19 October 2001

This was an invitational seminar sponsored by the Digital Preservation Coalition and the British National Space Centre. The seminar aimed to raise the profile of the Open Archival Information System Reference Model (OAIS) standard in the UK and share practical experience of digital curation in the digital library sector, archives, and e-sciences.

Presentations

Digital Curation: digital archives, libraries and e-science - report by Philip Pothen JISC

A report on an invitational seminar held on the 19th October at 75-79 York Rd, and sponsored by the Digital Preservation Coalition and the British National Space Centre

Neil Beagrie (JISC Digital Preservation Focus and Secretary, Digital Preservation Coalition) welcomed the guests to the seminar and outlined the main objectives of the event; it would attempt, he suggested, to

- raise the profile of relevant UK and international standards and "hands-on" initiatives in the UK;
- show their application in the sciences, libraries and archives, and in the educational sector and beyond;
- illustrate their role in securing and promoting access to the digital outputs of research and other activities for current and future generations.

Three developments have been key to the timing and organisation of this event:

- the imminent approval of the Open Archival Information Systems (OAIS) Reference Model as an ISO standard;
- the launch of the Digital Preservation Coalition (DPC), a major cross-sectoral coalition of

over 15 major organisations in the field;
- and the development of the e-science programme to develop the research grid.

Session 1. The OAIS Reference Model and digital archive certification

Lou Reich of NASA spoke about how NASA and the Consultative Committee for Space Data Systems (CCSDS) had been central to the development of the OAIS Reference Model, but how they had ensured widespread consultation and cooperation with the archive community, both in the US and internationally. The resulting model had been developed, therefore as an 'open' and public model and was already being widely adopted as a starting point in digital preservation efforts. Lou Reich explained that a new version of the OAIS Reference Model was delivered to the ISO and CCSDS Secretariats in July 2001 for a two month public review period, and a final standard should be produced in late 2001.

Lou Reich provided a summary of some of the main characteristics of the OAIS Reference Model: it can be applied to all digital archives, and their Producers and Consumers; it identifies a minimum set of responsibilities for an archive to claim it is an OAIS; it establishes common terms and concepts for comparing implementations, but does not specify an implementation; and it provides detailed models of both archival functions and archival information. Dr Reich concluded by outlining some of the use and implementation efforts of the OAIS Model by the science, library and archive communities, including the Networked European Deposit Library (NEDLIB), the National Library of Australia, CEDARS, NSSDC (National Space Science Data Center), the National Archives and Records Administration (NARA), and others.

Bruce Ambacher of the US National Archives and Records Administration spoke about the development of OAIS, particularly in the areas of Ingest, Identification and Certification. Through the October 1999 Archival Workshop on Identification, Ingest and Certification (AWIICS), Mr Ambacher was particularly involved in the area of Certification; the AWIICS Certification Working Group developed a preliminary checklist for certification that develops best practices and procedures for each aspect of the OAIS model, including legal issues, mission plans, compliance with relevant regulations, relationships with data providers, ingest procedures, data fidelity and life-cycle maintenance. The workshop also acknowledged that the full range of best practices was not yet in place.

Mr Ambacher summarised the current initiatives that are pointing the way towards developing a suite of approaches that will support certification, including the InterPARES reports, the RLG

and OCLC report *Attributes of a Trusted Digital Repository for Digital Materials*, the Global Electronic Records Association, the JISC's standards and guidelines for the DNER, and many others.

Robin Dale of RLG spoke on RLG and OCLC report *Attributes of a Trusted Digital Repository for Digital Materials* . She re-emphasised the importance of certification as a key component of a *trusted* digital repository; self-assessment, she said, will not always be adequate. There is a need, therefore, for certification practices to be formalized and made explicit. The AWIICS draft report had suggested the need for an official certifying body, for identifying the attributes to be measured and to define the conditions of revocation of certification. But many questions still remained to be answered, including, who will be on such a body, who will set up this body and which stakeholders will be represented on it.

David Ryan of the UK Public Record Office spoke about some of the challenges facing the PRO in terms of preservation of public records, technical obsolescence and physical deterioration. Once again he emphasised the importance of collaborating over these issues with a range of interested parties - scientists, users, archivists and technical staff. Mr Ryan outlined the PRO mandate to store and make available comprehensive 'born digital' public records and how its activities in this area were a core part of the PRO e-business strategy.

Part of the PRO's efforts in "hands-on" e-preservation would be development of a database PRONOM to support a technology watch function. This would aim to identify and warn of impending changes to hardware and software and risks of obsolescence.

The importance to users of *interaction* with preserved digital information, rather than simply using that information as an historical record (as is more likely the case with printed data) was made, and this gave an added importance to the work of the PRO in its e-preservation activities. The PRO is collaborating therefore not only with Government bodies, but is also a founder member of the DPC. Appraisal is crucial to preservation activities, said Mr Ryan, and this means talking to users ('talking to the cook and not the chef') in a meaningful and collaborative way.

An interesting discussion followed in which the relative costs of printed and digital storage were discussed. There was vigorous debate over whether preservation of digital materials would

actually be more expensive than original materials in other media. The duty of care and costs associated with traditional special collections and important archives was cited. With digital storage the costs of computer storage are diminishing constantly so costs would be related primarily to staff effort required for long-term preservation. The degree of automation which could be implemented for future migrations and preservation efforts, would therefore be critical to relative and absolute costs over the long-term. It was argued that issues such as appraisal and migration represented costs that were ongoing. It would be easy to underestimate the costs of long-term digital preservation where it was dependent on human intervention and perhaps could not be scaled across collections. Other issues discussed at this stage were the importance of metadata to long-term preservation, and the question of selection, particularly pertinent to the PRO and its efforts to secure the preservation of the UK public record.

Session 2. Data Curation and the Grid

Professor Tony Hey, Director of the UK e-science programme, began by stating that e-science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it. He quoted John Taylor, Director General of the Research Councils who said that 'e-science will change the dynamic of the way science is undertaken.' NASA's Information Power Grid has promoted a revolution in how NASA addresses large-scale science engineering problems by providing a persistent infrastructure for 'highly capable' computing and data management services. The Grid, by providing pervasive, dependable and inexpensive access to advanced computational capabilities will provide the infrastructure for e-science, said Professor Hey.

The UK e-science initiative represents £120m worth of funding over the next three years to provide next generation IT infrastructure to support e-science and business, £75m of which is for Grid applications in all areas of science and engineering, £10m for the supercomputer and £35m for the Grid middleware. It uses SuperJANET and all e-science centres will donate resources to form a UK National Grid. Professor Hey outlined some of the projects being funded under this programme, including the Comb-e-Chem project which will integrate structure and property data sources within a knowledge environment to find new chemical compounds with desirable properties; My Grid, and the PPARC e-science projects, such as GridPP and AstroGrid.

Peter Dukes from the Medical Research Council outlined the overall scope of the MRC's programme as well as its current work to develop an archiving policy. The MRC has a number of data sets, including genome databases, genetic databases and populations databases. It has made a considerable investment in its populations studies databases which has meant that a

management policy governing both archiving and access has been crucial. Outlining some of the access issues, such as rights and ownership issues, consent and ethical issues, Peter Dukes detailed some of the central concerns in the next stage of the MRC's development of an archiving policy, such as the use of case studies to examine current practices, skills and costs, the potential for a specialised data centre, as well as the specific issues of metadata, ownership and the management of access. It was clear that the research Grid would provide tremendous opportunities for advancing science and that work on research data policies and practice will also help unlock the potential of the Grid for collaborative scientific research.

David Boyd of the CLRC e-science Centre looked at how the Grid can help some of the problems involved in scientific data curation. The problems are of a rapidly increasing capability to generate data in many different formats in the physical and life sciences, the increasingly expensive facilities needed to generate this data, the irreplaceability of much of the data and the increasing need for access to be on a global scale.

The problems of data curation in this environment are immense too: the question of who owns the data is often unclear, which can mean that responsibilities for curation are unclear, resulting in it often not getting done; the lack of a culture of data curation, and the lack of recognition in some quarters that data is a significant long term asset which requires investment. The Grid can help with some of these problems by offering a mechanism for control and access (through authentication and authorisation, etc.), by making the location and the existence of data more visible, by providing easier access to data and by facilitating distributed collaborative working by sharing data. Dr Boyd concluded by outlining some of the CLRC's current activities, including the development of a portal for accessing scientific data, operating a large-scale data archive, as well as participating in global standards activities.

Paul Jeffreys, Director of the Oxford e-Science Centre, spoke about its recent launch, and the management structures that will enable managerial, technical and user concerns to be integrated within the activities of the Centre. He spoke too about the Oxford-wide collaboration that the centre is involved in, such as the work with the Oxford Digital Library, the Oxford Text Archive and Humbul. Although, Dr Jeffreys said, global science is driving the initiative, the interest is much wider, and these areas of collaboration suggest that the centre will become a core part of the University's life.

A discussion followed in which it was asked whether, given the nature of the collaboration that Paul Jeffreys had outlined, whether there was a place for greater cooperation between the Grid and the Arts and Humanities community, in particular the Arts and Humanities Research Board. Researchers in the Arts and Humanities do not generate the vast amounts of data that those in

the sciences did, but they have need of different data types (video, for example), a need to overcome certain traditional cultural barriers to the use of digital information. There was therefore a need for their involvement in wider developments.

Another key issue that came up in this discussion session was the need to look at data policies, archival models, and how to incentivize the submission of primary research in digital form and appropriate metadata. Ideas put forward included: financial incentives (perhaps linking part of the research grant funding to archiving as has already done by some research councils); through increasing and enhancing recognition of the value of digital resources in general among the research and scholarly community; to persuading funding councils, the RAE and publishers to take these matters more seriously and to build such considerations into their funding and reward processes. An interesting example was given of linkage between primary data and publication articles and how this had provided the incentive for researchers to complete and submit the primary research archive to a high standard. It was also recognised that project funding while valuable in targeting research on current needs and tight deliverables tended to ignore long-term data needs and the infrastructure to support it. This was something the research sector would need to address to ensure an appropriate balance.

Session. 3. Curation of Digital Collections

Maggie Jones and **Derek Sergeant** from the CEDARS project funded by JISC, explained how the project had so far delivered the CEDARS Demonstrator Archive as well as the CEDARS Preservation Metadata outline specification, both based on the OAIS Reference Model, which has been a significant influence on the project and one whose development the project had, in fact, contributed to. Plans for the forthcoming extension year including a major redesign of the CEDARS web site, the production of a series of guidance documents and the hosting of an invitation-only workshop in early 2002, which would involve all of the organisations that CEDARS has been collaborating with. Some of the lessons of the project were outlined, including the centrality of metadata to the preservation of resources, and the increasing consensus that is emerging about standards.

Deborah Woodyard, Digital Preservation Coordinator at the British Library, outlined the BL's main activities in the area of digital curation and preservation. At the moment the BL's digital collecting was on the basis of a voluntary deposit, along with purchases made by the BL, as well as created digital resources undertaken by the BL itself. Among its main priorities, Deborah Woodyard said, was to ensure improved coverage of the UK's National Published Archive, to increase the collection of digital materials and to continue to collaborate with other major players in the field. There was also a major consideration, in keeping with the BL's policy in

other areas, and that is to make the library's digital collections more accessible to users.

The BL's Digital Library Store aims to support the storage and long-term preservation of digital materials. It is based on the OAIS standard which, Deborah Woodyard said, had helped the understanding of the information objects that were being preserved, clarified functionality and had helped provide a common language for its curation and preservation activities. Once again, collaboration was seen to be central to these activities, with the BL being involved with a number of key organisations and projects in this field including the DPC, the National Library of the Netherlands, the OCLC-RLG Preservation Metadata and Attributes of a Trusted Digital Repository Working Groups, and the CEDARS project.

Kevin Ashley of the University of London Computing Centre (ULCC) and the National Digital Archive of Datasets (NDAD) spoke about ULCC's role under contract to the PRO and others for digital preservation and their practical experience of running a digital preservation service. He also covered the history of mass storage of information, and of the different forms of archival resources. He noted most discussion centres on digital forms of preservation metadata, but it is important too to recognise that some forms of metadata are in non-digital format - manuals, specifications, a person's individual expertise, etc that was also an important consideration in preservation of materials. Another important issue was the role of third parties such as the data creators and departments who have a different agenda and priorities and how this impacts on preservation.

Kevin Ashley also spoke about the OAIS model; it advantages were clear, he said in that it eases procurement of hardware and software, and interworking with compliant systems, as well as migration tasks, but there are question marks about interworking with traditional repositories, as well as its working with mixed-mode models, questions which will need to be looked at closely in the future.

A discussion followed on the potential value and limitations of the OAIS model. Its value in the early stages of system design and development was recognised but at the same time it would not provided the detailed implementations and practice. Documenting and sharing practical experience in this area will be vital.

The difficulties and importance for archives in working with disparate data creators and departments in either industry or the public sector were also discussed. It was recognised that this often would require a cultural change process and outreach to work with data creators. The

increasing role of spatial data and Geographical Information Systems (GIS) in organisations was cited as one factor which is giving increasing prominance to corporate datasets, archiving and standards.

Session. 4. The Way Forward

In the final session from the day Neil Beagrie, David Giaretta, and Tony Hey reflected on the seminar and ways forward.

David Giaretta (Rutherford Appleton Laboratory/BNSC and chair of CCSDS panel developing the OAIS model) noted the next international CCSDS meeting which would discuss OAIS and archive certification was being held in Toulouse in the following week. He would report on the UK seminar and its discussion. He felt the seminar had been exceptionally valuable and it would be important to continue the momentum and progress it had achieved. It was also important to continue co-ordination across sectors and the Coalition could be immensely valuable in achieving this. Tony Hey suggested collaborating with the e-science institute in Edinburgh to arrange further follow-on meetings.

Neil Beagrie highlighted a number of additional areas. He welcomed the presence of his colleague Louise Edwards from JISC at the seminar who was working on the primary research data in the JISC Collections Policy. Clearly the use of data by the sector and closer involvement between JISC and the research councils to support users of the Grid and primary research data would be important. The creation of a JISC research committee chaired by Tony Hey could clearly have an important role in this area. He also wished to thank everyone for their contribution and felt sure that members of the Digital Preservation Coalition would be willing to initiate and participate in future seminars. He hoped to see close links with e-science and a growing membership of the DPC amongst the research councils and data centres.

In developing digital archive certification it was suggested we may need a two track process: some rapid prototyping and implementations e.g. the RLG/OCLC attributes work or the JISC information environment, etc, and an evolving standards process -- hopefully getting both practice and theory in a feedback loop.

The topic of linkage between "digital libraries", data curation and preservation research and the

Grid touched on during the seminar was reviewed. Tony Hey noted he was open to discussion of possible projects in data curation or indeed other areas raised during the seminar but it was important to note the need currently for industry involvement and funding in such proposals.

Finally David Giaretta noted a meeting report was being prepared and would be made available on the Web shortly with the speakers' presentations. The seminar was concluded by all participants thanking the speakers and the DPC and BNSC for an extremely lively and stimulating day on a key topic of cross-sectoral interest.

Philip Pothen JISC.

End of Meeting Report

Participant Evaluation of Sessions (RTF 5KB)

Presentations

To open PDFs you will need Adobe Reader

Welcome and Introduction (RTF 9KB) - Neil Beagrie (JISC)

Session 1: OAIS and Digital Archive Certification
The OAIS Reference Model Lou Reich (NASA) (PDF 140KB)
Panel Presentations
Bruce Ambacher (NARA) (PDF 86KB)
Robin Dale (Research Libraries Group) (PDF 23KB)
David Ryan (UK Public Records Office) (PDF 125KB)

Session 2: Data Curation and the Grid
E-science and the Research Grid (PDF 684KB) - Tony Hey (Office of Science and Technology/EPSRC)
Panel Presentations

Peter Dukes (Medical Research Council) (PDF 56KB)
David Boyd (Central Laboratory of the Research Councils) (PDF 28KB)
Paul Jeffreys (Oxford University) (PDF 819KB)

Session 3: Curation of Digital Collections
Derek Sergeant & Maggie Jones (CEDARS Project) (PDF 259KB)
Deb Woodyard (British Library) (PDF 163KB)
Kevin Ashley (University of London Computing Centre) (PDF 19KB)

Delegate List (RTF 11KB)