservation Development Programme to support the first year of this three year strategy is indicative of the concerted effort to move digital preservation firmly beyond the realm of short-term projects to sustainable services. Several projects and studies have been funded by this programme:

- Archiving E-Publications investigated the implications of licensed e-journals, in particular those licensed under the JISC NESLI deals. This is an area where the blurring of boundaries between owners and creators of content and libraries is most evident. Publishers have had no preservation role to play in the print environment, now find themselves being asked to sign a Model Licence which gives them responsibility for ensuring continued access to licensed journal content. The draft report was circulated for consultation and peer review in May 2003. A key recommendation of the report was that JISC establish a post dedicated to acting as a negotiator and central point of contact for both libraries and publishers, to enable pursuit of appropriate archiving arrangement, without delaying current licence deals.

- Data Curation for e-science in the UK investigated the current provision and future requirements for the curation of primary research data within UK e-science. The DTI and Research Councils have committed £118m for a government-industry programme on e-science, employing grid technology. Grid technology (also known as cyberinfrastructure in the US) is regarded as the natural successor to the world wide web and is set to revolutionise the way scientists conduct their research. It is therefore critical that the huge volumes of data being produced are appropriately managed. The draft report *Data curation for e-science in the UK: an audit to establish requirements for future curation and provision,* was circulated for consultation and peer review in May 2003. A key finding of this project was that further action is urgently needed to support the curation of the anticipated huge increases in data resulting from this major investment.

- Feasibility and Requirements Study on Preservaton of E-Prints. The driving force behind the development of E-Prints and institutional repositories has been on encouraging researchers to deposit their materials in these emerging facilities. The focus on the JISC study was in assessing what needs to be done to maintain access to deposited e-prints into the future. The drat report *Feasibility and requirements study on preservation of e-prints* was circulated for consultation and peer review in May 2003. The report suggests that there is an opportunity to address preservation issues before the issue becomes urgent, without inhibiting the deposit of material into e-print repositories.

- Web archiving. UKOLN undertook a feasibility study into web archiving on behalf of JISC and the Wellcome Trust. The report, *Collecting and preserving the World Wide Web: a feasibility study,* was released in February 2003. A separate study was also initiated to focus on the legal implications of web archiving, recognising that this is a key concern of those wishing to preserve web based material. The report *Legal issues relating to the archiving of Internet resources in the UK, EU, USA, and Australia* was also released in February 2003. As the description of the project indicates, *Collaboration will be the key to any successful attempt to collect and preserve the web.* Digital preservation as a global issue is also reinforced here. Accordingly, further work is being undertaken collaboratively. The British Library and the National Archives UK (formerly the PRO and the Historical Manuscripts Commission, which merged in April 2003) have already begun work in this area and they will work with JISC and Wellcome and others to implement the PANDAS software developed by the National Library of Australia to harvest web-based materials it collects.

- Representation and rendering project. Both the work of Cedars and CAMiLEON fed into the work of a JISC funded project on file formats and rendering undertaken by the University of Leeds. The need for this arose from the dependence on understanding the formats in which the data is encoded before digital materials can remain accessible over time. This work will in turn be critically important for the proposed Digital Curation Centre, which is discussed in more detail below. The report, *Representation and Rendering Project, Survey and assessment of sources of information on file formats and software documentation* was released in May 2003.

- Revision of the Records Lifecycle was undertaken to update an earlier JISC study which provided guidance on good practice for the management of records and archives in UK HE institutions. The implementation of the Freedom of Information Act 2000 has provided an additional spur to provide guidance for records managers and archivists. The report, *Higher Education and Activity Model & Record Retention Schedule* was released in May 2003.

The Supporting Institutional and Records Management Programme has funded seventeen projects, including the development of an electronic records

management (ERM) training package. All projects are due for completion by December 2003. These are intended to support UK higher and further education institutions in the implementation of records management programmes.

At the time of writing, JISC and the e-science core programme has issued an Invitation to tender for a Digital Curation Centre. The impetus for this service is the recognition that there are certain activities required by all digital repositories which are most efficiently and effectively undertaken centrally. The Digital Curation Centre will not itself be a repository for digital data but will collect information on file formats and preservation planning tools to support those who are undertaking this role. This service will also include technology watch.

This considerable investment by JISC in such a wide range of activities is indicative of the increasing urgency in the need to accelerate progress towards implementation of digital preservation approaches. It implicitly acknowledges that while there are still significant challenges, it is necessary to embark on a concentrated effort to establish, in the words of the JISC Strategy:

> *An infrastructure of collective and institutional services and repositories.*

Both the British Library and the National Archives are major players in the UK digital preservation agenda and both have undertaken significant planning and preparation, as well as building repositories capable of dealing with large quantities of digital materials.

The British Library is focussing on four major categories of digital content. These are:

- Deposited material (a voluntary deposit scheme has been implemented, pending legal deposit legislation)

- Websites. The BL is participating in two collaborative projects, one is based on the JISC Wellcome proposal, mentioned above. The other is the Web Archiving Consortium, a collaboration of national libraries and the Internet Archive.

- Digitisation. The BL has undertaken a number of digitisation initiatives over the years to enhance access to its rich collections. A 'Collect Britain' is the BL's largest digitisation project to date and has been funded by a grant from the New Opportunities Fund. It is aimed that by the summer of 2004, 100,000 images and sounds selected from the BL's collections will be available online.

- Digital materials purchased for the provision of services

Legal deposit has always been a key enabling mechanism to support the preservation of a nation's printed published heritage and in the UK there has been an active campaign to extend this to non-print materials. A brief synopsis of the steps towards extension of legal deposit to non-print materials illustrates what a slow process this is.

A working party, including representatives from publishers and deposit libraries and chaired by Sir Anthony Kenny was established and reported in 1998, having been charged with drawing up a voluntary code of practice. However, the Working Party indicated that it "was convinced that only a system of legal deposit will be adequate to secure a comprehensive national published archive."

The voluntary code of practice was introduced in January 2000 and covered the deposit of UK non-print publications in microform and offline electronic media (e.g CD-ROM's, floppy disks etc.). This code of practice does not, as a general rule, include online digital publications and where there are parallel print and electronic publications, the former is usually regarded as the deposit copy. Film, sound, and Ordnance Survey digital maps fall outside the voluntary code of practice and are subject to separate voluntary schemes.

The Legal Deposit Libraries Bill achieved its second reading in the House of Commons in March 2003 and is scheduled to have its third reading in July 2003. In the meantime, the voluntary code of practice has enabled the British Library and other deposit libraries to prepare for what could potentially be a substantial influx of digital materials.

A report was commissioned by the Joint Committee on Voluntary Deposit to assess the cost implications of legal deposit for both legal deposit libraries and publishers. Using two different scenarios, based on the British Library receiving either 60% or 80% of unique e-serial publications within the first three years of legal deposit, with e-monographs remaining stable at c. 75%. This suggested that the cumulative costs of processing this material at the British Library over the first three years would be £2.6m and £2.7m respectively. The report suggested that publishers were not concerned about delivery costs of depositing publications to the deposit libraries but had a number of concerns relating to access/usage, metadata and rights issues.

There are still ongoing discussions between the British Library and publishers and this has been a steep learning curve for both in determining practice and procedures which are mutually acceptable and which will safeguard the digital published heritage.

The National Archives has been actively engaged in providing guidance to government departments in implementing good electronic records management practice. This has been given particular urgency with the introduction of the Freedom Information Act (2000) and also the Data Protection Act (1998). The compliance issues related to this legislation, combined with a Government initiative which had the target of all government

departments storing and retrieving all new records electronically by 2004 has made it imperative that good ERM takes place.

To facilitate this, the National Archives has worked with government departments to support the introduction of electronic management systems. In addition, the National Archives has built a Digital Archive capable of providing long-term storage for all electronic records produced by government departments and selected for archiving.

The National Archives has also developed the PRONOM database, which stores detailed descriptions of software applications, including file formats to which an application can read and write and any dependencies that may exist in order to run the application.

The National Archives website also includes a page of trials and pilots they are undertaking, to prepare for large volumes of data being accessioned into the Digital Archive. These include testing data integrity, and an Open Source Software Pilot.

As already mentioned, the National Archives is also engaged in preserving websites. The selection of government websites is currently the focus of a small working group established to consider the selection and preservation of government websites. As already indicated, the National Archives is also working collaboratively with others in web archiving.

With such intense activity, it is not surprising that a mechanism for co-ordinating digital preservation activity and fostering collaboration was seen to be very important. The Digital Preservation Coalition was formed following a summit in January 2001 and formally launched at the House of Commons in February 2002. From six founding members, the DPC had grown to twenty five members by June 2003. Membership is deliberately cross sectoral, recognising the intersecting interests between sectors and the need to harness individual efforts to address mutual concerns. While focussing primarily on developing the UK Digital Preservation agenda, it has recognised from the beginning, the fact that digital preservation is very much a global issue and it is critically important to establish good lines of communication with all those engaged in digital preservation efforts. A MOU with the National Library of Australia has provided a fruitful collaboration. Contact has also been established with the US National Digital Information Infrastructure and Preservation Program (NDIIPP) co-ordinated by the Library of Congress. It is hoped that this will also lead to a MOU between the DPC and NDIIPP.

The DPC resources are modest, and it is only since May 2003 that a full-time officer was appointed to the DPC to act as Co-ordinator. Nevertheless it has managed to achieve a great deal in its short existence. A PR campaign was launched in the belief that there needed to be much more awareness of the importance of digital preservation and the profile needed to be raised significantly. There was a certain feeling of a relatively small number of interested parties essentially preaching to the converted. Before the PR campaign, a search was done to establish the number of times digital

preservation was mentioned in the media. Only one example was found, contrasting with more than forty news items since the DPC launch to date.

The packaging and dissemination of relevant information has also been a key activity, recognising the difficulty of maintaining current awareness in such a rapidly developing area. There is now an electronic edition of the *Preservation Management of Digital Materials Handbook* and this will be updated. Quarterly updates of *What's New in Digital Preservation* are compiled by Michael Day of UKOLN, and Gerard Clifton, of the National Library of Australia, and provide an invaluable source of reference and current awareness.

A Digital Preservation discussion list was established and has almost one thousand members as of June 2003, further reinforcing the increasing importance of digital preservation and the need to forge and develop contacts with others working in this area.

Finally, training is seen as a key priority by many DPC members and so the DPC has held regular Forums focussing on specific themes, as one mechanism to increase knowledge and understanding. A series of DPC Training Workshops, based on the *Preservation Management of Digital Materials Handbook* has also been developed for DPC members. This is an area of activity which will need to expand to cope with increasing demand and the DPC will need to act as both provider and co-ordinator of training activities.

In summary, there has certainly been a great deal of progress since the first Warwick Workshop was held in November 1995. There is a much richer understanding of the range and complexity of the challenges and the need for a range of strategies and procedures to deal with them. There is more reliable information on costs beginning to emerge. More practical experience is beginning to inform debate and development of policies and strategies. There have been opportunities to learn from each other and to share what we have learned.

It is however, axiomatic to say that much more needs to be done. We still need a far greater level of awareness, particularly from funding agencies, combined with greater clarity on roles and responsibilities at the national, regional, and local level. We need to keep track of what material is being produced and its level of vulnerability. And we need to maintain and develop links with others so that we can all continue to learn and to benefit from the collective progress being made.

References

Archiving e-journals consultancy – final report, v.2.0, May 2003 <http://www.jisc.ac.uk/uploaded_documents/ejournalsdraftFinalReport.pdf>

British Library. Collect Britain.<http://www.bl.uk/collectbritain>

CAMiLEON project <http:www.si.umich.edu/CAMILEON/>

Cedars Project <http://www.leeds.ac.uk/cedars/>

Collecting and Preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust, 25 February 2003.
<http://library.wellcome.ac.uk/projects/archiving_reports.shtml>

A Continuing Access and Digital Preservation Strategy for the Joint Information Systems Committee (JISC) 2002-2005.
<http://www.jisc.acu.uk/index.cfm?name=pres_strategy>

Data Curation for E-science in the UK: an audit to establish requirements for future curation and provision. V.1.0, May 2003
<http://www.jisc.ac.uk/uploaded_documents/e-Sciencefinaldraft.pdf>

Digital Preservation Coalition
<http://www.dpconline.org>

Digital Preservation e-mail list
<http://www.jiscmail.ac.uk/lists/digital-preservation>

Feasibility and Requirements Study on Preservation of E-prints, v.1.0,9 May 2003
<http://www.jisc.ac.uk/uploaded_documents/eprints_report_1-0.pdf>

Feeney, M (ed) (1999). Digital Culture: maximising the nation's investment. London: National Preservation Office

Higher Education Institutions: Business Function and Activity Models & Retention Schedule [Revision of the Study of the Records Lifecycle], 2003
<http://www.jisc.ac.uk/index.cfm?name=srl_structure>

The Impact of the extension of legal deposit to non-print publications: assessment of cost and other quantifiable impacts. Study report prepared for the Joint Committee on Voluntary Deposit by Electronic Publishing Services Ltd.
<http://www.epsltd.com/ExtensionOfLegalDeposit.htm>

JISC Circular 6/03: Digital Curation Centre
<http://www.jisc.ac.uk/index.cfm?name=circular_06_03>

Legal Issues Relating to the Archiving of Internet Resources in the UK, EU, USA, and Australia: a study undertaken for the JISC and Wellcome Trust, 25 February 2002.
<http://library.wellcome.ac.uk/projects/archiving_reports.shtml>

Long Term Preservation of Electronic Materials: a JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib). 27-28 November at the University of Warwick. <http://www.ukoln.ac.uk/services/papers/bl/rdr6238/>

National Archives (formerly Public Record Office). Digital Preservation Trials and Pilots.
<http://www.pro.gov.uk/about/preservation/digital/trials.htm>

National Digital Information Infrastructure and Preservation Project.
<http://www.digitalpreservation.gov/ndiipp/>

Preservation Management of Digital Materials: a Handbook. London: British Library, 2001 <http://www.dpconline.org/graphics/handbook/index.html>

Representation and Rendering Project, survey and assessment of sources of information on file formats and software documentation: final report, 2003. <http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf>

Supporting Institutional Records Management Programme.<http://www.jisc.ac.uk/index.cfm?name=programme_supporting_irm>

Waters, D and Garrett, J (1996). 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. <http://www.rlg.ac.uk/ArchT/>

What's New in Digital Preservation <http://www.dpconline.org/graphics/whatsnew/>

Working Party on Legal Deposit (1999). Code of Practice for the voluntary deposit of non-orint publications <http://www.bl.uk/about/polocies/coedprac.html>

Working Party on Legal Deposit (1998). Report. <http://www.bl.uk/about/policies/workreplegdep.html>