The International Journal of Digital Curation

Issue 2, Volume 3 | 2008

The Fifth International Conference on Preservation of Digital Objects (iPRES 2008)

Alex Ball, Michael Day and Manjula Patel

UKOLN,

University of Bath

October 2008

Summary

The Fifth International Conference on Preservation of Digital Objects was held at the British Library on September 29–30, 2008, with the theme "Joined Up and Working: Tools and Methods for Digital Preservation". Topics ranged from the technical foundations of digital preservation through preservation system architectures to the organizational and policy issues facing the custodians of digital resources. There were also sessions dedicated to dealing with particular types of content, training needs, and methods for auditing needs and services.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

From its roots as a Chinese-European workshop on digital preservation, the iPRES conference series has grown to become a fully international forum for researchers and practitioners in the field to exchange ideas and expertise. Over 250 delegates from 33 different countries converged on the British Library in London for the Fifth International Conference on Preservation of Digital Objects, held on September 29–30, 2008. This was the first year that papers were collected, and so full proceedings are available from the conference Web site as well as copies of the presentation slides (British Library, n.d.).

The theme for the conference was "Joined Up and Working: Tools and Methods for Digital Preservation". The theme proved to be more aspirational than descriptive, but there were promising signs, not least the presence of commercial vendors who are putting the research into practice.

With 54 papers to fit in, the conference was split into a practitioner track and a technical track. On the first day, the conference alternated between plenary and parallel sessions, while on the second day all the sessions had parallel tracks.

Day One

The proceedings commenced with a keynote speech from **Lynne Brindley**, Chief Executive of the British Library, who portrayed digital preservation as a sort of jigsaw puzzle: we know the end goal (the picture on the box) and we have a lot of disconnected tools to help us achieve it (the pieces of the puzzle); the challenge is to join them all up, and it will require some considerable effort. Brindley also argued that it was time to drop the means-focussed term "digital preservation" and instead use something more end-focussed, such as "digital access forever"; though one imagines it will be some time before the programme committee changes the title of the conference to iPERM (for "permanent access").

Session One: Modelling Organizational Goals

The first session was dedicated to the organizational context. **Neil Beagrie** and **Najla Rettberg** (Charles Beagrie Ltd) discussed digital preservation policies in UK Higher Education Institutions (HEIs) – or rather, those institutional policies that happen to mention or have implications for digital preservation – while **Angela Dappert** (British Library) of the Planets Project presented a promising conceptual model for expressing the values, goals and requirements that underlie an organization's preservation activities. **Barbara Sierman** (Koninklijke Bibliotheek) and **Raymond van Diessen** (IBM Netherlands) explained how the National Library of the Netherlands (KB) had applied IBM's Component Business Modelling (CBM) method to identify the operational functions that support its strategic goals, and to map these functions to supporting services. This allows the KB to see at a glance where improvements in its IT infrastructure need to be made. Finally, **Susanne Lang** (Bayerische Staatsbibliotek) and **Michael Minkus** (Universität der Bundeswehr München) reported on a study by the Bavarian State Library and the University of the Federal Armed Forces Munich that aimed to produce a sustainable process-orientated business model for organizations involved with digital preservation.

Session Two: Disciplinary Contexts/Digital Preservation Formats

The practitioner track of the second session, *Disciplinary Contexts,* looked at issues of preservation in particular fields. **Sabine Schrimpf** (Deutsche Nationalbibliotek) introduced the field of electronic literature – online poetry/fiction, hypertext fiction, networked writing, computer-generated poetry, etc. – and described how nestor (Network of Expertise in Long Term Storage of Digital Resources), the German Literature Archive (DLA) and the German National Library (DNB) are working together to preserve it. The DLA harvests and catalogues materials after asking permission from authors; the DNB, with its legal deposit mandate, harvests materials from large publishers' Web sites, and relies on smaller publishers and individuals to register their works through an online form. Preserving links and connections within and between works is a particular problem.

Tim Au Yeung (University of Calgary) showed that digital art gives a whole new twist to the issue of deciding which aspects of a digital object to preserve. This was explored in the 2004 Guggenheim Museum exhibition, *Seeing Double* which compared original new media works with emulated versions. In some cases the art lies not in the digital output itself but in the way in which it was generated, so perhaps the answer is to provide sufficient documentation to allow a work to be recreated (Deprocas, 2001; Rinehart, 2007).

Louise Corti (UK Data Archives) described the JISC-funded DExT Project, which aims to devise a standard format for richly encoded qualitative data (interviews, field notes, diaries) and to develop tools to convert data from proprietary statistical formats to open formats. On the first count, the project has produced the QuDEx XML format, which is compatible with the XML output from two major CAQDAS (Computer Assisted Qualitative Data AnalysiS) vendors; packaging is provided by METS. On the second count, an extensible Java-based application has been written to convert SPSS (amongst other things) to and from plain text and Data Documentation Initiative (DDI) 3.0 XML.

Bradley Daigle (University of Virginia) provided advice on preserving digital scholarship outputs and, crucially, the interrelationships between them. He emphasized the need to involve staff and students early on, so that their outputs are created in a preservation-friendly way and the preservation infrastructure meets their requirements. He also encouraged preservation units to be realistic about what they could achieve, to consider the business models for their operation, and to establish memoranda of understanding and service level agreements with their institutions.

Lastly, **Jeremy Leighton John** (British Library) explained how the BL is ingesting eManuscripts: the digital equivalent of personal corpora of letters, diaries, photographs, etc. Three techniques are being examined. The first is computer forensics, where an image of the hard disc contents is created, and the significant files indexed and prepared for extraction to either an exact copy or an interoperable version. The second is ancestral computing, where specialist modern equipment is used to recover data from obsolete media and/or emulate obsolete hardware. The last technique is evolutionary computing – really a group of techniques, including DNA-based digital storage and bioinformatics-inspired text mining.

The technical track, entitled *Digital Preservation Formats,* began with **Oya Rieger** describing how Cornell University Library has been dealing with its rapidly expanding collection of digitized books. In terms of infrastructure, it uses MARC/XML aggregated in XMLTape for metadata, JPEG2000, ARC packages and the aDORe system for archival storage, and a Django-powered user interface for access. The Library had been relying on persistent IDs from Microsoft to link its catalogue records to the digital books, but since the closure of Microsoft's Live Search Books service, it must look elsewhere (probably Fedora) for access middleware. The most challenging aspects of the digitization programme were setting up sufficient storage, and the intellectual property rights (IPR) issues surrounding in-copyright material.

Kam Woods (Indiana University) looked at the technical challenges of providing virtual CD-ROM collections. Quite apart from issues with the original CD-ROMs themselves[1] – hard-coded absolute paths, platform-dependent executables, quirky implementations of ISO 9660 extensions (1988) – there were also problems with misreported volume sizes and Mac-specific resource/data forking in the ISO images. It was interesting that Woods' solution involved both the normalization of file formats (to PDF, HTML or XML) and an emulation technique: mounting the ISO images in virtual machines.

Brian Kelly (UKOLN, University of Bath) gave a brisk overview of JISC's Preservation of Web Resources (PoWR) project. The aim of the project was to produce guidelines for UK HEIs, and as part of the consultation process, two workshops were held: one on "Web 1.0" environments and one on "Web 2.0" environments. The first workshop emphasised the need for a clear driver for Web preservation – comparisons were made between preserving the University home page and preserving old copies of the prospectus. The second produced examples of good practice with respect to wikis, blogs, microblogs, online slides, voice-over-IP telephony and "amplified conferences" (Dempsey, 2007).

Adrian Brown (The National Archives) gave **Amanda Spencer**'s presentation on the approach of the [UK] National Archives (TNA) to collecting and archiving Government Web sites. The issue was raised because Government Web pages are increasingly being cited in parliamentary proceedings, yet have poor persistence. Through a contract with the European Archive, TNA are now undertaking comprehensive snapshots of Government Web sites and adding them to a database. When someone requests a missing page, the servers are now configured to search the archival database for an archived copy of the page, which is presented to the user with a banner explaining that the page is no longer current.

Stephen Abrams (California Digital Library) rounded off the session by introducing the second generation of the JHOVE tool. The new version is now more modular and easier to use as a plugin service, and is no longer restricted to a "one object is one file in one format" data model. It does not, however, support all the formats that the previous version did.

[1] Readers may be interested in *Migration Performance for Legacy Data Access* by Kam Woods and Geoffrey Brown in this same issue.[IJDC]

Session Three: Preservation Planning

The third session comprised two theoretical and two practical papers on preservation planning. **Dirk von Suchodoletz** (University of Freiburg) outlined a strategy for managing emulation-based preservation methods, using the concept of *view paths:* routes through a set of preservation layers such as applications, operating systems or emulated hardware (van Diessen, 2002). **Gareth Knight** (Centre for e-Research, King's College London) reported on the InSPECT Project and other recent research into the significant properties of digital objects – loosely speaking, the characteristics that need to be preserved for the objects to retain their meaning and utility. Knight illustrated the difficulties of consistently judging which properties are significant across a range of object types, and noted that the relationship between significant properties and (OAIS) representation information (Consultative Committee for Space Data Systems, 2002) needed clarification. **Alex Ball** (UKOLN, University of Bath) applied the concept of significant properties to Computer-Aided Design (CAD) models, and showed how they might be used not only to determine adequate preservation strategies but also to devise ways of tailoring CAD information for different roles within engineering firms. Finally, **Mark Guttenbrunner** (Vienna University of Technology) used the Planets preservation planning approach with respect to console video games, and found that while emulation was the most promising technique, finding a suitable emulator involves judicious sampling, careful testing and quite often patching as well.

Session Four: Understanding Costs and Risks/Preservation Metadata

The practitioner track of the final session of the first day was dedicated to understanding the costs and risks of preservation. **Paul Wheatley** (British Library) introduced the LIFE and LIFE² Projects, which aim to provide a methodology for monitoring and ultimately predicting the costs of preservation in both the short and long terms. The methodology is supported by a model of preservation stages (high-level processes) and elements (low-level functions and activities), which is used to break down and identify relevant costs. **Rory McLeod** recounted how the British Library's Digital Preservation Team conducted a risk assessment for the Library's digital collections, and found that almost all the material held on optical media and magnetic tape was at high risk, particularly but not exclusively from media degradation. As a result, the Library must make thoroughgoing changes to the way it handles digital material. **Richard Wright** (BBC) argued that more should be done to mitigate against the costs of small-scale data losses. In particular, archives should take advantage of the falling costs of storage to keep files in uncompressed formats, though the full benefit of this will not be apparent until storage systems and rendering software become better at recovering damaged files. **William LeFurgy** (Library of Congress) and **Adrienne Muir** (Loughborough University) presented the results of the International Study on the Impact of Copyright Law on Digital Preservation (Besek, 2008). There is little international consistency on copyright exemptions for the purposes of digital preservation, leading to confusion among those institutions with digital collections, and stunting the development of systematic approaches to preservation. The Study outlines the legal reform needed to ensure preservation institutions have sufficient rights to preserve their digital holdings.

The technical track concerned preservation metadata. **Arwen Hutt** (University of California, San Diego Libraries) described the metadata requirements that were identified for the Chronopolis Digital Preservation Framework, a grid-based system that, in its first phase, will maintain the integrity of heterogeneous content mirrored at several locations, and deliver content back to the depositor. Not only are these requirements different for archival and dissemination information packages, there are distinctions between manifestation- and copy-level metadata to consider, and corresponding differences in how the metadata are stored and mirrored.

Markus Enders described how the British Library is using METS (Library of Congress, 2008a) to package e-journals in its Digital Library System, extended with PREMIS preservation metadata (PREMIS Working Group, 2005) and MODS (Library of Congress, 2008b) descriptive metadata. The three metadata schemes overlap to some extent, and this is dealt with pragmatically; for the most part, metadata are duplicated between the METS and PREMIS metadata sections, while MODS is not used at the file level, but is the only place where rights information is stored (as this is not used programmatically). **Tobias Steinke** (Deutsche Nationalbibliotek) considered the use of METS for packaging the results of archival crawls of the Web, and concluded that the metadata capabilities of the new WARC standard (ISO/DIS 28500, n.d.) would suffice for emulation purposes and for providing migrated versions, but that a METS structural wrapper would still be of use for managing the WARC files in archival storage.

Ronald J. Murray (Library of Congress) presented a critique of the FRBR conceptual model (International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records, 1998) from the point of view of modern data modelling practice. Having argued that FRBR is too beholden to legacy data models, over-simplistic and under-abstracted, he presented what FRBR might have looked like had it been constructed by an experienced data modeller. The discussion at the end of this session included a call for fuzzy logic to be applied to significant properties: that is, metadata that listed the properties of an information object and recorded its salience for a particular type or group of users, rather than imposing a strict significant/insignificant distinction.

Day Two

The second day of the conference opened with a keynote speech from **Horst Forster**, Director for Digital Content and Cognitive Systems at the European Commission’s Directorate General for Information and Media. Forster pointed out that new methods for document production have largely ignored digital preservation; indeed, while 80% of businesses claim a need to keep documents for fifty years or more, less than 20% have a digital preservation strategy, with the remainder relying on print-outs. IPR issues are a particular concern, as digital formats and media are rendered obsolete long before copyright expires, and current copyright exemptions are not sufficient for digital preservation. Mandates would appear to go some way towards resolving these problems, whether from national governments (in the case of Web archiving) or funding bodies (in the case of peer-reviewed papers).

Session One: National Initiatives/Grid Storage Architecture

The practitioner track of the first session was a panel discussion centred on national and international digital preservation initiatives. **Neil Grindley** (JISC) discussed JISC-funded projects in the areas of eJournal archiving, repositories and preservation, records and asset management, digital object properties, data curation, legal/economic policies and collaboration, and the preservation of web resources. **Martha Anderson**

(Library of Congress) described the [US] National Digital Information Infrastructure and Preservation Programme (NDIIPP), which involves a network of partners working towards the preservation of digital text and datasets, images, audiovisual presentations, geospatial data and Web-based content. The focus is not just on research and development, but also setting up the organizational and technical infrastructure for a co-ordinated and distributed preservation system. **Steve Knight** of the National Library of New Zealand talked about how the Library has implemented its digital preservation system, the National Digital Heritage Archive, in order to preserve all sorts of digital objects, from archival crawls of the Web through born-digital and digitized publications to datasets. Lastly, **Natascha Schumann** (Deutsche Nationalbibliotek) reported on the work of nestor, Germany's network of expertise in the long-term storage of digital resources. The network consists of only seven partners, but the five working groups – Trusted Repository Certification, Media, Long-Term Preservation Standards, Grid/eScience and Long-Term Preservation, and Cooperative Long-Term Preservation – have many external members. Funding for nestor runs out at the end of August 2009, but the partners are looking at ways to see it continue, as they have found the benefits of collaboration to be invaluable.

The technical track, entitled *Grid Storage Architecture,* began with **Steve Hitchcock** (University of Southampton) describing how the Preserv 2 Project sees repositories using "smart storage": reliable, large-scale storage on top of which risks are profiled and appropriate services automatically invoked. Examples included using a calendar server to schedule actions, OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) to monitor the storage platform for new items, and DROID to classify the formats of these items. The theme of large-scale storage was taken up by **Keith Rajecki** of Sun Microsystems, who outlined some of the key architectural features of Sun's latest products. In the questions following the presentation, Rajecki revealed that the "Honeycomb" code written for the now-obsolete X2100 servers was so hardware-specific that Sun could not easily rewrite it for other server models; this is why the code has been made open source, allowing others to make the necessary adaptations. **Douglas Kosovic** (University of Queensland) explained how the PresSRB project implemented preservation services similar to those developed in the PANIC and AONS projects, for objects stored within the Storage Resource Broker (SRB) system developed by the San Diego Supercomputer Center. Format identification, obsolescence detection and migration and preview services were implemented as SRB shell scripts, while a Java application was written to update a local database with information from the PRONOM and Library of Congress Sustainability of Digital Formats (LCSDF) registries; the standard Unix cron programme was used for scheduling. The SHAMAN project, described by **Claus-Peter Klas** (FernUniversität in Hagen), provides a unified front end to various different preservation and storage systems – including SRB, iRODS, Dspace, Fedora and Kopal – by means of abstraction layers and wrappers. The idea is to be able to integrate multiple legacy systems into a single distributed system; there are of course policy and management issues to work out in addition to the technical ones. The panel at the end of the session threw up some interesting points, such as the power and physical implications of storing tens of petabytes of data, and how off-line tape storage could be integrated into a service-orientated architecture: one suggestion was to use a HTTP Retry-After header.

Session Two: Establishing Trust in Service Providers/Service Architecture for Digital Preservation

The practitioner track of the second session, entitled *Establishing Trust in Service Providers,* began with **Tyler Walters** (Georgia Institute of Technology) discussing the creation of trust relationships for distributed digital federations in terms of prior theoretical work that had modelled institutional trust systems and had explored the dynamics of trust between institutions and people. The presentation teased out some lessons from three trust models developed by organisational science researchers. The first of these three focused on trust factors affecting the formation of virtual organisations and highlighted the importance of subjective trust based on past experience (trust antecedents) influenced by the free flow of performance information. The second model was primarily concerned with intra-organisational relationships, suggesting that such relationships often involved an evolving dialectic between formal and informal trust markers. In this model, it was important to maintain the correct balance between informal processes like personal links (here known as “psychological contracts”) and more formal processes like contracts or audit-based feedback mechanisms like the TRAC (Trustworthy Repositories Audit & Certification) criteria and checklist (Center for Research Libraries, n.d.). The third model concentrated on the dynamics of institutional trust, and emphasised that trust was both a rational and an emotional process. This model emphasised the importance of establishing trust on a personal or emotional level, e.g. by means of shared experiences and emotional honesty.

The remaining presentations in this session returned to dealing with more practical matters. **Suzanne Dobratz** (Humboldt-Universität zu Berlin) spoke on the use of quality management standards in trustworthy digital archives making the point that - to date - such standards have largely tended to be used as guidelines. From a small survey undertaken by Germany’s network of expertise in the long-term storage of digital resources (nestor), Dobratz commented that very few digital repositories had implemented formal quality assurance procedures. **Sara Jones** (University of Glasgow) followed this with an introduction to the Data Audit Framework, a Digital Curation Centre project that had developed a methodology and toolkit that would help institutions identify digital research assets that may be at risk and improve their data management procedures. Finally, **Henk Harmsen** (Data Archiving and Networked Services) gave an overview of the data seal of approval that had been recently developed by Data Archiving and Networked Services (DANS) on behalf of its founders, the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO). The DANS data seal of approval is focused primarily on research data, and considers the quality of the data to be the responsibility of the data producer; the quality of the data storage and its management to be the responsibility of the repository and the quality of the data reuse to be the responsibility of the data consumer.

The technical track followed on neatly from that of the previous session, with papers on service-orientated architecture for preservation. Three of the presentations were concerned with comprehensive repository architectures at various stages of implementation. **Randall Fischer** (Florida Center for Library Automation) reported on progress converting DAITSS, dark archive software, from a monolithic Java application to a RESTful Web services architecture (Fielding, 2000, chap. 5). **Christian Saul** (Fraunhofer Institute of Digital Media Technology, Ilmenau) proposed a theoretical architecture using Service Component Architecture (SCA) and Business Process Execution Language (BPEL). **Jose Carlos Ramalho** (University of Minho) announced the imminent release of RODA, software for a Repository of Authentic Digital Objects; RODA has a Fedora back end, but also incorporates a set of preservation services collectively known (and separately developed) as CRiB. The final presentation was from **Maurizio Lunghi** (Fondazione Rinascimento Digitale) on the National Bibliographic Number (NBN): a URN scheme roughly equivalent and complementary to the Digital Object Identifier (DOI) used by many publishers. The resolution mechanism for NBNs uses the same registry delegation philosophy as Domain Name Servers (DNS). When asked why NBNs do not *actually* use DNS, as OpenID does, Lunghi replied that the chosen system helped to promote national library infrastructures and allowed greater flexibility.

Session Three: International Approaches to Web Archiving/Training and Curriculum Development

The practitioner track of the third session, a panel discussion entitled *International Approaches to Web Archiving,* was dedicated to the work of the International Internet Preservation Coalition (IIPC) and its members: mainly national and state libraries. **Thorsteinn Hallgrímsson** (National and University Library of Iceland) started by introducing the IIPC, which is now in the second year of its second phase. Its primary role is to support its members and co-ordinate their efforts as they attempt to collect snapshots of the Web and preserve them for future study. The most notable outputs of the IIPC have been a range of tools (the enhanced Heritrix crawler, Web Curator Tool, Netarchive Suite, NutchWAX indexer, the open source version of the Wayback Machine, etc.) and the WARC file format for packaging the results of crawls. Hallgrímsson went on to describe the efforts of the National and University Library of Iceland in collecting Icelandic Web materials. The remainder of the session saw representatives from different countries give their own national perspectives: **Helen Hockx-Yu** (British Library), **Birgit Henriksen** (Royal Library, Denmark), **Gildas Illien** (National Library of France) and **Colin Webb** (National Library of Australia).

The principle of legal deposit has to be applied to Web resources in a different way from print resources: instead of putting an extra responsibility on Web editors to forward their work to the national library, it gives the national library the right to take copies of Web content without the site owner’s permission. Three of the countries represented had legal deposit laws – Iceland, Denmark and France – and each of them had taken a threefold approach: a bulk harvest of the national domain three or four times a year, a weekly (or thereabouts) harvest of a select group of Web sites, and event-based harvesting, where all Web content about a major election, sporting event or disaster is collected. Because of the personal data that such mass harvesting would archive, in all three cases access to the material is restricted.

The other two countries have been collecting Web sites on a consensual basis: Australia since 1996, the UK since 2004. This severely limits the amount of information that can be collected – the UK Web Archiving Consortium only received permission to harvest from a quarter of those it asked – but does mean that the results can be accessed with fewer restrictions, and makes it feasible for them to be catalogued. In addition, the National Library of Australia has also been performing annual bulk domain harvests since 2005, but due to the lack of a mandate, does not allow any access to the material.

The impression was that the legal issues with Web archiving would take longer to solve than the technical issues, which is not to say there are not severe technical issues: database-driven sites are a particular headache. To date there has not been a need to provide emulated environments for viewing archived sites – browsers are incredibly tolerant – but this was not expected to last for long. On the subject of Web editors asking for their sites to be archived, Illien recalled a political party asking to see a copy of its Web site from the election three years prior, as it had forgotten what it had promised to do!

The technical track, meanwhile, hosted a panel discussion on training and curriculum development for digital preservation. **Nancy McGovern** and **Aprille McKay** (Inter-University Consortium for Political and Social Research) shared their experiences running the Digital Preservation Management Workshop, initially at Cornell University but now at the University of Michigan. More than 500 participants from 200 institutions in 20 countries have attended the workshop over the five years it has been running, primarily from academic libraries and government institutions. The curriculum is now being expanded to include training for trainers, and in future phases will link up with training provided by the Digital Preservation Coalition, the National Archives of Australia and courses run by other American universities. **Kevin Ashley** of the University of London Computer Centre talked about their Digital Preservation Training Programme, which was inspired by both the Digital Preservation Management Workshop and Digital Preservation Coalition workshops. The feedback from attendees has been overwhelmingly positive so far, but some have suggested that at five days it is too long and too expensive. In response, the course has been slimmed down to two-and-a-half days, and is increasingly tailoring its content to its audience.

Rachel Frick of the Institute of Museum and Library Services (IMLS) gave an overview of some IMLS-funded projects that are either solely concerned with training or have a training element. These included Twenty-First Century Library Professional, a funding programme for developing or enhancing library and information science courses concerned with digital collections, and the Education for Digital Stewardship programme, which helps to fund such activities as the Digital Curation Curriculum conference held by the University of North Carolina, Chapel Hill, and internships and post-doctoral fellowships in the field of moving image archiving and preservation, run by New York University.

The last panellist was **Joy Davidson** (HATII, University of Glasgow), who performed a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis of education, training and careers in digital preservation in the UK, and described some projects that the Digital Curation Centre (DCC) is involved with to improve the situation. These include the DCC Digital Curation 101 residential course and the DCC Digital Curation Manual; the DCC is also heavily involved with the International Digital Curation Education Action Group (IDEA) and WePreserve.

Session Four: Digital Preservation Services/Foundations

For the final session of the day, the practitioner track comprised four presentations on the practical aspects of building digital preservation services. This opened with a presentation by **David Minor** of the San Diego Supercomputer Center on collaborative infrastructure, focusing primarily on the role of the national supercomputer centres funded by the National Science Foundation (NSF) in the

United States. In addition to hosting key projects like Chronopolis and developing major cyberinfrastructure tools like SRB/iRODS, these centres had also provided an important focus for multi-disciplinary research. The presentation then provided slightly more detail on some of the key components of a distributed preservation cyberinfrastructure, including: grid-based environments for the distribution and replication of data, software tools that support data management and security, data grid technologies (like iRODS), the provision of long-term storage facilities, and the further development of high-performance networks for access.

Adam Rusbridge (EDINA, University of Edinburgh) then gave an overview of the UK LOCKSS (Lots of Copies Keeps Stuff Safe) Pilot Programme, through which a number of UK Higher Education institutions had collaborated in an attempt to maintain long-term access to e-journal content purchased from publishers. While the technical infrastructure in the pilot had worked well, there had still been a number of outstanding problems relating, for example, to publisher awareness and licensing regimes. Rusbridge noted that the pilot programme had recently evolved into a membership-based UK LOCKSS Alliance, open to all UK higher education institutions, with annual fees ranging from £1,800 to £5,000. The next presentation, by **Marcel Ras** (Koninklijke Bibliotheek), provided an overview of how the development and implementation of the e-Depot had been an important driver of innovation and organisational change within the National Library of the Netherlands (KB). The need constantly to rethink long-term digital preservation requirements justified the KB's continued investment in research and development activity. In his slide presentation, Ras included a striking image of part of e-Depot's original hardware standing abandoned outside the library – a testament to the speed of hardware development and obsolescence. The final presentation in the session, by **Filip Boudrez** (Stadsarchief Antwerpen), described the development and implementation of a fully integrated digital recordkeeping system for the City Archives of Antwerp. As part of the project, the archives had tried to develop a system based on generic recordkeeping principles that would be suitable for both digital and paper records. The archives had then developed tools to support core repository functions like ingest, management, migration and storage.

Meanwhile, the technical track went back to basics. **David Rosenthal** (Stanford University) gave a particularly entertaining talk debunking claims that bit preservation is a solved problem. Even if the mean time to data loss (MTTDL) statistics that vendors provide are accurate for discs – which is doubtful – there are many other components in a storage system that can cause data loss, much of which can go unnoticed: a server farm of 1.5m drives was recently found to be suffering on average one silent bit change event every hour, despite being managed in a RAID array (Corbett et al., 2004). **Yan Han** (University of Arizona Libraries) demonstrated how a continuous time Markov chain could be used to model the reliability of systems such as RAID and LOCKSS. On this model, a three-copy system (two discs and one tape) has a reliability rate of about 68% over a millennium, whereas a four-copy system works out to have a reliability rate of about 98% over a millennium. The perhaps surprising conclusion was that it is more reliable and cost-efficient to use lots of cheap consumer drives in a redundant array than to have a few expensive enterprise drives.

Christian Keitel (Staatsarchiv Ludwigsburg) argued that digital preservation was becoming so complex that only large organizations have the resources to cope with it.

In response, he suggested ways in which smaller organizations could simplify their strategy to make it more manageable: concentrating on a limited number of objects or object types; using a small set of actions to do most operations; automating process; and using tools and expertise developed by others. In this way, organizations could recognize patterns from existing archival workflows in their digital preservation activity.

Lastly, **Michael Hartle** (Technische Universität Darmstadt) presented a data description language, Bitstream Segment Graph (BSG), as a viable alternative to human-readable format specifications. BSG describes how bit sequences should be transformed, decoded and interpreted, using a classification of six bit segment types and three transformation types. Applying this to a subset of PNG and testing against a set of 156 sample images of various levels of complexity, Hartle found that eight images provided 53 BSG rules, which fully described a further 56 images and described the rest with an average coverage of 79%. Further tests suggested that 80% of the images would be described by 20% of the rules required for full coverage of the entire set. When asked how it compared to other data description languages, Hartle said the others he had seen were either too high-level or too general.

Closing Remarks

The conference concluded with a talk from **Steve Knight** (National Library of New Zealand), who gave some stark warnings about complacency. He felt that talk of significant properties shows a mentality of “acceptable loss” and “good enough” preservation – but he asked whether “good enough” was really good enough. After answering the conference theme, “Joined up and working,” with, “Not really,” Knight went on to suggest there is too much fragmentation – JHOVE, DROID and other tools ought to be consolidated – and argued that preservation needs to get on a commercial footing if it is to make a real impact.

Conclusions

The conference acted well as a showcase for the various projects working on digital preservation issues around the world. It was good to see how the groundwork of previous years – standards such as METS and PREMIS, services and tools such as PRONOM and JHOVE – is being built on and how preservation systems are beginning to be used in earnest. That said, the conference showed plenty of areas where the state of the art of digital preservation is still in its early stages, and demonstrated that even with all the progress that has been made, there is a lot more research and development to be done before the various systems around the world can be described as joined up and working.

References

Besek, J.M., LeFurgy, W.G., Rasenberger, M., Weston, C.D., Muir, A., Atkinson, B. et al. (2008). *International study on the impact of copyright law on digital preservation.* Brisbane, Australia: Queensland University of Technology Printing Services. Retrieved October 29, 2008, from http://www.digitalpreservation.gov/library/resources/pubs/docs/digital_preservation_final_report2008.pdf

British Library. (n.d.). *IPRES 2008: Programme.* Retrieved October 23, 2008, from http://www.bl.uk/ipres2008/programme.html

Center for Research Libraries. (n.d.). *Trustworthy Repositories Audit and Certification.* Retrieved October 23, 2008, from http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91

Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS)* (Blue Book No. CCSDS 650.0-B-1). Washington, DC: Author. (Also published as ISO 14721:2003) Retrieved October 20, 2008, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Corbett, P., English, B., Goel, A., Grcanac, T., Kleiman, S., Leong, J., & Sankar, S. (2004). Row-Diagonal Parity for Double Disk Failure Correction. In *3rd USENIX Conference on File and Storage Technologies. San Francisco, CA* (pp. 1–14). Berkeley, CA: USENIX Association. Retrieved October 22, 2008, from http://www.usenix.org/events/fast04/tech/corbett.html

Dempsey, L. (2007, July). The amplified conference. *Lorcan Dempsey's weblog.* Retrieved October 17, 2008, from http://orweblog.oclc.org/archives/001404.html

Deprocas, A. (2001). Digital preservation: recording the recoding the documentary strategy. In G. Stocker & C. Schopf (Eds.), *Takeover: Who's doing the art of tomorrow* (pp. 334–345). Vienna, Austria: Springer-Verlag Telos. Retrieved October 30, 2008, from http://www.fondation-langlois.org/e/activites/ArsElectronica/imprimante.html

van Diessen, R.J. (2002). *Preservation requirements in a deposit system.* Number 3, IBM/KB Long-Term Preservation Study Report Series. Retrieved October 20, 2008, from http://www.kb.nl/hrd/dd/dd_onderzoek/reports/3-preservation.pdf

Fielding, R.T. (2000). *Architectural Styles and the Design of Network-based Software Architectures.* Unpublished doctoral dissertation, University of California, Irvine. Retrieved October 21, 2008, from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

International Federation of Library Associations and Institutions Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: Final report* (UBCIM Publications, New Series Vol. 19). Munich, Germany: Saur. Retrieved October 20, 2008, from http://www.ifla.org/VII/s13/frbr/frbr.pdf

ISO 9660:1988. *Information processing – Volume and file structure of CD-ROM for information interchange.*

ISO/DIS 28500. (n.d.). *Information and documentation – WARC file format.* (Under development)

Library of Congress. (2008a). *Metadata Encoding and Transmission Standard (METS) official Web site.* Retrieved October 20, 2008, from http://www.loc.gov/standards/mets/

Library of Congress. (2008b). *Metadata Object Description Schema: MODS (Library of Congress)*. Retrieved October 20, 2008, from http://www.loc.gov/standards/mods/

PREMIS Working Group. (2005). *Data dictionary for preservation metadata: Final report of the PREMIS Working Group*. Dublin, OH: OCLC. Retrieved October 20, 2008, from http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Rinehart, R. (2007). The Media Art Notation System: Documenting and preserving digital/media art. *Leonardo: Journal of the International Society for the Arts, Sciences and Technology, 40*, 181–187.