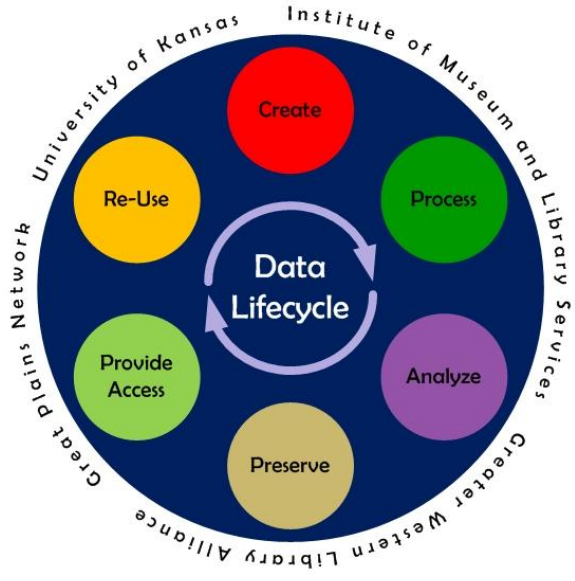
THE REPORT

PLANNING FOR THE LIFECYCLE MANAGEMENT AND LONG-TERM PRESERVATION OF RESEARCH DATA: A FEDERATED APPROACH

June 18, 2013

IMLS 51-12-0695



UNIVERSITY OF KANSAS LIBRARIES & INFORMATION TECHNOLOGY

GREAT PLAINS NETWORK

GREATER WESTERN LIBRARY ALLIANCE

INSTITUTE OF MUSEUM AND LIBRARY SERVICES

PRELIMINARY REPORT
PLANNING FOR THE LIFECYCLE MANAGEMENT AND LONG-TERM PRESERVATION OF RESEARCH DATA: A FEDERATED APPROACH

A report to the Advisory Council, IMLS Grant 51-12-0695

June 18, 2013 Version 1.1

Grant Staff

Deborah M. Ludwig, Principal Investigator
Scott R. McEathron, Investigator
Bob Lim, Investigator
Paul K. Farran
Joni M. Blake, Investigator
Gregory E. Monaco, Investigator
Nicole A. Potter. Project Coordinator

Acknowledgements: We wish to acknowledge the efforts and contributions to this report from many individuals, including the members of the committees and research teams listed in Appendix A. We also thank Lars Hagelin of the Greater Western Library Alliance for the development of graphics and technical advice.

This project is made possible by a grant from the U.S. Institute of Museum and Library Services

---

TABLE OF CONTENTS

Section C. Review of Key Projects and Technologies

This section provides a useful reference with guide to various projects and technologies as exemplars representing five categories of data management services.

1. Curation of Data
2. Preservation of Digital Materials
3. Archiving & Repository Services
4. Storage Systems
5. Enabling Technologies

1. Curation of Data

> *Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.* [1]

In this section, we share collaborative or federated efforts to curate data including:

1. Australian National Data Service (ANDS)
2. Data Conservancy
3. DataONE
4. Dataverse
5. Digital Curation Centre
6. Dryad
7. The Interuniversity Consortium for Political and Social Research (ICPSR)
8. Texas Advanced Computing Center TACC
9. The Digital Archaeological Record (tDAR)
10. UK Data Archive

[1] Graduate School of Library and Information Science, The iSchool at Illinois. http://www.lis.illinois.edu/academics/programs/specializations/data_curation

Australian National Data Service (ANDS) **CURATION**

Deborah Ludwig

http://www.ands.org.au/

Brief Description of the Project

The Australian National Data Service (ANDS) is a project arising out of the need for platforms for collaboration under a plan by the Australian National Collaborative Research Infrastructure Strategy (NCRIS). It is one of several discrete services that addresses collaboration.[2] It is lead by Monash University.

According to the website:

ANDS aspires to build Australia's Research Commons, as a "cohesive collection of research resources from all research institutions, to make better use of Australia's research data outputs." The data commons component registers data, which resides in repositories across a network of institutions.

ANDS is transforming Australia's research data environment to:

- *make Australian research data collections more valuable by managing, connecting, enabling discovery and supporting the reuse of this data*
- *enable richer research, more accountable research; more efficient use of research data;*
- *and improved provision of data to support policy development*

Monash University[3] leads the ANDS partnership with the Australian National University[4] (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO)[5].

The scope of ANDS as a collaborative effort is extremely broad. ANDS includes a rich mix of public sector partners such as the Australian Bureau of Statistics, the National Archives of Australia, the Australian Institute of Health and Welfare working alongside university partners and NCRIS organizations. There are several communities of practice for metadata, software infrastructure and tools, data managers, public sector, and specialist communities. The lengthy list of projects is on the ANDS website.[6]

Specific services of ANDS include:

- *Cite My Data* for assignment of persistent Digital Object Identifiers

[2] Treloar, Andrew. "Design and Implementation of the Australian National Data Service." *International Journal of Digital Curation.* 4, no. 1 (2009): 125-137.
[3] http://www.monash.edu.au/
[4] http://www.anu.edu.au/
[5] http://www.csiro.au/
[6] https://projects.ands.org.au/getAllProjects.php?start=all

- *Publish My Data* for registering descriptions of data collections. ANDS does not store the data itself.
- *Register My Data* is a more complex metadata registry for data collections which also allows metadata to be exposed and included in other discovery services
- *Identify My Data* is a service to attach Handles (Corporation for National Research Initiatives or CRNI service) to data.

Reviewer's Analysis

ANDS is a project to create a data commons by registering metadata for data held in various repositories at a wide group of institutions. It seems similar to DataONE in terms of providing a metadata registry with links to distributed nodes. A difference is in scope with DataONE being focused on earth science data and ANDS being focused on many disparate types of research data. More would need to be discovered about local or institutional data sources and expectations for the treatment of those data sources for permanence.

Considerations and Recommendations

If GWLA and GPN member institutions are interested in creating a research commons approach to create more knowledge about available data, this would be an interesting model to further digest. Note that it was built initially on three years of external funding.

Key Contacts: contact@ands.org.au
Sponsors: Australian Commonwealth Department of Education, Science, and Training (DEST).
Funding: 3 years of initial funding to develop the project. Ongoing is not clear from sources consulted.
Inception: 2007
Geographic Location: Australia

Data Conservancy **CURATION**

http://dataconservancy.org/

Philip Konomos, Arizona State University

Brief Description of the Project

The Data Conservancy (DC) is an initiative to support the preservation, management, and re-use of scientific data, particularly data that results from grant funded research projects. The project started as a collaboration between the Sheridan Libraries at Johns Hopkins University and the Sloan Digital Sky Survey (SDSS). Sheridan was tasked with curating the resulting astronomical data from the SDSS and developed a data curation framework to complete the work.

Over time, using funding from the 2007 NSF DataNet solicitation as well as the Institute for Museum and Library Services, the DC has expanded its focus to four areas of science information research:

- conducting ongoing needs assessment among science researchers to develop new digital data curation tools and services;
- researching and developing the cyberinfrastructure needed to curate, preserve and make science data accessible;
- working with Library and Information Schools on data curation education and professional development; and,
- examining strategies for the long-term sustainability of repositories and data curation centers (business model).

Reviewer's Analysis

The Data Conservancy is still in development, but aims to include an open, digital repository of science data as well as tools and services to enhance the re-use of data. One example of a research tool is the Feature Extraction Tool - the ability to find and integrate disparate data sets in the network of DC repositories, by querying using taxonomic, spatial or temporal terms. There is no public interface allowing querying or access to data currently, however. Instead, staff and affiliated scientists are working on analyzing the needs of the science communities and building appropriate curation and analysis tools.

Considerations and Recommendations

DC is actively seeking partner institutions to download and install an "instance" of their software. For GWLA members who have strengths in astronomy, earth science, life sciences, or social sciences, this might be a worthwhile partnership. However, institutions would have to be willing to commit significant technical staff time to installing, configuring and developing the DC software.

Key Contacts: Sayeed Choudhury, Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center Sheridan Libraries, Johns Hopkins University

Sponsors: The Data Conservancy began at the Sheridan Libraries at Johns Hopkins University, but the current model is to create a community of partners (universities, research centers, institutions) with instances of the cyberinfrastructure and data, who help with continued research and development of new tools and services.

Funding: In 2007 NSF issued a call for proposals to create cyberinfrastructure that would curate and make accessible the increasing amounts of grant funded science research data. Two proposals were funded starting in 2008: DataONE, a repository covering ecology, evolutionary, and earth science data; and the Data Conservancy, a repository network focused on astronomy, earth science, life sciences, and social science data. Additional funding for DC has also come from the Andrew W. Mellon Foundation and the Institute of Museum and Library Services (IMLS).

Inception: 2007-2008

Geographic Location: Located at Johns Hopkins University Library, but including a network of other universities with instances of the infrastructure.

DataONE (Data Observation Network for Earth) **CURATION**

http://www.dataone.org/

Michael Bolton, Texas A&M University, Sterling C. Evans Library

Brief Description of the Project

From the web site, DataONE is one of the original DataNets supported by the U. S. National Science Foundation grant #OCI-0830944. It provides a distributed framework and sustainable cyber infrastructure to provide open, persistent, robust and secure access to well-described and easily discovered Earth observational data. In short, DataONE provides a comprehensive set of tools and a repository for collecting metadata on datasets dealing with Earth, environmental, atmospheric and ecological sciences. The infrastructure is composed of THREE (3) coordinating nodes to provide the network-wide services. A growing number of member nodes provide data to the coordinating nodes where it is indexed and replicated across all 3 nodes. In addition to the repository service, DataONE also provides a rather long list of tools referred to as the Investigator Toolkit. These tools cover the full spectrum of data management, from preparing the data management plan to archiving metadata in the ONEShare repository. DataONE also has a strong outreach component that includes educational services as well as a User's Group that helps promote the service. DataONE is a fairly mature service, well documented and accessible. Its governance model is equally mature with an Executive Team, a Leadership Team and a fairly large number of working groups.

Reviewer's Analysis

DataONE provides a very good model for DataNet, either regional or focused on a particular discipline, such as Earth Sciences. The architecture lends itself to a consortia approach in that a series of super-nodes, Coordinating Nodes for DataONE, aggregate data for a larger number of member nodes. Multiple, geographically dispersed coordinating nodes are key to the preservation model used in DataONE.

It should be noted DataONE does not store data, just the metadata. The actual datasets reside at the originating member node. DataONE facilities an integrated search and discovery service and also provides replication services to the member nodes. So I do not see DataONE as a real preservation network at this time, which puts a premium on deploying something like DPN [the Digital Preservation Network] for deep archiving of data.

It includes a number of nice tools, which we can elect to add to our offering. Some of these tools have been in existence for some time while others were developed as part of the DataONE grant. The use of Mercury as the search engine, along with the harvesting into a centralized index makes for a very strong search interface.

Considerations and Recommendations

This would be a good place to start in considering a model. In looking at the Digital Preservation Network model, the idea of coordinating and contributing nodes seems to be a common model. I think we should investigate how we could use the DataONE architecture for our project. The DataONE Architecture documents are online at http://mule1.dataone.org/ArchitectureDocs-current/index.html

Key Contacts: PI – William Michener, Executive Director – Rebecca Koskela
Sponsors: University of New Mexico
Funding: DataONE is supported by NSF grant #OCI-0830944 – part of the initial DataNets project. The grant is estimated to expire July 31, 2014.
Inception: From the NSF grant, August 1, 2009 is the start date.
Geographic Location:

Current configuration has THREE (3) Coordinating Nodes located at;

- The University of New Mexico
- The University of California Santa Barbara
- The University of Tennessee (collaboration with Oak Ridge National Laboratory)

Member nodes, of which there are 10 at this time, are all over the country.

The Dataverse Network **CURATION**

http://thedata.org/

Jon Wheeler, University of New Mexico

Brief Description of the Project

The Dataverse Network is an open source web application for hosting research data and provides flexible tools for the management of institutional branding, user accounts and use/access requirements. Additional features include dynamic analysis and visualization tools, versioning, reformatting of some statistical file types, and use tracking. Depending on the original file type, some data can be converted to preservation format on upload, with DDI and Dublin Core metadata stored in XML format. Additional metadata profiles can be specified through templates. Each Dataverse Network is composed of one or more individual Dataverse, and may be distributed across multiple departments or organizations. Individual Dataverse can be defined at various levels – per institution, per researcher, per project, etc.

Reviewer's Analysis

The Dataverse Network is a feature rich application with regard to publication and post-publication data management activities. In particular, the facets of data sharing and reuse are well supported through granular access management and user comment features. Many frequently cited concerns regarding data publication are addressed - user permissions can be specified at the file level, allowing for multiple access options even within a single project. Access terms can be customized, and the guestbook features allow researchers and administrators to require information from users prior to authorizing data download. Organizations running a Dataverse Network can enable online analysis and visualization by installing an optional R server package, allowing users to interact with the data for validation or selection purposes. Formatted data citations and handles are provided. Finally, archival and preservation processes are supported through scheduled XML metadata exports as well as LOCKSS and OAI harvesting utilities. (Metadata profiles for the physical sciences and humanities can be specified in addition to the default DDI.)

Other phases dealing with data creation and analysis workflow are less well represented. User comments and versioning features apply to static, posted data sets, but robust collaboration and version control tools would have to be managed by other means.

With regard to implications for a collaborative or consortial approach to research data management, the flexible networking and granular access controls are significant. Making use of these features, a lead institution could centrally manage the Dataverse Network application while contributing institutions remotely manage their individual Dataverses. Use and access policies can be configured locally, so a one-size-fits-all approach would not be required unless agreed upon by consortia members.

Considerations and Recommendations

The project does merit further review and analysis, as it is under active development and has already achieved a sizable user base within the social sciences. Additionally, a Dataverse Network plugin for the Open Journal System is in development, which may broaden the adoption of this resource.

Key Contacts: Sonia Barbosa, Eleni Castro
Sponsors: Institute for Qualitative Social Science at Harvard University, Massachusetts Institute of Technology
Inception Date: Development began in 2006

Digital Curation Centre (DCC) **CURATION**

http://www.dcc.ac.uk/

Susan Matveyeva, Wichita State University

Brief Description of the Project

From the web page:

> The Digital Curation Centre (DCC) is a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management across the UK's higher education research community.

DCC provides expert advice, training, support and consultancies for the UK higher education's research community.

The U.S. works with DCC in international organizations, for example in the Committee on Data for Science and Technology CODATA (http://www.codata.org). The U.S. National Research Council's Board on Research Data and Information represents US in CODATA. Francine Berman and Clifford Lynch among members of US/CODATA.

Reviewer's Analysis

DCC is a comprehensive reference source on research data curation. The Center website offers well-structured information, resources, and tools to everyone interested in data curation: briefing papers, how-to guides, curation reference manual, curation lifecycle model, policy and legal resources, data management plans, sets of tools including Collaborative Assessment of Research Data Infrastructure and Objectives CARDIO (http://cardio.dcc.ac.uk/), Data Asset Framework DAF (http://www.data-audit.eu/index.html), Digital Repository Audit Method Based on Risk Assessment DRAMBORA Interactive toolkit (http://www.repositoryaudit.eu/); case studies, standards, training courses, and research and development resources.

DCC is one of major reference sites, so there is always something useful, including training materials; standards watch; tools, programs, and reports:

- Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) tool can be useful. http://www.dcc.ac.uk/projects/cardio
- Neil Beagrie, JISC Benefits from the Infrastructure Projects in the JUSC Managing Research Data Programme. Final Report, version 5.0, Sept. 2011 http://www.jisc.ac.uk/media/documents/programmes/mrd/RDM_Benefits_FinalReport-Sept.pdf
- Disciplinary metadata standards: http://www.dcc.ac.uk/resources/metadata-standards
- Catalog of resources for curators and researchers: http://www.dcc.ac.uk/resources/external/tools-services

Considerations and Recommendations

The DCC Web Site can be included to the list of useful resources. DDC training materials and tools can be used especially when similar U.S. materials have not yet developed. In more direct approach, specific training and assessment materials, tools, and recommendations may be selected and adapted as needed during planning and implementation of the GWLA/GPN initiative.

Key Contacts: Kevin Ashley, Director kevin.ashley@ed.ac.uk +44 131 651 3823; Sarah Jones, Senior Institutional Support Officer sarah.jones@glasgow.ac.uk;
Twitter: sjDCC Phone: +44 141 330 3549

Sponsors: HATII, UKOLN and STFC, with the University of Edinburgh (from March 2004 through February 2010). From March 2010 the DCC has reorganized into a three-cornered consortium, led from Edinburgh, with the following Principal Partners: Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow and UKOLN, a center of expertise in digital information management, based at the University of Bath.

Funding: JISC

Inception: 2004-03-01

Geographic Location: Edinburgh

Dryad **CURATION**

http://datadryad.org/

Susan Matveyeva, Wichita State University

Brief Description of the Project

From the Dryad's web page at NCSU:

> *Dryad is a joint project of NESCent and the UNC Metadata Research Center, with North Carolina State University participating as a development partner, along with the University of New Mexico, and Yale University, focusing on creation of a repository for data underlying scientific publications, with an initial focus on evolution, ecology, and related fields. Dryad allows investigators to validate published findings, explore new analysis methodologies, repurpose the data for research questions unanticipated by the original authors, and perform synthetic studies such as formal meta-analyses."*

Reviewer's Analysis

Dryad addresses all key facets of data lifecycle management. It is focused on sustainability, preservation, reuse of data, and automation of data flow and exchange. Developers continue adding new functionality. The Dryad UK Project (2010-2011) created a mirror of the Dryad repository and developed a sustainability plan to help Dryad become established as an international not-for-profit organization empowered to ensure long-term preservation and accessibility of its data holdings.

Software: DSpace Manakin XMLUI customized by @mire (http://atmire.com/website/) with the following features: dataset embargo; dataset security; Discovery SOLR Search System customization; enhanced submission and workflow; configurable workflow; item versioning; integration with EZID, DOI, DataCite and PubMed. Some customizations are now standard features of DSpace.

Search: SOLR Discovery faceted browse and search. Includes indexes:

authority -- terms for auto-completion using controlled vocabularies, including HIVE
dataoneMNlog -- log of accesses through the DataONE API
dryad -- local storage of DOIs
search -- primary search index
statistics -- log of accesses to item pages and bitstream downloads

Access to Data: Authors release their data in public domain under the terms of a Creative Commons Zero (CC0) waiver to emphasized data sharing and increase data reuse. Embargo is an available option.

Content submission: Data is submitted as part of the publication process. "Journal integration with Dryad is available at no cost for any journal that wishes to implement low-burden data archiving and enhance their published articles with links to data." No restrictions on file format. Data packages include data files and "ReadMe" file. Each data package is assigned DOI and linked to the journal article. Submission is simple: (see 2-min. video: http://www.youtube.com/watch?v=RP33cl8tL28&feature=youtu.be)

Content sharing: Dryad uses EZID service from the California Digital Library that manages DOI registration with DataCite. Dryad is a node of DataONE.

Reusability: LinkOut functionality. PMID in Dryad metadata. Citation sharing technology (Cite and Share in RIS and BibTex). Handshaking (the process of coordinating submission between Dryad and specialized repositories in order to (a) lower user burden by streamlining the submission workflow and (b) allow Dryad and specialized repositories to exchange identifiers and other metadata in order to enable cross-referencing of the different data products associated with a given publication (integration with GenBank and TreeBASE).

Integration: Integration with ORCID/DataCite is in a planning stage.

Metadata: Metadata profile: Dublin Core with addition of few fields from DDI, Darwin Core and PRISM (see http://www.cendi.gov/presentations/11-17-09_cendi_nfais_Greenberg_UNC.pdf). Metadata generation is automatic or semi-automatic.

Dryad implemented an automated integrated submission from multiple publishers' sites: http://wiki.datadryad.org/Submission_Integration. The permanent link between the article and data packages is created during submission. A similar approach can be used to link records and/or full text/ or data in the local repository/storage and the consortium's central hub.

Workflow: Dryad's workflow practices of assigning DOI to data packages is another process to look at during the planning step of the regional consortium development.

Considerations and Recommendations

I like Dryad's careful approach to the enhancement of qualified Dublin Core records with fields from other metadata. These enhancements do not destroy the integrity of a Dublin Core record; inclusion of the elements of other schemas is as minimal as possible. Only few necessary elements are included: scientific name of a plant (Darwin Core dwc:ScientificName); journal name (Prism prism:publicationName). I would recommend similar approach for multidisciplinary repository: have major schema (e.g. Dublin Core) and few fields from the metadata schema of the appropriate discipline (if it was developed).

I recommend looking more closely at Dryad's best practices in metadata automation, use of identifiers, integration with multiple publisher's platform during submission, data packages sharing with other repositories (http://wiki.datadryad.org/BagIt Handshaking), versioning, curation practices and reports, experience of partnership with DataONE, DataCite, CLOCKSS, ORCID. Dryad's site has good documentation: http://wiki.datadryad.org/Main_Page See also Curation Manual: http://wiki.datadryad.org/wg/dryad/images/8/85/Curation_man_2012-12-21.pdf The Dryad practices analysis and usage would be especially useful if the committee choose DSpace as a software platform.

Key Contacts: Jane Greenberg, metadata management (Metadata Research Center; University of North Carolina at Chapel Hill) Phone: 919-962-8066 ; Fax: 919-962-8071; Email: janeg@email.unc.edu
Hilmar Lapp, technical management (NESCent); Todd Vision, project director (NESCent/University of North Carolina at Chapel Hill)

Sponsors: Dryad developed by the National Evolutionary Synthesis Center and the University of North Carolina Metadata Research Center, in collaboration with several Partner Organizations..
U.K. development partners: JISC (the Joint Information Systems Committee), Oxford University, and the British Library (see the DDC project DryadUK: http://www.dcc.ac.uk/projects/dryaduk)
Dryad is governed by a twelve member Board of Directors elected by members, representing publishers, societies, research and educational institutions (http://wiki.datadryad.org/wg/dryad/images/9/94/Dryad_ByLaws_April2012.pdf

Funding: NSF grant DBI-0743720 (2008-2012);
NSF grant 2012-2016 (Abstract DBI-1147166);
NESCent, the NSF-funded DataONE;
IMLS grant (LG-07-08-0120-08)

Business plan and Sustainability: Dryad is currently applying for status as a 501(c)3 not-for-profit to be incorporated in North Carolina. It also plans to charge membership and submission fees (http://wiki.datadryad.org/Business_Plan_and_Sustainability). Staff will continue applying for R&D grants.

Inception: 2008
Geographic Location: hosted by the North Carolina State University

Inter-university Consortium for Political and Social Research (ICPSR) CURATION

http://www.icpsr.umich.edu
Deborah Ludwig

Brief Description

The Inter-university Consortium for Political and Social Research is a consortium of over 700 institutions. Membership includes research institutions, colleges and universities. ICPSR's mission, according to the website is to provide "leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community. "

The website notes that "ICPSR maintains a **data archive** of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields." In addition to acting as a data archive, ICPSR includes a focus on training and education for data professionals and researchers.

Unlike some models that seek to unify data collections held in disparate locations, ICPSR is a centralized data archive. Data are deposited in the ICPSR archive and held there with an advanced emphasis on long-term preservation.

Reviewer's Analysis

ICSPR celebrated its 50th anniversary this year, so they have delved into issues such as metadata and data preservation at a level that not many other data archives have approached. They utilize a subscription based membership model. They focus on the social science domain and are a key disciplinary repository.

Considerations and Recommendations

If a centralized repository approach were envisioned for GWLA and GPN, a closer look at their operational and business model could inform our work and help us develop our success criteria. Noting their strategies for preservation could also inform our work.

Key Contacts: http://www.icpsr.umich.edu/icpsrweb/content/membership/contact-people.html
Sponsors: University of Michigan
Funding: Membership-based model
Inception: 1962

Texas Advanced Computing Center **CURATION**

http://www.tacc.utexas.edu

Deborah Ludwig

Brief Description

The Texas Advance Computing Center is support by the National Science Foundation the University of Texas, Austin, and through other grants. It is one of 11 national centers for advanced computing to support computational research. Project partners include the Department of Energy, the National Oceanic and Atmospheric Administration (NOAA), and the National Archives and Records Administration (NARA).

The website states:

> *Computational science has become the third pillar of scientific discovery, complementing theory and physical experimentation, allowing scientists to explore phenomenon that are too big, small, fast, or dangerous to investigate in the laboratory. Thousands of researchers each year use the computing resources available at TACC to forecast weather and environmental disasters such as the BP oil spill, produce whole-Earth simulations of plate tectonics, and perform other important research.*

Within TACC, there is a Data Management and Collections Group (DMC)[7] that was established in 2008. The DMC group "builds and maintains large data-management and storage resources and consults with collections' creators in all aspects of the data lifecycle, from creation to long-term preservation and access. The DMC group actively seeks out research and grant proposal collaborations with researchers and institutions with collections of interest." Data management is built on the iRODS platform. A recent presentation at the Preservation and Archiving Special Interest Group[8] by TACC's Dan Stanzione states that there are 20 true data "collections" mixed in with "user storage" for data and projects.

TACC has two funding models: pay annually or pay once, store forever (which works well for researchers on grant funding.) Forever is about 5 years for most projects.

Data collections are hosted in a system called Corral, which includes storage for structure and unstructured data, lots of different interfaces to the data and iRODS-based data management services underlying the collections. Corral had over 700 TB of data as of 2012. The Corral services include hierarchical storage with multiple disk speeds, multiple encryption and data mechanisms for HIPPA and other secure or confidential data. Preservation of collections is also under development. TACC Deputy Director Dan Stanzione is part of The Digital Preservation Network (DPN) leadership group.

[7] http://www.tacc.utexas.edu/tacc-projects/dmc

[8] PASIG 2012, Austin http://lib.stanford.edu/preservation-archiving-special-interest-group/presentations-pasig-meetings-austin-texas-january-2012

Reviewer's Analysis

Corral data collections as implemented by TACC offers a grand view of what can be done at the intersection of high performance computing, research, and building collections of reusable research data.

Consideration and Recommendations

While building the level of infrastructure housed at TACC might be out of range for GWLA and GPN, perhaps there are opportunities to work with one or two high performance computing centers on a regional basis as part of a next-phase grant project. This should be explored.

Key Contacts: Dan Stanzione, **dan@tacc.utexas.edu** http://www.tacc.utexas.edu/staff/dan-stanzione
Sponsors & Funding: NSF & University of Texas
Inception: 2001
Geographic Location: Austin Texas

tDAR – (the Digital Archaeological Record) **CURATION**

http://tdar.org

Philip Konomos, Arizona State University.

Brief Description of the Project

tDAR, **t**he **D**igital **A**rchaeological **R**ecord, is an international repository for digital archaeological data, reports, and other files. Developed by faculty and staff at Arizona State University, tDAR's mission is to make accessible, and preserve in perpetuity, the digital files (data as well as text) that are generated as a result of archaeological investigations. Physical materials recovered during archaeological projects (artifacts, organic material, paper forms, etc.) are typically curated in museums but digital records from these projects are vulnerable to loss or neglect. Often the CD or DVD media that contain the digital files are treated as additional "artifacts" from the project and boxed with the pottery and stone tools. tDAR provides a stable preservation environment that makes these digital files discoverable and accessible.

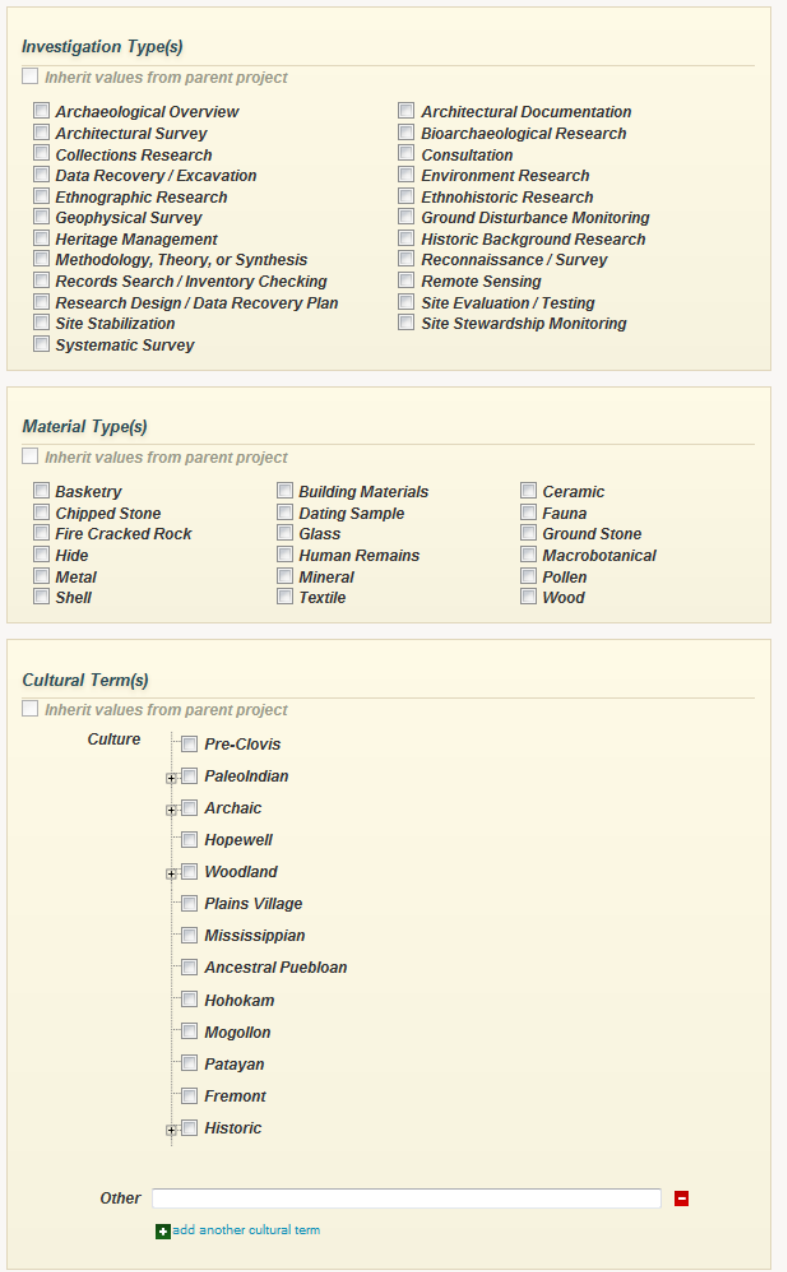
Reviewer's Analysis

This is one of several archaeological disciplinary repositories that have been developed over the past decade. One strength of tDAR is that it accepts a wide variety of file types (GIS data, tabular data sets, spreadsheets, text files, images, 3D scans, etc.) from any user (university faculty member, Federal agency archaeologist, state historic preservation officer, etc.).

Note: 1 Geographic Search Screen

Users can browse and search tDAR for free, without creating an account. Downloading files requires an account, which is free and simply involves providing a name, email address, user ID and password. Fees are charged to upload files to the repository.

tDAR is primarily an open repository but a variety of accommodations can be made for different user needs. Sensitive information in tDAR (such as exact archaeological site locations) can be redacted from reports and specific files can be marked confidential with access limited to specified individuals only. Metadata fields are extensive and customized to support the professional vocabulary common to archaeology. Users can search by drawing a box on a map as well as by using key words related to culture area, time period, material type, or archaeological site type :



Note: 2 Other metadata fields.

tDAR also provides a database integration tool for advanced research. Users looking for specific information (show me all of the database tables that contain information on fish remains) can query the database files in tDAR and get results that only include the rows and columns specified in the query.

Considerations and Recommendations

tDAR contains a large, varied collection of text and data files useful for teaching and research. Staff are developing example curriculum modules for use in university classes. tDAR provides a

helpful research tool for undergraduate and graduate students beginning research on a particular archaeological topic or geographic area. Librarians, particularly anthropology subject specialists, will find this a useful resource.

Key Contacts: The tDAR email is comments@tdar.org . The Digital Antiquity email is info@digitalantiquity.org . Francis Pierce-McManamon, Executive Director, Center for Digital Antiquity
Arizona State University
PO Box 872402
Tempe, AZ, 85287-2402

Adam Brin, Director of Technology, Center for Digital Antiquity
Arizona State University
PO Box 872402
Tempe, AZ, 85287-2402

Sponsors: tDAR is managed by the non-profit organization Digital Antiquity, at Arizona State University. Digital Antiquity is run by a staff of five, and is governed by a Board of Directors and a Professional Advisory Panel.

Funding: tDAR, through Arizona State University faculty and staff, has received over 4.3 million dollars in support from the National Science Foundation, the Andrew W. Mellow Foundation, the National Endowment for the Humanities, and the UK-based Joint Information Systems Committee (JISC).

Inception: Software development began in 2004 and the public website (supporting ingest, search, browsing and download) was launched in 2009.

Geographic Location: Offices are located at Arizona State University, but the repository contains international archaeological data.

UK Data Archive **CURATION**

http://data-archive.ac.uk

Greg Monaco, Great Plains Network (GPN)

Brief Description of the Project

The UK Data Archive is an all-in-one portal to create and store/deposit, manage, and find shared data. This project *curates the UK's largest collection of digital social and economic research data*, including *data from government departments, researchers and research institutions, public organisations and companies.* The archive may be searched using the ***Economic and Social Data Service*** website at http://www.esds.ac.uk/Lucene/Search.aspx. The steps of this project's data curation process are located at http://data-archive.ac.uk/curate/process.

Reviewer's Analysis

The project provides a one-stop portal for all aspects of the lifecycle management process for a slice of research data: "Our collection encompasses a significant range of data relating to society, both historical and contemporary, covering the social sciences, economics and humanities, as well as the societal aspects of environmental and medical data." Rather than distributed data, it appears that all data is centrally located and centrally curated.

This is an example of a complete approach to lifecycle management, and I suggest that we explore what works and what the limits to this approach may be. (One question is whether this is scalable to other domains of knowledge/data?)

Considerations and Recommendations

This appears to be a high visibility project. I recommend that we do a further review of this project in order to address similarities and differences in approach between this project and our ultimate proposed project.

This project, in association with JISC, uses federated access management, likely Shibboleth (see Shibboleth review). We should probably determine this, for sure. (http://data-archive.ac.uk/about/projects/identity-management)

Key Contacts:
Sponsors: University of Essex, JISC, European Commission
Funding: http://data-archive.ac.uk/about/projects/past
Inception: Website states that it was established over 40 years ago.
Geographic Location: United Kingdom

2. Preservation of Digital Materials

For the purpose of this report, we consider the practice of preserving digital research data to be aligned with the preservation of other types of digital information and refer to "digital preservation" rather than "data preservation."

> *Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time*.[9]

EXAMPLES OF DIGITAL PRESERVATION SERVICES AND FRAMEWORKS

1. Archivematica
2. Dark Archive in the Sunshine State(DAITSS)
3. DuraCloud
4. Digital Preservation Network (DPN)
5. Ex Libris' Rosetta
6. OAIS Reference Model for Preservation Services
7. TRAC Certification Process

[9] Definition prepared by the ALCTS Preservation and Reformatting Section, Working Group on Defining Digital Preservation. ALA Annual Conference, Washington, D.C., June 24, 2007. http://www.ala.org/alcts/resources/preserv/defdigpres0408

Archivematica **PRESERVATION**

*https://www.**archivematica**.org*

Nicole Potter and Deborah Ludwig

Brief Description of the Project

Archivematica is free, open-source Linux/MySQL software for digital preservation system. It is installed locally on an institution's servers. Archivematica provides a web-accessible dashboard that allows users to ingest digital objects. It also allows for administrative operations so that the user can modify preservation plans, storage locations, configuration of micro-services and user access levels. Archivematica also permits the user to enter metadata using various standards including METS, PREMIS, Dublin Core and other metadata standards.

Archivematica is compliant with the Open Archives Information System model, which is an ISO standard that defines the functions of digital preservation and uses a micro-services approach, which relies on an integrated suite of software applications to handle various tasks. Archivematica addresses three preservation strategies: emulation, migration, and normalization, to help create a standard technology platform for institution to ensure compatibility now and in the future.

Archivematica was originally developed for the city of Vancouver, BC to store records after the Olympics. The municipal archive of the City of Vancouver has a blog containing more information about the tool. [10] The Library of Congress has also mentioned Archivematica in it's a blog post. [11]

Reviewer's Analysis

Archivematica is essentially a wrapper around a set of software tools that create a pipeline to preservation for digital objects. It is a really interesting platform that continues to gain ground because it is reasonably lightweight in terms of installation and start up.

Considerations and Recommendations

If GWLA / GPN institutions are interested in a preservation service associated with any type of digital content, including research data perhaps held in a accessible open repository, Archivematica merits a closer look. It is still very early in release cycles. Alpha 0.9 was just released this fall.

Key Contacts: Peter Van Garderen, President, Artefactual Systems, email: info@artefactual.com
Sponsors: The UNESCO Memory of the World's Subcommittee on Technology, the City of Vancouver Archives, the University of Alberta Libraries, the University of British Columbia Library, the Rockefeller Archive Center, Simon Fraser University Archives and Records Management, Yale University Library
Funding: NA
Inception:
Geographic Location: NA

[10] http://opensourcearchiving.org/content/archivematica-city-vancouver-archives
[11] http://blogs.loc.gov/digitalpreservation/2012/10/archivematica-and-the-open-source-mindset-for-digital-preservation-systems/

Dark Archive in the Sunshine State (DAITTS) **PRESERVATION**

http://daitss.fcla.edu/

Deborah Ludwig, University of Kansas

Brief Description of the Project

DAITSS is a dark archive focused is on preserving "digital masters" and reconstituting those masters for access upon request. Access is handled either outside of the archive through other delivery mechanisms or by creation and delivery of a "dissemination information package" from the "archival information packet." (Refer to the review of the OAIS reference model if these terms are not familiar.)

DAITSS was developed by the Florida Center for Library Automation (FCLA) for use by the Florida Digital Archive (FDA) which is a digital repository shared by the eleven universities in the Florida public university system. DAITSS is strictly modeled on the Reference Model for an Open Archival Information System (OAIS). DAITSS can accept a Submission Information Package (SIP), transform the SIP into a stored Archival Information Package (AIP), and transform the AIP into a Dissemination Information Package (DIP) on request. To do so, DAITSS directly implements four of the six OAIS functional entities: Ingest, Data Management, Archival Storage, and Access. FDA staff performs functions of the remaining two OAIS entities, Administration and Preservation Planning, with support from DAITSS reporting and data management functions. DAITSS is unique among repository applications in that it was designed to ensure the long-term render-ability of authentic digital materials. DAITSS maintains standardized preservation metadata including digital provenance, and performs continuous fixity checking on multiple stored copies. The preservation protocol implemented by DAITSS combines bit-level preservation, format normalization, and forward format migration. [12]

Considerations and Recommendations:

How does the project address, or potentially address, key facets of lifecycle management? DAITSS is available for use through a GPLv3 license. The DAITTS website provides links to access to a fully configured VM version of DAITSS that can be downloaded to run under any VM manager, along with sample SIPs (submission packages) and documentation.

DAITSS was written for a multi-user environment and supports consortial as well as institutional preservation repositories. If a solution of interest was to provide a centralized point for multiple institutional deposit of research data coupled with a catalog for access and a service to request research data sets and to deliver those "on demands", DAITSS might provide a viable solution.

DAITSS could possibly be a component in a curation approach; however there may be other emerging choices to consider as well. DAITSS is certainly a quality open software solution available at no cost and built by a team that has been doing this work for a number of years for a consortium. There are several published articles that have been written about DAITSS as well.

[12] Caplan, Priscilla. "**DAITSS, an OAIS-based preservation repository" in** Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop. Article #17.

Key contacts: Pricilla Caplin
Sponsors: Florida Center for Library Automation, Florida Digital Archive (11 universities)
Funding: Florida Institutions and IMLS
Inception: Spring 2007
Geographic Location: Florida

Digital Preservation Network (DPN) **PRESERVATION**

www.dpn.org
Deborah Ludwig

Brief Description

Verbatim from the website:

The Digital Preservation Network (DPN) was formed to ensure that the complete scholarly record is preserved for future generations. DPN uses a federated approach to preservation. The higher education community has created many digital repositories to provide long-term preservation and access. By replicating multiple dark copies of these collections in diverse nodes, DPN protects against the risk of catastrophic loss due to technology, organizational or natural disasters.

Reviewer's Analysis

DPN is an effort to develop a group of preservation nodes that are not accessible accept to carry out preservation functions. The idea is that content can be restored to an access repository if lost and that data replication coupled with messaging between nodes will underpin preservation efforts. Members have paid into the development of the network ahead of its actual design and implementation to further these efforts.

Recommendations

Key Contacts: Email: inquiry@dpn.org / Phone: (434) 286-3436. Steven Morales is the director of DPN.
Steven Morales, steven.morales@dpn.org
DPN Program Director
434-286-3436

DuraCLOUD **PRESERVATION**

http://www.duracloud.org

Deborah Ludwig

Brief Description of Project

DuraCloud is a storage and preservation solution that is hosted or "in the cloud." DuraCloud allows storage of redundant copies at multiple storage provider sites based on a simple dashboard approach. The website advertises:

> *…Replication and backup activities, preservation and archiving, repository backup, and multimedia access. DuraCloud also acts as a mediation layer between you and cloud storage providers, therefore eliminating the risk of vendor lock-in. … [DuraCloud] does not address fine-grained policy and access control considerations. It can be used to house entire collections of confidential data, and/or support a system which provides granular controls, but it does not do so itself. DuraCloud does support basic authentication; and you can make spaces within DuraCloud dark or light.* [13]

DuraCloud is a service of DuraSpace[14] the support and development organization for DSpace and FedoraCommons repository software in common used around the globe.

Reviewer's Analysis

DuraCloud is an attractive option for cloud-based preservation. There are various storage options, including Amazon (and Glacier), Rackspace, and EMC. Archives can be dark or light (accessible). Media can be shared or streamed from DuraCloud if the archive is light. The Colorado Alliance of Research Libraries (Alliance) conducted a pilot in 2010 as did the BioDiversity Heritage Library, and the WGBH Media Library and Archives. [15]

Considerations and Recommendations

Consortial pricing is available. DuraCloud coupled with a discovery layer could serve some research data purposes and require limited local infrastructure. More information could be obtained from current customers or those who have undertaken pilots.

Key Contacts: info@duracloud.org; http://www.duracloud.org/contact
Sponsors: DuraSpace
Funding: not-for-profit
Inception: October 2009, first pilot projects
Geographic Location: N/A

[13] http://www.duracloud.org/faq
[14] http://www.duraspace.org/
[15] http://www.digitalpreservation.gov/meetings/documents/ndiipp10/NDIPP-DuraCloud-Panel-Final.pptx

The Digital Preservation Network (DPN) **PRESERVATION**

http://www.dpn.org

Deborah Ludwig

Brief Description of Project

The Digital Preservation Network (DPN) is building a proof-of-concept system to demonstrate digital preservation of the scholarly assets of higher education. There is no existing system yet. Members are contributing funding toward development. There is a data partnerships sub-group working on an environmental scan of research data preservation. DPN envisions university repositories as contributing nodes to federated replicating nodes. The replicating nodes are dark archives, used to restore access to content to contributing nodes in the event of data loss.

Reviewer's Analysis

Several GWLA and GPN institutions are among the 50+ contributing members to DPN. Based on a recent presentation at the Coalition for Networked Information (CNI), there is an assumption that institutions will affiliate with some content node and that different nodes may offer different services. The California Digital Library and the Texas Digital Library are members as well who represent large-scale digital collections.

Considerations and Recommendations

This merits our attention as a future preservation strategy to undergird efforts to make data available in access repositories.

Key Contacts:
Sponsors:
Funding:
Inception:
Geographic Location:

Ex Libris Rosetta **PRESERVATION**

http://www.exlibrisgroup.com/category/RosettaOverview

Deborah Ludwig

Brief Description of Project

Rosetta is a commercial software solution for digital preservation. It was designed in response to the needs of the National Library of New Zealand. Other customers include the ETH-Bibliothek in Sweden, which is using it as a platform for preserving research data. SUNY Binghamton University and Brigham Young University are also U.S. higher education customers. There is good information available on the website.

Reviewer's Analysis

Ex Libris Rosetta is the only major commercial software solution for digital preservation. In addition to the National Library of New Zealand, the Church of the Latter Day Saints is another large customer with sizeable large digitization efforts.

Considerations and Recommendations

Ex Libris has extensive experience building software solutions for libraries. Products include Voyager, Aleph, and Alma integrated library systems. While Rosetta could preserve research data as part of a solution for digital library collections that could include research data, its purpose is much broader than research data.

Key Contacts: http://www.exlibrisgroup.com/category/ContactUs
Sponsors: Ex Libris
Funding: For-Profit-Company
Inception: --
Geographic Location: --

LOCKSS **PRESERVATION**

www.lockss.org
Deborah Ludwig

Brief Description of Project

LOCKSS stands for "Lots of Copies Keep Stuff Safe." Private LOCKSS networks are collaborative communities that work together to preserve their institutional assets. There are a number of LOCKSS communities, including the Data Preservation Alliance for the Social Sciences (Data-PASS), which is focused on research data. Data-PASS[16] is a consortium with the initial goal "to create a sustainable partnership model for preserving 'at risk"\' social science data." LOCKSS was funded by the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP)[17] and the effort is lead by Harvard's Institute for Quantitative Social Science, which is noted in the literature review on federated and collaborative approaches to data management.

LOCKSS lists eleven collaborative communities on its website, including CLOCKSS which is focused on preservation of scholarly journal literature as an effort of publishers working with librarians.

Key Contacts: http://www.lockss.org/contact-us/
Sponsors: LOCKSS is a not-for-profit organization.
Funding: LOCKSS has received funding from Andrew W. Mellon Foundation, the National Science Foundation, and the Library of Congress.
Inception: 1999 at Stanford University with participation from Indiana, Emory, and the New York Public Library. Released into production, 2004.
Geographic Location: Global adoption

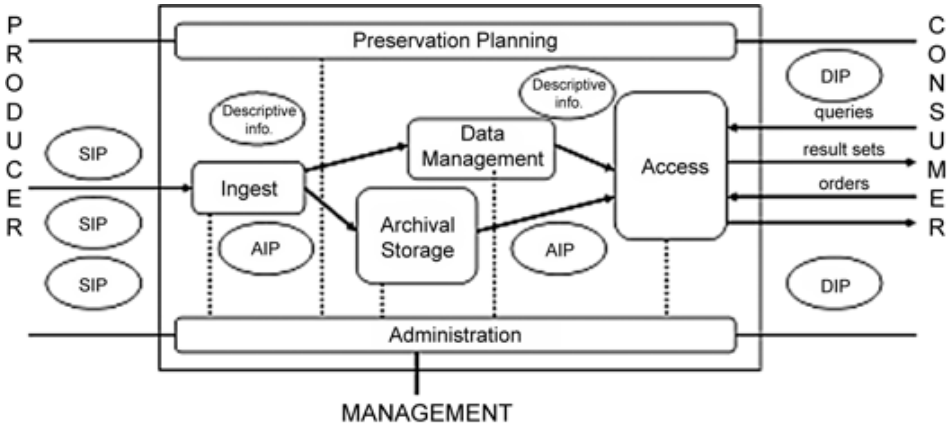
[16] http://www.data-pass.org
[17] http://www.digitalpreservation.gov/index.php

OAIS **PRESERVATION**

(Open Archival Information Systems Reference Model —ISO 14721:2003)
http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284
Deborah Ludwig

Brief Description of the Project

OAIS is not a computer system. OAIS is a reference model (ISO 14721:2012) that defines an open archival information system (OAIS), an archive, with people and systems responsible for to preserve information and making it available for a designated community. OAIS is not a particular system or repository, but a standard, which provides a model for understanding what must be in place for long-term preservation of digital information. The classic illustration of an OAIS below identifies the components for preserving information bracketed between the roles of information producer and consumer.



Note: 3 OAIS Functional Entities Standard Diagram

Information packages within respect to the archive include the SIP (submission information package) the AIP (archival information packet) and the DIP (dissemination information packet). Functions of the archive include ingest, storage, administration, planning, and provision of access. For access, descriptive information is required.

Reviewer's Analysis

The Digital Preservation Coalition defines digital preservation as the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary."[18] OAIS creates a framework within which these managed activities are carried out. OAIS specifies a set of mandatory responsibilities that the people and systems comprising the archive must undertake to account for preservation of the data and its long-term access for the intended consumer. A repository utilizing an OAIS framework must:

[18] http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

- Negotiate for and accept appropriate information from information producers.
- Obtain sufficient control of the information provided to the level needed to ensure long-term preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the designated community and, therefore, should be able to understand the information provided.
- Ensure that the information to be preserved is independently understandable to the designated community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.
- Follow documented policies and procedures, which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated.

The most important implication of OAIS for a shared data management project is likely the understanding that long-term access to research data will involve more than establishing technologies. A system will require policies, agreements with data producers, designated stewards for data, specialists in developing the specifications and workflows for the "packages" of information that must accompany a research data set as it is ingested into a repository, stored, and disseminated to the intended consumer

Considerations and Recommendations

OAIS provides a high-level model for what needs to happen for long-term preservation of research data. It provides a reference to help us understand what happens in a preservation system. The advice of Chris Rusbridge (with the Digital Curation Centre) may provide a reasoned approach to thinking about how to implement a reasonable OAIS framework:

> *Investment in digital preservation is important for cultural, scientific, government and commercial bodies. Investments are justified by balancing cost against risk; they are about taking bets on the future. The priorities in those bets should be: first, to make sure that important digital objects are retained with integrity, second to ensure that there is adequate metadata to know what these objects are, and how they must be accessed, and only third to undertake digital preservation interventions.*

Key Contacts: ISO, the International Standards Organization
Sponsors: Consultative Committee for Space Data Systems (CCSDS)
Funding: NA
Inception:
Geographic Location: NA

Additional Resources

- Allinson, Julie. *OAIS as a reference model for repositories.* November 21, 2006 accessed online at www.ukoln.ac.uk/repositories/.../**oais**.../Drs-**OAIS**-evaluation-0.5.pdf

- Digital Preservation Coalition website http://www.dpconline.org
- Digital Preservation Coalition handbook with definitions. http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts
- Rusbridge, Chris. *Excuse Me … Some Digital Preservation Fallacies?* Ariadne, 46, February 8, 2006. http://www.ariadne.ac.uk/issue46/rusbridge

Trusted Repositories Audit & Certification (TRAC) **PRESERVATION**

http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories

Deborah Ludwig, University of Kansas Libraries

Brief Description of the Project

Claims of trustworthiness are easy to make but are thus far difficult to justify or objectively prove. As Clifford Lynch has stated, 'Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate' (2003) Establishing more clear criteria detailing what a trustworthy repository is and is not has become vital. [19]

TRAC is not a computer system for digital preservation. TRAC is a set of criteria applicable to a range of digital repositories and archives, from academic institutional preservation repositories to large data archives and from national libraries to third-party digital archiving services. TRAC is now under the management of the US Center for Research Libraries In general terms, TRAC:

- Provides tools for the audit, assessment, and potential certification of digital repositories
- Establishes documentation requirements required for audit
- Delineates a process for certification
- Establishes appropriate methodologies for determining the soundness and sustainability of digital repositories

TRAC provides tools for the audit, assessment, and potential certification of digital repositories; establishes the documentation requirements required for audit; delineates a process for certification; and establishes appropriate methodologies for determining the soundness and sustainability of digital repositories. Only a few digital repositories have pursued the certification process to date, including: *Portico*, *HathiTrust*, *Chronopolis*, the US National Archives and Records Administration (*NARA*), and Michigan's *Interuniversity Consortium for Political and Social Research* (*ICPSR*). Many more have utilized the TRAC checklist as a planning tool for building trusted repositories of digital information.

Reviewer's Analysis

How does the project address, or potentially address, key facets of lifecycle management? The TRAC checklist is used by institutions for planning long-term preservation of cultural heritage and other digital resources and has been used in combination with the OAIS reference model as a digital preservation planning tool. TRAC audit specifications consider 3 major areas:

- **Organizational infrastructure, including governance, accountability, and staffing**
- **Digital object management**
- **Technologies, technological infrastructure, & security**

[19] TRAC Certification Checklist. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

An approach to data management that is federated across a group of partners represents an opportunity to share the work and develop as partners to implement standards and best practices. At the same time, it represents a level of complexity that can only be helped by using time-tested approaches and tools for planning and implementation that have worked for a variety of other organizations.

Considerations and Recommendations

Used in conjunction with the OAIS reference model, the *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) can provide a useful guide to understanding identifying the organizational and the technical components required for developing an approach to long-term access to research data.*

Sponsors: Center for Research Libraries (http://www.crl.edu/).
Inception: The TRAC checklist was published in 2007 by the National Archives and Records Administration, Research Libraries Group, and Center for Research Libraries.

Additional Resources

- Digital Curation Centre Web Site on Trusted Repositories: http://www.dcc.ac.uk/resources/repository-audit-and-assessment/trustworthy-repositories
- Center for Research Libraries Web Site on Archiving and Preservation: http://www.crl.edu/archiving-preservation

3. Archiving & Repository Services

Data archiving definitions vary widely. In the simplest terms and in accordance with the National Science Foundation Data Archiving Policy [20], we consider data archiving to be the sharing of data through an appropriate archive or library to encourage data sharing with other researchers. Data archiving is something more than simply storing data and backing it up and may include some common elements with digital preservation.

For the purposes of this report, we have focused more narrowly on systems for making data available, most often through some type of repository software. Initiatives to curate data will generally rely upon some repository system as a basis for making data available to appropriate consumers.

Examples of Data Archiving Platforms:

1. The DataFlow Project (with DataBank and DataShare)
2. DataSpace
3. DuraSpace
4. Hydra
5. OneShare (with DataUP)

[20] http://www.nsf.gov/sbe/ses/common/archive.jsp

DataFlow Project [DataFlow, DataStage, DataBank] **ARCHIVING**

Greg Monaco, Great Plains Network (GPN)

http://www.dataflow.ox.ac.uk/

Brief Description of the Project

DataFlow is a two-stage data management infrastructure: locally, the user installs ***DataStage*** and the user's institution installs ***DataBank***. The user then saves a dataset, locally, to her/his hard drive. That dataset will be copied to the cloud (similar to dropbox). The software is open source.

From the website:

> Rather than storing datasets on external hard drives in the lab, ***DataFlow*** lets researchers save their work in institutional memory banks. The system will be lightweight (nothing for researchers to install; just save data to a mapped drive on their computer), with best-practice standards to make sure data is well looked after.
>
> ***DataStage*** is a secure personalized 'local' file management environment for use at the research group level, appearing as a mapped drive on the end-user's computer. It can be deployed on a local server, or on an institutional or commercial cloud. Once the software has been installed on the server, there is no additional software for the end-user to install.
>
> Users save files to ***DataStage*** just as they would on ordinary C: drive -- but with added extras:
>
> - Private, shared and collaborative directories, with password-controlled access
> - Web access – work with stored files over the web, anywhere in the world
> - Users can add richer metadata via the web interface, using free-text "notes" fields
> - All files can be automatically backed up via your usual backup service
> - Users can invite colleagues to access group files, via password control
> - Repository submission interface makes it easy for researchers to define data packages, enter minimal metadata, and deposit them in a repository of choice. The minimal metadata is in RDF format; additional (non-RDF) metadata can be added via free-text fields at the submission stage
> - Packaging done using BagIt file packaging specification, soon to be SWORD-2 compliant
> - Flexibility to dynamically invoke additional cloud storage as required

DataBank is a scalable data repository designed for institutional deployment that is designed to

- provide a definitive, sustainable, reference-able location for (potentially large) research datasets
- allow researchers to store, reference, manage and discover datasets

DataBank instances will expose both human- and machine-readable metadata describing their datasets, and will assign Digital Object Identifiers (DOIs) to hosted datasets, obtained automatically using the DataCite API, to aid discovery and citation…By default, all objects are assigned a DOI and a cc-zero Open Data Waiver, and all RDF-format metadata is visible to the outside world, but other licensing/secrecy arrangements can be accommodated. Users can define an optional embargo period (making metadata visible but withholding the underlying data), add richer metadata to make their data easier to find (when searching within DataBank, or via web crawlers like Google), and users can revise datasets that

have already been submitted (new DOI issued for each version, all versions kept in perpetuity). DataBank can also be run as a "dark" archive with metadata and data invisible to the outside world.

- Institutions can have their own ***DataBank*** instances hosted within an external cloud (e.g. Eduserv), or can choose to deploy ***DataBank*** on local hardware, at institutional, departmental or individual research group level.
- ***DataBank*** can be used together with ***DataStage***, or separately.
- DataBank is a virtualized, cloud-deployable version of the databank created by Oxford's Bodleian Libraries. We are actively pursuing a variety of sustainability options for ***DataBank***, but at minimum, the software will be maintained and developed for use by the Bodleian Libraries, with their code made available open-source under an MIT license.

Reviewer's Analysis

This project presents a different approach to management of research data, and encompasses all stages from creation to archiving and curation to sharing and reuse. This projects attempts to tackle the consortial/collaborative approach and promises to be quite flexible.

Considerations and Recommendations

Merits further review and possible download and testing.

Key Contacts:
Sponsors: **University of Oxford**
Funding: JISC (see http://www.dataflow.ox.ac.uk/index.php/about/51-project-funding)
Inception: Version 0.1 of the software packages were released on March 2, 2012. The project appears to have started in 2011.
Geographic Location: UK

DataSpace **ARCHIVING**

http://dataspace.princeton.edu/jspui/

Michael Bolton, Texas A&M University, Sterling C. Evans Library

Brief Description of the Project

DataSpace is a digital repository meant for archiving and publicly disseminating digital data that are the result of research, academic or administrative work. DataSpace has a cost model for Princeton users; however, it is billed as a one-time charge for long-term storage of digital data. The repository also provides persistent URLs, which aid in dissemination of works. DataSpace accepts a wide range of datasets from research data to student projects and reports, conference and workshop proceedings, technical reports and digital collections of images and other digital assets. It is an Open Access repository and users can subscribe to news feeds that will keep them informed of new submissions to Communities in DataSpace. The web site says future plans do include making metadata available to services such as OAIster.org to help with discovery services. Princeton has developed a LibGuide for this service (http://libguides.princeton.edu/content.php?pid=211802&sid=1763060) and I have seen presentations on the service (Web Seminar for EDUCAUSE).

Reviewer's Analysis

DataSpace is built on DSpace, which would be convenient for DSpace-base institutions. I believe this is a separate instance of DSpace, intended solely for dataset management. It has the same concept of communities and collections, just as with an Institutional Repository. For a DSpace shop, the paradigm would be easy to follow and has appeal since it would leverage existing knowledge and expertise. The **About** page includes information on the licenses as well as a cost model for storing data.

DataSpace is a data repository and does not by default include a preservation model. That would have to be added. However, if you add preservation to your DSpace IR, you now have preservation for the datasets as well. In this case, the DPN project would work well for us.

How does the project address, or potentially address, key facets of lifecycle management? This model leverages existing expertise in DSpace and an institution's repository. It can be easily linked to research or back to items in the IR. It does not have a preservation model as yet and that would have to be added. It does support easy data sharing via the persistent URLs; however, it must be coupled with a stronger discovery service for broader access.

Does this project have implications for a collaborative or consortial approach to defining a regional (GWLA/GPN) strategy for lifecycle management of research data? If so, what does this project offer in terms of approaches that might be incorporated or services that might be acquired? Given the number of DSpace instances in use this would seem to be a good place to start in testing technologies. DSpace is very robust when it comes to sharing data (via the OAI interfaces). With SWORD deposits possible, deploying and using DataSpace could be fairly straightforward and easy.

Considerations and Recommendations

What I like about this project is it represents a quick way to get into dataset management for existing DSpace shops. I think it would need a metadata management tool since DSpace is not particularly strong in that area. If coupled with DataUP, and DataUP was able to export to DSpace, this would be a quick win.

Key Contacts: Serge Goldstein, Associate CIO and Director of Academic Services at Princeton
Sponsors: Princeton initiative
Funding:
Inception: 2010?
Geographic Location: Princeton

DuraSpace (DSpace & Fedora) **ARCHIVING**

http://www.duraspace.org/

Philip Konomos, Arizona State University.

Brief Description of the Project

DuraSpace is a not-for-profit umbrella organization that manages, develops, coordinates and supports three important open source digital repository projects:

- DSpace (http://www.dspace.org/)
- Fedora (http://www.fedora-commons.org/)
- DuraCloud (http://www.duracloud.org/)

DSpace is a free, open source digital repository application. The software allows institutions to manage, preserve, and provide long-term access to many common types of digital files. DSpace is probably one of the most commonly used digital repository / institutional repository applications at research universities around the world because it is configured to work "out-of-the-box" with minimal need for custom programming. A large number of universities use DSpace including MIT, University of Texas, and the University of Illinois (click here for a comprehensive list).

Fedora is also a free, open source, digital repository application, although it is much more complex than DSpace to configure. Fedora is an acronym for Flexible, Extensible, Digital, Object Repository Architecture, which indicates that this is really a sophisticated framework, rather than an out-of-the-box repository package. Fewer universities use Fedora, but those include Cornell, Indiana University, University of Virginia, and Rutgers. Fedora offers much more robust functionality and complete customization but at the cost of significant programming staff expertise.

DuraCloud is an outgrowth of a 2009 initiative between the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and DuraSpace. It is designed to offer a hosted solution (Software as a Service, or SAAS) for institutions that want to ensure perpetual access to their digital collections using cloud computing.

DSpace and Fedora were originally separate initiatives, but in 2009 the two organizations saw the wisdom of joining forces to leverage development ideas and technical expertise under the rubric of the DuraSpace Foundation. DuraCloud was one of the first new initiatives begun by the newly formed DuraSpace Foundation.

Reviewer's Analysis

These are important repository options for universities to consider since selection of the appropriate digital repository platform is a critical decision. Fedora and DSpace have a long history of use by a wide variety of institutions and repository managers are well aware of the strengths and weaknesses of each

platform. DuraCloud, as a more recent service, has less of a track record for interested users to use for evaluation.

Considerations and Recommendations

Since both DSpace and Fedora have been widely adopted by American universities, it might be worthwhile to invite GWLA members who use one or the other option to present a demonstration and evaluation.

Key Contacts: The project email is info@duraspace.org . Michele Kimpton, Chief Executive Officer; DuraSpace, 28 Church Street, Unit #2, Winchester, MA 01890

Sponsors: Fedora was originally developed (1997) by Professor Sandy Payette and graduate student Carl Lagoze at Cornell University. By 2007, as it's community of users grew, faculty at several of the universities that used Fedora formed the Fedora Commons not-for-profit organization. DSpace was originally developed (2002) by MIT Libraries and HP Labs, and later (2007), as the number of users grew, supported by the non-profit DSpace Foundation. In both cases, a large community of universities and other institutions committed significant time and programming expertise to support the growth of each software project. The non-profit organizations were a way to organize and manage development and support user institutions. Currently, DuraSpace is run by a staff of nine and governed by a Board of Directors of seven.

Funding: Fedora was funded by grants from the Andrew W. Mellon Foundation and the Gordon and Betty Moore Foundation. Original funding for **DSpace** came from the Andrew W. Mellon Foundation and HP (Hewlett Packard). **DuraCloud** was developed with support from the Gordon and Betty Moore Foundation, the Andrew W. Mellon Foundation and the Library of Congress NDIIPP program.

Inception: Fedora was developed by computer science faculty at Cornell University in 1997. DSpace was started in 2002, and remained a separate project until 2009. DuraCloud was begun by DuraSpace as a hosting option in 2009-1010.

Geographic Location: N/A

Globus Online **ARCHIVING**

www.globusonline.org

Greg Monaco, Great Plains Network (GPN)

Brief Description of the Project

The ***Globus Project***'s original focus was on middleware tools to enable grid computing. ***Globus Online*** is the Globus Project's response to the issue of researchers needing to manage huge datasets. Globus Online provides the researcher with a secure method to organize, access, move and share with collaborators data that is located at multiple, distributed sites. Access management interfaces with Shibboleth and InCommon (see separate reviews). Rather than an institutional solution, Globus Online provides the individual researcher with a way to organize all the datasets to which s/he needs access from the Globus Online portal.

Globus Connect and ***Globus Connect Multi-User*** are software versions that allow one to make local storage (desktop, laptop) and shared storage (servers, data repositories) resources available to a potential community of users via Globus Online.

Note: The Globus Project team has historically been interested in collaborating with others making novel use of their tools.

From the website:

> *Globus Online is a fast, reliable file transfer service that makes it easy for any user to move any data anywhere. Recommended by HPC centers and user communities of all kinds, Globus Online automates the time-consuming and error-prone activity of managing file transfers, so users can stay focused on what's most important: their research.*

Reviewer's Analysis

Globus Online provides individual researchers with a way to manage access to their personal data sets, from data creation, processing and analysis to data sharing and reuse. The ability to create groups and share data means that users may create teams who initially add to and reuse data.

With Globus Connect Multi-User it is possible to make test bed resources available to potential users via Globus Online. This offers the advantages of providing

- a common interface for access to resources without having to reinvent it,
- an interface that can be branded for this project,
- integration with other key technologies (Shibboleth, InCommon)

Considerations and Recommendations

I recommend contacting the Globus Project Team to gauge their interest in participating in the further development of this project.

Key Contacts: Steve Tuecke, Ian Foster
Sponsoring entities: University of Chicago, Argonne

Funding Source, if Known: DOE (Energy), NSF, NIH
Inception Date, if Known: Globus Project (1995), Globus Alliance (2003)
Geographic Location, if Applicable: Headquartered in Chicago, IL

Hydra **ARCHVING**

http://projecthydra.org/

Jason Stirnaman, University of Kansas Medical Center

Brief Description of the Project

Hydra adds interfaces for discovery, workflow, and access control to Fedora-based data through a Ruby on Rails framework, Solr, and Blacklight. Enables powerful use of Fedora's capabilities through a familiar lightweight, Ruby-based toolkit. "Hydra's flexibility means we use the same infrastructure, but can generate individual solutions."

Hydra is not just a repository software solution. Rather, it is three complementary components:

- A vibrant, highly active community supporting the work of the project which shares an underlying philosophy behind all that it does
- Design (and other) principles involved in constructing a successful Hydra "head" for use with compatible digital objects, and,
- The software components, the Ruby gems, that the Hydra community has constructed which are combined together to provide a local installation

Reviewer's Analysis

How does the project address, or potentially address, key facets of lifecycle management? [Data Creation, Data Processing, Data Analysis, Data Preservation, Data Access for Others (data sharing), and Data Reuse] Hydra, although extensible, is primarily concerned with the facets of preservation, access, and reuse of digital content objects. Hydra expects that objects in the repository:

- follow a model pattern asserted by the objects themselves.
- have an associated rights schema that Hydra may enforce.
- have accompanying metadata.
- may have digital content for delivery.

Ruby on Rails provides models, controllers, and interfaces to create, read, update, and delete objects and their associations. Fedora stores objects in one of two standard models. Apache Solr indexes objects and their metadata, and Blacklight provides a discovery interface to the index of objects and metadata.

Examples of research data implementations:

- History DMP Project, University of Hull
  - Interaction between Hydra and DataCite has been explored to enable the additional benefits of this widely used citable standard identifier
  - "The work thus far has enabled us to exploit our institutional repository's flexibility to use it for datasets. We shall also be using it as a data catalogue. We know we will need to exploit this flexibility further as different research data needs emerge." [1]

- Datasets in Penn State's Scholarsphere repository. https://scholarsphere.psu.edu/catalog?f%5Bgeneric_file__resource_type_facet%5D%5B%5D=Dataset

Does this project have implications for a collaborative or consortial approach to defining a regional (GWLA/GPN) strategy for lifecycle management of research data? If so, what does this project offer in terms of approaches that might be incorporated or services that might be acquired? Yes, as a repository solution, Hydra has implications for a collaborative or consortial strategy for data management. The Hydra/Fedora approach to Rights Enforcement and Access Control may be of particular interest. A major advantage of Hydra is the commitment to sharing solutions that can be easily adopted by other Hydra sites. Hydra is a fully open framework built on familiar, lightweight tools and supported by a growing community. Hydra repositories are scalable, flexible, and modular.

Recommendations

Does this project merit further review as part of a site visit, phone interview, or further analysis? (and, if so, what do you recommend?) Yes, this project merits further review if a repository solution or support is considered. I recommend a phone interview with core partners or further analysis that addresses specific GWLA/GPN use cases.

Additional Resource

- Using Hydra's flexibility to manage datasets. 2012. https://hydra.hull.ac.uk/catalog/hull:6010

Key Contacts: Matt Zumwalt, Mediashelf; Bess Sadler, Stanford University; various communication channels listed on Duraspace wiki.

Sponsors: Stanford University, University of Virginia, University of Hull, DuraSpace, MediaShelf LLC, University of Notre Dame, Northwestern University, Columbia University, Penn State University, Indiana University, London School of Economics and Political Science, The Royal Library of Denmark. Governance model

Funding: "Hydra is *not* (and has never been) grant funded. It *is* distributed, robust and open. Any single developer could walk away. Any single institution could walk away. People ask what's your sustainability plan? We say we've already passed the first hurdle—more than four years of self-funded productivity, and a growing code, contributor and user base, not dependent on a transition plan."

Inception: 2008

Geographic Location: Distributed.

Islandora Open-Source Digital Asset Management Framework ARCHIVING

http://islandora.ca/

Susan Matveyeva, Wichita State University

Brief Description of the Project

Islandora is an open source framework developed by the University of Prince Edward Island's Robertson Library. Islandora combines the Drupal and Fedora open software applications to create a robust digital asset management system that can be fitted to meet the short and long term collaborative requirements of digital data stewardship. Additional open source applications are added to this core stack to create what we call Solution Packs. Islandora operates under a GNU license. Islandora may be used to create large, searchable collections of digital assets of any type and is domain agnostic in terms of the type of content it can steward. (from the Islandora website)

Reviewer's Analysis

Islandora supports data lifecycle. Islandora implemented by the Colorado Research Alliance Repository[21]. he infrastructure is described on the website.[22]

Features of the Colorado Research Alliance Digital Repository:

Fully Hosted Service: ADR Services centrally manages all the hardware, software, updates, and backups for the repository.

Customization with Drupal: Customize the look and feel of the repository front end with the Drupal web content management system.

Access and Authentication: Manage restricted, embargoed, or dark archive content with user accounts and security metadata.

Content Loading: Add content to the repository with easy web forms for adding metadata and attaching files.

Search: Solr indexes metadata and the full text of PDFs and other documents, which can be searched with simple and advanced search functions.

Streaming and Viewing: Embedded viewers and players display common formats of document, image, audio, and video content without users having to download anything from the repository.

Cloud Storage: Back up objects in the cloud and perform fixity checks with DuraCloud.

Content Sharing: Share OAI metadata for harvesting and aggregation into other sites. Every object is assigned a Handle link for persistent identification.

[21] http://adrresources.coalliance.org/

[22] http://adrresources.coalliance.org/?page_id=13

Reusability: FedoraCommons open-source repository software supports management, reuse, migration, and transformation activities on digital objects for access and preservation.

Hardware: ADR hardware is stored at a collocation facility to meet power, cooling, and security needs. Data is backed up to disk and tapes. For more information about the ADR hardware, please contact adr@coalliance.org.

Software: The ADR Basic repository platform runs on Islandora, an open-source Drupal-based repository system, and uses Fedora Commons as its core repository software.

Cloud Storage: In 2010, ADR Services participated in a pilot project for DuraCloud, cloud repository management software from DuraSpace, the parent organization of Fedora Commons. In fall 2011, the ADR joined the DuraCloud service and will begin using its cloud services for remote backup and bit integrity in 2012.

Descriptive Metadata: All the objects in the repository have a basic set of descriptive information in MODS (Metadata Object Description Schema), which is a set of metadata designed to describe resources commonly found in libraries. The ADR and its members follow best practices for MODS as set out by the Digital Library Federation's Implementation Guidelines for Shareable MODS and MODS Levels of Adoption.

Many ADR members do not have pre-existing MODS metadata for their records. Members can create a MODS record for an object by filling out a web form in the repository software, which then builds MODS. Alternatively, members can send metadata to ADR Services in MARC, Dublin Core, or even a spreadsheet and we will transform the metadata into MODS that can be used by the repository. The MODS metadata is currently being used in the search index and object display, and is being transformed to Dublin Core for OAI harvesting. ADR Basic is highly flexible with what metadata it can accept and display. Avenues for future development include objects described in VRA Core 4.0 or Darwin Core.

Security Metadata: The ADR uses XACML security metadata to control access to collections, objects, and data streams in the repository.

Considerations and Recommendations

I believe, the project and its implementation by a consortium worth a closer look

Key Contacts: Mark Leggott, University Librarian and developer of Islandora islandora@upei.ca and the founder of a company DiscoveryGarden Inc. that provides services around Islandora software http://www.discoverygarden.com/ Mark Leggott website: http://loomware.typepad.com/about.html
Sponsoring entities: University of Prince Edward Island
Funding Source, if Known: multiple: 2,3 mini grants (2010); over 750,000 research projects, donations, library operational budget (see: Leggott presentation: http://loomware.typepad.com/docs/S2I2_Islandora.pdf)
Inception: 2008
Geographic Location: Charlottetown, Prince Edward Island

ONEShare Repository [& DataUp] **ARCHIVING**

http://www.cdlib.org/cdlinfo/2012/10/02/california-digital-library-and-partners-launch-dataup/

Michael Bolton, Texas A&M University, Sterling C. Evans Library

Brief Description of the Project

ONEShare is a repository build specifically for DataUP. ONEShare is built on the Merritt Repository software developed by Stephen Abrams at CDL.[23] Merritt is used extensively by CDL and a special instance was developed to support the DataUP project. DataONE in turn harvests metadata from ONEShare via API.

Reviewer's Analysis

DataUP and ONEShare should be viewed as a single entity for our purposes since developing a generic export facility to support other repositories could take a while. In this case, CDL relied heavily on technologies it knew well and used as part of their infrastructure, which is a smart and safe move. It does mean we would have to adapt the tools for the general use case we are developing. Merritt is an open source project with the code available on Bitbucket.

Considerations and Recommendations

If we were to test DataUP, ONEShare would need to come along for the ride.

Key Contacts: California Digital Library (CDL)
Sponsors: CDL UC3
Funding:
Inception: Launched with DataUP, October 2, 2012
Geographic Location: Open source software

[23] http://www.cdlib.org/services/uc3/merritt/index.html

4. Storage Systems

Storage for digital data is foundational to any program of curating, preserving or archiving data. Storage systems are becoming more sophisticated and able to do more than act as simple file directories with backup. The line between storage and archives is blurring. iRODS stands out as something more than storage. With a rules engine and metadata catalog, it shares characteristics with repositories.

Three storage platforms are reviewed in this section.

1. iRods
2. Quantum StorNext
3. SGI DMF & LiveArc

iRODS (Integrated Rule-Oriented Data System) **STORAGE+++**

http://www.irods.org/

Michael Bolton, Texas A&M University, Sterling C. Evans Library

Brief Description of the Project

Based on the web site, iRODS is a data grid software system developed by the Data Intensive Cyber Environments (DICE) research group, which developed the Storage Resource Broker (SRB). iRODS is scalable, released under an open source license, freely available and has a very active user community. The Service Resource Broker (SRB) was the forerunner of iRODS and was described as a one-size fits all system in that policies used to manage the data were hard-coded into the system. In the development of iRODS, a more adaptive approach was used. Now policies could be applied to data based upon rules.

The rule engine is not only very powerful; it is equally flexible as well. This allows for automated data management such as replicating data across zones. iRODS is accessed via a set of uniform APIs and GUIs. This allows for iRODS storage to be accessed through a number of mechanisms such as a web frontend or it can be integrated into more complex systems as the storage subsystem. For instance, iRODS can be the backend storage for repository software including DSpace and Fedora Commons. iRODS employs a data catalog to not only track where data is stored, but to collect a rich set of metadata for the object. Schemas of various sorts can be incorporated into the catalog to provide a more controlled vocabulary and the rules engine can be used to ensure metadata is added when objects are stored in the file system.

iRODS has gained popularity in areas such as numerically intensive computing and big data, primarily because of its ease of access. Like SRB, iRODS abstracts certain aspects of the storage subsystems by providing an API that can be used to access the system. Access to the data is via the MCAT, or iRODS metadata catalog, which tracks the movement of data through the system. Users do not need, nor care to know, where or how the data is stored. The iRODS server can find the dataset, read it from whatever media and transfer it to the client using standard protocols. A growing collection of micro-services help the iRODS server carry out its automation tasks, applying policies and rules as defined by the user and system. iRODS servers can easily be linked together to form large, geographically dispersed clusters. And the rule engine can be used to control and manage replication of data across the various nodes in a cluster.

The iRODS user group meets annually and develops a list of requested features and services, which is used by the DICE team to enhance and further develop the product. The user group meetings also showcase customers who use the service, along with their use cases and applications.

Reviewer's Analysis

I have always thought of iRODS as an infrastructure service, that is, the file system used by higher-level applications. What it offers is very attractive for us in that it abstracts the real storage layer. iRODS provides a robust set of APIs and interfaces that allow administrators to use whatever physical storage is necessary and appropriate for the service without entangling the application in the mechanics of using that storage. On the customer side, a number of different clients can interact with the iRODS servers to gain access to data. The iRODS server uses an authentication system to protect the data and manage

access privileges. iRODS also has a very well developed JAVA-based SDK (Jargon) to allow developers to integrate iRODS into their applications. So we could actually develop applications that have UIs customized for our service while exploiting the features of iRODS. DPN (the Digital Preservation Network) will be using iRODS as one of its preservation services and the Texas Advanced Computing Center (TACC) is a large iRODS deployment.

Considerations and Recommendations

While standalone iRODS implementations would be feasible for this project I believe our effort would be better served by integrating iRODS with easier to use frontends. This would include web interfaces, possibly iDROP or WebDAV. Integrating iRODS into a broader service offering, such as that delivered in the iPlant Data Store would be better still.

I do like the rules-based management of data provided by iRODS. As we develop our service, we could write rules to ensure data is replicated across geographically dispersed nodes. I also like the metadata model supported by the iRODS catalog. I am not in favor of supporting multiple metadata management tools so we would have to develop a process to populate the MCAT with data from our primary metadata management too. We then augment that metadata with information collected and managed by iRODS, such as checksums, file locations, number of copy, age of replicas, etc.

Additional Notes

Information related to the various grants awarded for SRB and iRODS

- NSF ITR 0427196, Constraint-Based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives (2004–2007)
- NARA supplement to NSF SCI 0438741, Cyberinfrastructure; From Vision to Reality—Developing Scalable Data Management Infrastructure in a Data Grid-Enabled Digital
- NARA supplement to NSF SCI 0438741, Cyberinfrastructure; From Vision to Reality—Research Prototype Persistent Archive Extension (2006–2007)
- NSF SDCI 0721400, SDCI Data Improvement: Data Grids for Community Driven Applications (2007–2010)
- NSF/NARA OCI-0848296, NARA Transcontinental Persistent Archive Prototype (2008–2012)

Some key sites or organizations using iRODS as part of their data management plan or system.

- iPlant Data Store - https://pods.iplantcollaborative.org/wiki/display/start/Storing+Your+Data+with+iPlant+and+Accessing+that+Data
- University of Michigan Office of Research – Cyberinfrastructure - Data Management - http://orci.research.umich.edu/resources-services/data-management/

Key Contacts: PI – Reagan Moore – Director of the Data Intensive Cyber Environments Center (DICE)
Sponsors: National Science Foundation and the National Archives and Records Administration[24]

[24] Sponsors of iRODS - https://www.irods.org/index.php/Sponsors

Funding: NSF and NARA

- First NSF grant awarded Fall 2004 for SRB
- Latest NSF Grant awarded September 28, 2011 for prototype national data management infrastructure[25]

Inception: iRODS began in 2006 with release 0.5 released December 20, 2006. It is a continuation of the SRB project, which started in 2004. The first iRODS users group meeting was held in 2009.

Geographic Location: Widely used in a number of systems and applications.

[25] Press release - http://www.renci.org/news/releases/nsf-datanet

Quantum StorNext **STORAGE**

http://www.quantum.com/products/software/stornext/index.aspx
Toby Axelsson, University of Kansas, Information Technology

Brief Description of the Project

Quantum StorNext is a multi-tier (tape & disk) archival system providing access to a clustered file system via NFS, CIFS or direct mount using provided software. StorNext is well suited for storing large amounts of data (Petabytes).

Reviewer's Analysis

How does the project address, or potentially address, key facets of lifecycle management? Quantum StorNext provides means of storing and archiving large amounts (Petabytes) of data.

- Access can be provided through common file access protocols (NFS & CIFS) as well as direct network mount using provided software (Distributed LAN Clients). The Distributed LAN Client software supports both Windows and Linux.
- Data can easily be accessed after it has been archived, with delays if the data needs to be re-hydrated from tape (transparent to user).
- StorNext can be configured to handle moderate HPC workloads.
- The system is agnostic of what type of disk subsystem (SAN attached) and tape library is used.

Does this project have implications for a collaborative or consortial approach to defining a regional (GWLA/GPN) strategy for lifecycle management of research data? If so, what does this project offer in terms of approaches that might be incorporated or services that might be acquired? Not sure.

Considerations and Recommendations

Does this project merit further review as part of a site visit, phone interview, or further analysis? (and, if so, what do you recommend?) This is an interesting platform for long time storage of large amounts of data. I don't believe it does data lifecycle management to the degree that we are interested in. Coupled with a product for data lifecycle management, it may be a worthy candidate for long-term storage of research data, in particular if the total expected amount of data would exceed a Petabyte. If a solution like this fits well in the overarching vision, this is something that should be investigated further.

Key Contacts: Brian Morsch Brian.Morsch@quantum.com Sr. Account Executive
Sponsors: Quantum
Funding: Commercial
Inception: At least since the 1990's.
Geographic Location: N.A

SGI DMF & LiveArc **STORAGE**
http://www.sgi.com/products/storage/software/dmf.html
Toby Axelsson, University of Kansas Information Technology

Brief Description of the Project

SGI DMF is a multi-tier (tape & disk) archival system providing access to a clustered file system via NFS, CIFS or direct mount using their kernel modules. DMF is well suited for storing large amounts of data (Petabytes).

Reviewer's Analysis

How does the project address, or potentially address, key facets of lifecycle management? [Data Creation, Data Processing, Data Analysis, Data Preservation, Data Access for Others (data sharing), and Data Reuse]

- SGI DMF provides means of storing and archiving large amounts (Petabytes) of data.
- Access can be provided through common file access protocols (NFS & CIFS).
- Data can easily be accessed after it has been archived, with delays if the data is stored only on tape.
- DMF can be configured to handle moderate HPC workloads.
- The system is agnostic of what type of disk subsystem (SAN attached) and tape library is used.

LiveArc provides Digital Asset Management by classifying data and creating searchable metadata based on file content (API exists for adding file types not already supported).

Does this project have implications for a collaborative or consortial approach to defining a regional (GWLA/GPN) strategy for lifecycle management of research data? If so, what does this project offer in terms of approaches that might be incorporated or services that might be acquired? Not sure.

Considerations and Recommendations

Does this project merit further review as part of a site visit, phone interview, or further analysis? (and, if so, what do you recommend?) This is an interesting platform for long time storage large amounts of data as well as the classification of such (metadata collection). I don't believe it does data lifecycle management to the degree that we are interested in. Coupled with a product for data lifecycle management, it may be a worthy candidate for long-term storage of research data, in particular if the total expected amount of data would exceed a Petabyte. If a solution like this fits well in the overarching vision, this is something that should be investigated further.

Key Contacts: Timothy Belcher *tbelcher@sgi.com* (Federal Channel Sales Manager, West)
Sponsors: SGI
Funding: Commercial
Inception: Not sure, but oldest still running implementation is 21 yrs. old (going through multiple generations of hardware and software revisions).
Geographic Location: N/A

5. Enabling Technologies, Services, and Components for Data Management

The final group of technical reports on data management include some examples of various technologies or services that reviewers looked at for potential to contributing to data curation, either directly or conceptually. These have been sorted into 5 categories as follows.

1. Access Management
   - InCommon
   - Shibboleth
2. Discovery
   - Blacklight
   - Databib
   - Linked Data
   - Mercury
   - Researcher Networks: VIVO / Profile / Bibapp
   - VuFind
3. Identifiers for Digital Objects and Researchers
   - ARK
   - DOI
   - ORCID
4. Licensing for Data
   - Creative Commons
5. Planning for Data Management
   - Data Curation Profiles
   - DMP Tool

InCommon **ENABLERS: Access Mgt.**

http://www.InCommon.org

Greg Monaco, Great Plains Network (GPN)

Brief Description of the Project

InCommon provides certificates that can be used in a trusted identity and service infrastructure in conjunction with Shibboleth for secure single sign on by users to web services in a trusted/federated identity framework. It is scalable and can work in partnership with universities, schools, libraries and government agencies. There are currently 215 higher education members.

From the website:

> *The mission of InCommon is to create and support a common trust framework for U.S. education and research. This includes trustworthy shared management of access to on-line resources in support of education and research in the United States. To achieve its mission, InCommon will facilitate development of a community-based common trust fabric sufficient to enable participants to make appropriate decisions about the release of identity information and the control of access to protected online resources. InCommon is intended to enable production-level end-user access to a wide variety of protected resources.*

Reviewer's Analysis

The InCommon Certificate Service provides the necessary layer for secure access to shared (inter-institutional/federated) web services via Shibboleth. A successful project will need to resolve the issue of access to various web services at multiple institutions without creating an administrative nightmare. InCommon and Shibboleth provide a secure method for single sign on to web services across multiple administrative domains.

Considerations and Recommendations

InCommon and Shibboleth (see separate review) are central to a streamlined federated access management approach. They have been developed by and for the higher education and research communities and will be critical for the GWLA/GPN project. It is recommended that all participants to the project undertake implementation of Shibboleth and join InCommon as entry criteria to becoming part of the project.

Shibboleth **ENABLERS: Access Mgt.**

http://shibboleth.net

Greg Monaco, Great Plains Network (GPN)

Brief Description of the Project

Shibboleth was originally conceived as an institutional solution (with the library as a focus) to the problem of sharing information via the Internet among institutions of higher education without creating an overwhelming burden on users to have multiple credentials for multiple web services and on organizations to manage separate identity providers for each web service.

From the website:

The Shibboleth System is a standards based, open source software package for web single sign-on across or within organizational boundaries. It allows sites to make informed authorization decisions for individual access of protected online resources in a privacy-preserving manner.

The Shibboleth software implements widely used federated identity standards, principally OASIS' Security Assertion Markup Language (SAML), to provide a federated single sign-on and attribute exchange framework. Shibboleth also provides extended privacy functionality allowing the browser user and their home site to control the attributes released to each application. Using Shibboleth-enabled access simplifies management of identity and permissions for organizations supporting users and applications. Shibboleth is developed in an open and participatory environment, is freely available, and is released under the Apache Software License.

Reviewer's Analysis

Shibboleth allows users from trusted organizations to access web services provided by those organizations for data storage, data access and so forth using sign on credentials (user name and password) from their home organization. This greatly reduces the administrative overhead associated with managing a web service and it reduces the burden on the user to manage multiple IDs, one for each web service.

A successful project will need to resolve the issue of access to various web services at multiple institutions without creating an administrative nightmare. Shibboleth is a standards-based approach that offers a proven method for single sign on to web services across multiple administrative domains.

Recommendations

Shibboleth and InCommon (see separate review) are central to a streamlined federated access management approach. They have been developed by and for the higher education and research communities and will be critical for the GWLA/GPN project. It is recommended that all participants to the project undertake implementation of Shibboleth and join InCommon as entry criteria to becoming part of the project.

Key Contacts:

Sponsors: The Shibboleth Consortium is sponsored by Internet2, JISC and SWITCH (US, UK, Swiss Higher Ed consortiums).

Funding: Originally funded by NSF Middleware Initiative and Internet2
Inception: Began as an Internet2 Middleware project in 2000.
Geographic Location:

Blacklight **ENABLERS: DISCOVERY**

http://projectblacklight.org

Jason Stirnaman, University of Kansas Medical Center

Brief Description of the Project

Blacklight is an open source Ruby on Rails gem that provides a discovery interface for any Solr index. Blacklight provides a default user interface that is customizable via standard Rails (templating) mechanisms. Blacklight accommodates heterogeneous data, allowing different information displays for different types of objects. Community-contributed add-ons offer additional features.

Reviewer's Analysis

Blacklight primarily addresses the facet of data access by providing faceted search and filtering to indexed content and metadata. Blacklight adds an elegant and customizable user interface on top of a Solr search index. Blacklight is also used as the default discovery interface in Hydra (see separate review of Hydra)

It should be noted that Solr includes robust tools for text analysis and can be employed as a tool for data analysis, statistics, and normalization. For example, see Erik Hatcher's presentation on rapid prototyping Data.gov with Solr. In that sense, Blacklight + Solr may address data processing or data analysis needs.

Blacklight + Solr offer a pure and modern discovery layer for data and metadata.

Recommendations

I recommend further analysis that addresses specific GWLA/GPN use cases for: large-scale data analysis, data discovery, metadata discovery.

Key Contacts: Bess Sadler, Stanford University; Jonathan Rochkind, Johns Hopkins University
Sponsors: The University of Virginia, Stanford University, Johns Hopkins University, and WGBH are the principal contributors to the code base and use it heavily at their institutions. There are dozens of sites worldwide that use Blacklight.
Funding: Blacklight is supported by community members through their technical leadership and contributions.
Inception: 200x? Blacklight was originally developed at the University of Virginia Library and is made public under an Apache 2.0 license.
Geographic Location: Distributed.

Databib **ENABLERS: DISCOVERY**

http://databib.org/

Philip Konomos, Arizona State University

Brief Description of the Project

Developed by Purdue University Libraries, Databib is a free, online database that contains short bibliographic records describing disciplinary repositories that hold / accept research data. Their website states that Databib is "a searchable catalog of research data repositories." Databib is designed to help faculty, librarians, and others answer questions such as:

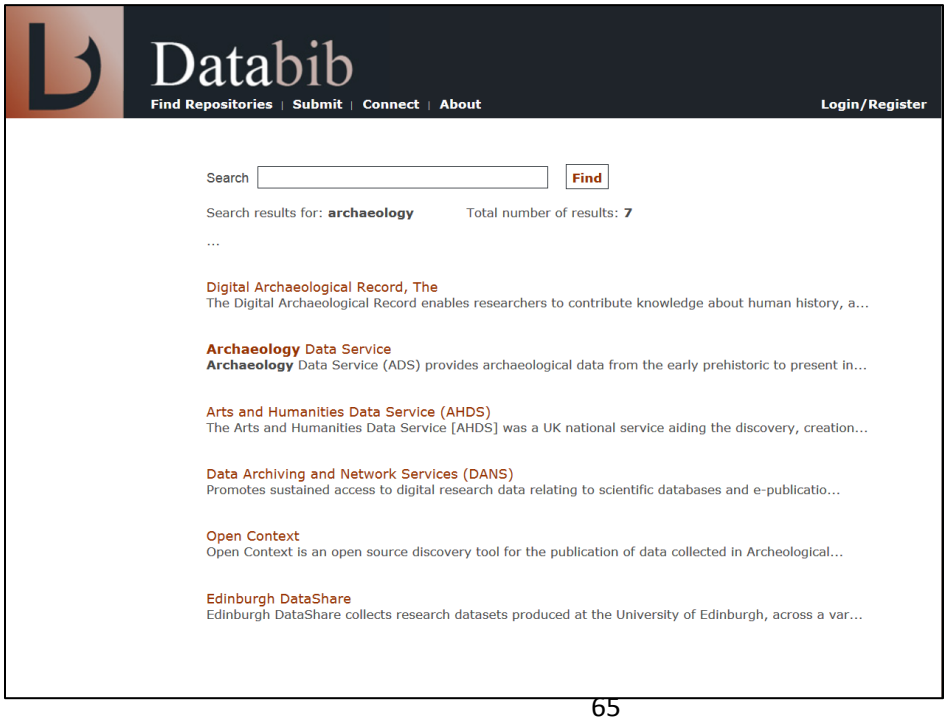
- What repositories are appropriate for a researcher to submit his or her data to?
- How do users find appropriate data repositories and discover datasets that meet their needs?
- How can librarians help patrons locate and integrate data into their research or learning?

The entries are international, though much more representative of the U.S. than other countries. Currently (December, 2012) Databib includes 501 repository descriptions. Additional entries are added as new information becomes available. Users can submit the names of new data repositories for consideration. Members of the Purdue University Library Databib editorial board then review each suggestion and create the new entry if appropriate. Users may also volunteer to be editors for a subject field (or fields), curating existing Databib records and suggesting new entries.

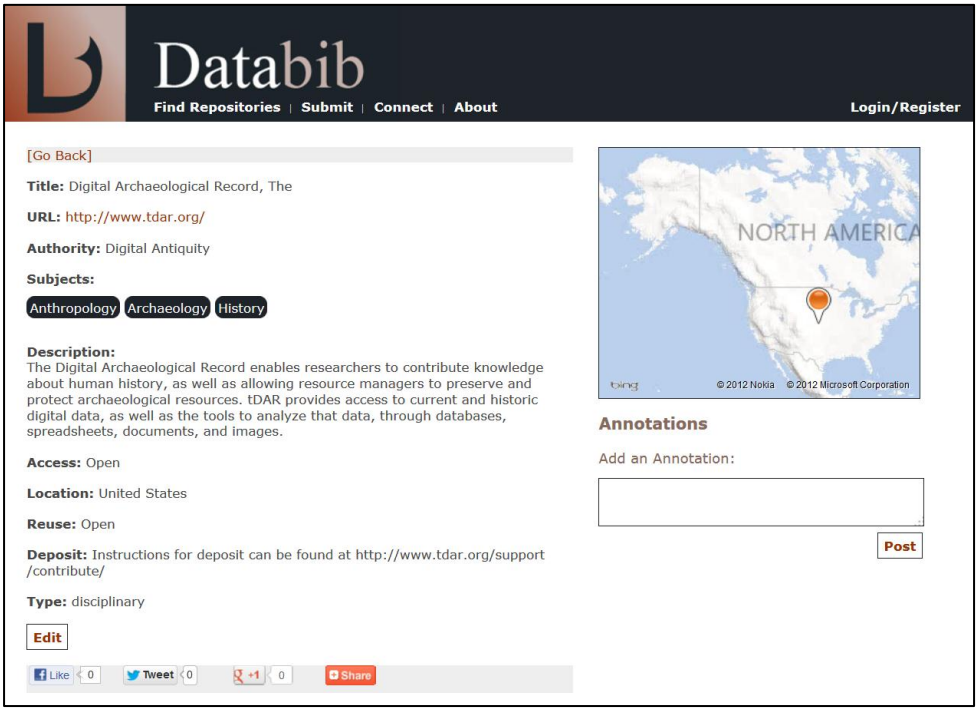
Databib is guided by an international advisory board with representation from Europe, Australia, Asia, Africa, and North America. An editorial board is being assembled to increase the coverage and continue the curation of records in Databib, according to their web site.

Reviewer's Analysis

This is a valuable service because it provides up-to-date information about available online digital data repositories across a broad spectrum of academic disciplines. Databib is easy to use, with clear, intuitive search and browse functions. Users searching Databib from the home page can enter a search term(s) in the search box, or browse an alphabetical list of all repositories. For example, a search for "archaeology" returns a list of 6 international repositories:



Selecting the first one (tDAR) opens a new page with standard descriptive information and the URL for the repository:



Considerations and Recommendations

Library staff among GWLA members will benefit from becoming familiar with Data**bib.** This is a valuable resources that is easy to use. The fact that Purdue staff and affiliates actively search for new content and manage the citations ensures that information in the catalog is up-to-date. While most universities now have an in-house institutional or digital repository, disciplinary repositories provide a number of advantages. Disciplinary repositories often use richer metadata, frequently containing elements specific to that particular field, for example.

Key Contacts:

The project email is databib@gmail.com The key contact is:

Michael Witt, Interdisciplinary Research Librarian
Assistant Professor of Library Science
Purdue University Libraries 765-494-8703 mwitt@purdue.edu

Sponsors: Data**bib** was developed by Purdue University Libraries' Distributed Data Curation Center (D2C2), and is managed by Michael Witt.

Funding: Creation and development of Data**bib** was funded through a 2011 grant from the Institute for Museum and Library Services (IMLS).

Inception: 2011

Geographic Location: Located at Purdue University but containing international information.

Linked Data

ENABLERS: Discovery

Jason Stirnaman, University of Kansas Medical Center

http://www.w3.org/standards/semanticweb/data, http://linkeddata.org, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Brief Description of the Project

Linked Data lies at the heart of what Semantic Web is all about: large scale integration of, and reasoning on, data on the Web.[4] Linked Data describes a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." [1]

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data. "[2]

Linked Data functions through links between arbitrary things described by RDF. The URIs identify any kind of object or concept, but regardless of HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the established standards (e.g. RDF, SPARQL)
4. Include links to other URIs, so that more things can be discovered

Reviewer's Analysis

As of 2010, the so-called Linked Open Data cloud covers more than an estimated 50 billion facts from many different domains like geography, media, biology, chemistry, economy, energy, etc. The data is of varying quality and most of it can also be re-used for commercial purposes.

Data Management programs should consider Linked Data from the perspective of Provider as well as Consumer. Careful attention to established standards, while also allowing different subject domains to dictate which standards (e.g. vocabularies) apply to their work, is crucial to making Linked Data work. Recognized standards:

- RDF
- Microdata (embeddable data) standards: Schema.org, RDFa

Linked Data has significant implications for discoverability and reuse of data. Researchers will often be both providers and consumers of Linked Data. Linked Data affords researchers and data providers an opportunity to increase the impact and reuse of their data. Linked Data enables consumers to more efficiently analyze data, recognize correlations between datasets, and make new discoveries.

Providing Linked Data: See "Ingredients for high quality Linked (Open) Data" by the W3C Linked Data Cookbook [5]

Consuming Linked Data: Linked Data consumers can integrate and provide high quality information and data collections to mix their own data with. Such integration enables better decision making, disaster management, knowledge management and/or market intelligence solutions. Tools such as ontology reasoners and SPARQL enable LD consumers to analyze semantically-enriched data.

Further examples and tools on the LinkingOpenData wiki. WebSchemas group provides a vocabulary and proposal for extending schema.org for dataset description. Tools for publishing, consuming, and integrating semantic data are readily available for most modern programming languages and web frameworks.

Considerations and Recommendations

I recommend phone interviews with current practitioners or further analysis that addresses specific GWLA/GPN use cases for: data discovery, data reuse, data analysis.

Sources:

1. http://linkeddata.org
2. Berners-Lee, T. Design Issues: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html
3. Bauer, F and Kaltenbock, M. Linked Open Data: The Essentials. http://www.semantic-web.at/LOD-TheEssentials.pdf
4. http://www.w3.org/standards/semanticweb/data
5. http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook#Ingredients_for_High_Quality_Linked_Data

Key Contacts:
Sponsors: W3C
Funding:
Inception: 2007
Geographic Location: Distributed.

Mercury (Oak Ridge National Laboratory) **ENABLERS: DISCOVERY**

http://mercury.ornl.gov/ &
http://www.dataone.org/software-tools/mercury-metadata-editor
Michael Bolton, Texas A&M University, Sterling C. Evans Library

Brief Description of the Project

Borrowing heavily from the ORNL (Oak Ridge National Laboratory) website, Mercury is a web-based system for searching metadata and retrieving the associated data. It is open source software and is based on a Service Oriented Architecture. It appears to be fairly flexible in how the service is provided supporting RSS, Geo-RSS, OpenSearch, Web Services and JSR-168 Portlets. This allows for easy integration into just about any interface or application. Mercury has the ability to extract or harvest metadata from HTML pages or XML files which makes participating in the service very easy. All the data provider needs to do is post the content on a Web server and Mercury will pick it up. The data is then incorporated into a centralized index where users can search the metadata either via a simple search mechanism or more intricate advanced search services. Mercury supports a number of metadata standards including XML, Z39.50, FGDC, Dublin-Core, Darwin-Core, EML and ISO-19115. Presentation slides are available at http://mercury.ornl.gov/slides/Mercury_presentation_05152012.pdf. ONEMercury is a specialized Mercury server for the DataONE network.

Reviewer's Analysis

In the area of finding data, Mercury seems to be a strong tool and service. The harvesting capabilities make the service very easy to use by researchers – there are no special processes to run or servers to install. Just publish on a web site that Mercury can access. How a researcher publishes this metadata may vary depending on discipline. That is, some communities provide metadata management tools, or cataloging services to be used by researchers wishing to publish their work.

A number of agencies and services are already using Mercury, in particular, the DataONE project where it is included in the software tools catalog. Continuing to use the DataONE model, contributors would use a service such as DataUP to create and edit metadata. It would then be published to a repository, such as ONEShare, which is a DataONE member node. The metadata is automatically harvested and added to the Mercury index. This same model could work for our consortia. Within our consortia we would have a number of contributing nodes that would feed our Mercury-based central index. I think this would be one of the real strengths of Mercury for us, that is, all the consortia members contributing to a centralized search index.

One option would be to join DataONE as a contributing member. DataONE is currently focused on the Earth, environment, atmospheric and ecological sciences. Repositories that fall into this general area are welcome to join as member nodes. While that would be good for researchers in those disciplines, it would leave other disciplines in a lurch. Maybe our niche could be to cover what DataONE does not.

I have found several references to a Mercury Metadata Editor tool; however, I never could find much information on the product. I did see links to the ORNL Online Metadata Editor (OME) but access to the page is restricted by a login.

Recommendations

Testing of individual tools will be difficult without a fairly robust test bed incorporating a metadata editor, repository and search engine. We can use an existing test system, such as the one being developed and deployed by DataONE or develop one of our own. From my research, I have found a number of "centers" that focus on a particular discipline, such as Earth Sciences. While that would be acceptable for initial testing, I can foresee our needs being much broader and we will need to support a cross-discipline service. That will mean we need to customize tools, schemas and best practices. Deploying our own test service appears to be about the only way we can control all the variables and make a truly informed decision on the suitability of the toolkits.

Key Contacts: mercury-support@ornl.gov
Sponsors: NASA, USGS, Office of Science – U. S. Department of Energy
Funding: Ongoing support is through the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC).
Inception: First reference seems to be 2008 but it may be older than that. There are a number of deployments and great many data providers indicating this software is fairly mature.
Geographic Location: Open Source software

Research Networking: VIVO, Profiles, BibApp **ENABLERS: Discovery**

http://vivoweb.org, http://bibapp.org,

Jason Stirnaman, University of Kansas Medical Center

Brief Description of the Project

Generally, *Research Networking* systems are web-based tools for discovering and using research and scholarly information about people and resources. Though underlying architectures vary widely, most "RN" systems share a high-level set of features:

- Compile data about research-related works (book citations, article citations, grants, presentations) and resources (labs, equipment).
- Provide tools to import, harvest, curate, and enrich the data, e.g. harvesting data from enterprise sources or PubMed, providing access to full-text through localized OpenURL resolution.
- Perform some amount of machine learning to disambiguate personal names in citation data.
- Attribute this data, unambiguously, to the people, often referred to as "experts", who are responsible for the works.
- Provide a single discovery interface for the data,
- Show collaborations and relationships among experts, e.g. through co-authorship, shared subjects, shared lab space.
- Provide data useful in measuring research output or decision-making.
- Provide data output using common format standards, e.g. use VIVO Ontology and RDF to express data and relationships through the VIVO Ontology, export styled citations, expose data objects through additional APIs as JSON, XML, etc.

Reviewer's Analysis

Research Networking systems address the facets of data access and data reuse. They aggregate and network data that may otherwise be siloed by database vendors or enterprise systems. They offer a discovery user interface for experts, collaborators, and their works. They may display metrics or some measure of impact for a work. They provide Linked Data/Semantic Web data sources and endpoints for scholarly works.

VIVO may currently be the most widely used RN system. VIVO is a semantic web, n-store (triples and quads)-based approach to gathering and sharing data about research activity. VIVO is developed by a consortium in the US. The project is based on the code base, VIVO, and the VIVO ontology for describing research.

VIVO is modeled to provide for a federated hub-node network of metadata about research activities. As a semantic web application, VIVO data is accessible as standard RDF, enabling the possibility of repurposing or sharing of data over the web. Each VIVO instance, e.g. at the institution-level, provides structured Linked Data that can be harvested and aggregated into a broader network. VIVO Searchlight is a practical application of this. Designed to accommodate data about multiple kinds of resources, typical bibliographic and enterprise data, as well as metadata about physical spaces and equipment. VIVO provides a Harvester Framework to facilitate ingesting data from external systems.

VIVO is primarily a Java application consisting of a Java-based UI, an underlying Jena SDB semantic store implemented with a MySQL database, and Solr for searching.

BibApp is used by a small, but growing number of institutions. BibApp specializes in the collection of bibliographic data and displaying the output of experts and groups. One unique feature of BibApp directly related to repositories is that it comes packaged with a SWORD client for allowing a user to archive a copy of a work directly into any SWORD-compliant repository, e.g. DSpace. BibApp uses the Sherpa/RoMEO API to allow an organization to monitor, which articles are open access and could be archived. BibApp exposes VIVO-compliant RDF Linked Data using the VIVO ontology, allowing an institution's BibApp data to be discoverable alongside VIVO nodes.

BibApp is a Ruby on Rails application. It is compatible with PostgreSQL, MySQL, and likely most other RDBMS. BibApp also uses Solr for searching.

Reviewer's Analysis

Research networks provide a useful discovery interface for associating research output with people, institutions, and cross-disciplinary groups. Most repository systems are concerned with digital objects, not people, as "first-class" objects. A research network can provide a useful and attractive complementary interface to a repository system.

Considerations and Recommendations

Any or all of these projects merit further review if a repository solution or support is considered. I recommend further analysis that addresses specific GWLA/GPN use cases.

Key Contacts: VIVO: ?, Profiles: ?; BibApp: Sarah Shreeves, UIUC & Jason Stirnaman, KUMC
Sponsors: VIVO: Various partner institutions, DuraSpace; Profiles: Harvard U. ; BibApp: University of Illinois, Urbana-Champagne
Funding: VIVO: NIH grant; Profiles: ; BibApp: University of Illinois, Urbana-Champagne and various institutions contributing development resources
Inception:
Geographic Location: Distributed.

VuFind

ENABLERS: Discovery

http://vufind.org
Jason Stirnaman, University of Kansas Medical Center

Brief Description of the Project

VuFind is an open source PHP + Solr application that provides a discovery interface for a Solr index. The goal of *VuFind* is to enable users to search and browse through all of your library's resources by replacing the traditional OPAC to include:

- Catalog Records
- Digital Library Items
- Institutional Repository
- Institutional Bibliography
- Other Library Collections and Resources

VuFind is completely modular, so you can implement just the basic system or all of the components. A wide range of configurable options allows extensive customization without changing any code.

Reviewer's Analysis

VuFind, like Blacklight (see Blacklight review), primarily addresses the facet of data access by providing faceted search and filtering to indexed content and metadata. VuFind adds an elegant and customizable user interface on top of a Solr search index.

VuFind was originally developed specifically to replace clunky vendor online catalog software for libraries and continues to be more oriented toward bibliographic and library collections. VuFind includes tools for harvesting metadata and content as well as importing metadata into Solr. Examples for harvesting and importing various "Open Data" metadata sources can be found at http://vufind.org/wiki/open_data_sources. However, no research data or dataset examples were listed at the time of viewing.

It should be noted that Solr includes robust tools for text analysis and can be employed as a tool for data analysis, statistics, and normalization. For example, see Erik Hatcher's presentation on rapid prototyping Data.gov with Solr. In that sense, VuFind + Solr may address data processing or data analysis needs.

Reviewer's Analysis

VuFind + Solr offer a pure and modern discovery layer for data and metadata, however VuFind's default toolset is more oriented toward repurposing library (MARC) and Dublin Core metadata.

Considerations and Recommendations

Does this project merit further review as part of a site visit, phone interview, or further analysis? (and, if so, what do you recommend?) I recommend further analysis that addresses specific GWLA/GPN use cases for: data discovery, metadata discovery. However, a more generic, less bibliocentric solution like Blacklight may be preferable.

Key Contacts: Demian Katz, Villanova University
Sponsors: Falvey Memorial Library, Villanova University
Funding: VuFind development is funded by Villanova University. Additional technical contributions provided by various institutions.
Inception: 2007
Geographic Location: Distributed.

ARK [Archival Resource Key] Identifiers **ENABLERS: Identifiers**

https://confluence.ucop.edu/display/Curation/ARK

Susan Matveyeva, Wichita State University

Brief Description of the Project

ARK identifier was developed by the National Library of Medicine and California Digital Library in March 2001. It supports a long-term access for information objects, both tangible and intangible. It is based on the idea that service or curation is the main condition of longevity of digital objects. An ARK is a URL created according to special rules. It answers the questions: who, what, when, and where (e.g. author – title – year – location). "The ARK (Archival Resource Key) naming scheme is designed to facilitate the high-quality and persistent identification of information objects.

A founding principle of the ARK is that persistence is purely a matter of service and is neither inherent in an object nor conferred on it by a particular naming syntax. The best that an identifier can do is to lead users to the services that support robust reference. The term ARK itself refers both to the scheme and to any single identifier that conforms to it. An ARK has five components:
[http://NMAH/]ark:/NAAN/Name[Qualifier]

(1) An optional and mutable Name Mapping Authority Hostport (usually a hostname)
(2) the "ark:" label
(3) the Name Assigning Authority Number (NAAN) [Institutional Identity.-S.M.]
(4) the assigned Name [information object. –S.M.], and
(5) an optional and possibly mutable Qualifier supported by the NMA.

The NAAN and Name together form the immutable persistent identifier for the object independent of the URL hostname.

An ARK is a special kind of URL that connects users to three things: the named object, its metadata, and the provider's promise about its persistence. When entered into the location field of a Web browser, the ARK leads the user to the named object. That same ARK, inflected by appending a single question mark (`?'), returns a brief metadata record that is both human- and machine-readable. When the ARK is inflected by appending dual question marks (`??'), the returned metadata contains a commitment statement from the current provider. Tools exist for minting, binding, and resolving ARKs.[26]

Reviewer's Analysis

The DOIs are more popular than the ARKs. However, the latter has around 100 registered users including the Internet Archive, Portico, MIT, DCC, the National Library of France, Google and many others.

[26] (Kunze, J. and R. Rogers, The Ark Identifier Scheme. May 22, 2008.) See also Kunze's 2012 presentation: http://www.slideshare.net/jakkbl/the-ark-identifier-scheme-at-ten-years-old

University of Kansas is a registered NAA (Name Assigning Authority) (NAAN: 25031)
http://www.cdlib.org/uc3/naan_registry.txt

UC Curation Center provides EZID services for both identifiers: DOI and ARK. ARK is less expensive than DOI and can be hosted locally.

The ARK structure includes Name Assigning Authority (NAA) as prefix; each NAA has a unique number (NAAN). Practically, these numbers are the institutional identifiers. This feature may be useful in consortial environment: (1) NAANs provide an easy efficient way to maintain an institutional identity of member-institutions; (2) the institution (NAAN) and its contribution (unique ID of data) connected as parts of the ARK. "Each organization is identified by a unique NAAN, which can be used as a prefix for the object identifiers that it assigns. For example, CDL assigns ARK identifiers that begin with its NAAN, 13030, as a prefix." (see: Identity Service: Name Assigning Authority Numbers http://www.cdlib.org/services/uc3/naan_table.html). Ability to maintain the inextricable connection between a contributing organization and its contribution may help to overcome some members' anxiety that in a large consortium environment they may be "lost," or be "absorbed", or "disconnected" from their contribution.

Considerations and Recommendations

The discussion about possible advantages of the ARK identifiers in consortial environment might be useful. John Kunze, the expert in ARK, is the best person to talk.
http://www.cdlib.org/contact/staff_directory/jkunze.html

Key Contacts: John A. Kunze, California Digital Library Curation Center
Sponsors: University of California Curation Center's IZID service
Funding: UC Curation Center operates EZID on a cost recovery basis.
Inception: 2001
Geographic Location: University of California

DOI (Digital Object Identifier) ENABLERS: Identifiers

http://www.doi.org/ (see also: http://www.datacite.org/whatisdoi)

Susan Matveyeva, Wichita State University

Brief Description of the Project

From the *DOI handbook*, http://www.doi.org/hb.html:

> DOI is an acronym for "digital object identifier", meaning a "digital identifier of an *object". A DOI name is an identifier (not a location) of an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks. A DOI name can be assigned to any entity — physical, digital or abstract — primarily for sharing with an interested user community or managing as intellectual property. The DOI system is designed for interoperability; that is to use, or work with, existing identifier and metadata schemes. DOI names may also be expressed as URLs (URIs). .. The DOI System provides a ready-to-use system of several components: a specified numbering syntax, a resolution service (based on the Handle System), a data model system (including the indices Data Dictionary), and policies and procedures for the implementation of DOI names through a federation of Registration Agencies.*

Reviewer's Analysis

DOI is widely used in publishing industry for identification of research articles. In 2005, the German National Library of Science and Technology (TIB) has started to assign DOI to research data, and in 2009, the global consortium DataCite was founded with the goal to manage the DOI system for research data. There are three DOI Registration Agencies in U.S.: California Digital Library, Purdue University Library, and U.S. Dept. of Energy Office of Scientific and Technical Information (OSTI). The University of California Curation Center (UC3) at CDL and Purdue University Libraries offer "DataCite DOIs and other identifiers via the EZID service (http://n2t.net/ezid), developed by UC3 to support easy identifier creation and maintenance for educational, non-profit, governmental and commercial clients. .. Through the OSTI Data ID Service, DOIs are assigned to research datasets, and then registered with DataCite to establish persistence. OSTI offers this service for researchers performing U.S. Department of Energy (DOE)-funded research activities carried out at DOE labs and facilities nationwide and grantees at universities and other institutions, as well as to other U.S. federal agencies and thereby other federal government-funded researchers. .. (see: http://datacite.org/DataCiteUS).

Data publishers (e.g. data centers, institutional or subject repositories) at first should register for an account with a DataCite member. To obtain an account, data publisher should meet certain requirements (contact@datacite.org -- membership enquires; tech@datacite.org – technical questions). DOI structure: prefix / client ID / unique string of characters.

DOI is an ISO standard assigned to research data. Major data services use DOIs to identify research data. DataCite also offers another persistent identifier: ARK (see a separate review).

Considerations and Recommendations

It is recommended that the regional data service assign DOIs for data it will host, which requires registration with one of the DOI Registration Agency and subscription to the DOI service. For pricing see: http://n2t.net/ezid/home/pricing .

Currently, DOIs and ARKs are offered as standard identifiers for research data; further analysis may clarify the choice of the most appropriate identifier for the regional data service as well as the possibility to use both identifiers if needed.

Key Contacts: inquiries@us.datacite.org
Sponsors: The International DOI Foundation (IDF)
Funding: IDF is an independent, not-for-profit, open membership organization funded by its members
Inception: 1998
Geographic Location: NA

ORCID [Open Researcher and Contributor ID] **ENABLERS: Identifiers**

Susan Matveyeva, Wichita State University
http://orcid.org

Brief Description of the Project

ORCID ID is a persistent unique identifier and a method for linking to digital research object. According to the web site:

> *"ORCID is an open, non-profit, community-based effort to create and maintain a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. ORCID is unique in its ability to reach across disciplines, research sectors, and national boundaries and in its cooperation with other identifier systems. ORCID works with the research community to identify opportunities for integrating ORCID identifiers in key workflows, such as research profile maintenance, manuscript submissions, grant applications, and patent applications. ORCID provides two core functions: (1) a registry to obtain a unique identifier and manage a record of activities, and (2) APIs that support system-to-system communication and authentication. ORCID makes its code available under an open source license, and will post an annual public data file under a CCO waiver for free download. The ORCID Registry is available free of charge to individuals, who may obtain an ORCID, manage their record of activities, and search for others in the Registry. Organizations may become members to link their records to ORCID identifiers, to update ORCID records, to receive updates from ORCID, and to register their employees and students for ORCID identifiers."*

Reviewer's Analysis

The author name ambiguity is an old problem. There is the number of initiatives that attempt to solve it. For example, VIAF (Virtual International Authority File (http://viaf.org/), the national libraries project, currently includes over 12 million authority records; another example is the OCLC WorldCat Identities service (http://www.oclc.org/developer/services/worldcat-identities), which uses OpenURL technology to provide information on author's works and the works about him/her. ResearcherID (http://www.researcherid.com/) is the Thomson Reuther service integrated it with the Web of Science. Scopus also offers author ID numbers. In 2012, ISO published ISO 27729 ISNI (http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=44292) to uniquely identify contributors of media content, such as books, TV programs, and newspaper articles.

The ORCID project is the newest and the most promising method of author identification. It is built on ResearcherID source code that was donated by Thomson Reuther. It is open source and relay on open source philosophy. It has strong support of major publishers, societies, universities, and research community. It is a global in score and discipline agnostic. Researchers can control privacy of their records, as they wish from public, to protect and be completely private. The ORCID ID format (16-digit

number) is compatible with the ISO 27729 ISNI. ORCID ID is easy to integrate with other systems'[27] research articles with DOI through CrossRef; the documents that do not have a DOI need to be entered manually. The intent is to use ORCID for grant applications, works of any format, and for research data. ORCID will also link all metrics and relevant sources related to scholarly contributions of a given author, including citations, usage data, etc.

Reviewer's Analysis

Researchers can delegate control of the ongoing management of their profiles to their institutions. This feature is very useful for the future consortium. Another important detail is a business model of the ORCID, which ensure its sustainability. The service is free for individual researchers, but not for organizational members. The size of membership fees will depend on the type and size of the organization, large commercial publishers will pay more than academic institutions. Pay may depend on the number of organizational members; when ORCID will have more members, fee may be decreased. Currently, ORCID offers two membership/subscription categories for organizations, basic and premium, and provides a 20% discount on the annual fee for non-profit organizations. Basic membership: $5,000 per year. See more at: http://about.orcid.org/about/membership

Considerations and Recommendations

My recommendation is to implement ORCID in the GWLA/GPN consortium.

[27] http://about.orcid.org/about/community/launch-partners

Creative Commons for data and databases **ENABLERS: LICENSING +**

http://creativecommons.org/science

Jason Stirnaman, University of Kansas Medical Center

Brief Description of the Project

Creative Commons provides licenses, tools, metadata, and best practices that facilitate the sharing of content. CC provides tools for integrating license selection and metadata into asset or content management systems (see http://wiki.creativecommons.org/Web_Integration).

From http://wiki.creativecommons.org/CC0_use_for_data :

> *(1) We do recommend CC0 for scientific data — and we're thrilled to see CC0 used in other domains, for any content and data, wherever the rights holder wants to make clear such is in the public domain worldwide, to the extent that is possible (note that CC0 includes a permissive fallback license, covering jurisdictions where relinquishment is not thought possible).*
>
> *(2) However, where CC0 is not desired for whatever reason (business requirements, community wishes, institutional policy…) CC licenses can and should be used for data and databases, right now (as they have been for 8 years) — with the important caveat that CC 3.0 license conditions do not extend to "protect" a database that is otherwise un-copyrightable.*

Real world uses of CC for data can be found at http://wiki.creativecommons.org/Data

Reviewer's Analysis

Creative Commons addresses the facets of Data Access (sharing and re-use). Reasons to share data include:

- fulfilling funder requirements,
- some journals require data archiving,
- raising interest in the research conducted,
- facilitating and increasing speed of research,
- establishing priority and providing public record.

Facts alone are not copyrighted but their arrangement may be sufficient original expression to merit copyright. For databases, there may be a mix of copyright and data for a research project to consider.[1] Before making a database available under a CC license, a database provider must first make sure she has all rights necessary to do so. Often, the database provider is not the original author of the database contents, which may mean the database provider needs separate permissions from third parties before publishing the database under a CC legal tool.

Also, the database provider must consider what elements of the database she wants to be covered by the CC legal tool and identify those elements in a manner that re-users will see and understand.

A big part of the potential value of data, in particular its society-wide value, is realized by use across organizational boundaries. What are the legal mechanisms for this? Many sites give narrow permission

to use data via terms of service. Much ad hoc data sharing occurs among researchers. And increasingly, open data is facilitated by sharing under public terms to manage copyright restrictions that might otherwise limit dissemination or reuse of data, e.g. CC licenses or the CC0public domain dedication.

The current recommended application of Creative Commons to data and databases is discussed at http://creativecommons.org/weblog/entry/26283.

Does this project have implications for a collaborative or consortial approach to defining a regional (GWLA/GPN) strategy for lifecycle management of research data? If so, what does this project offer in terms of approaches that might be incorporated or services that might be acquired? Yes, research data and any inclusive management system should be accompanied by policies, expectations, or licenses for data re-use. Those policies, expectations, or licenses should be clearly expressed and readable by both humans and machines. Most modern content and asset management systems offer some of this functionality by integrating the services at http://wiki.creativecommons.org/Web_Integration

Considerations and Recommendations

This project merits further review if a repository solution or support is considered.

Sources:

1. http://library.uoregon.edu/datamanagement/ip.html
2. Key Contacts: Puneet Kishor, Project Coordinator for Science and Data, http://creativecommons.org/staff#puneetkishor

Sponsors: https://creativecommons.net/supporters/
Funding: Non-profit. Private and corporate donors.
Inception: 2001. CC version 1.0 released 2002. Science Commons launched in 2005 and was re-integrated with Creative Commons 2011.
Geographic Location: Distributed. Offices located at MIT.

Data Documentation Initiative (DDI) **ENABLERS: Metadata**

http://www.ddialliance.org/

Larry Hoyle, University of Kansas, Institute for Policy and Social Research

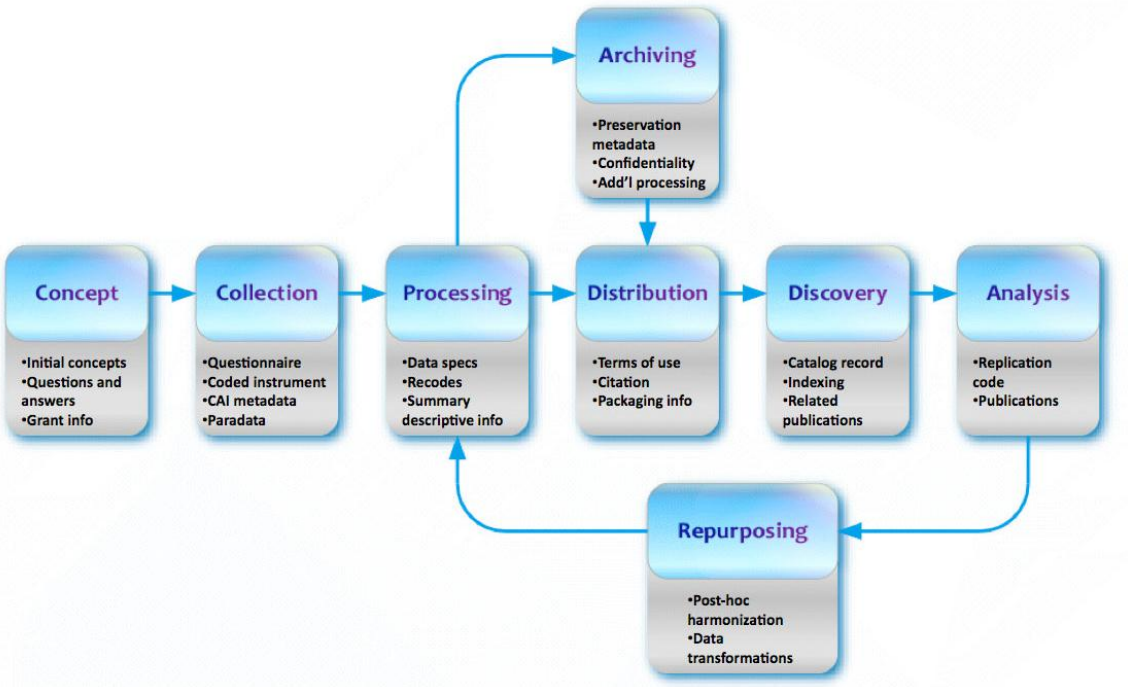
Brief Description of the Project

The current description from the DDI web site is: "The **Data Documentation Initiative (DDI)** is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving." This is the current description of the next generation of DDI:

> *The Data Documentation Initiative (DDI) is an international standard for describing data related to the observation and measurement of human activity. With origins in the quantitative social sciences, researchers in other disciplines are increasingly using DDI. The DDI specification is also being used to document other data types, such as social media, biomarkers, administrative data, and transaction data."* (from: Developing a Model-Driven DDI Specification)

Reviewer's Analysis

DDI has two separate ongoing branches, DDI Codebook, and DDI Lifecycle. The latter aims to cover the data lifecycle as shown in the figure below. DDI Lifecycle is currently and XML based standard, defined by a set of XML schemas. A future version of DDI Lifecycle will be model-based with bindings into XML, RDF, and relational schemas.



Considerations

and Recommendations

DDI is one of the major standards for Social Science research. With growing use by data archives (especially in Europe) and national statistical agencies, I think it is an important standard.

Also note that we will be having the first North American DDI Conference here at KU April 1-3, 2013 http://www.ipsr.ku.edu/naddi/

Key Contacts:
Mary Vardigan, ICPSR, DDI Alliance Director
DDI users listserv http://www.ddialliance.org/community/listserv

Sponsors: The DDI Alliance http://www.ddialliance.org/alliance
Funding: Alliance membership
Inception: The DDI (Data Documentation Initiative) began in the mid-1990s as a project to create a structured metadata standard for the social sciences.
Geographic Location: Worldwide

Metadata Encoding & Transmission Standard (METS) **ENABLERS: Metadata**
http://www.loc.gov/standards/mets/
Larry Hoyle

Brief Description of the Project

The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

Mets includes elements to wrap representations of digital objects, organize content into hierarchical structures, and associate executable behaviors with content. Mets can also carry metadata regarding file groupings, provenance (preservation-related actions) , rights, and related parties and their roles. Mets can point to other metadata in external formats.

Reviewer's Analysis

Mets plays a management role for digital objects in an archive.

Considerations and Recommendations

Merits further attention.

Key Contacts:

Network Development and MARC Standards Office
Library of Congress
LS/OPS/NDMSO (4402)
Washington, DC 20540-4402
ndmso@loc.gov
http://www.loc.gov/help/help-desk.html

Sponsors: Network Development and MARC Standards Office of the Library of Congress ndmso@loc.gov
Funding:
Inception:
Geographic Location:

Metadata Object Description Schema (MODS) **ENABLERS: Metadata**

(see also MADS)

http://www.loc.gov/standards/mods/

see also: http://www.loc.gov/standards/mods/registry.php (implementation registry)

and http://www.loc.gov/standards/mads/

Larry Hoyle, University of Kansas, Institute for Policy and Social Research

Brief Description of the Project

Mods is a schema for metadata at the collection and object (item) level. The MODS main page describes it as:

> *"Metadata Object Description Schema (MODS) is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users. -- More about MODS"*

A companion schema (MADS – Metadata Authority Description Schema) is described as:

> "The Metadata Authority Description Schema (MADS) is an XML schema for an authority element set that may be used to provide metadata about agents (people, organizations), events, and terms (topics, geographics, genres, etc.). MADS serves as a companion to the Metadata Object Description Schema (MODS) to provide metadata about the authoritative entities used in MODS descriptions. The standard is maintained by the MODS/MADS Editorial Committee with the Network Development and MARC Standards Office of the Library of Congress and input from users. "

Quoting from the Mods Schema Outline (http://www.loc.gov/standards/mods/mods-outline.html):

Top Level Elements:

titleInfo	note
name	subject
typeOfResource	classification
genre	relatedItem
originInfo	identifier
language	location

physicalDescription	accessCondition
abstract	part
tableOfContents	extension
targetAudience	recordInfo

Reviewer's Analysis

The MODS registry lists 34 projects currently using MODS. Some University repositories. More than half of the projects are archiving digitized objects (which might be classed as qualitative, or unstructured data). Examples include sheet music, photographs, digitized physical objects, buildings, and archived web sites. In several cases MODS is used with Fedora or DSpace. Some use MODS within a METS framework. Some use MODS as an intermediate format between other metadata schemas.

Considerations and Recommendations

There may be usable tools built on MODS.

Key Contacts: ndmso@loc.gov
Current MODS/MADS Editorial Committee Membership

- Rebecca Guenther, Library of Congress, Chair
- Jan Ashton, British Library
- Ann Caldwell, Brown University
- Reinhold Heuvelmann, German National Library
- Bill Leonard, Library and Archives Canada
- Sally McCallum, Library of Congress
- Betsy McKelvey, Boston College
- Jon Stroop, Princeton University
- Robin Wendler, Harvard University

Sponsors: Library of Congress
Funding:
Inception:
Geographic Location:

PREservation Metadata: Implementation Strategies **ENABLERS: Metadata**
(PREMIS) - http://www.loc.gov/standards/premis/
Larry Hoyle

Brief Description of the Project

"The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation."

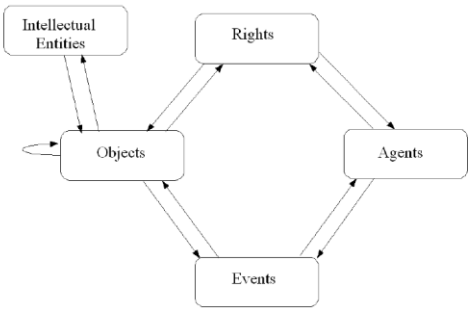


Figure 1: The PREMIS Data Model

Figure 1 at the right is taken from http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

Objects in PREMIS are "described as a static set of bits. It is not possible to change a file (or bitstream or representation); one can only create a new file (or bitstream or representation) that is related to the source Object. "

Key Contacts:
Sponsors: Library of Congress
Funding:
Inception: June 2003 OCLC and RLG sponsored the formation of the initial working group in May 2005 with the release of *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group.*
Geographic Location:

Data Curation Profiles **ENABLERS: Planning**

http://datacurationprofiles.org/

Susan Matveyeva, Wichita State University

Brief Description of the Project

Data Curation Profile is a free online resource for Library and Information Science professionals, Archivists, IT professionals, Data Managers, and others who want information about the specific data generated and used in research areas and sub-disciplines that may be published, shared and preserved for re-use. Data Curation Profiles capture requirements for specific data generated by a single scientist or lab, based on their reported needs and preferences for the data.

Each Data Curation Profile is essentially an outline of the "story" of a data set or collection, describing its origin and lifecycle within a research project. The website includes the ***Directory*** of completed Data Curation Profiles on a variety of subjects; the downloadable toolkit, and a list of resources. In 2011-2012, the profile developers conducted 12 workshops, funded by the Institute of Museum and Library Services. The Guidelines for Authors[28] provides information on definition, structure, and sequence of core and optional modules of the Data Curation Profile. Profile creators are encouraged to submit the completed profiles for publication in the Directory. Publication process and requirements are clearly described. The published profiles are assigned DOI; indexed by Google Scholar, major library discovery tools, and preserved with CLOCKSS and Portico.

A Data Curation Profile is a valuable tool for any data curation project. The profile addresses the data lifecycle and helps data curators to interview researchers, to become familiar with data in different disciplines and subject areas, to identify possible data services, and to plan data curation projects. A Data Curation Profile can be included into a documentation package of regional consortia as a standard tool for data curators working with researchers and planning data curation projects.

Considerations and Recommendations

This project can be included to a list of useful resources for data curators, project managers and librarians.

Key Contacts: Jake Carlson, Associate Professor of Library Science / Data Services Specialist, Purdue University (jrcarlso@purdue.edu) and D. Scott Brandt, Associate Dean of Research, Professor of Library Science, Purdue University (techman@purdue.edu)
Sponsoring entities: Purdue University Library; Distributed Data Curation Center; IMLS
Funding Source: The Institute of Museum and Library Services
Inception: 2007
Geographic Location: Purdue University, West Lafayette, Indiana

[28] http://docs.lib.purdue.edu/dcp/guidelines.html

DMPTool **ENABLERS: Planning**

https://dmp.cdlib.org/

Deborah Ludwig, University of Kansas

Brief Description of the Project

The DMPTool is a freely available web service. The primary goals of the tool are to allow researchers to quickly and easily produce a quality data management plan, and to inform researchers of relevant resources and support services across the community and within their institution. Features include:

- Create ready-to-use data management plans for specific funding agencies
- Meet funder requirements for data management plans
- Get step-by-step instructions and guidance for your data management plans as you build it
- In many cases, get data management advice and resources for your specific institution

The tool identifies those elements that specific funders want grant applicants to address, and it allows users to edit, save, share (if desired), print and download their data management plans. [source: CNI program, http://www.cni.org/pbs/dmptool/] There is also a video demo (http://dmp.cdlib.org/help/video_demo) that shows how local resources and services appear to researchers using the DMPTool.

The DMPTool site is gaining momentum. The web site reports "October [2012] was our biggest month ever. 375 new users logged into the DMPTool in October, and 336 plans were created. There [are] now a total of 3,466 users, and they've created almost 3,000 plans total. 2 more universities customized the DMPTool for their researchers, bringing the total to 28. 65 have configured their campus single-signon for the DMPTool." There is a map of our participating organizations: http://bit.ly/L85sKj

Reviewer's Analysis

This project addresses the overall need for planning a data management strategy at project inception, not only in response to funder requirements, but also in alignment with institutional resources if the local institution chooses to become a member institution and to customize the tool for its researchers so that the institutions choices and options are reflected in the researchers data management plan. As such, the tool potentially addresses aspects of data management across the full lifecycle of data. The ability for researchers to save, re-use, and to share (if desired) their data management plans through the plan's repository makes this a particularly advantageous approach for the researcher and possibly for the home institutional.

Specifically, the sections of a DMP-generated plan addresses: 1) Data generated by the project, 2) Period of data retention, 3) Data format and dissemination 4) Data storage and preservation of access, and 5) additional possible data management requirements.

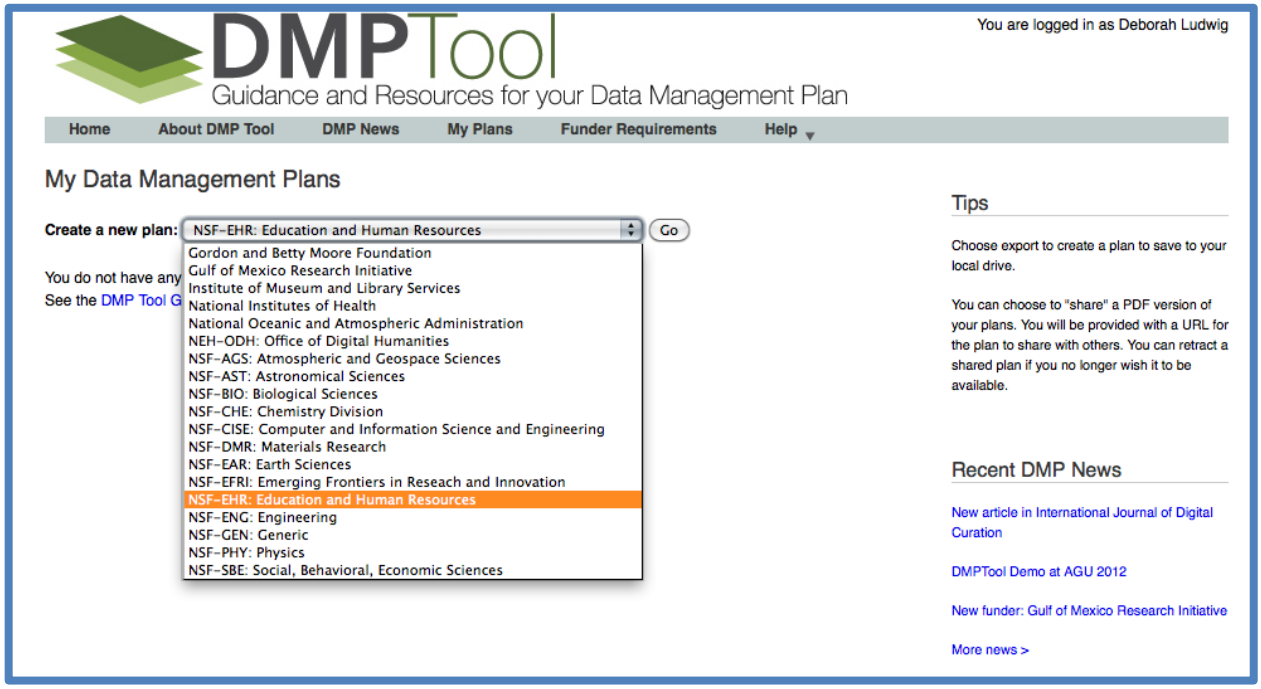
The implication of this project for a collaborative formed by GWLA and GPN might be the need to provide workshops to introduce this tool to member institutions, to train user or customer services staff at institutions work with researchers who are creating in the context of their local institutions, and to

perhaps gather information from members about needed enhancements to the tool. Because the tool is hosted, there is no need for additional infrastructure to support the tool outside of member institutions' development of shibboleth.

Only a few GWLA and/or institutions are currently listed as institutions that have customized the tool for their researchers.

Considerations and Recommendations

Perhaps a DMP Tool workshop to future GWLA / GPN member meetings? Compare with the Data Curation Profiles developed by Purdue.



Key contacts: uc3@ucop.edu; Andrew Sallans, Head of Strategic Data Initiatives, Library; Co-Lead on DMPTool Project, University of Virginia; Carly Strasser, Data Curation Specialist, California Digital Library. See http://www.cni.org/pbs/dmptool/

Sponsors: DMPTool is a collaborative effort of eight institutions: the California Digital Library, DataONE, the Digital Curation Centre, the Smithsonian Institution, the University of California at Los Angeles Library, the University of California at San Diego Libraries, the University of Illinois at Urbana-Champaign Library and Office of Cyberinfrastructure, and the University of Virginia Library. Other institutions can join this effort by becoming a contributing organization. Member institutions can:

- Enable Shibboleth login ("single sign-on") as InCommon members: https://dmp.cdlib.org/help/dmp_shibboleth
- Add links to local resources, help text, suggested answers, contact information

Funding: Part of the funding is from member institutions that contribute.

Inception: 2011
Geographic Location: Located physically on servers at UC3, California Digital Library