DCC
Digital Curation Centre

5th International Digital Curation Conference
“Moving to Multi-Scale Science: Managing Complexity and Diversity”

Poster 001

Achieving long term data integrity from unreliable storage technology and services

Matthew Addis, Richard Lowe, Lee Middlteon, Nicola Salvo
IT Innovation Centre, University of Southampton

The UK TSB supported AVATAR-m[1] and EC supported PrestoPrime[2] projects are both investigating what it really takes to preserve and make accessible huge volumes of digital file-based audiovisual data (Petabytes and above) over long periods of time (50 years or more). In the broadcast and AV archive sectors targeted by these projects, digital audiovisual archives are rapidly becoming 'embedded' as services within wider networked infrastructures and content-centric production and distribution processes. This transition is typically accompanied by a move to online (network accessible) storage of digital content using commodity technology e.g. disk-servers and tape-robots, along with conventional solutions for safety, e.g. backup and disaster recovery.

But are these solutions safe? Can they assure the data integrity needed for long-term preservation of Petabyte volumes of data? The answer is no. Field studies, e.g. by CERN [3] and [4], reveal that data corruption can take place silently without detection or correction including in 'enterprise class' systems explicitly designed to prevent data loss.

Our approach [8,9] recognises that loss does occur in storage, despite what vendors may say, and provides new tools for archivists to control and manage this loss. Policy-based replication of content is used across multiple, distributed and heterogeneous storage locations to provide control over how many copies to make, where to put them and what file-formats to use. Automated integrity checking and repair is used to check for corruption. Large AV assets are deconstructed into smaller files, each of which can have different preservation policies applied to them. This allows differential strategies to be used, e.g. for the audio, video and metadata components of an MXF object, depending on the relative needs of each part of the asset for safety, accessibility, longevity. Most importantly, this allows us to take into account the sensitivity of the specific data formats used for AV [6] to the various failure modes of the technology used to storage them[5].

The tools and techniques we have developed have been disseminated to the media and broadcast community (e.g. NEM, NAB, IBC), but much less so to the wider digital curation and preservation community. There is potential for use in other domains, e.g. scientific, medical, environment, where similar challenges exist, and also where there could also be misconceptions on the reliability of standard IT mass storage technologies for storing very large volumes of digital content.

[1] www.avatar-m.org.uk
[2] www.prestoprime.org
[3] Silent Corruptions, KELEMEN Péter. CERN IT. LCSC 2007, Linköping, Sweden.
[4] Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. Weihang Jiang, Chongfeng Hu, and Yuanyuan Zhou, University of Illinois at Urbana-Champaign; Arkady Kanevsky, Network Appliance, Inc. FAST '08 pp. 111–125 of the Proceedings.
[5] Bit Preservation: A Solved Problem? David H Rosenthal, Stanford University. In proceedings of iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, British Library, London.
[6] Heydegger, V (2008) Analysing the Impact of File Formats on Data Integrity. Proceedings of Archiving 2008, Bern, Switzerland, June 24-27; pp 50-55
[8] RELIABLE AUDIOVISUAL ARCHIVING USING UNRELIABLE STORAGE TECHNOLOGY AND SERVICES, M. Addis, R. Lowe, L. Middleton, N. Salvo. In proceedings of IBC2009.
[9] Wright, R; Matthew Addis; Ant Miller (2008) The Significance of Storage in the ‘Cost of Risk’ of Digital Preservation. Proceedings of iPRES 2008

Poster 002

An Emergent Approach to Data Curation

Stephen Abrams, Patricia Cruse, John Kunze,Tracy Seneca, Perry Willett California Digital Library, University of California

Abstract

Information technology and resources have become indispensible to the pedagogic mission of the University of California. Members of the UC community routinely produce and utilize a wide variety of digital content in the course of teaching, learning, and research. These assets represent the intellectual capital of the University; they have inherent enduring value and need to be managed carefully to ensure that they will remain available for use by future scholars. Within the UC system the California Digital Library (CDL) has a broad mandate to provide curation and preservation services to ensure the long-term usability of the University's digital assets. These efforts are complicated by three factors:

- The ever increasing number, size, and diversity of content.
- Content originating in new contexts of academic departments and research groups unfamiliar with traditional library workflows and practices.
- The inevitability of disruptive change in technology and user expectation.

Much of this content now comes from new initiatives undertaken by the CDL with campus and external partners dealing with anthropological, biological, environmental, and other experimental and observational sciences. In order to be responsive to the specialized needs of these diverse data communities, the CDL has moved to an emergent approach in which infrastructure functionality is devolved into a set of orthogonal and granular micro-services. Since each is small and self-contained, they are more easily developed, deployed, maintained, and enhanced; yet at the same time, complex curation function emerges from the strategic combination of individual, but interoperable services. An important principle in this new approach is that data no longer need be transferred to a central repository for effective curation. Instead, the curation micro-services can be usefully deployed and operated on content *in situ*, for example, in a research laboratory computing cluster or on a scholar's desktop. This poster will highlight CDL's efforts in using its new micro-services architecture for scientific data curation.

Poster 003

Analyzing Data Curation Job Descriptions

Melissa H. Cragin, Carole L. Palmer, Virgil E. Varvel Jr., Aaron Collie, Molly A. Dolan Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, USA

Introduction

There is rising interest on the part of academic and research institutions to implement strategies to preserve and provide access to their digital assets. At the individual level, scientists and scholars are increasingly engaged in sharing the products of their work through administered repositories or directly via the Web. Meeting the technical and organizational requirements ensuing from of these trends will require a workforce skilled in data management, organization and representation, access and preservation, technology implementation, project management, and often domain expertise. In light of this, recent reports have identified a need for trained data librarians and data scientists (e.g. Swan & Brown, 2008), or called for new, organized programs to train LIS professionals and others to meet the anticipated demand for skilled professionals (e.g. Association of Research Libraries, 2006). However, defined career paths for library and information science professionals interested in data curation or data science are not yet firmly established (Interagency Working Group on Digital Data, 2009). As yet, studies of the emerging job market are rare (for an exception, see Lee, 2008), but these are needed for the design of educational programs intended to meet the needs of an emergent labor force, as well as to support the flow of newly-trained professionals into the market . This poster will present analysis of an initial set of data curation-oriented job postings, and characterize the DC employment landscape.

In developing the Data Curation Education Program (DCEP) at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, we have been engaged in a continuous, multi-faceted needs assessment of the range of data service needs for scientists and scholars across domains to inform curriculum development. The DCEP is training LIS professionals for employment across a range of information-oriented institutions, including data centers, libraries and institutional repositories, museums, archives, and private industry. As the data curation field matures, it is essential to assess required skill sets in relation to current jobs, as well as emerging or anticipated roles. This study uses current job postings as a source for analyzing the types of jobs emerging, the kinds of institutions looking for professionals with data curation skills, education requirements for current positions, and expectations for different categories of skill sets.

Methods

Seventy-five (75) online job postings within the Sciences and Social Sciences were harvested from 12 online sites from 8/27/08 - 5/01/09 and pooled with pertinent job postings received by DCEP staff. Job postings were retrieved using keyword queries using the following terms: data; digital; research; librarian; archivist; data scientist; data curator; digital curator; data librarian; digital librarian; data archivist; digital archivist. These terms were selected in order to return broadly relevant job descriptions (i.e. data; digital; research, etc) and narrowly relevant job descriptions (i.e. digital curator; data librarian, etc). Sites without search functionality were both manually browsed and searched via Google site searches. Online sites were chosen to retrieve job announcements across several organization types, including library and information science sites (JobLIST, ARL, LISjobs, iSchool), science and technology sites (National e-Science Center, Society of Systematic Biologists, SAS Google Group, Science-Jobs, Science Careers, SDLP), and a sample of general job hosting sites (e.g. Simply Hired). Inclusion of postings was selective, and relevance was determined by manual examination of the posted job description.

Findings and Discussion

Job postings clustered in a few states, primarily California, New York, Texas and Massachusetts. There is great variation in job duties, and the average years experience required is 4.4 years. More than half of the jobs required at least a master's degree, and 36 (48%) requiring a master's in library science. With respect to general job functions, non-library jobs were more likely to be administrative, and those posted by corporate or research organizations were more likely to require domain expertise, most often in science. It is notable that only 17% of the jobs require some computer programming experience (see Table 1). Several additional characteristics will be reported in the poster, including patterns across employer type (i.e. organization), distribution of required skills across job type, and required computer programming languages.

The analysis of jobs data presented in this poster covers the first stage of this DCEP needs assessment activity. As more LIS professionals prepare to enter the data curation field, it is incumbent on educators to understand and monitor employer needs. It is clear that this initial sample of job postings has some limitations; there is, for example, bias toward library-based jobs because of the relatively narrow scope of job sites searched. While this constrains some parameters of analysis at this point (such as the range of employer type analysis), this sample is suggestive of trends to be assessed in a larger job post data set. Our current approaches to data collection and analysis are being refined in preparation for an extended, comparative analysis of a much larger pool of curation-oriented jobs being drawn from across the science, social science, and humanities domains.

Acknowledgements
We thank Xinrong Lei for her help in data processing and analysis. This work is part of the Needs Analysis activities funded as part of the IMLS grant award # RE-05-06-0036-06.

References

[report] Association of Research Libraries (2006). *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. Washington, DC: ARL. . Retrieved August 6, 2009, from http://www.arl.org/info/events/digdatarpt.pdf.

[report] Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society.* Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, January, 2009. Available at http://www.nitrd.gov/about/Harnessing_Power_Web.pdf. Accessed June 25, 2009.

[Conference Presentation] Lee. C. (2008). What do Job Postings Indicate about Digital Curation Competencies? Society of American Archivists Research Forum Research Forum, August 26, 2008, San Francisco, CA. Retrieved August 7, 2009, from http://www.archivists.org/publications/proceedings/researchforum/2008/CalLee-SAA-ResearchForum2008.pdf.

[report] Swan, A., & Brown, S. (2008). Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs. Retrieved August 6, 2009, from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf.

Poster 004

Archiving scientific blogs with ArchivePress

Richard Davis, ULCC (r.davis@ulcc.ac.uk)
Edward Pinsent, ULCC (e.pinsent@ulcc.ac.uk)
Maureen Pennock, British Library (maureen.pennock@bl.uk)
University of London Computing Centre, 20 Guilford Street, London, WC1N 1DZ
The British Library, Boston Spa, Wetherby, West Yorkshire. LS23 7BQ1

Description

ArchivePress is a JISC-funded rapid-innovation project, developing a cost-effective solution for archiving and managing blog content in academic and research environments as part of the institutional record.

People have been keeping diaries and journals for centuries, and early forms of diaries have been found that date back as far as the 2nd century AD. Their long term value is indisputable, and many archives hold the diaries and correspondence of notable scientists within their collections (Linnaeus, Darwin, Rutherford, Curie). It's become increasingly apparent that the modern equivalent is the blog – or online web log - which uses a dated post/comment structure, and provide web feeds to alert users of new posts or content.

Scientists in particular represent a user group for which blogs have become an integral element of knowledge communication, development, and exchange of ideas, and the value of this often unique content cannot be disputed. Peter Murray Rust and Michael Nielsen are just two eminent scientists of our generation who passionately espouse the value of blogging to their work.

The common and familiar web-archiving scenario is for a web crawler to capture copies of web content and provide subsequent access to the web site as an integral whole. This is perfectly acceptable if the requirement is that the site is presented as an integral whole (though not immune to preservation issues such as obsolescence and persistence). ArchivePress is based upon the premise that blogs are a distinct class of web-based resource, in which the *post*, not the *page*, is atomic, and certain properties, such as layouts and colours, are demonstrably superfluous for many (if not most) users.

ArchivePress exploits the existing functionality of WordPress and the associated technology of newsfeeds (Atom, RSS) to create repositories of blog posts and comments. WordPress is arguably the most widely used blogging application for self-hosted blogging and the engine behind many prominent free and commercial blogging services. It is Open Source, GPL-licensed, uses well-defined open data schemas, and is written in PHP, a language widely used by web developers. These features make it eminently suitable for digital preservation applications.

The ArchivePress team proposes to present a poster at iDCC describing their work and conclusions, and illustrate use of the tool through a number of small case studies. An on-screen demonstration will be provided, facilities permitting. We hope that the presence of ArchivePress at iDCC will contribute to awareness of the value of blog content as part of the academic, institutional, and scientific record, the importance of managing it effectively, and development and discussion of technical solutions to support archiving and preservation of online research in the scientific community, and beyond.

Poster 005

BCEES evaluating cost-effectiveness of eResearch Tools

Troy Sadkowsky[1], Deborah Glass[2], Jennifer Girschik[3], Lin Fritschi[4]
[1]Data Scientists Pty Ltd, Innovation centre, Sunshine Coast, Australia, troy.sadkowsky@datascientists.com
[2]Monash Centre for Occupational and Environmental Health, Depmiment of Epidemiology and Preventive Medicine, Monash University, Melbourne, Victoria, Australia, deborah.glass@med.monash.edu.au
[3]Western Australian Institute for Medical Research, University of Western Australia, Perth, Western Australia, girschik@waimr.uwa.edu.au
[4]Western Australian Institute for Medical Research, University of Western Australia, Perth, Western Australia, fritschi@waimr.uwa.edu.au

Introduction

This paper details the initial fIndings and outcomes discovered by the piloting of available eResearch tools and services in conjunction with the running of an epidemiological research study on breast cancer. The Breast Cancer, Environment and Employment Study (BCEES) is a three year study funded by NHMRC. It is being carried out at Westem Australian Institute of Medical Research (WAIMR) and the University of Western Australia (UWA), in collaboration with Monash University looking at the identifying envirol1lllental and employment risk factors that could contribute to breast cancer. The study's aim is not only to identify these risk factors, but also to explore the use of new technologies to improve the quality and efficiency of the data collection process. The study has adopted a number of eResearch products and services to assist in meeting this aim. Cost-effectiveness of available eResearch tools is measured by performing individual benefit analysis of each product along with detailed time tracking involved in the installation, training and maintenance of the product. The piloting of these eResearch products will serve as a practical example of how effective the adoption ofeResearch can be for a research study in a discipline where these have not been used to a great extent previously. Products and Services that are being piloted include:

Federal server supplied by Australian Research Collaboration Services (Centos Server Viliual Machine)

Collaborative environment (Open Source Content Management System, Plone)

- Issue tracking system
- Meeting workspace
- News and events
- Public interface

Occupational Integrated Database for Exposure Assessments (OccIDEAS)

- Participant tracking
- Online questionnaire development
- Online questionnaire data entry
- Automatic exposure assessments

Visualization tools (Google Charts)

In addition to the products and services listed above, the study has also adopted the concept of a data scientist to assist with the evaluation and integration of the products and services. The role of the data scientist is to be the first port of call in the management of all eResearch tools and to establish an effective plan for managing digital-data.

Analysis

Initial findings show that considerable advantages are being achieved throngh the adoption of these tools. The main disadvantage is the thue (and therefore cost) required in adopting and leaming the product or service. Another disadvantage is the degree of trial and error in using these tools. It has been difficult to find other epidemiological research studies that have used such tools and can advise on their long term benefits. However, it was found that with the addition of the data scientist role; the time required to evaluate the tools was reduced; performance of the tool was increased; the effort required by other investigators was decreased; and the adoption rate of new tools was increased with a specific person available for assistance with system difficulties.

Federal server supplied by Australian Research Collaboration Services
Benefits
The federal server supplied by the Australian Research Collaboration Services (ARCS) is provided free of charge and has provided the study with state of the mi in system admhlistration (security, backups and network maintenance), hardware and software. The ability to collaborate is increased by having the server on a federal academic network and not behind one institution's firewall.
Costs
It has taken time (90 hours) to coordinate the establishment of the server. More time will also be required for documentatiou and cataloging how to gain access to the server as well as how to access the data stored withhl it. More time will be required to formalize ownership and responsibilities of the data and its management.

Collaborative Environment
Benefits
A collaborative web environment has been established on top of the federal server. This has allowed the study team to communicate effectively regarding meetings, issues and news despite the large geographical distances between investigators (Perth, Melbourne, Canada and Germany).
Costs
It has taken time to establish the collaborative environment (100 hours) and requires ongoing support and maintenance (estimated at 52 hours per year).

OccIDEAS
Benefits
OccIDEAS improves the accuracy and objectivity of exposure assessments. OccIDEAS provides a collaborative environment for investigators to establish questionnaires and predetermined rules which ultimately could save months of effort in exposure assessments.
Costs
The OccIDEAS application has required time in installation and configuration (20 hours) and also requires ongoing maintenance and integrity checks.

Visualization tools
Benefits
Presenting the data in a graphical representation increases the scope of the audience. Google charts are provided fl'ee of charge and provide advanced graphical representations of data.
Costs
Time has been invested in acquiring the skills in using these tools and integrating them into the public web site (20 hours).

Poster 006

Curation & Preservation of Crystallography Data

Manjula Patel[1], Simon Coles[2], Liz Lyon[1]
UKOLN & DCC, University of Bath, UK
[2]EPSRC NCS, University of Southampton, UK

Abstract

This poster will describe several issues as well as the work undertaken in relation to the effective curation and preservation of crystallography data within the context of the eCrystals Federation Project. The aim of the Project is to enhance the management of crystallography data at the institution level, incorporating data generated in departments, laboratories and by individual researchers or practitioners; consequently, the work described is concerned with the development of approaches to the preservation and curation of crystallography data in open repositories. For the crystallography community, the long-term provision of data is particularly important since structure determination can only be truly repeated or verified when the raw data is available. In addition, the availability of raw and derived data is very useful for reanalysis and reprocessing as improved methods for performing these tasks emerge.

Crystallography Data

Crystallography is the sub-discipline of chemistry concerned with determining the structure of a molecule and its 3D orientation with respect to other molecules in a crystal through the analysis of diffraction patterns obtained from X-ray scattering experiments. This normally involves several stages, which can be characterised as: data collection; data processing; data workup and publication. We examine the workflows and the characteristics of crystallography datasets with reference to the EPSRC National Crystallography Service based at the University of Southampton, UK.

Preservation Planning

Today, repository managers are under increasing pressure to provide expertise in IT skills, domain knowledge and preservation issues. Expertise in all three areas is essential for the effective management of digital research data. We attempt to break down the process of preservation planning into various practical aspects, including: analysis of data and workflows; evaluation of preservation requirements; defining a preservation policy; formulating a preservation strategy; recording preservation metadata; modeling costs; planning for sustainability; and regular evaluation or self-assessment[1].

Preservation Metadata

Preservation metadata is information that supports and documents digital preservation processes; it includes information relating to provenance, authenticity, preservation activity, technical environment and rights issues. This strand describes on-going community consensus work relating to the development of a Metadata Application Profile for crystallography data[2].

OAIS Representation Information

According to the Reference Model for an Open Archival Information System (OAIS), *Representation Information (RI)* is any information required to render, process, interpret, useand understand data. The work described here, is concerned with an investigation of RI for crystallography data and its role in the curation, maintenance and management of such data[3].

[1] Manjula Patel, Preservation Planning for Crystallography Data, WP4, eCrystals Federation Project, 25th June 2009, http://wiki.ecrystals.chem.soton.ac.uk/images/8/82/ECrystals-WP4-PP-090625.pdf

[2] Manjula Patel, Preservation Metadata for Crystallography Data, WP4, eCrystals Federation Project, September 2009. http://homes.ukoln.ac.uk/~lismp/eCrystals/eCrystals-WP4-PM-090907.pdf

[3] Manjula Patel, Representation Information for Crystallography Data, WP4, eCrystals Federation Project, 19th May 2009, http://wiki.ecrystals.chem.soton.ac.uk/images/e/e1/ECrystals-WP4-RI-090519.pdf

Poster 007

An Information Commons Approach to Long-term Accessibility and Preservation of Polar Data

Kathleen Cass, CODATA

The polar regions are changing rapidly with dramatic global effect. Ensuring the accessibility and preservation of a diverse range of polar data and information is essential not only to advance scientific understanding of the polar regions, and global-scale changes more broadly, but also to support wise management of resources, improved decision support, and effective international cooperation on resource and geopolitical issues.

The fourth International Polar Year 2007-2008 has generated a wealth of new observations and other information and knowledge about the polar regions, cutting across the natural, social, and health sciences. Although most of these data will be shared in accordance with the IPY's data policy, the long-term preservation of these important data resources is not guaranteed. For example, much IPY data remain primarily in the hands of the data originators, often individual scientists or distributed project teams, who may not have the capacity or institutional longevity to ensure long-term preservation.

To address the issue of long-term accessibility and preservation of polar data, an international team led by CODATA, the Committee on Data for Science and Technology of the International Council for Science (ICSU), has begun developing the concept of a Polar Information Commons (PIC). The PIC would serve as an open, virtual repository for polar data and relatedinformation and resources. Data originators and owners would digitally label their data and information resources as part of the PIC and, in the case of data, apply the Creative Commons legal waiver CC0 to place the data in the public domain. This removes any ambiguity about the usability of the data, but it does not eliminate all obligations for the data user. The PIC would establish a set of community norms on appropriate and ethical data use, including acknowledgement of data sources.

Open access to polar data would allow a wide range of stakeholders to contribute to improving the quality, accessibility, usability, and long-term sustainablility of polar data. Communities of experts and data managers would be able to appraise data for long-term data curation, and focus their efforts on improving necessary documentation and preparation of suitable discovery and archival metadata. New tools and services could be developed to mine the large and diverse holdings in the PIC, or even to preserve all of the "bits" in the PIC. Certain data centers, digital libraries, scientific bodies, or other organizations could take on responsibility for data stewardship of important PIC resources of a particular type or theme or region.

We believe that an information commons approach has the potential to blend the "bottom up" contributions of a wide range of stakeholders with the "top down" coordination provided by key polar institutions and stakeholders. Designing a practical implementation that can grow organically over time yet still provide a stable framework for long-term data stewardship remains a key challenge.

Initial development of the PIC concept is being supported by ICSU and CODATA. Other partners include the World Meteorological Organization, the International Arctic Science Council, the Scientific Committee for Antarctic Research, the International Union of Geodesy and Geophysics, the Royal Netherlands Academy of Science, and the Science Commons.

Poster 008

Data Seal of Approval - Simplicity fits all

Henk Harmsen, Chair of the International Editorial Board of the DSA, Vice director of Data Archiving & Networked Services (DANS)

Abstract

The data seal of approval consists of 16 guidelines that may be helpful to an archiving institution striving to become a trusted digital repository (TDR). The guidelines have been formulated in such a way that they are easily understandable and leave sufficient room for a broad interpretation. Standardization was not the objective as the point of departure was that the data seal of approval would remain dynamic during its first years. The seal of approval does not express any views regarding the quality of the data to be archived, but does regarding the provisions an archive has made to guarantee the safety and future usability of the data.
The seal of approval mentions 4 stakeholders: the financial sponsor, the data producer, the data consumer and the data repository, which share an interest and are responsible for a properly functioning data infrastructure. The sponsor is advised to use the guidelines as a condition for financing of research projects. The remaining three stakeholders are addressed in the 16 guidelines. For example, the data producer is expected (three guidelines) to place its data in a TDR and to provide the research data as well as the metadata in the format requested by the data repository. The data consumer must, if it has access to or uses the information in a TDR, respect (inter)national legislation, (scientific) codes of behavior and the applicable licenses (three guidelines). The data repository, in its turn, must ensure that the archive is equipped in such a way that data producer and data consumer are able to meet their obligations. In addition, there are eleven more guidelines for the data repository, regarding organization (mission, dealing with legal regulations, quality management, long-term planning and scenarios), processes (transfer responsibility, data references, integrity and authenticity) and technical infrastructure (OAIS and automated processes). In other words, the data repository is the stakeholder of which most is expected.

In the poster session I will give more information about the assessment procedure and the Editorial Board. I will present the DSA roadmap from 2008 to 2011 and the role of the European Commision and the DSA community.

The term Trusted Digital Repository (TDR) occurs in almost all seals of approval. However, it is unclear what a TDR is exactly. At the time of writing, Wikipedia does not yet have a description of the concept. Main point of such a repository is ‘trust’. It is the basis of the data seal of approval.

Poster 009

DataONE: enabling data-intensive environmental research through cyberinfrastructure

William Michener , University of New Mexico
Patricia Cruse, Presenter. California Digital Library – University of California

Abstract

DataONE addresses four key challenges: (1) data loss; (2) data dispersion; (3) data deluge; and (4) poor practice. DataONE will enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it. The system focuses on preserving at-risk data and is designed around a nucleus of three existing data centers (i.e., coordinating nodes) and a broad array of data holdings maintained by libraries, research networks, and academic and governmental organizations (i.e., member nodes). The cyberinfrastructure promotes the discovery and access of data by providing the one-stop shopping for data and metadata (information about the data that enables its use) about Earth's biota and environments, covering a range of spatial and temporal scales, and scales of biological organization. DataONE provides an "investigator's toolbox" that includes metadata management, scientific visualization, and other tools that empower scientists and organizations to more easily and effectively manage, analyze, and synthesize data. Innovative training and outreach to scientists and students include best-practice videos, podcasts, on-line certificate programs, downloadable best practice guides and exemplars of data management plans. Through working group meetings, computer and information scientists are engaged in developing and promulgating ontologies that will facilitate data integration and simplify creation of complex scientific workflows. The DataONE portal simplifies the process of acquiring and using appropriate scientific workflow software like Kepler and Taverna, as well as publishing and sharing new workflows via mechanisms such as myExperiment that allows workflows to be re-used and possibly adopted for other uses. This poster will highlight DataOne’s approach to data-intensive environmental research through cyberinfrastructure.

Poster 010

The Dryad Repository: Designing a Curation Workflow

Jane Greenberg, University of North Carolina at Chapel Hill, USA
Hilmar Lapp, National Evolutionary Synthesis Center, USA
Ryan Scherle, National Evolutionary Synthesis Center, USA
Todd Vision, National Evolutionary Synthesis Center, USA
Hollie White, University of North Carolina at Chapel Hill, USA
Sarah Carrier, University of North Carolina at Chapel Hill, USA
Peggy Schaeffer, National Evolutionary Synthesis Center, USA

The Dryad Repository

Dryad (http://datadryad.org) is a digital repository for data underlying published works in ecology, evolution, and related fields. Dryad is supported by a collaboration involving the National Evolutionary Synthesis Center (NESCent); the Metadata Research Center at the School of Information and Library Science, University of North Carolina at Chapel Hill (UNC/CH); North Carolina State University; University of New Mexico; and Yale University. Additional partners include major societies and journals in the field of evolutionary biology. Dryad allows investigators to validate published findings, explore new analysis methodologies, and repurpose the data for research questions unanticipated by the original authors. An effective curation procedure is significant for Dryad's success. This poster reports on several studies providing results helpful to developing the repository's curation workflow.

Research Supporting Curation Plans

The Dryad development team has undertaken a series of analyses to develop an effective curation plan:

A *survey* of 400 evolutionary biologists provides empirical data on their data sharing attitudes and behaviors. Results documenting biologists' interactions with existing data archives; their data sharing practices; and their dependency on digital media for research and reporting indicate various curatorial skills and limitations of prospective Dryad depositors.

Intensive semi-structured interviews with 17 evolutionary biologists provide examples of how scientists describe data (White, 2008), and point to system features and functionalities that can aid depositor curation.

A *metadata content analysis* of eight schemes and a *vocabulary mapping study* including nearly 600 terms (Greenberg, 2009) indicate where automatic metadata generation techniques can expedite Dryad's curation workflow.

Conclusions

Findings from the studies outlined above, and results from Dryad stakeholder meetings, support a curation workflow integrating automatic and human metadata generation techniques, and leveraging depositor/scientist and professional curator expertise. This poster will highlight key findings informing Dryad's curation workflow.

Acknowledgements

This work is supported by NSF EF-0423641 and the IMLS/NLG program.

References

[journal article] Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly*, 47, 3: 380-402

[proceedings] White, H.C. (2008). Exploring evolutionary biologists' use and perceptions of semantic metadata for data curation. *International Conference on Dublin Core and Metadata Applications*. *Berlin, Germany*

Poster 011

Addressing multi-disciplinary and cross-agency data management and curation

Luis Martinez-Uribe, University of Oxford

Background

The University of Oxford produced an Information and Communication Technology (ICT) Strategic Plan back in Spring 2007 to ensure a coordinated and coherent approach to the development and deployment of ICT services across the University to support teaching, learning, research and administration. The development of services for the management and curation of digital data generated by Oxford researchers was one of the key priorities set on the strategic plan. The Oxford Digital Repositories Steering Group, established in April 2007 with a coordinating role for digital repository activities, soon took responsibility and secured funding from the Oxford University Press' John Fell Fund to start scoping institutional data curation services.

A scoping study was undertaken throughout 2008 to gather researchers' data management requirements as well as to document existing services and gaps in service provision. In addition to this, the DISC-UK DataShare project supported the piloting of the Data Audit Framework (DAF) in Oxford. Findings of these exercises revealed that researchers across the University felt that more support and infrastructure was needed. Moreover, the study helped realising that institutional data management and curation ought to be a cross-agency activity working with scholars early in the research lifecycle. At the same time, the UK Research Data Service (UKRDS) feasibility study was taking place to assess the viability of national data service and Oxford participated as one of four case study institutions. The final UKRDS report stated that a national data service was feasible and proposed a two-year Pathfinder phase with the case study Universities.

A multidisciplinary and cross-agency collaboration

In Autumn 2008 and to continue with the development of institutional data management services, a group of several service units and two research groups, which had participated in the scoping study, came together bringing their complementary areas of expertise to the service of a new data curation activity in Oxford.

The Embedding Institutional Data Curation Services in Research (EIDCSR) project is funded by the Joint Information Systems Committee (JISC) and aims to scope the data management requirements of two collaborating research groups and address these requirements by bringing together the expertise and services of a range of service units working across the data curation lifecycle.

The participating research groups, one in Life Science and the other in Medical Sciences, create and share data from a variety of laboratory instruments as well as GRID computing simulations as part of a Biotechnology and Biological Sciences Research Council (BBSRC) funded project. Their collaboration represents an extraordinary exemplar of multi-disciplinary data re-use. Data generated by one group feeds the creation of the data of the next group in a pipeline that combines expertise and knowledge in various and different domains such as histology, image processing or magnetic resonance.

The service units involved in EIDCSR include Research Services to develop policy and guidance, Library Services to oversee the metadata management aspects, Computing Services to embed their back-up services and the e-Research Centre to advice on user requirements gathering and GRID computing. In addition to this, IBM contributes to the project investigating ways to provide federated data storage and access management for large datasets.

This poster is aimed at providing a visual description of the EIDCSR project as well as engaging and encouraging discussion with conference attendees interested in institutional models for data management and curation. It will include:

- a description of the EIDCSR project with aims and objectives;
- findings of the EIDCSR audit and requirements analysis using the Data Audit Framework (DAF) methodology and
- lessons learned on initial work on metadata standards to describe and preserve the datasets and also on preliminary work on development of data management cost models.

Poster 012

Model Migration Approach for Database Preservation

Arif Ur Rahman[1,2], Gabriel David[1,2], Cristina Ribeiro[1,2], [1]Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto [2]Instituto de Engenharia de Sistemas e Computadores, Porto

Abstract

Strategies developed for database preservation in the past include technology preservation, migration, emulation and universal virtual computer strategy. In this paper we present a new concept of “Model Migration for Database Preservation”. Our proposed approach involves two major activities. First, migrating the database model from conventional relational model to dimensional model and second, calculating the information embedded in code and preserving it instead of preserving the code required to calculate it. This will affect the originality of the database but improve two other characteristics. The information considered relevant is kept in a simple and easier to understand format and the systematic process to preserve the dimensional model is independent of the DBMS details and application logic.

Organizations are increasingly relying on databases as the main component of their recordkeeping systems. However, at the same pace the amount and detail of information contained in such systems grows, also grows the concern that in a few years most of it may be lost, when the current hardware, operating systems, database management systems (DBMS) and actual applications become obsolete and turn the data repositories unreadable. The paperless office increases the risk of losing significant chunks of organizational memory and thus harming the cultural heritage. In our poster we present the ‘model migration approach for database preservation’.

The design of a database preservation process requires a proper balance of the aspects of originality, integrity, accessibility, intelligibility and authenticity in each major problem to be solved. Two of the problems are the complexity of the relational model of real size information systems and the embedding of important knowledge from the application domain in code.

The complexity of the relational model may prove to be a serious stumbling block for preserving databases. Part of it comes from the requirement of redundancy elimination that transaction oriented databases must follow in order to be efficient and consistent in capturing facts.

The strengths of the dimensional model make it better for long term preservation and access of the information. Report writers, query tools and user interfaces can all make strong assumptions about the dimensional model which makes the processing more efficient. Other features of dimensional model include explicit separation of structure and contents, and hierarchies in the dimensions. The separation of the structure and contents help in making it DBMS independent which is crucial for database preservation. The hierarchies in dimensions help in aggregating the data and results in faster access. In the past dimensional modeling had not been considered for database preservation.

Following this approach the five characteristics of databases must be preserved. They include context, content, structure, appearance and behavior.

Conclusions

For migration of the relational model of the operational system, a thorough understanding of the same is required. All the steps in the migration process should be properly documented and the results coming from the data in the dimensional model must be compared with the operational system so that it is ensured that the process is done without loss of information. This will result in keeping the database authentic in the new model.

The approach presented in the poster does not affect the ‘appearance’ aspect of a database as the same rendering method can be used in the future.

We present an alternative to preserving code (functions, triggers, stored procedures) in the process of database preservation i.e. execute the code and preserve the results in a simpler and easily accessible model for the future.

Before migration it should be decided what is to be kept for the future and what can be discarded. This work is similar to the evaluation, elimination and description work an archivist must perform before archiving a set of documents.

Poster 013

National e-Infrastructure for Social Simulation

Mark Birkin[1], Rob Allan[2], Sean Beckhofer[3], Iain Buchan[4], June Finch[5], Carole Goble[3], Andy Hudson-Smith[6], Paul Lambert[7], Rob Procter[5], David de Roure[8], Richard Sinnott[9]

[1] School of Geography, University of Leeds; m.h.birkin@leeds.ac.uk (corresponding author) [2] STFC, Daresbury [3] School of Computer Science, University of Manchester [4] School of Medicine, University of Manchester [5] School of Social Sciences, University of Manchester [6] Centre for Applied Spatial Analysis, UCL, [7] Applied Social Science, University of Stirling [8] Electronics and Computer Science, University of Southampton [9] NeSC, University of Glasgow

NeISS is a unique project aiming to build simulation services for social scientists and policy-makers. The project is characterised by an interest in the deployment of simulation tools for data integration, modelling, analysis and simulation. These tools are manipulated in an Information Environment (portal) which facilitates other activities in the research lifecycle, such as the composition and enactment of workflows, publication and sharing of analytical results.

In this poster, we will discuss the design of an architecture for the NeISS project, consisting of four interacting layers (see figure). The service layer comprises the fundamental analysis components on which simulation depends. Above the composition layer sits an architecture layer. This provides tools and methods needed to provide portal access to simulation services and workflows, and to combine these into domain-specific exemplars which are mobilised through a deployment layer. The project team has established interests in housing, transport, health and social care, education and demographic planning. Engagement with inexperienced users accessing services which are integrated within domain-specific service users will provide a clear demonstration of the impact of this work.

NeISS is therefore a project which is concerned with the management and re-use of secondary data from a wide variety of sources, but also with the generation of intelligence from data through modelling and analysis processes, and with the curation, analysis and exchange of both inputs and outputs from the social simulation process.

Poster 014

Policies and planning for data curation

Sarah Jones,DCC and HATII preservation researcher, University of Glasgow

Abstract

As reliance on digital information increases, so too does the need for appropriate data management policies and guidance to ensure that these resources are created and managed in a way that ensures they remain accessible and meaningful over time. The DCC is working to increase researchers' capacity to better understand the data curation lifecycle and how it relates to the management of their data. We believe that better equipped researchers will be able to proactively inform and influence the development and workable implementation of data repository policies and higher-level, institutional strategies for digital preservation. The DCC and HATII at the University of Glasgow have been conducting various strands of research in these areas as outlined below.

Research funders' curation requirements

DCC Associates Network members requested advice and guidance on the curation policies of UK funding bodies. We have collated this information and pulled together links to useful references and tools to help enable researchers and institutions to comply with these requirements. Detailed information is presented in a policy report[4] and an associated quick-glance overview is presented online.[5]

Data management plans

In response to increasing curation requirements from UK Research Councils, the DCC is working to develop a range of data management plan templates to assist researchers better manage their data. Having analysed requirements, we have provided a content checklist that draws together all details researchers may be asked to provide.[6] This is currently being developed into an online tool to allow users to select relevant sections for their context and adapt best practice examples to aid completion.

Digital preservation policy scoping studies

A Digital Preservation Advisory Board has been established at the University of Glasgow, which commissioned several scoping studies to explore the feasibility of implementing an institution-wide digital preservation policy. The studies mapped existing practices and support mechanisms to identify gaps in provision and areas requiring additional support. An institutional roadmap for digital preservation has been established and is currently being reviewed.[7]

Enhancing Repository Infrastructure in Scotland

As part of the JISC-funded ERIS project, the DCC will be surveying existing approaches to curation adopted by Scottish HEIs, developing a curation policy framework and establishing requirements for machine-readable preservation metadata at an object level to facilitate automated ingest into preservation repositories.

[4] Jones, Sarah, *A report on the range of policies required for and related to digital curation*, version 1.2, (DCC, March 2009), http://www.dcc.ac.uk/docs/reports/DCC_Curation_Policies_Report.pdf

[5] See: http://www.dcc.ac.uk/resource/curation-policies/

[6] Donnelly Martin, & Jones, Sarah, *Data Management Plan Content Checklist*, (DCC, June 2009), http://www.dcc.ac.uk/docs/templates/DMP_checklist.pdf

[7] Jones, Sarah, *Findings and recommendations from preservation policy scoping studies at the University of Glasgow,* (HATII, August 2009)

Poster 15

Sarcomere: A System for Data Interoperability

Nassib Nassar, Renaissance Computing Institute, University of North Carolina at Chapel Hill,
John A. Kunze, California Digital Library, University of California
Gregory B. Newby, Arctic Region Supercomputing Center, University of Alaska Fairbanks, USA
Kevin Gamiel, Renaissance Computing Institute, University of North Carolina at Chapel Hill, USA

Abstract

Sarcomere is a server that provides access to data and user-defined services through a protocol and a set of simple models for data interoperability, for the purpose of improving curation, sharing, and integration of research data. The system offers a conceptual data abstraction (Sarcomere schema) which maps onto the relational model and promotes curation of normalized data. The system also utilizes a new data management protocol, THUMP-DL, which is an extension of The HTTP URL Mapping Protocol (THUMP) and defines URLs that can be resolved to data sets and subsets. Sarcomere currently runs as a thin layer on top of database systems and offers users and software developers several primitives for sharing and integrating data. We are exploring per-user virtual machines for running user code on servers to help with data interoperability, such as for data conversion and automating the deployment of code to computing resources.

We suggest that high quality curation, reuse, and integration of data are subtle challenges that require diversity of experimentation rather than a monolithic solution. This system design is composed of simple models and primitives in order to facilitate such experimentation. In addition, the use of URLs is proposed as a way of encouraging participation of all users in the exchange of interoperable research data, analogous to the common exchange of web content.

Sarcomere is available as open source software and supports various relational database systems via ODBC. The purpose of this poster is to describe the design and implementation of the Sarcomere system.

Poster 016

Science and Scholarship in the Blogosphere: Blog Characteristics, Blogger Behaviours and Implications for Digital Curation

Carolyn Hank, School of Information and Library Science, University of North Carolina at Chapel Hill

Abstract

Blogs are widely recognized as a key part of contemporary online culture. The information and library science community has responded to this new form of communication both as participants in the blogosphere – as authors, readers, and blog hosts – and as advocates for blog preservation, through calls in the literature as well as through emerging blog preservation programs and research initiatives. This poster reports on research into a specific genre of blog content – science blogs – with the overall objective to inform current and future blog preservation initiatives. It shares preliminary findings from a current study of blogs and bloggers aligned with ScienceBlogs, a network of seventy-five blogs self-described as an "experiment in science communication." ScienceBlogs represents a range of disciplines, comprised of blogs from the domains of life science, physical science, environment, humanities education, politics, medicine, brain and behaviour, technology, and information science.[8] The intent of this research is three-fold. One, it allows for further investigation into a methodological framework for studying this diverse and ever-growing group of content creators and the characteristics of this dynamic, near ubiquitous form of communication, building from earlier work on the digital preservation preferences of bloggers, regardless of domain.[9,10] Second, it will inform current and future blog curation and preservation initiatives through identification of blog attributes and blogger behaviours and preferences impacting the acquisition, identification, storage, use and viability for long-term stewardship. Third, since examining a particular class of blogs and bloggers, it will contribute to on-going dialogue and investigation into the transformative nature of scholarship, and the role of blogs as a new form and process of scholarly communication. These research goals are accomplished through two data collection techniques.[11] First, a document analysis of blogs from the ScienceBlogs domain is currently underway, examining particular blog elements, including posts, comments, use statements, and "about" and contact information, quantifying publishing behaviors (e.g., frequency of updates) and publication impact (e.g., blog rolls, linkages, and commentary). Second, a web-based survey of ScienceBlogs bloggers will be administered, to assess perceptions on preservation, including bloggers' support, or non-support, of preservation of their own blog; opinions on what to preserve (e.g., entire blog or specific elements); perceptions on access to preserved blogs (e.g., open, closed, or mediated); and reflections on responsibility for blog preservation (personal and/or organizational). Additionally, the survey gathers reflections on how bloggers' perceive their blogging activities in relation to scholarly communication.

[8] ScienceBlogs: http://scienceblogs.com/

[9] [proceedings] Hank, C., Choemprayong, S., & Sheble, L. (2007). Blogger Perceptions on Digital Preservation. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries: Building and Sustaining the Digital*

[10] [proceedings] Sheble, L., Choemprayong, S., and Hank. C. (2007). Preservation in Context: Survey of Blogging Behaviors. In *Proceedings of the Third International Digital Curation Conference: Curating our Digital Scientific Heritage: A Global Collaborative Challenge*

[11] Note on study: Study is currently underway; anticipated date for data collection and analysis is September 30, 2009. If poster accepted, this will be reflected in final poster abstract, due November 13, 2009.

Poster 017

The CNRS project SIDR

Magali Roux, CNRS research director, SIDR project leader
Institute of Scientific & Technical Information (INIST), Vandoeuvre-lès-Nancy, France

Abstract

Systems biology studies mechanisms that underlie various biological functions that can occur at any integration level: from unique cell to a whole organism. Progress in systems biology relies on production, collection and annotation of complex quantitative data that can be used in mathematical models in order to explain and predict the essential biological functions and interactions. Research communities that produce resources in this field have to face major issues related to data sharing and exchange, including that of using standards, ontologies, controlled vocabularies, exchange languages, etc.

The CNRS Standard-based Infrastructure with Distributed Resources (SIDR) initiative aims at building a platform of international scope that will address these issues by providing research teams with a well-structured access point and that will enable identification, dissemination, adding value and sharing of resources for which quality will be specified and guaranteed.

The poster will describe:
the SIDR background
the SIDR data resources
the SIDR functions
the SIDR member laboratories collaboration policy
the SIDR architecture

Poster 018

Smart Migration Service: The MIXED project

Dirk Roorda, René van Horik , Data Archiving and Networked Services, Royal Netherlands Academy of Sciences - http://www.mixed.dans.knaw.nl

Abstract

Migration is a strategy for repositories to maintain the usability of their assets over time. Earlier research has recommended migration to an intermediate XML data format as a cost-effective method. The MIXED project is implementing this method. This poster reports the progress and points to a demo.

1 Introduction

There is a growing interest to publish scientific and scholarly data alongside with the textual publications that are based on them. Therefore it is important that repositories are able to store these data and provide meaningful access to them for a long time to come. A major concern is the file formats in which the data are encoded. These formats tend to be application-oriented instead of information-oriented. When applications become obsolete, these formats follow, and finally the data are trapped in unusable representations. There is a concrete, cost-effective solution to this problem: smart migration. This poster discusses the method, the project and the progress.

2 Smart Migration

Smart migration is the idea of optimizing migration by using an intermediate format and developing a scenario to exploit the benefits of it. The XML data format was advocated as a suitable preservation format for at least spreadsheets and databases. By carefully designing XML schemas for several kinds of data, it is possible to encode the data in an information-oriented format that is much less subject to change than application-oriented formats. The advantages of XML are so profound, that it is advisable to convert data to XML as soon as possible, rather than to wait till the original format is nearly obsolete. This leads to a scenario where format migrations are applied at ingest and at dissemination time. Whenever a data producer submits data to a repository, an XML version of the data will be created alongside the original. Whenever a user of the repository wants to access data, it will be converted from the XML version to the user's format of choice. The scenario for the long term is as follow: At first, say in 2009, datasets in various formats are ingested, and converted to XML. Later, say in 2020, this intermediate format will be superseded by a new version. By providing migrations on the fly between old and new XML, it is still possible to view old content by new applications, and new content by old applications. Eventually, say in 2050, the 2009-XML becomes obsolete, in the sense that it is not feasible anymore to maintain the software that uses it. Then a bulk migration is needed to migrate datasets out of the 2009-XML.

3 The MIXED implementation

In order to implement a smart migration service, two things must be done: the intermediate XML must be defined, and conversion software must be developed. The MIXED project at DANS is actually in the process of doing both. MIXED limits itself to tabular data, in particular data in file oriented databases such as Microsoft Access, in spreadsheets, and in statistical applications such as SPSS. MIXED considers these as three kinds of data, and it defines an intermediate XML format for each of them. The collection of these XML formats is called "Standard Data Formats for Preservation" (SDFP). The rationale for being a data kind is: whenever there are dominant applications around that handle certain kinds of information, it makes sense to call that information a data kind. Examples are: databases, spreadsheets, editable documents, vector images, drawings, etc. The dominant application tends to promote its file format as the de facto standard for that information, which leads to lack of transparency and vendor lock-in. There exist considerable efforts to make file formats more transparent. Examples are the standardization of Open Document Format, Office Open XML and PDF. MIXED aims at bringing the best candidates for the data kinds together. So SDFP is selecting rather than defining XML representations. The intention is that more data kinds are added to SDFP. The task that the MIXED software must perform is: a set of conversions to and from the SDFP format. Because SDFP is essentially open, and because there will be an open set of applications that deal with any given data kind, the MIXED software itself must be easily extensible. MIXED as software consists of a framework plus conversion plugins, that can be used for two simple things: (i) add/remove conversion plugins and (ii) perform file format conversions. It is quite easy for third parties to develop their own conversion plugins for MIXED (the SPSS plugin e.g. is provided by UKDA). We expect that MIXED has the potential to grow into a gateway to trusted format conversions, once repositories start to focus their migration efforts into a unified methodology as MIXED provides.

The framework itself is open source. It can be downloaded and locally installed. An easier option is to use the MIXED web console to experiment without changing local systems. MIXED will also be published as a webservice.

References

1. The MIXED project. "Migration to Intermediate XML for Electronic Data". Carried out at "Data Archiving and Networked Services" (DANS). 2007-2009 http://mixed.dans.knaw.nl

2. Migration to Intermediate XML for Electronic Data. White Paper. Data Archiving and Networked Services. 2007. http://mixed.dans.knaw.nl/node/114 (also available in PDF from there).

3. Migration to Intermediate XML for Electronic Data. iPres 2007 special issue. 2007

4. Standard Data Formats for Preservation. XML schema developed in the MIXED project. Documentation and schema files. http://mixed.dans.knaw.nl/node/431

5. Web interface to the MIXED software, provided by DANS. http://mixed11.dans.knaw.nl:8080/mixed-web-console/conversion.jsf

Poster 019

Social Science Data Reuse and Institutional Review Boards:
A Pilot Study of Stakeholders' Opinions

Samantha Guss, Data Services Librarian, New York University

This poster reports on a pilot study conducted at a large research university, which assessed the opinions of IRB members and social science researchers about data sharing and archiving. The study utilized several methods, including a survey of social and behavioral science researchers, interviews with IRB members, and an analysis of the language used in blank Informed Consent documents from previously approved studies related to potential data reuse. The poster will describe the results of these inquiries as well as suggest future steps toward understanding and clarifying possible ethical issues surrounding the preservation and reuse of social science datasets.

Building on the work of Davis & Connolly (2007), Foster & Gibbons (2005), and the IMLS-funded MIRACLE (Making Institutional Repositories in A Collaborative Learning Environment) Project (Markey, et al., 2007), among others, this research examines potential users' attitudes towards digital repositories and potential barriers to their full utilization. Its findings reinforce the notion of stewardship by librarians and archivists early in the data life cycle (see Green & Gutmann, 2007, and many others). In a presentation entitled "IRBs and Data Sharing," given at the 2009 IASSIST (International Association for Social Science Information Service & Technology) Conference, Witkowski & Alter identified ideas for future research in this area--notably gathering "nuanced information about current policies, practices, and training initiatives by [c]onducting more in-depth interviews of administrators of IRB programs and researchers who are IRB members." The research presented in this poster is a natural addition to the body of literature on IRB-related data archiving issues.

Recommendations include early intervention by repositories, open discussion within Institutional Review Boards on data archiving in social and behavioral science, increased guidance for IRB applicants, continued education, and increased transparency and communication among IRBs and researchers. Combined, the data gathered provides multi-faceted insight into these two stakeholder groups, especially regarding IRB application language and subjects' informed consent. Additionally, it identifies appropriate methodology, a knowledge base, and some "hot topics" to lay the groundwork for a more comprehensive study.

The author completed this study while a Carolina Digital Curation Fellow (2007-2009) at the School of Information and Library Science at the University of North Carolina at Chapel Hill, supported through IMLS Grant # RE-05-06-0044.

References

[Internet journal]Davis, P. M., & Connolly, M. J. L. (2007, March/April). Institutional repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine, 13,(3/4)*. Retrieved August 4, 2009, from http://www.dlib.org/dlib/march07/davis/03davis.html

[Internet journal]Foster, N. F., & Gibbons, S. (2005, January). Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine, 11,(1).* Retrieved August 4, 2009, from http://www.dlib.org/dlib/january05/foster/01foster.html

[journal article]Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems and Services, 23(1), 35-53*. Bingley, UK: Emerald Group Publishing Ltd.

[report]Markey, K., Rieh, S. Y., St. Jean, B., Kim, J. & Yakel, E. (2007). *Census of institutional repositories in the United States: MIRACLE Project research findings.* Council on Library and

Information Resources. CLIR Publication No.140: Washington, D.C. Retrieved August 4, 2009, from http://www.clir.org/pubs/reports/pub140/contents.html

[proceedings]Witkowski, K. & Alter, G. (2009). IRBs and Data Sharing. *IASSIST/IFDO 2009: Mobile Data and the Life Cycle. Tampere, Finland.* Retrieved August 4, 2009, from http://www.fsd.uta.fi/iassist2009/program.html#Thursday

Poster 20

Value-Enabled Digital Curation: Toward Ethically Responsible Acquisition

Christopher A. Lee, University of North Carolina, Chapel Hill USA

Abstract

When acquiring, managing and providing access to materials, professionals in collecting institutions must consider various norms, laws, codes of ethics, policies, procedures and personal values. As they address curation of digital collections, they will increasingly discover that established sources of guidance suffer from what Lawrence Lessig has called "latent ambiguities." Digital resources are composed of interacting components that can be considered and accessed at different levels of representation (e.g. bitstream, through a filesystem; files as rendered through specific applications; records composed of multiple files; abstract "works"; larger aggregations such as web sites). To ensure integrity and future use, digital curation professionals must make decisions regarding treatment of materials at multiple levels of representation.

Digital curation professionals are faced with many new decisions, which require an understanding of underlying digital representations, in order to appropriately enact professional values. For example, when acquiring a disk as part of a collection, should an archivist create a bit-level image of the disk, in order to ensure the potential to recreate not only all the data but also system and program files? Should she retain "hidden" data in a Word document or only retain what she assumes to be the text that the author intended? If the disk includes a Microsoft Outlook .pst file (including saved and sent messages, calendar items, draft and deleted messages, address book, and possibly viruses), should she retain the whole .pst file, or simply extract messages and attachments that were sent and received? If a collection documents the life of an individual, what should be the scope of collecting information associated with that person's online presence (e.g. postings, affiliations, profiles, micro-contributions)?

There are many stakeholders who a legitimate interest in the ways that particular digital objects are managed, preserved and disseminated. This poster presents the concept of "curatorial intent" as a way of operationalizing the values of stakeholders. In order to enable specific digital curation practices, curatorial intent must ultimately be expressed in terms of policies, rules and services for the (1) production of significant properties of digital objects and (2) control over access to particular components or levels of representation.

This poster presents a framework for ethics of digital curation that is designed to reflect both well-established professional values and the representational complexity of digital collections. It also explores implications and potential applications of the framework for the acquisition of digital materials, including approaches (policies, procedures and interfaces) for eliciting, recording, implementing and testing the curatorial intent of Producers at the point of donation or submission.

Demo 001

DataStaR: An Institutional Repository to Promote Sharing and Publication of Digital Research Data

Dianne Dietrich (presenter) Gail Steinhart , Brian Lowe, Brian Caruso, Jon Corson-Rikert, Ann Green, and Kathy Chiang
Albert R. Mann Library, Cornell University, Ithaca, NY, USA

DataStaR, a Data Staging Repository developed and in use at Cornell University, is a platform that researchers can use to share data with selected colleagues, document research data to the standards most in demand by Cornell researchers, and publish data and metadata to external repositories or to Cornell's own institutional repository. The intended users of DataStaR are primarily the creators of "small science" data, that is, data sets that can typically be distributed in their entirety as a file or set of files, and do not require specialized infrastructure to access or manipulate the data. By offering local support for documenting and publishing data sets, we aim to empower researchers to share more widely data that may otherwise never be made public. Our experience so far has shown that there is local demand both for infrastructure to support sharing of data while research is in progress, as well as distributing completed data sets to a broader audience.

The DataStaR platform itself consists of a Fedora-based repository for storage of data sets, while the infrastructure for managing metadata is based on Vitro, an integrated ontology editor and semantic web application developed at Mann Library[12]. Our assumption in taking a semantic web approach to metadata management is that scientific communities will increasingly use formal semantics and ontologies for describing scientific data, indeed, several examples of this already exist. An advantage of this approach for users of the DataStaR system is the ability to reuse information, thus reducing manual input. At a higher level, this approach lays the foundation for future semantic interoperability.[13] This demonstration will show a user's view as well as an administrator's view of the DataStaR platform

[12] http://vitro.mannlib.cornell.edu/

[13] Lowe, Brian. 2009. DataStaR: Bridging XML and OWL in Science Metadata Management. 3rd International Conference on Metadata and Semantics Research. Sept.30-Oct. 2, 2009. Milan, Italy.

Demo 002

A Demo of The Digital Curation Exchange: An Interactive Space for All Things Digital Curation

Heather Bowden, Creator/Site Manager
University of North Carolina at Chapel Hill

Introduction

Digital curation is a diverse area of discourse which has been emerging and evolving in different locations, at different rates, and in different forms. As digital curation practice, research, and education have become more prevalent, calls for spaces to connect its disparate occurrences have been on the rise. Responses to these calls have taken the form of conferences , journals, blogs, wikis, and a few online discussion forums.

Each of these methods of communication and information sharing have served their function of bringing much needed cohesion to digital curation. There exist, however, some gaps between the digital curation entities which still need to be filled. These entities such as the members of the research, education, and practice arenas of digital curation can all benefit from shared tools, knowledge, and experiences. A space where all of these groups can exchange ideas and share their experiences, regardless of geographic or temporal position, would serve to further strengthen the foundations of digital curation.

The Digital Curation Exchange

The Digital Curation Exchange [DCE] (www.digitalcuratinexchange.com) has been created to serve as a space for conversation, sharing and interaction among practitioners, researchers, educators, and students of digital curation. The site has been designed to take advantage of the social networking capabilities which have been made available through Drupal, an open source content management platform which uses a combination of PHP and MySQL.

These social networking tools have been configured in such a way as to enable a member to create an individual profile, create blog posts, post information in discussion forums, upload files, share links and videos, create groups, and comment on other members' posts. All of these capabilities enable the users to maintain distinct, personal identities while also participating in and contributing to the larger community. The purpose of the site then is to provide a platform for the development of this community which will then facilitate the sharing of knowledge across the spectrum of parties who participate in digital curation.

The DCE brings added value on several fronts. It is a free space which is open to all variety of digital curation participants and it provides an environment for exchanges across all perspectives. It also provides a wide variety of tools for information sharing which can be used in multiple ways to facilitate both individual and group experiences.

Acknowledgements
The Digital Curation Exchange is supported by funds granted by the IMLS [RE:05-08-0060-08] to DigCCurr II at the University of North Carolina at Chapel Hill, School of Information of Library Science.

Demo 003

Managing Research Assets for the Long Term: End to End Preservation Solutions Demonstrated Using Digital Curation Centre Tools

Esther Conway, Brian McIlwrath and Matthew Dunckley, Science and Technology Facilities Council

Abstract

The challenge of digital preservation of scientific data lies in the need to preserve not only the data itself but also the ability it has to deliver knowledge to a future user community. A true scientific research asset allows future users to reanalyze the data within new contexts. Thus, in order to carry out meaningful preservation we need to ensure that future users are equipped with the necessary information to re-use the data. This demonstration will present an overview of a preservation analysis methodology which was developed in response to that need. We intend to place it in relation to other digital preservation processes discussing how they can interact to provide archives caring for scientific data sets with the full arsenal of tools and techniques necessary to rise to this challenge. We present a fully worked end to end example which utilises Preservation Analysis, Preservation Network Modelling, Registry Repository of Representation Information, Repinfo Toolkit and Packaging Tools developed by the Digital Curation Centre.

The analysis method facilitates modelling of information networks based on the archival information package solution. Using illustrative examples we intend to show how these network models are a representation of the digital objects, operations and relationships which allow a preservation objective to be met for a future designated community. The model provides a sharable, stable and organized structure for digital objects and their associated requirements. They expose the risks, dependencies and tolerances within an archival information package. This allows for the automation of event driven or the periodic review of archival holdings by knowledge management technologies. We will demonstrate how the clear definition of relationships also facilitates the identification of reusable solutions which can be deposited within the Registry Repository of Representation information, thus sharing preservation efforts within and across communities.

Research assets are likely to be held by and transferred between institutional repositories. These repositories need to be designed planned and managed, competing for resources within complex organizational structures. In this paper we intend to conclude by touching upon how preservation analysis can inform audit and certification processes such as DRAMBORA and TRAC or planning activities for new repositories such as PLATTER. By doing this we allow preservation analysis at the data set level to be placed within the context of institutional planning and operations

In conclusion this approach seeks to apply the practices of quality assurance, risk/knowledge management, analysis, planning and modeling in a structured controlled manner to optimize beneficial re-use of institutions scientific research assets for the community it s

Demo 004

Demonstration of OECD iLibrary
Application of bibliographic and citation standards to Datasets and Data Tables
Terri Mitton, Project Manager, Statistics Dissemination, OECD Publishing
Organisation for Economic Co-operation and Development (OECD)

In mid-2009 OECD launched a new online platform, OECD iLibrary, which makes datasets and tables as discoverable and citable as the Organisation's published books and papers[14]. This innovative online library provides a complete integrated knowledge base for all of the Organisation's data and analysis.

In April 2009, OECD published a white paper titled "We Need Publishing Standards for Datasets and Data Tables", which examines the problems with current data discoverability and citations and proposes a remedy in creating industry standards for bibliographic dataset metadata and linking[15]. Following the release of this paper, OECD implemented the proposed bibliographic and citation standards for dataset and data tables. The metadata schema is generic and could be applied to any dataset or data table, not just statistics. The Organisation's publishing catalogue was enhanced to manage the bibliographic metadata for datasets and tables alongside other publications. Statistical editors quality assured added a layer of publishing metadata and catalogued 361 datasets, 42 key tables and over 5000 static tables and graphs in analytical books.

The OECD iLibrary search results are granular, providing integrated access to chapters, tables, graphs, books, articles and databases. Every publication is displayed in its own context, linking to the previous editions and various available file formats. Users of the OECD iLibrary can easily download citations for datasets and tables in a form compatible with popular bibliographic management systems. All of the datasets and tables have digital object identifiers (DOIs) deposited with Crossref registration agency so that they can be referenced. Every published dataset or table in a given language has its own home page with the title, DOI, abstract, publication date, periodicity, links to view and download the published data, and links to related publications. Librarians will be offered standard MAchine Readable Cataloguing (MARC21) records for datasets, alongside records for books and periodicals. Users will be able to discover, cite and link to datasets and tables as easily as any other published output of the Organisation. This integrated platform was built in response to users' needs and it confirms the central importance of the statistical datasets and tables to OECD's publishing programme.

This demonstration will highlight the application of bibliographic metadata and citation standards on datasets and tables in the OECD iLibrary. It will show the benefits of linking analytical publications, such as books and journals, to datasets and

[14] [Internet Library] OECD iLibrary, www.oecdilibrary.org

[15] [White paper] Green, T (2009), "We Need Publishing Standards for Datasets and Data Tables", *OECD Publishing White Paper*, OECD Publishing. doi: 10.1787/603233448430, http://dx.doi.org/10.1787/603233448430