Joint US-UK Digital Preservation Workshop, Washington, D.C., May 7-9, 2006

Michael Day,
UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
m.day@ukoln.ac.uk
http://www.ukoln.ac.uk/

Helen Hockx-Yu
JISC Executive, King's College London, Strand Bridge House, 138-142 The Strand, London
WC2R 1HH, United Kingdom
http://www.jisc.ac.uk/

Joint US-UK Digital Preservation Workshop, Washington, D.C., May 7-9, 2006

Michael Day,
UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
m.day@ukoln.ac.uk

Helen Hockx-Yu
JISC Executive, King's College London, Strand Bridge House, 138-142 The Strand, London
WC2R 1HH, United Kingdom
h.hockx-yu@jisc.ac.uk

Michael Day and **Helen Hockx-Yu** report on an invitational workshop held in Washington, D.C. on the 7-9 May 2006; organised by the Joint Information Systems Committee and the US National Information Infrastructure and Preservation Program.

1. Introduction

On May 7-9, 2006, the Sofitel Lafayette Square Hotel in downtown Washington, D.C. played host to a digital preservation workshop jointly organised by the Joint Information Systems Committee (JISC; http://www.jisc.ac.uk/) and the US National Digital Information Infrastructure and Preservation Program (NDIIPP; http://www.digitalpreservation.gov/). The workshop had two main goals. Firstly to provide an opportunity for people involved in JISC and NDIIPP research projects to meet together and to exchange information, stimulating ideas for ongoing collaboration. Secondly, to try to identify those aspects of digital preservation that require more attention. This report provides an overview of the workshop proceedings.

2. Overviews of digital preservation activities

The workshop opened on the evening of the 7th May with a reception and poster session focused on the activities of research projects funded by the National Science Foundation (NSF) as part of the Digital Archiving and Long-Term Preservation (DIGARCH) program (NSF, 2004), a key part of the first phase of NDIIPP. A corresponding poster session on the following evening focused on activities supported by the JISC, including the Digital Curation Centre (DCC; http://www.dcc.ac.uk/), the Arts and Humanities Data Service (http://www.ahds.ac.uk/), the UK Data Archive (http://www.data-archive.ac.uk/) and the various projects funded as part of the Supporting Digital Preservation and Asset Management in Institutions programme (http://www.jisc.ac.uk/index.cfm?name=programme_404).

The workshop itself commenced on the morning of the 8th May with a welcome by **William LeFurgy**, Digital Initiatives Project Manager at the Library of Congress. He reminded delegates about the workshop's two main goals. Firstly, he considered that it would provide a useful forum for sharing information about the US and UK's respective programmes and practices, recognising that digital preservation is still a relatively new enterprise and noting the importance of trust, sustainability and developing communities of practice. Secondly, he hoped that the workshop would utilise the combined intelligence of participants to help identify significant gaps in current activities, which will help NDIIPP and JISC plan and prioritise future work. **Stephen Griffin** of the National Science Foundation (NSF) then provided some information on upcoming funding opportunities in the US.

2.1 UK activities

The session then turned to general overviews of UK and US digital preservation activities. **Sarah Porter**, Head of Development at JISC, first provided some background information on the work of the Joint Information Systems Committee in fostering digital preservation activities in the UK.

Highlighting the importance of the pioneering Cedars (CURL Exemplars in Digital Archives) project and its successors, Porter explained that digital preservation was now a core part of all JISC research and development programmes. Ongoing activities included collaboration with the UK research councils' e-Science Core Programme on the Digital Curation Centre and additional funding available through a new Capital Programme worth £80 million; £14 million of which had already been earmarked for the support of digital repositories and preservation (http://www.jisc.ac.uk/capital.html).

Helen Hockx-Yu, Programme Manager at JISC, followed this by providing a more detailed outline of digital preservation activities in the UK, highlighting the importance of digital preservation and curation in the context of continued government investment in research through the Science & Innovation Investment Framework, 2004-2014 (HM Treasury, Department of Trade and Industry & Department for Education and Skills, 2004). The presentation built on Sarah Porter's comments by emphasising the role of JISC as a key driver of initiatives focused on embedding digital preservation in UK higher and further education institutions. Examples provided included the JISC's influential Continuing Access and Digital Preservation Strategy (Beagrie, 2002), a series of feasibility and scoping studies that had helped to inform and prioritise JISC's ongoing research and development activities, and the co-funding of national services like the Arts and Humanities Data Service, the UK Data Archive and the DCC. JISC also collaborated with other organisations when required, the presentation highlighting participation in the UK Web Archiving Consortium and the Digital Preservation Coalition (DPC; http://www.dpconline.org/), as well as the development of a formal partnership with the British Library. There followed brief pointers to digital preservation activities currently being undertaken by a range of organisations, including: the DPC, the British Library, The National Archives, the research councils and the British Broadcasting Corporation. Other important work was being undertaken through the participation of UK organisations in key projects funded by the European Commission as part of the Sixth Framework Programme. The presentation concluded with an acknowledgement that while there was a growing awareness of digital preservation issues at various levels, the practice of preservation had still not been embedded as an integral part of most organisational workflows and there was little in the way of commonly agreed best practice. With reference to the DPC's recently published UK Needs Assessment (Waller & Sharpe, 2006), the presentation argued that there also needed to be more focus on the strategic level, including the clarification of organisational roles and responsibilities for a task that, it is assumed, can ultimately only be successful as a collaborative activity.

2.2 US activities

Laura Campbell, Associate Librarian for Strategic Initiatives at the Library of Congress, then provided an overview of activities undertaken as part of the US National Digital Information Infrastructure and Preservation Program (NDIIPP). This initiative was created by federal legislation in December 2000 and is worth up to $175 million, including funds matched from non-federal resources. Campbell explained that the vision of NDIIPP is "to ensure access over time to a rich body of digital content through the establishment of a national network of committed partners." Specific NDIIPP goals include helping to identify and preserve at-risk content, supporting the development of improved tools, models and methods for preservation, and the development of a national collection and preservation strategy. Reflecting the decentralised nature of US digital preservation activities, the program also aims to work in co-operation with a wide range of other organisations, including federal agencies, libraries, research institutions, etc. There is also some co-operation on an international level, e.g. through the Library of Congress's membership of the International Internet Preservation Coalition (IIPC) and NDIIPP links with the UK's DPC and British Library. NDIIPP itself is a portfolio of activities focused on three main areas. Firstly, the program has begun to create a network of preservation partners, making co-operative agreements with a range of institutions dealing with a variety of content types, including digital television, Web sites and geospatial data (http://digitalpreservation.gov/partners/project.html). Secondly, NDIIPP has developed the principles of a modular, upgradeable architecture for digital preservation. Projects associated with this have included a test of archive ingest and handling that compared digital preservation systems in four universities (e.g., Shirky, 2005). A third focus

of NDIIPP has been on research, primarily through the NSF's funding of the DIGARCH program. Additional activities have included a Library of Congress initiated study group looking at the implications of digital technologies for the exceptions applicable to libraries and archives in Section 108 of the US Copyright Act (http://www.loc.gov/section108/). Phase II of NDIIPP commenced this year and includes the further expansion of partnerships to include commercial content and technology companies, the encouragement of state or local repositories and the funding of another series of NSF research grants.

2.3 European activities

Towards the end of the workshop - but fitting conceptually into this opening session - **Carlos Oliveira** of the European Commission's Information Society and Media Directorate-General gave a short summary of European Union (EU) activities relating to digital preservation. After an introduction to the role of the directorate-general and an explanation of why digital preservation was important for the Commission, the presentation introduced various initiatives funded or otherwise supported by the Commission. He mentioned a range of Research, Technology and Development projects funded as part of the Information Society Technologies thematic priority under successive framework programmes, including two new integrated research projects - PLANETS and CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) and a co-ordinating activity (Digital Preservation Europe), all recently funded as part of the EU's Sixth Framework Programme and led by UK organisations. The Commission was also working at the policy level with things like the "i2010 digital libraries" initiative, which is attempting to boost digitisation activities in Europe. Oliveira commented that digital preservation was an important part of the EU's Lisbon Agenda - focused on the emergence of the EU as a competitive knowledge-based economy - and outlined a range of future short and longer-term activities, including research opportunities in the forthcoming Seventh Framework Programme.

3. Shared challenges, gaps, and needs

Donald Waters of the Andrew W. Mellon Foundation gave a keynote presentation reflecting on digital preservation developments a decade on from the publication of *Preserving digital information*, the final report of the Task Force on Archiving of Digital Information (Garrett & Waters, 1996). The task force, of which Waters was co-chair, provided a potent reminder in the mid-1990s of the critical challenges posed by the long-term preservation of digital information and identified some components of the 'deep infrastructure' it considered necessary to support preservation activities. The strategic importance of the task force's report has been widely acknowledged, not least in the UK, where the release of the draft report in 1995 resulted in the commissioning of a series of studies into digital preservation topics (summarised in Feeny, 1999) and later to early JISC initiatives like the Cedars project.

After a quick walk-through of the key findings and recommendations of the task force's report, Waters focused on four grand challenges that remained to be addressed ten years on.

3.1 Intellectual property, preservation and access

Waters first revisited the vexed question of intellectual property (IP) rights, namely the perception that copyright law makes digital preservation problematic. He thought that at least part of the difficulty was the popular understanding that preservation equals access, an equation formulated by those responsible for preserving out-of-copyright brittle books. Waters felt that rights holders would resist changes in copyright law if preservation became a potential 'backdoor' for the redistribution of "at risk" materials. The task force itself had focused on the need for 'aggressive rescue,' arguing that no distributed system of preservation services would be effective "unless it provides for a powerful rescue function allowing one agency, acting in the long-term public interest of protecting the cultural record, to override another's neglect of or active interest in abandoning or destroying parts of that record" (Garrett & Waters, 1996, p 23). As evidence of such neglect, Waters cited the recent court judgement against US investment bank Morgan Stanley for failing adequately to ensure the retention of e-mails (e.g., Day, 2006). He also noted

that publishers, for a variety of reasons, sometimes removed content from their e-journal services, referring to the discussion on Elsevier's controversial 'takedown' policies on the liblicense-l mailing list in 2003. In that discussion, Jim O'Donnell of Georgetown University argued that there could be similar concerns with the preservation of content in institutional repositories. He asked that, "where the author retains control over the copyright of his/her material - what protection do we then have to assure us that articles will remain archived, unchanged, in perpetuity?" (O'Donnell, 2003). Looking towards solutions, Waters first mentioned that Jane Ginsburg and June Besek of Columbia Law School were currently engaged in a Mellon-funded study of legal strategies for protecting archives from takedown demands. He also argued that initiatives like CLOCKSS (Controlled LOCKSS - Lots of Copies Keeps Stuff Safe; http://www.lockss.org/clockss/Home) and Portico (http://www.portico.org/) were shedding better light on the potential role of 'dark archives' in preserving content. For example, Portico's policies on access reflect the fact that many publishers are only prepared to participate if the e-journal archiving service is seen not to challenge their current business models. Waters concluded his comments on intellectual property issues by arguing for a more nuanced understanding of preservation *vis-à-vis* access, e.g. on the exact triggers that might initiate aggressive rescue.

3.2 Networks of trusted institutions and the question of certification

Waters then turned to another issue raised in the task force's 1996 report, the importance of *trust* within a distributed system of preservation services. The report argued that, in order "to ensure that no valued digital information is lost to future generations, repositories claiming to serve an archival function must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification" (Garrett & Waters, 1996, p. 9). Waters noted that, while the subject of certification has (so far) attracted a great deal of attention, trust - or its absence - is the central issue that needs to be considered. In this context, the presentation identified two main trust-building features. Firstly, a repository must have the technical ability to maintain and to be able to demonstrate the authenticity and integrity of preserved objects. Secondly, the organisation must itself be able to demonstrate a commitment to the preservation mission include well-defined preservation services. After a brief mention of the various working groups and initiatives that have looked in detail at the attributes of repositories and their certification (RLG/RLG Working Group on Digital Archive Attributes, 2002; RLG-NARA Task Force on Digital Repository Certification, 2005), Waters noted that both the report by the US National Science Board (NSB) on *Long-Lived Data Collections* (NSB, 2005) and the DPC's UK Needs Assessment (Waller & Sharpe, 2006) suggested that the requirements for repository certification might, in practice, be fairly complex. For example, the *Long-Lived Data Collections* report (NSB, 2005, p. 14) identified at least three functional categories of data collection - project, community and reference collections - each with very different requirements in terms of longevity, the use of standards, sustainable funding, and much else. In response, Waters suggested that there was an emerging consensus that certification needed to be community-driven, rather than mandated centrally for all repositories.

3.3 The interoperability gap

A third challenge was the need for interoperability, enabling repositories to co-operate better with each other and with other preservation services. Waters noted that there was scope for extensive co-operation between repositories, with potential efficiency gains from organisations focusing on particular tasks (division of labour). In addition, the bulk transfer of content across and among repositories could meet some of the needs for remote backup and replication. A number of standards and services relevant to interoperability were already being developed, examples being the PREMIS data model and data dictionary for preservation metadata (PREMIS Working Group, 2005) for preservation metadata, the Journal Archiving and Interchange Document Type Definition (DTD) developed by the National Center for Biotechnology Information (http://dtd.nlm.nih.gov/), and prototype registry services like the Global Digital Format Registry (GDFR; http://hul.harvard.edu/gdfr/). However, Waters argued that there was a need for more practical experimentation and testing, noting the importance of exemplars like the NDIIPP archive ingest and handling test. The presentation then included a short summary of a recent invitational

meeting on "Augmenting Interoperability across Scholarly Repositories," an event sponsored by Microsoft, the Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation, and the JISC (http://msc.mellon.org/Meetings/Interop/). This meeting had focused on the needs of complex digital objects, looking at appropriate data models and the key service functions that need to be represented in repository interfaces.

3.4 Business models and models of co-operation

The final challenge that Waters addressed was the need for sustainable business models for digital preservation. He first noted that preservation was dependent on both commitment and the availability of sustained resources. He then raised some key issues relating to the generation of such resources. Firstly, he commented that preservation, as a public good, is subject to the free-riding problem. As Waters (2002, p 84) has written elsewhere:

> *A special property of archiving is that if one invests in preserving a body of information and that information is eventually lost to others who did not take out the insurance policy, the others are not excluded from the benefits, because the information still survives. Because free riding is so easy, there is little economic incentive to take on the problem of digital preservation, and this partly explains why there has been so little archive building other than that funded by governments. Potential investors conclude that "it would be better for me if someone else paid to solve the archiving problem." In fact, one of the defining features of a public good - and think here of other public goods such as parks or a national defense system - is that it is difficult and costly to exclude beneficiaries.*

As with other free-riding situations, government funding can help to address the problem, but may not - in itself - be enough. Waters also introduced a second issue related to economies of scale. While traditional cultural heritage organisations might be able to take responsibility for digital preservation as part of their mission, achieving economies of scale might be a problem. Third party preservation services might be able to operate at the relevant scale, but are subject to the problems of working in a two-sided market, i.e. one that needs to take account of the needs of both producers and consumers (or their intermediaries). While many established two-sided market operations (e.g., the credit card industry) are able to raise revenue from either one or both sides of their operations, it is far from clear how this might work for preservation services. Waters explained that the start-up negotiations and market analysis for Portico resulted in an abandonment of its initial interest in access provision and a lowering of costs, producing a service that might be characterised as being based on an insurance model. A final issue raised in the presentation related to the nature of necessary interaction with commercial entities. Traditional approaches tended to compartmentalise the differing preservation requirements of commercial, not-for-profit and government-funded organisations, but it could be argued that longer-term sustainability would ultimately depend on the existence of a range of different funding streams. Waters argued that there was a need for innovative approaches to developing a shared vision for preservation and for mutual support across all sectors, including a much deeper interaction with the commercial sector.

4. Breakout groups

In the afternoon, the workshop divided into four breakout groups, each looking at a different aspect of current digital preservation requirements. Groups reported back the following morning and the following sections summarise the reporting back sessions.

4.1 Basic preservation tools and methods

First to report back was the breakout group looking at basic preservation tools and methods. Group leaders Adam Farquhar (British Library) and Bob Horton (Minnesota Historical Society) summarised the discussion of the group, identifying things that they thought were currently being done fairly well, but also highlighting areas that they thought may require additional effort. One major need identified by the group was for appropriate preservation tools or services to be made

available to those organisations (or individuals) with limited resources, capacity and expertise, and in particular for those for whom digital preservation is not their primary mission. Examples of such organisations include public libraries, local authority archives, historical societies, educational institutions and individual content creators (writers, photographers, etc.). The group had split into two smaller groups. One sub-group identified various ways in which awareness and understanding of the digital preservation problem had already been raised and proposed a five-step programme for working with particular communities, recognising that each will have its own priorities. For each audience, the programme would first identify potential partners, evaluate their needs and help to produce business cases for preservation. The final two steps involve the embedding of preservation within established business routines and the ongoing effort needed to deal with changes in technologies, personnel or knowledge levels. The second sub-group was asked to identify solutions, first highlighting the range of preservation tools that already exist - including PDF/A, LOCKSS (Lots Of Copies Keeps Stuff Safe; http://www.lockss.org/), institutional repository software, file identification and validation tools, Web harvesting tools, preservation metadata - then proposing areas that need more work. One major need was for clear relevant guidance, e.g. for developing policies, for deciding what content to keep, for decisions on formats and metadata. A related need was for a body of evidence that would support decision-making processes, e.g., evaluating and benchmarking preservation activities and tools. A third and more specific need was for methods for preserving dynamic content. Finally, the group again emphasised that preservation requirements are not homogenous. They, therefore, were sceptical about the emergence of monolithic solutions, but instead encouraged communities of practice to develop around shared needs.

The discussion that followed - like that which followed all breakout group reports - was lively. It is impossible to capture every single point that was made, but important points that were raised included:

- In a discussion on responsibility for preservation, several speakers made the point that the primary initiative for preservation needs to come from the local level. Martha Anderson (Library of Congress) cited the example of the International Internet Preservation Consortium, in which tools are developed first at a local level, and then shared with the broader group.

- Margaret Hedstrom (University of Michigan) noted problems with building momentum and was sceptical about the inclusion of preservation capability into business processes or software. She argued that our assumptions were predicated on the idea that people will be prepared to co-operate.

- David Rosenthal (Stanford University) made the point that organisations responsible for preservation had to be upfront about the economic and technical limitations on what could be preserved and not to encourage unrealistic expectations that they could deal with whatever is supplied by producers. In this context, Helen Tibbo (University of North Carolina at Chapel Hill) stressed the importance of appraisal tools, matching institutional missions with projected costs and benefits.

- In considering what JISC and NDIIPP could do to address some of these concerns, Hedstrom encouraged an initial focus on common services that are needed by everyone, e.g. bit-level preservation services. Neil Beagrie (JISC and British Library) highlighted areas where there had already been successful collaboration, e.g. on training programmes, harvesting tools and registries of file format information. Several speakers commented on the importance of information exchange. Others noted the new challenges thrown up by international large-scale scientific collaborations and by compliance agendas related to things like freedom of information legislation and the Sarbanes-Oxley Act.

4.2 Institutional stewardship and life cycle management

The second group to report was the one looking at institutional stewardship and life cycle management, chaired by Fran Berman (San Diego Supercomputer Center) and Sheila Anderson (Arts and Humanities Data Service). The group's report, given by Melanie Wright (UK Data

Archive), explained that group members had been encouraged to complete a questionnaire prior to attending the meeting. This had revealed several interesting things about the repositories that participants represented. Firstly, that most repositories were relatively new, having been established within the last five years. Secondly, there was a tension between the short-term (or project-based) funding available for repository development and the need to develop these into sustainable services in the longer-term. A third set of issues related to the multiplicity of content types held in repositories and to differences in organisational purpose, e.g. whether the primary focus was on preservation or access. Other points raised included the importance of standards, protocols and workflow tools, the need for the education of content creators, the problems of dealing with multiple formats or unstructured data, and the need for different ways of working in the digital age.

The group discussed a variety of different challenges - e.g. how to determine the value of information and decide what to save, how to make economic arguments for funding preservation, and debating who should have ultimate responsibility for stewardship - and came up with a number of concrete recommendations. Firstly, supporting the principle of sharing information more widely, the group proposed the creation of several Web-based resources. These included:

- A directory of skills and hardware relating to data recovery techniques, e.g. recording the existence of obsolete hardware

- Registries of tools, approaches, methods, case studies, business cases, cost-benefit studies, etc.

- An interactive environment (e.g. a wiki) that can be used for the evaluation of tools or for sharing experiences.

Secondly, the group proposed the funding of individuals (or working groups) to develop templates of the "significant properties" of different types of objects, to evaluate standards and develop consensus, and to facilitate staff-exchanges across national and disciplinary boundaries.

In the discussion that followed, major issues raised included:

- Several delegates, starting with Margaret Hedstrom, said that we needed to address the myth that we are able to save everything, noting that not even Google - with its massive resources - could afford to keep everything. Others supported this point of view, e.g. Fran Berman noted that not all scientific data needed to be kept, e.g. in some cases the costs of resimulation would actually be cheaper than storing data. Responding a comment by Kristine Hanna (Internet Archive) that we could not know exactly what content would be of value to the future, Hedstrom noted that even the collection of static Web pages only - on the Internet Archive model - was already producing massive amounts of content. David Rosenthal added that current Web harvesting techniques failed to deal with the range of technologies categorised as Web 2.0. It is timely to engage the Web 2.0 community to start thinking about the long-term availability of the content they create.

- Other comments referred to the need to liaise better with those who create (or fund) content. Part of this related to advocacy, e.g. being better able to articulate the value of digital resources to society, Fran Berman commenting that around $1.5 of research depends on the existence of the Protein Data Bank (http://www.rcsb.org/pdb/). Paul Ayris (University College London) stressed the importance of advocacy and thought that there could be lessons from the debates on open access (OA), an issue that has been successfully brought to the attention of policy makers and funding bodies. With regard to the role of creators, Jessie Hey (University of Southampton) noted that skills and knowledge differed widely. Abby Smith (Library of Congress) added that in the real world it was not always possible to influence creators in the desired way, citing the management aphorism, "culture eats strategy for breakfast," a phrase that recognises that business strategies need to take into account the underlying culture of organisations. Sheila Anderson stressed the potential role of funding bodies and argued that they needed to take more responsibility for the preservation of the data that they fund, e.g., on the lines of the Arts and Humanities Research Council in the UK.

- Another subject that came up for discussion was the exact role of "significant properties" *vis-à-vis* costs, Caroline Arms (Library of Congress) and others being keen to see some identification of the different cost factors that would apply to preserving those characteristics of objects that are deemed essential.

- A final suggestion was that adoption of the X-PRIZE (http://www.xprizefoundation.com/) funding model - which fosters innovation through competition - could facilitate fast progress in areas identified as needing more immediate attention.

4.3 Distributed infrastructure interoperability models and services

The third group to report back had been looking in more detail at distributed infrastructures and interoperability and had been led by Martha Anderson (Library of Congress) and Seamus Ross (University of Glasgow). The group's report, given by Kevin Ashley (University of London Computer Centre), first provided some random definitions of what interoperability might mean in the preservation context. Points raised in the presentation largely focused on the technical aspects of interoperability, for example the development of common interfaces that enable repositories (or other services within a preservation network) to work together, or for supporting the seamless interchangeability of components (or tools) within preservation systems. It was agreed that interoperability was to some extent related to the development and adoption of common standards, but Ashley argued that it was as much concerned with the development of shared ways of thinking, citing the experience of the JISC-funded LIFE (Life Cycle Information For E-Literature) project in developing a generic approach to modelling preservation costs (McLeod, Wheatley & Ayris, 2006). Thinking about Donald Waters's interoperability gap, the group emphasised the importance of shared understandings, even where standards existed and had been implemented. Another crucial point raised was that interoperability has costs, in that a system designed to support interoperability will not always be optimal in other contexts. In some senses, interoperability can be a trade off between short-term optimality and potential longer-term benefits. In making decisions about this, therefore, it is important to consider carefully the exact reasons why systems might differ, and to evaluate whether facilitating interoperability would have a beneficial role in reducing costs or in facilitating sharing data or technologies with other communities. Distributing specialised tasks between organisations in a distributed interoperable network may be one way of driving down the costs of preservation through economies of scale, but the group noted that, in practice, many organisations have not (thus far) been willing to give up things deemed essential to their preservation mission. Finally, the group proposed a tentative five-layer model of interoperability, encompassing the levels of: bits, metadata, applications and performance (behaviour), policy, and social contexts. Some final discussion points raised various issues about the role of distributed infrastructures, about trust models, the problems of monoculture, and scalability.

The following discussion, convened by Helen Tibbo, raised a number of important topics, including the following:

- There were a number of comments on the difficulty of achieving interoperability over time (temporal interoperability), e.g. the difficulty of planning 'plug and play' functionality for systems that had not been designed yet. It was argued that if effective interoperability cannot be achieved in the present, it would probably not be successful in the future. Joe Kopena (Drexel University) noted the problems associated with longer-term changes in semantics and metadata structures.

- There were a number of comments on the trust models. While acknowledging the importance of trust, some participants suggested instead the adoption of threat models, a technique used in software engineering to help identify and respond to security threats. While some commented on the importance of contractual agreements, others argued that legal agreements had no value unless they could be verified. David Rosenthal, eager to avoid over-optimism in engineering, suggested that the attitude should be to "trust, but verify." Chris Rusbridge (Digital Curation Centre) made the more general point that within a distributed

network infrastructure, organisations will import the risk profiles of all the organisations they are dependent on, making such risks much more difficult to evaluate.

- There was quite a lot of discussion about whether distributed infrastructures were actually necessary. Fran Berman noted that distributed systems were now widely used in scientific research, but things like replication and security could be costly. Kevin Ashley suggested that we needed to characterise when 'distribution' was good and when it was bad. David Giaretta (Council for the Central Laboratory of the Research Councils) thought that systems should be distributed when necessary, but noted the importance of appropriate identifiers. Several speakers raised the topic of institutional repositories, e.g. introducing the JISC-funded Arts and Humanities Data Service project SHERPA DP (http://ahds.ac.uk/about/projects/sherpa-dp/), which has been exploring a disaggregated model that defines the respective requirements - e.g. with regard to metadata - of institutional repositories and the third party services to which they could supply content for long-term preservation.

- Margaret Hedstrom suggested that one area where we can already work together is with the preservation planning entity defined in the Reference Model for an Open Archival Information System (OAIS). Adrian Brown (The National Archives) commented that the PLANETS project was already committed to looking at preservation planning and how it can be tailored to particular organisational needs.

4.4 Social, economic, and policy issues

The final breakout group looked at socio-economic and policy issues, and was led by Abby Smith and Chris Rusbridge. The report of the group's discussions, by Keith Johnston (Stanford University), largely focused on the importance of articulating the value of digital preservation and the related issue of trust. In this regard, it was noted that value is situational - i.e. dependent on context - and can also change over time. There is a need to find ways of articulating value and, in particular, the differences between use value and re-use value. Trust was seen as the key to maintaining the value of information over time, prompting a discussion on the nature of trusted institutions, i.e. those we can trust to do things correctly. The group argued that a key feature of trust was transparency. In helping to spread awareness about digital preservation into the wider culture, it was suggested that certain topics - e.g. the potential loss of personal information - could be used to help people understand the wider context. It was also suggested (by Sayeed Choudhury of Johns Hopkins University) that digital preservation advocates could perhaps learn lessons from the development of the environmental movement, citing the long-term influence of Rachel Carson's 1962 book *Silent Spring*.

The discussion focused on some of the same issues.

- It was noted that disasters like Hurricane Katrina were a 'wake-up call' for organisations for which records - both paper and digital - are essential for continuity of operations. Abby Smith argued that funding bodies (and others) needed to articulate the economic value of investing in and preserving information, e.g. by producing use cases that demonstrate how information can be reused. Chris Rusbridge challenged organisations to articulate the value of digital information by paying for it, not by waiting for additional funding to become available, but by building sustainability by changing institutional priorities and practices.

- There was some discussion of organisational models, especially for shared infrastructure initiatives like the Global Digital Format Registry (GDFR), where issues of long-term hosting and control are of concern. Kevin Ashley raised again the free-rider problem, commenting that the UK Government might legitimately ask why it alone - through The National Archives - funds the PRONOM registry (http://www.nationalarchives.gov.uk/pronom/), when it is of wider benefit. David Rosenthal said that the information in registries needed to be both authoritative and up-to-date and wondered if this could be achieved in a distributed model.

- With regard to policy frameworks, several participants commented that it would be useful to have a common way of expressing local policies, e.g. for collection development or access control, although this is likely to be complex. There was some debate as to whether these

frameworks needed to be immediately 'machine-expressible' - as with rights expression languages (e.g., Coyle, 2004) - but at the very least they should be able to help organisations develop policies of their own. As a first step, William LeFurgy suggested that we should look in more detail at existing workflows and practice and use the knowledge gained to inform the development of policies.

5. Summing-up

Following Carlos Oliveira's brief introduction to European Union activities (for summary see section 2.3, above), **Clifford Lynch**, Director of the Coalition for Networked Information, provided a short summing-up of the workshop.

He started with some comments on Donald Waters presentation, which had reminded delegates of the questions being asked a decade ago, noting where progress has been made and where it has not. With regard to aggressive rescue, Lynch noted the experience of the Internet Archive in undertaking rescue on a heroic scale of the surface Web. He noted that, in doing this, the archive had reframed traditional IP rights practice by means of an innovative "opt-out" model, whereby robots exclusion tools are used to enable content owners to prevent material being accessed. On IP barriers, Lynch reported that progress had been mixed. There had been a number of relevant court cases, but few were explicitly concerned with preservation. He suggested that there had been some progress on developing legislation, citing Section 108 revision in the US and the approach of some national libraries with regard to the harvesting of Web content in lieu of legal deposit. In thinking about the trust issue, Lynch noted that there had been progress on one level - e.g. in terms of defining what constitutes trusted (i.e. competent) repositories - but that that there were a number of additional challenges.

Lynch then introduced a number of additional themes that he felt were important.

His first comments were on the components of a common infrastructure, noting that bit-level preservation was at the very bottom level but that pathetically little progress had been made in this area. The need for survivable bits is one that extends well beyond the preservation communities - as evidenced by the experience of business post 9/11 and Hurricane Katrina. More research was needed into whether there could be economic and technical tradeoffs between business and preservation requirements in this area. Lynch noted that there was a potential opportunity to take this forward in the next few years as the current academic network backbone in the US (Abilene) is due to be replaced. He said that we will need a geographically distributed, highly survivable, bit management and storage infrastructure, arguing that these could even make use of international networks, noting the existence of good networks in the UK and relatively abundant fibre between the two countries.

A second set of comments related to ingest. Lynch commented that the experience of things like NDIIPP's ingest and handling test underscored the importance of work on repository interoperability. On the robustness of ingest, he felt that we were probably mistaken in our assumptions that ingested materials would be well structured and documented. Instead, we needed to be realistic and should insist on the development of *robust* software tools that can check object validity and, where necessary, undertake error correction or recovery.

Another comment related to the range of content types being considered by the organisations involved in the workshop, e.g. scientific datasets, scholarly communication, multimedia, and the Web. Lynch felt that during the workshop discussions there were a number of conversations that were going past each other, partly because the specific needs of each genre of material were quite different (above the bit level). He argued that some areas were getting mature enough to be taking on their own character and thought that it might be time to move towards discussion in more specialised groups.

Lynch identified a potential shortcoming in that there was little effort being spent on identifying content that is in imminent danger, i.e. those 'train wrecks' unravelling right now. One specific area he mentioned was the urgent need for personal papers to enter larger collections at the end

of their lives, noting that there was little work in this area at present, except the JISC-funded paradigm (Personal Archives Accessible in Digital Media) project (http://www.paradigm.ac.uk/).

He then returned to the topic of trust, noting that the breakout discussions were painful because delegates were talking about a vast number of different things. He was keen that the community should separate the discussion of trust in engineering or the design of distributed systems - which are areas in which we cannot generally have trust - from the need for institutional interdependence in the cultural memory sector. He said that we could not predicate institutional trust only on the ability to audit colleagues. More widely, the public needs to trust that cultural memory organisations will be good stewards, and that governments will keep the records that it needs to keep. Lynch added that he thought that it would be useful to come up with a framework to help us from getting lost in the morass of things that we call trust.

Lynch's final points related to the challenge of facilitating public understanding of the importance of digital preservation. He raised again the analogy previously made with the growth of the environmental movement - noting that the publication of *Silent Spring* was accompanied by a range of related legal and public policy work (e.g. by the lawyer Joseph L. Sax). Something similar needs to be done with regard to educating the public about the importance of cultural memory and how it interacts with current IP law. One potential opening may be to focus on what might happen to the personal digital collections of individuals, currently mostly photographs, music and files, but likely to get more complicated in the future. This may eventually lead us to reconsider fundamental principles, e.g. about the standing of the historical record in law.

6. Conclusions

In conclusion, **Neil Beagrie** thanked those who had organised the workshop and made it run so smoothly, acknowledging the good interaction at meals and the poster sessions. After providing some practical information on the future publication of workshop materials, he finished with some personal reflections. Firstly, he thought that we needed to think in more detail about how we should work with those in government or industry whose preservation (or records management) needs are shorter term but share some of the same problems, e.g. 'intermediate preservation' for between 5 and 75 years. Secondly, echoing Clifford Lynch, he said that we should focus on the impact of digital preservation on individual citizens, noting a growing investment in digital photographs, medical records and life-long learning spaces. Thirdly, he thought that some type of 'X-Challenge' competition for digital preservation might be an idea worth scoping and exploring further, especially for areas where we want things to happen quickly. He concluded with another reminder of Donald Waters's four grand challenges, adding another one related to skills and training.

The workshop agenda, presentation slides, notes from the breakout sessions and other workshop materials are now available from the JISC Web site: (http://www.jisc.ac.uk/index.cfm?name=ndiip_jisc).

References

Beagrie, N. (2002). *A Continuing Access and Digital Preservation Strategy for the Joint Information Systems Committee (JISC), 2002-2005.* London: Joint Information Systems Committee, October. Retrieved August 17, 2006, from: http://www.jisc.ac.uk/index.cfm?name=pres_continuing

Coyle, K. (2004). *Rights expression languages: a report for the Library of Congress.* Retrieved August 17, 2006, from: http://www.loc.gov/standards/relreport.pdf

Day, K. (2006). "Morgan Stanley will pay fine over e-mails." *Washington Post*, 11 May, p. D3.

Feeny, M., ed. (1999). *Digital culture: maximising the nation's investment.* London: National Preservation Office.

Garrett, J., & Waters, D., eds. (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Washington, D.C.: Commission on Preservation and Access;

Mountain View, Calif.: Research Libraries Group. Retrieved August 17, 2006, from: http://www.rlg.org/legacy/ftpd/pub/archtf/final-report.pdf

HM Treasury, Department of Trade and Industry, & Department for Education and Skills. (2004). *Science & Innovation Investment Framework, 2004-2014*. Norwich: The Stationery Office. Retrieved August 17, 2006, from: http://www.hm-treasury.gov.uk/spending_review/spend_sr04/associated_documents/spending_sr04_science.cfm

Keller, M. A., Reich, V. A., & Herkovic, A. C. (2003). "What is a library anymore, anyway?" *First Monday*, 8(5), May. Retrieved August 17, 2006, from: http://www.firstmonday.org/issues/issue8_5/keller/

McLeod, R., Wheatley, P., & Ayris, P. (2006). *Lifecycle information for e-literature: full report from the LIFE project*. Retrieved August 17, 2006, from: http://eprints.ucl.ac.uk/archive/00001854/

National Science Board. (2005). *Long-lived data collections: enabling research and education in the 21st century*. Retrieved August 17, 2006, from: http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf

National Science Foundation. (2004). NSF Digital Archiving and Long-Term Preservation (DIGARCH) program solicitation (NSF 04-592). Retrieved August 17, 2006, from: http://www.nsf.gov/pubs/2004/nsf04592/nsf04592.htm

O'Donnell, J., (2003) "Re: vanishing act." E-mail to liblicense-l list <liblicense-l@lists.yale.edu>, 29 January. Cited in: Keller, Reich & Herkovic (2003).

PREMIS Working Group. (2005). *Data dictionary for preservation metadata*, Dublin, Ohio: OCLC Online Computer Library Center, 2005. Retrieved August 17, 2006, from http://www.oclc.org/research/projects/pmwg/

RLG/OCLC Working Group on Digital Archive Attributes. (2002). *Trusted digital repositories: attributes and responsibilities*. Mountain View, Calif.: Research Libraries Group, May. Retrieved August 17, 2006, from: http://www.rlg.org/longterm/repositories.pdf

RLG-NARA Working Group on Digital Repository Certification. (2005). *An audit checklist for the certification of trusted digital repositories: draft for public comment*. Mountain View, Calif.: RLG, August. Retrieved August 17, 2006, from: http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Shirky, C. (2005). *Library of Congress Archive and Handling Test (AIHT): final report*. National Digital Information Infrastructure and Preservation Program, June. Retrieved August 17, 2006, from: http://digitalpreservation.gov/technical/aiht.html

Waller, M., Sharpe, R. (2006). *Mind the Gap: Assessing Digital Preservation Needs in the UK*. York: Digital Preservation Coalition, February. Retrieved August 17, 2006, from: http://www.dpconline.org/docs/reports/uknamindthegap.pdf

Waters, D. (2002). "Good archives make good scholars: reflections on recent steps toward the archiving of digital information." In: *The State of Digital Preservation: An International Perspective*, Washington, D.C.: Council on Library and Information Resources, pp. 78-95. Retrieved August 17, 2006, from: http://www.clir.org/pubs/abstract/pub107abst.html

Web references

Arts and Humanities Data Service: http://www.ahds.ac.uk/

Augmenting Interoperability Across Scholarly Repositories meeting, New York, April 20-21, 2006: http://msc.mellon.org/Meetings/Interop/

CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval: http://www.casparpreserves.eu/

Center for Research Libraries, Certification of Digital Archives: http://www.crl.edu/content.asp?l1=13&l2=58&l3=142

CLOCKSS - Controlled LOCKSS: http://www.lockss.org/clockss/Home

Digital Curation Centre: http://www.dcc.ac.uk/

Digital Preservation Europe: http://www.digitalpreservationeurope.eu/

European Commission Information Society and Media Directorate-General: http://ec.europa.eu/comm/dgs/information_society/index_en.htm

European Commission, i2010 Digital Libraries Initiative: http://europa.eu.int/information_society/activities/digital_libraries/index_en.htm

Global Digital Format Registry (GDFR): http://hul.harvard.edu/gdfr/

International Internet Preservation Coalition: http://netpreserve.org/

JISC Capital Programme: http://www.jisc.ac.uk/capital.html

JISC Repositories and Preservation Programme (Phase Two): http://www.jisc.ac.uk/rep_pres.html

JISC Supporting Digital Preservation and Asset Management in Institutions programme: http://www.jisc.ac.uk/index.cfm?name=programme_404

Joint Information Systems Committee: http://www.jisc.ac.uk/

National Center for Biotechnology Information, Archiving and Interchange DTD: http://dtd.nlm.nih.gov/

National Digital Information Infrastructure and Preservation Program: http://www.digitalpreservation.gov/

National Science Foundation: http://www.nsf.gov/

NDIIPP partnerships: http://digitalpreservation.gov/partners/project.html

Paradigm (Personal Archives Accessible in Digital Media) project: http://www.paradigm.ac.uk/

Planets project: http://www.planets-project.eu/

Portico: http://www.portico.org/

Protein Data Bank (PDB): http://www.rcsb.org/pdb/

Section 108 Study Group: http://www.loc.gov/section108/

SHERPA DP project: http://ahds.ac.uk/about/projects/sherpa-dp/

The National Archives, PRONOM technical registry: http://www.nationalarchives.gov.uk/pronom/

UK Data Archive: http://www.data-archive.ac.uk/

X-PRIZE: http://www.xprizefoundation.com/index.asp