

Purdue Libraries
Libraries Research Publications

Purdue Libraries

Year 2006

Enabling Interaction and Quality in a
Distributed Data DRIS

D. Scott Brandt*

James L. Mullins[†]

Michael Witt[‡]

Enabling Interaction and Quality: Beyond the Hanseatic League, 8th International Conference on Current Research Information Systems (pp. 55-62). The Netherlands: Leuven University Press, 2006.

*Purdue University, techman@purdue.edu

[†]Purdue University, jmullins@purdue.edu

[‡]Purdue University, mwitt@purdue.edu

This paper is posted at Purdue e-Pubs.

http://docs.lib.purdue.edu/lib_research/1

Enabling Interaction and Quality in a Distributed Data DRIS

D. Scott Brandt, James L. Mullins, Michael Witt
Purdue University Libraries, West Lafayette, Indiana, USA

1 Summary

Purdue University Libraries has undertaken the development of a distributed data institutional repository in response to needs of researchers. The result will be a Distributed Research Information System (DRIS) for the university, composed of multiple repositories which may be formal or informal Current Research Information Systems (CRIS). This paper reviews strategic approaches, architectures, roles and interactions of all the players and parts involved.

The goal of this research is to develop a proof-of-concept system that provides access and high-level discovery to a variety of research data. The plan of work to accomplish this included:

- Developing an operational model for a multi-CRIS
- Identifying multiple datasets that vary in type (amount of metadata associated with them and their location)
- Constructing a local data repository and interfacing with remote repositories that house the datasets
- Implementing OAI-PMH with a harvester and metadata repository to retrieve and aggregate metadata
- Reconciling and enhancing metadata to facilitate discovery and improve description
- Building an effective end-user interface

2 Background

According to a recent review of the state of repositories in the United States, the success of the “if-you-build-it-they-will-come” approach to research information systems and institutional repositories in academic settings has been lackluster (Lynch 2005). This is emphasized in a study which recognizes that faculty do not see document repositories as useful or needed facilities for their involvement (Foster 2005). It appears that some efforts at universities in the US include building repositories as library-driven projects which anticipate a problem or propose a solution without determining consumer need. However, the TARDIS Institutional Repository route map at

Southampton recognizes and accounts for integrating with the natural processes of researchers (Hey 2004). It is recognized that efforts in Europe have been more successful in developing national repositories (Vattulainen 2004).

Taking its cue from Southampton, Purdue University Libraries set out on a dual “top-down” and “bottom-up” approach to identifying user needs before attempting to build a DRIS for the university. It tackled this effort first by making high-level contact with every college, school and department head, as well as several center directors, to announce an initiative of interdisciplinary collaboration on research within the university, explaining librarians’ expertise in organizing information and collections. The Dean of the Libraries initiated a new program of interdisciplinary research which sought to partner librarians as co-principal investigators on sponsored funding proposals to NSF, NIH, etc. Four of such proposals with librarians as co-PIs were submitted within six months of the program’s start.

3 Multi-CRIS approach

Along with collecting, preserving, and providing access to traditional materials such as preprints, postprints and reports, libraries must deal with other bodies of work of the university--datasets and the related content of various networked information systems. The Purdue Libraries initiated a multi-CRIS project named the Distributed Institutional Repository (DIR) to meet the needs of researchers to help them organize, store, disseminate, and provide an opportunity to repurpose their data and research outputs in new and interesting ways.

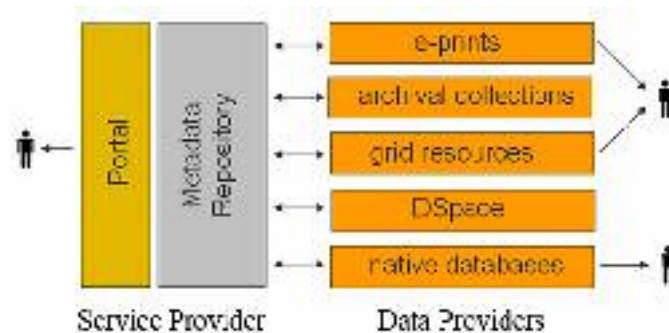


Figure 1

A unified portal acts as a high-level gateway to search and access the DIR by harvesting metadata and provides various levels of descriptions and linkages to digital objects. At a minimum, the description is unqualified Dublin Core; the metadata transport involves the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH); and the linkage is a URL. Enhanced metadata and delivery of information improves levels of both descriptions and linkages. The unified portal is not intended to replace the native interfaces of participating information systems; instead, it serves to increase the discovery of resources and provide an architecture to support

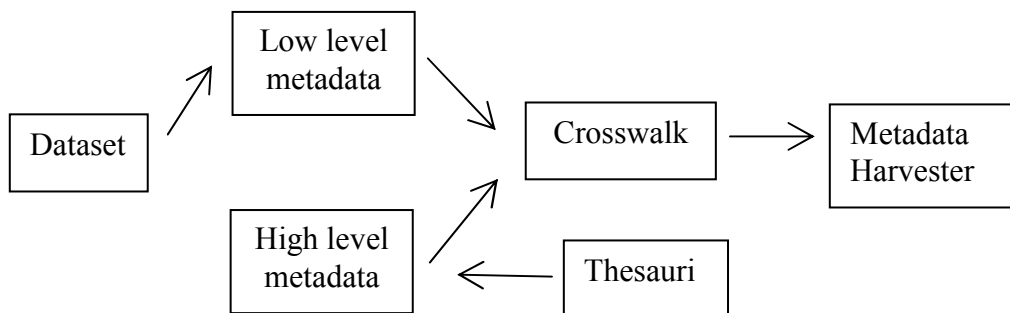
future interoperability among them. While some information systems may be operated locally by the Libraries, many will be hosted remotely and/or administered elsewhere on-campus or on the Internet.

4 System model for reconciling metadata

A second goal was to “reconcile” low and high level metadata for the files to enhance access and discovery. A testbed of three disparate data sets from different areas was used to investigate, integrate and apply metadata schemas. Where possible, relevant discipline specific schemas were reviewed to identify whether and how we could enrich and extend the metadata to make the data more useful and provide opportunities for it to be repurposed. The objectives for investigating metadata addressed reconciling low and high level metadata, and determining a method to incorporate metadata from many disparate places into a metadata harvester which sits behind the DIR portal.

For purposes of this research, reconciling is defined as comparing and integrating low level system metadata and high level descriptive metadata to facilitate greater discovery. It is recognized that distinctions between low and high levels can be fuzzy. For instance, a metadata element “creator” of an electronic file may be filled in automatically by the operating system user account preferences, when in fact the files are being manipulated by a graduate student or part time staff member for a research group. Thus, a standard definition is needed which best allows extraction of low level metadata, identification of higher level descriptors, comparison of the two, and enhancement to fulfill more robust discovery.

The underlying approach for this project was to develop a system model of reconciliation and enhancement which will inform eventual automation of the process. While this model primarily would be a manual process, it would identify elements and process that could be automated. For instance, as noted below, librarians constructed either low or high level metadata (or both) “by hand” to build an appropriate set of records.



Initial system model for reconciling metadata

Figure 2

5 Reconciling metadata

Three different datasets were chosen for review: water quality logs and reports, aerial and satellite imagery, and climate modeling data. These datasets differed in how much (or what kind of) metadata was assigned to them, where they were hosted, and which university departments maintained them. The amount or kind of metadata provided was varied to test librarians' ability to reconcile different levels of metadata. The location was varied to test the ability of the information systems and harvester to expose and gather metadata from diverse places. The departments were varied to test librarians' ability to work with heterogeneous groups (e.g., different research groups which used different terminology and had different purposes for their data).

The first dataset comprised water quality data collected from approximately fifty field testing stations that have been monitored and maintained by Purdue's Agronomy Department since 1993. Flow, temperature, and mineral information is logged into ASCII flat files in a proprietary format and periodically summarized in Excel spreadsheets. Previously, the spreadsheets and log files were written to CD-R media and stored on shelves in the principal investigator's office. A sample of five years' worth of data was ingested into the Libraries' DSpace repository and described as a part of this project.

The second dataset comprised multi- and hyper-spectral images maintained by Purdue's Laboratory for Applications of Remote Sensing (LARS). The LARS data spans approximately thirty years and had been marked up by a technician using Federal Geographic Data Committee (FGDC) metadata standards. The data is indexed on a website maintained by LARS and accessible using SRB and from TeraGrid. For our project, a dataset covering images of Indiana was selected.

The third dataset comprised climate data from researchers in Purdue's Earth & Atmospheric Sciences Department. The data is derived from recent Community Climate System Model (CCSM) experiments and is used to analyze climate systems and trends. The data resides in NetCDF format, and presently only native metadata extracted from the file headers is available. It is also accessible using SRB and from TeraGrid.

The primary objective was to use Dublin Core (DC) and find appropriate schemas and standards to bolster it. The hypothesis for reconciling metadata was that "low" level metadata (file name, type, date of origin, etc.) would need to be enhanced with "high" level (keywords, subjects, etc.) to facilitate discovery across or outside of the disciplines. A first pass at the literature and available websites for standards and schemas revealed several watershed, geospatial and climate projects which had standards or thesauri to describe data, as well as guidelines or using them. In addition, there were several metadata crosswalks or standards for converting from one schema to another. For instance, at the NASA Goddard Space Flight Center Global Change Mastery Directory (gemd.nasa.gov/Aboutus/standards) there is a crosswalk between FGDC to Directory Interchange Format (DIF). Elsewhere there is a crosswalk between DIF and DC.

For the water quality dataset the objective was to create a collection level metadata record in DC using DIF as a standard (DIF is less extensive than FGDC) and utilizing the US National Agriculture Library (NAL) thesaurus for subject and keyword terms. The NAL thesaurus was used because it was found to provide an extensive and highly relative controlled vocabulary. Identification of NAL terms to apply to the collection metadata record was based on a text analysis by a cataloging librarian of a published article on the water quality research.

For the remote sensing dataset the objective was to create a collection level metadata record in DC following the FGDC crosswalk. As with the DIF to DC crosswalk, translation requires repeating DC elements to account for the larger, more extensible schema. For instance, DIF fields for project, location or sensor name were converted to keywords, and latitude or longitude listed under the coverage element. NAL terms were also added to the subject element based on a text analysis of the LARS website. Specialists working with the LARS data previously had performed comprehensive application of FGDC, which was considered “well marked up.”

For the climate data the objective was to extract metadata for the dataset across the SRB. This work was done in close collaboration with Purdue’s IT group (ITaP), whose members built data management programs to extract metadata from the data files in their native NetCDF format (Zhao 2006). This proved to be the most challenging work, as export and extraction of metadata required considerable standardization.

6 Metadata harvesting

One of the primary functions of the DIR is to automate the function of harvesting metadata from disparate information systems. The harvester was coded in the java programming language, drawing heavily from the Online Computer Library Center’s OAICat open source software. The first time the harvester visits an OAI-compliant CRIS, it checks the schemas of the available metadata formats and creates the necessary relational tables in a MySQL database before harvesting and populating them. The database acts as the metadata repository that serves as the backend for a web-based portal that provides basic and advanced search interfaces. The portal was written in Coldfusion and connects to the metadata repository using Open Database Connectivity (ODBC).

As noted above, the project aimed to build on previous work. An early initiative by the San Diego Supercomputing Center (SDSC) addressed the problem of locating and managing large data sets identified the need for large data and information handling systems to deal with outputs of ever expanding computational science and automated data generation and acquisition. Their work described the integration of meta-computing systems and digital libraries, which led to the development of the Storage Resource Broker (SRB) to provide “attribute-based access” to remote data sets (Moore 1998). An SDSC/MIT/UCSD collaboration developed a “prototype persistent archive” which provided a user-friendly interface and supported “digital asset life-cycle management” using two different “preservation architectures” to benefit from both (Moore 2005). DSpace has a “front end for digital content life-cycle management” which includes “content ingestion, search and discovery, content management, dissemination services and preservation.”

SRB has an infrastructure for data management that “enables diverse collections of storage devices to operate as a single device” across networks such as TeraGrid. Outcomes included increasing the flexibility and scalability of DSpace by adding remote storage access via the SRB, and exchange of data between the two to support search and discovery, as well as infrastructure independence. Collections used in the project included still and moving images, theses, and web site files, which included streaming media, and took up nearly 6 TB of space.

Because DSpace has native support for OAI-PMH, harvesting metadata for the water quality datasets for the DIR was straight-forward. However, both remote sensing and climate data resided on the SRB as primarily grid-oriented resources. In order to access their metadata, an OAISRB protocol translator was developed. It converts OAI-PMH verbs to SRB calls and returns the metadata as XML. OAISRB is coded in java using the JARGON toolkit from San Diego Supercomputing Center.

7 Conclusion

Outcomes of this work were facilitated using available technologies to support collection management processes for research data in this distributed digital environment (e.g., SRB, DSpace, etc.). The investigation focused on identifying different environments and pathways depending on how data is utilized (e.g., for discovery, archive or use) and presenting a portal front-end to a distributed architecture which stores or points to data depending on its type, location or format. The project has not yet employed a comprehensive CRIS framework, and that is on a list of further objectives to be achieved.

Next the researchers want to move beyond proof-of-concept to develop processes and tools which assist researchers in assessing their data needs. Next steps will include:

- Scaling up the number of datasets, working with both local and distributed data
- Setting up “live” and archived datasets to explore management issues further
- Incorporating extended metadata to address rights (who gets access), or preservation (how long the data should sit in the repository), etc.
- Explore ontological applications for automated metadata generation

Additionally, the Libraries seeks to identify sustainability models to scale the DIR into a fully functioning production facility. We foresee investigating several issues; obvious ones include large system and support requirements, scalability of operation, growth of collections, maintenance of infrastructure, and the management of content and users. We are looking to work with others to approach the topic of exploring and/or developing such a sustainability model.

Last, to be successful, tools must be put in the hands of the data generators and managers to select the type of organization and management solutions they need (i.e., archiving, curating, accessing or discovering). Librarians, working with others, need to further identify and resolve data organization and management problems by building tools, applications and processes which can

be easily acquired and used. Specifically, Purdue Libraries will pursue a matrix that can be used for assessing datasets for collection management, and develop tools which can walk users through protocols to facilitate and automate metadata generation.

References

- Foster, N. & Gibbons, S. (2005): Understanding Faculty to Improve Content Recruitment for Institutional Repositories. In: *D-Lib Magazine*, 11, 1.
- Hey, J. (2004): Targeting Academic Research with Southampton's Institutional Repository. In: *Ariadne*, 40.
- Lynch, C. and Lippincott, J. (2005). Institutional Repository Deployment in the United States as of Early 2005. In: *D-Lib Magazine*, 11, 9.
- Moore, R. (1998): Data-Intensive Computing and Digital Libraries. In: *Communications of the ACM*, 41, 11, 56-62.
- Moore, R. (2005): NARA Supplement to the NPACI Collaboration: Integrating Data Management with Data Grids. In: *SDSC Technical Report 2005-4*, Final Report, July 2005.
- Vattulainen, P. (2004): National repository initiatives in Europe. In: *Library Collections Acquisitions & Technical Services*. 29, 39-50.
- Zhao, L. et al. (2006) : Purdue Multidisciplinary Data Management Framework Using SRB. In: *SDSC SRB Workshop, February 2-3, 2006, San Diego, CA*.

Contact Information

D. Scott Brandt
Associate Dean for Research
Purdue University Libraries
504 W. State Street
West Lafayette, IN, USA 47907
techman@purdue.edu

James L. Mullins
Dean of Libraries
Purdue University Libraries
504 W. State Street
West Lafayette, IN, USA 47907
jmullins@purdue.edu

Michael Witt
Senior Research Systems Administrator
Purdue University Libraries
504 W. State Street
West Lafayette, IN, USA 47907
mwitt@purdue.edu