The International Journal of Digital Curation
Issue 1, Volume 2 | 2007

Report from the DigCCurr 2007 International Symposium on Digital Curation, Chapel Hill, NC, April 18-20, 2007

Michael Day,

UKOLN, University of Bath

June 2007

Summary

This is a report from the DigCCurr (Digital Curation Curriculum) 2007 symposium held at the University of North Carolina at Chapel Hill on April 18-20, 2007. The event was organised as part of the project "Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum," funded by the Institute of Museum and Library Services (IMLS).

Introduction

The DigCCurr 2007 International Symposium on Digital Curation began on the evening of 18th April with a reception held in the Louis Round Wilson Library on the campus of the University of North Carolina at Chapel Hill (UNC-Chapel Hill). During this, Helen Tibbo of UNC's School of Information and Library Sciences and Richard Szary, Director of the Wilson Library, welcomed delegates to Chapel Hill. Professor Tibbo also introduced delegates to some of the key people involved in organising the symposium and noted with pleasure the large number of delegates, with around 300 attendees.

The serious work began at 8.30 the following morning in the conference centre of the William and Ida Friday Center for Continuing Education. Professor Tibbo again welcomed delegates and handed over to José-Marie Griffiths, head of the School of Information and Library Sciences, for a short overview of the School's activities on digital preservation and stewardship. Then Bernadette Gray-Little, Provost of UNC-Chapel Hill, gave a short presentation that highlighted the importance of managing digital assets in universities.

Research Library Perspectives

This introduction was immediately followed by four concurrent sessions designed to address the core questions being posed at the symposium, i.e. what do digital curators do and what do they need to know? The four sessions were designed to introduce the specific perspectives of: funding bodies; research libraries; consortia and professional organisations; and researchers. I attended the session on research libraries, chaired by Joanna Eustace of Case Western Reserve University. The opening presentation, by Stephen Chapman of Harvard University Library, first highlighted the importance of collaboration across the organisation. He argued that research libraries needed to develop reciprocal relationships between collection managers and repository staff, each of which would deal with a distinct - but linked - set of service obligations across content lifecycles. Then, using practical examples of heterogeneous discovery systems linked to Harvard's own digital collections, Chapman emphasised the importance of conveying both the context and meaning of digital content. On the topic of what curators needed to know, he concluded that curators needed a knowledge of information theory and management as well as a more specialised understanding of users and their needs, of all aspects of content and object management, and of the capability of all systems in use locally. He finished with a warning that the environment for curation in research libraries was inherently complex, often involving multiple systems for discovery and storage, heterogeneous content, diverse and demanding users, and the obligation to keep content for a very long time. He commented that there was a continued role for domain and technical specialists, e.g. people with specific knowledge of legal contexts or system design, but that generalists would be needed to facilitate communication between domains.

Chapman was followed by Tim DiLauro of Johns Hopkins University, who provided perspectives on the identity of curators through a review of some of the research activities in which the University is involved. For example, a technical analysis of repository software and tools had revealed the importance of matching functionality with requirements. Other projects highlighted the potential for increased

collaboration with researchers and other end-users, e.g. with medievalists and art historians in the *Roman de la Rose* Project, with astronomers in the US National Virtual Observatory. In his concluding observations, DiLauro argued that collections rarely stand alone and that curators need to plan for the repurposing of content. He emphasised that while curators need to be open to external expertise and collaboration, e.g. from data scientists or humanities scholars, research libraries remained good places for undertaking at least some curation activities because they retain the trust of academics and researchers.

Anne Kenney of Cornell University gave the final presentation in the session. She began by noting some general trends, e.g. noting that much library content (e.g. e-journals) is increasingly available in digital form only and that a growing proportion of it is leased from rights owners rather than owned outright. Moreover, increasingly vast amounts of content were being made available through the deployment of new kinds of service providers (e.g., YouTube) or through mass retrospective digitisation programmes like those run by Google and the Open Content Alliance. This new type of content has obliged collection developers to consider access and usability issues as well as those of content. There has also been the need to think about preservation as a collaborative effort. Kenney noted some of the new roles and responsibilities being taken on by institutions, e.g. with regard to intellectual property rights and scholarly communication. Citing a training gaps analysis for library staff produced for the Canadian Cultural Human Resource Council (8Rs Research Team, 2006), she argued that library professionals should have good communication and interpersonal skills, emphasising that leadership potential and managerial skills are increasingly seen as vitally important by employers. The skills required should not be overly centred on technology, but orientated on subject expertise, scholarly processes, user needs and lifecycles. Kenney then identified specific requirements for various job positions, including senior library administrators, collection developers, public service managers, as well as within technical services and information technology administration. With regard to special collections, she commented that curation ceases to be institution-specific in the digital realm. She concluded that research libraries could accommodate digital curation by building a professional staff of experts from a range of domains, but that all needed to share a common set of values. She added that both organisational and employment structures needed to be flexible, recognising the value of providing opportunities for training and collaboration. Summing up her presentation, Kenney noted that digital curation is as much about curation as it is about things digital.

Identifying Digital Curation Services and Functional Requirements

Following a short coffee break, it was time for the next set of concurrent sessions, this time on the identification of digital curation services and functional requirements. Under this general heading, there were five sessions covering the subjects of metadata creation, repository architectures, the management of data sets, user services and the selection and appraisal of content. I attended the session on metadata creation, which contained three presentations. Howard Besser of New York University provided an overview of a NDIIPP (National Digital Information Infrastructure and Preservation Program) project focused on the preservation of public television content, specifically the part of this project dealing with the capture of metadata created during the production process, information which is often discarded or otherwise lost. Realising

that creating preservation-specific metadata at the ingest stage would not scale, the project had explored the possibility of mining the metadata generated by production processes, thus helping to build useful partnerships with producers and creators. In the second presentation, Barbara Sierman of the Koninklijke Bibliotheek, the National Library of the Netherlands, explored how the library had used the PREMIS (Preservation Metadata: Implementation Strategies) Data Dictionary as a checklist to review existing metadata usage in the e-Depot system and to inform the development of an improved data model. The third presentation, by Mary Beth Weber and Sharon Favaro of Rutgers University Libraries, introduced workflow tools for metadata creation and management that had been designed for the New Jersey Digital Highway repository and portal.

What is Digital Curation?

After lunch, delegates gathered in the Grumman Auditorium for a plenary session on digital curation chaired by Cal Lee of UNC-Chapel Hill. In the first of two presentations, Peter Buneman of the University of Edinburgh and research director of the Digital Curation Centre talked about "databases and digital curation." He opened by talking about the dynamic nature of databases, noting that the old model of curation, where objects would be transferred at the end of their lifecycle, was not viable for most databases. He outlined research work on building models of provenance, on preserving the former states of active databases, and on database citation. He also described an experiment that converted a relational database into XML, providing potential benefits in terms of preservation (i.e., less representation information required) but also for repurposing in other ways. He concluded by arguing that digital curation curricula needed to include information on databases and data formats.

William LeFurgy, Digital Initatives Project Manager at the Library of Congress, gave the second presentation in this session, entitled "digital curation and sustainability." He first outlined various aspects of sustainability, defining it as meeting the needs of the present without compromising the ability of future generations to meet their own needs. It involved solving questions about resources (technology, staff, finance, etc.) and infrastructure. LeFurgy noted that one recent object lesson in sustainability had been the US Congress's recent $47 million cut in funding for NDIIPP. He explained the context of this cut and was hopeful that at least some of this funding would be restored in the 2008 budget. LeFurgy argued that sustainability was an important issue because the amount of digital material was still growing, because existing infrastructures were largely geared to non-digital objects, and because digital preservation funding was still largely project based. However, there remained a very large set of open questions, not least about the potential for organisational change, and the all-important matter of costs. LeFurgy argued that there were some basic things that needed to be taken forward. Firstly, we needed to find ways of convincing policy makers that digital materials were valuable, even when much of this value is intangible. Secondly, he suggested that we needed to work on business cases that address risks but include credible cost estimates. With regard to the this last point, LeFurgy referred to work on developing cost models undertaken by the LIFE (Life Cycle Information for E-Literature) Project at the British Library and University College London (McLeod, Wheatley & Ayris, 2006). Finally, LeFurgy noted the importance of developing business models that would enable digital preservation to be dealt with on an ongoing basis. In doing this, he argued that

collaboration would be key, as the digital preservation problem is too complex and expensive to be solved by any single organisation. This meant that there was a role for shared infrastructures and "service-oriented" networks from which different repositories could obtain common tools, services and best practices.

Mechanisms for Influencing Data Curation Practices

The final part of the day was given over to four concurrent sessions on mechanisms for influencing data curation practices. This included sessions dealing with models for education, ingest processes, repository architectures and the design and implementation of repositories across institutional boundaries. I attended the latter, chaired by Mark Conrad of the National Archives and Records Administration (NARA). Mike Smorul of the University of Maryland gave the first of three presentations, an overview of how the PAWN (Producer-Archive Workflow Network) could be used to support multiple ingest workflows. Following this, William E. Underwood of Georgia Tech Research Institute described experiences with using the PERPOS (Presidential Electronic Records Pilot System) electronic records repository and archival processing system, a set of tools designed to support archivists at the George H. W. Bush Presidential Library. Thirdly, Richard Marciano of the San Diego Supercomputer Center (SDSC) gave a presentation on the development of distributed repositories, specifically on collaboration between curators and technologists in the PAT (Persistent Archives Testbed) Project funded by the National Historical Publications and Records Commission (NHPRC). In this, SDSC had collaborated with several curatorial institutions on the design and implementation of a distributed repository infrastructure that was used for electronic records management by participating institutions. Marciano suggested that the PAT Project had demonstrated the potential of a community model for preservation, noting that long-term sustainability was often beyond the capacity of many individual archival repositories.

Views from National Libraries and Archives

The second day of the symposium commenced with a plenary session on the perspectives of national libraries and archives, chaired by Fynette Eaton of NARA. Peter Bruce, Director General and Chief Technology Officer of the Information Technology Branch of Libraries and Archives Canada (LAC), kicked off the session with a brief overview of activities in Canada. He provided an update on progress in several areas. He began with the LAC's ambition to develop a suite of Trusted Digital Repository services, starting with a 'Virtual Loading Dock' for the ingest of legal deposit material and government records. Secondly, he noted that national legislation (Library and Archives of Canada Act, 2004, c. 11) explicitly permitted the harvesting of Web content and that harvesting the Canadian Web had already commenced. Thirdly he introduced the Canadian Digital Information Strategy (CDIS), an attempt to co-ordinate activities in Canada, specifically focused on strengthening the production of content, maximising access, and the successful preservation of content considered to have enduring value (McDonald & Shearer, 2006). With regard to the main themes of the symposium, Bruce concluded that digital curation needs both professional and domain-specific expertise, e.g. in cross-disciplinary teams, but also that all curators would need some level of digital competency.

Adrian Cunningham of the National Archives of Australia (NAA) next provided an archival perspective on the subject of digital archives and digital curation. His paper

(Cunningham, 2007) articulated three key messages, the first two mainly focused on terminological issues. Firstly, he argued for a distinction to be made between "digital curation" and "digital archiving", feeling that, while archivists should be able to work within broad collaborative cross-domain environments, the curation of digital records should be seen as a specific archival activity. Secondly, he insisted that digital archives were not just digital libraries, emphasising the professional role that archivists have in ensuring that preserved records remain authentic evidence of administrative and other activities. He commented that records derive meaning and value from their context and the multiple relationships that surround their creation and use and noted that the metadata required for this is "infinitely more complex" than that needed for mere resource discovery and preservation. The third key message was that digital archiving requires active archival intervention across the whole records continuum, noting that the ever-popular Reference Model for an Open Archival Information System (OAIS) failed to take account of the fact that many organisations have no idea exactly what digital records they have, let alone being able to generate any submission information packages worth ingesting! Having made these three points, Cunningham went on to outline developments in Australia from the mid-1990s, including the development of a national standard for records management (later the basis for ISO 15489) and the "e-permanence" suite of standards and guidelines. In the early 2000s, the NAA began to develop an approach to preservation (based on capturing the user's experience of records) and tools for the normalisation of proprietary file formats into XML. More recent activities outlined in the presentation included MADIRA (Managing Digital Records for Access), which is focused on delivering meaningful digital records to researchers, and the collaborative Australasian Digital Recordkeeping Initiative (ADRI).

The final presentation in the session, by Kenneth Thibodeau of NARA's Electronic Records Archives (ERA) programme, concentrated on the core competencies needed for digital curation, based on thirty years of experience with managing digital records at various levels. His first point was that ongoing change is here to stay, suggesting that technical innovations in the next twenty years are likely to be as novel and as significant as anything we have seen in the past twenty. He also cited a recent IDC white paper on *The Expanding Digital Universe* (IDC, 2007) that had estimated that 161 exabytes of digital information had been created, captured and replicated in 2006, commenting that by 2010 the generation of new technical information would be so great that it would be difficult for anyone to keep up with it. Thibodeau then outlined four core competencies that would be required for digital curators. The first of these was *abstraction*, enabling curators to integrate the expectation of change into their approach. There is never going to be a permanent solution to the challenge of digital preservation, so the key will be "to step back and analyse problems and requirements in the abstract." Thibodeau's second core competency was *application*, the need to apply knowledge and skills in practical situations. The third was agility, the ability to deal with the challenges of new technologies and standards, but also the ability to acquire new knowledge and skills, when required. The final one was the need for *professional expertise*, specifically, but not exclusively, archival science, which has provided a substantial body of knowledge that has stood the test of time. The presentation ended with some comments on the need to operate with a solid understanding of costs and finance, and the ongoing need for collaboration.

Building Capabilities for Digital Curation Repositories

After a coffee break, it was time for more concurrent sessions. Their general theme was building capabilities for repositories, with four sessions covering the design and implementation of repositories within institutions, the definition of capabilities, the users of repositories, and requirements for education and training. I attended the session on repositories and their users, chaired by Patricia Cruse of the California Digital Library. The first two presentations covered different aspects of the DAITSS (Dark Archive in the Sunshine State) repository software developed by the Florida Center for Library Automation. Firstly, Randy Fischer provided a general overview of the Florida Digital Archive (FDA) and examined ingest and dissemination workflows within the DAITSS software. The FDA uses a model of shared operation, and in the second presentation, Stephanie Haas of the University of Florida Libraries explored some of the issues discovered when developing agreements with affiliated institutions. The third presentation, by Patricia Cruse and Kirsten Neilsen of the California Digital Library, provided an overview of the development of a Web Archiving Service (based on the University of California Libraries Digital Preservation Repository) to preserve Web content collected as part of the NDIIPP-funded Web-at-Risk Project.

Digital Curation in Practice

The final set of concurrent sessions concerned digital curation in practice. Again there were four sessions, this time covering social science data, scientific and biomedical data, collection development, and perspectives from the private sector. I attended the session on collection development and gave the opening presentation on collaboration in co-operative networks of institutional repositories, focused on the subject of collection development. This was followed by a presentation on the selection and management of Web content by Janice Ruth of the Library of Congress. This focused on Web capture activities in the Library of Congress and highlighted the different kinds of skills needed, e.g. with regard to appraisal and selection, securing permissions, and reviewing quality. Kathleen Murray and Mark Phillips of the University of North Texas Libraries gave the next presentation on best practice for collection development with specific reference to the NDIIPP Web-at-Risk Project and several other initiatives in which the UNT Libraries' Digital Projects Unit had been involved. They concluded that digital curators would need a range of expertise, including more generic skills in management, collaboration and negotiation, as well as more specific technical knowledge. The final presentation, by Victoria Reich of Stanford University, was an appeal for collaboration on collection building and preservation based on the LOCKSS (Lots of Copies Keeps Stuff Safe) distributed digital preservation infrastructure. She opened with a spirited defence of the idea that libraries, as memory organisations, should continue to be the physical custodians of content. She then outlined the main features of LOCKSS as a community organisation and explained how it could provide a framework for collaborative collection development and preservation. The presentation ended with an outline of collaborative activities, both with the largest STM publishers in the CLOCKSS (Controlled LOCKSS) initiative and with others through individual participation in the LOCKSS programme.

Lessons Learned

The final plenary session was an attempt to bring the many different strands of the symposium together. Firstly, Cal Lee of UNC-Chapel Hill synthesised some of the comments made by symposium delegates by means of a questionnaire. Challenges identified included the more generic need to influence the beliefs and perceptions of decision makers as well as more specific issues relating to resource allocation and collaboration. Technical knowledge of curation at various levels was felt to be important, but there was also an emphasis on good communication and project management skills, service orientation and professional values.

Finally, Clifford Lynch of the Coalition for Networked Information (CNI) provided a synthesis of the symposium and some final thoughts on digital curation and related topics. He began with a consideration of terminology, specifically of the problematic term "digital curation" and its dependence on concepts of "data curation" developed in the sciences. He observed that while the role of information technology had grown ever larger in almost every part of life, investment in traditional archives and records management activities had thus far failed to keep pace. The need now was for new ways of doing things, as reflected in the way that things like teaching, scholarship and healthcare had evolved in the past few decades to take account of new technologies. Lynch had just been at a repositories workshop organised by the National Science Foundation (NSF) and the Joint Information Systems Committee (JISC), where the key challenges were to develop new ways to do scholarship in the presence of vast amounts of data. He suggested that the way "data curation" was currently understood in the sciences was analogous to the traditional practice in the humanities of creating critical editions of texts. In this regard, he wondered whether data curators should instead be known as editors, and suggested that editing databases should count in discussions about tenure. Turning to "digital curation", Lynch thought that the term was perhaps too specific. Currently, much content happens to be in digital form, but the more important concept was that of curation, i.e. the need to acquire content and make it part of a managed collection. He thought that training curators in schools of information made a lot of sense, but the focus should be less on "digital curation" per se, than asking what curation means in the twenty-first century, in a world that has been reshaped by information technology and the abundance of data.

Turning to other issues, Lynch first highlighted the potential for confusion in developing a curriculum for curation. He commented that an organisation hiring a "digital curator" would not currently have a clear idea of what particular set of skills they were acquiring. Building on the list of required skills outlined in Cal Lee's presentation, Lynch added that there was also a need for people with skills and knowledge of legal, socio-economic and organisational issues. Building on ideas developed by Joseph Sax in *Playing Darts with a Rembrandt* ([1999](#)), Lynch argued for a focus on promoting the sustainable stewardship of digital content as a fundamental societal good. He added a couple of small points on learning from financial industries on quantifying value and the need to define acceptable loss rates.

In closing, Lynch echoed Kenneth Thibodeau's comments on the challenge of teaching the next generation how to do something that we do not properly know how to do ourselves. With the current level of uncertainty, e.g. on how to preserve content appropriately, he suggested that the understanding of fundamental principles would be

vital. This was another reason why he thought we should perhaps stick with the notion of curation rather than create a new discipline of digital curation.

Summing up

The symposium was well attended, excellently organised and very timely. It is extremely difficult to produce a definitive summary of an event with multiple concurrent sessions, but I felt that the symposium raised three main points. The first was the need for collaboration, e.g. the sharing of expertise, policy frameworks, tools and infrastructure services, both within individual institutions and as part of regional, national or supra-national networks. Collaboration has been recognised as being fundamentally important to digital preservation for some time. We should not underestimate, however, how difficult it can be to collaborate successfully, given fragmented (and uncertain) funding frameworks and the risks sometimes associated with dependence on services controlled by third parties. Secondly, while there would remain a need for specific technical expertise, many presenters emphasised the importance of more generic knowledge and skills, e.g. those relating to collection development, rights negotiation, the identification of user needs and the development of business cases. Thirdly, several speakers argued that in an ever-changing and uncertain world, traditional professional values (e.g. from the archives and information domains) would remain important in providing a consistent conceptual framework for dealing with future preservation needs.

The symposium programme, including abstracts, papers and presentation slides, and other materials are available from the DigCCurr 2007 Web site (2007).

References

8Rs Research Team. (2006). *Training gaps analysis: librarians and library technicians*. Ottawa, Ontario: Cultural Human Resources Council. Executive summary retrieved June 29, 2007, from http://www.cultureworks.ca/research/default-e.asp

Cunningham, A. (2007). Digital curation/digital archiving: A view from the National Archives of Australia. *DigCCurr 2007 International Symposium on Digital Curation, Chapel Hill*. NC, USA, April 18-20, 2007. Retrieved June 29, 2007, from http://www.ils.unc.edu/digccurr2007/papers/cunningham_paper_7.pdf

DigCCurr2007. (2007). DigCCurr2007: An international symposium on digital curation. April 18-20, 2007. Retrieved June 29, 2007, from event Website: http://www.ils.unc.edu/digccurr2007/

IDC. (2007). *The expanding digital universe: A forecast of worldwide information growth through 2010*. Retrieved June 29, 2007, from http://www.emc.com/about/destination/digital_universe/

McDonald, J., & Shearer, K. (2006). *Toward a Canadian digital information strategy: Mapping the current situation in Canada*, v. 2.0. Retrieved June 29, 2007, from http://www.collectionscanada.ca/cdis/012033-700-e.html

McLeod, R., Wheatley, P., & Ayris, P. (2006). *Lifecycle information for e-literature: Full report from the LIFE Project*. Retrieved June 29, 2007, from http://eprints.ucl.ac.uk/archive/00001854/

Sax, J. L. (1999). Playing darts with a Rembrandt: Public and private rights in cultural treasures. Ann Arbor, Mich.: University of Michigan Press.