4-30-2012

Databib: IMLS LG-46-11-0091-11 Final Report (White Paper)

Michael Witt
Purdue University, mwitt@purdue.edu

Michael J. Giarlo
Pennsylvania State University - Main Campus, michael@psu.edu

Final Whitepaper Report: Databib

IMLS Project Number: LG-46-11-0091-11

1. Administrative Information

Institution:	Purdue University
Project title:	Databib
Award amount (total project cost):	$24,594 ($28,161)
Project start and end dates:	August 1, 2011 - April 30, 2012
Project director:	Michael Witt (mwitt@purdue.edu)
Formal project partner organization:	The Pennsylvania State University

2. Project Summary

A collaboration between the Purdue University Libraries and Penn State Information Technology Services has created a new reference resource and software platform called Databib (http://databib.org) that sparks engagement of librarians in data services by providing them with an online, community-driven, annotated bibliography and catalog of research data repositories.

Needs and Rationale

The Fourth Paradigm: Data-Intensive Scientific Discovery[1] describes the current paradigm shift in science that is transforming the research process to focus on the capture, curation, and analysis of digital data. With the advent of e-Science, data are being created at a rapid rate, resulting in a "data deluge" widely reported in both scholarly literature and the popular press. A workshop convened by the National Science Board in 2005 produced a report on "Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century"[2] that recognized the importance of research data and underscored the need for the development of government policies to ensure the stewardship of datasets and preserve their value for improving and advancing science.

Sharing research data is an important ethic for many scientific disciplines, and repositories play a key role in this scholarly exchange. Having access to datasets is critical to validating reported research findings. Data can also be reused to advance the original research or new lines of inquiry. Moreover, preserving and sharing existing datasets in repositories avoids the cost of generating new data from scratch. In the case of government-sponsored research, such repositories make research data available to the taxpayers who funded the research as well as to citizen-scientists, students, and other researchers.

A role for libraries in digital data stewardship was articulated by an Association of Research Libraries (ARL) workshop report to the National Science Foundation in 2006.[3] This forecast was substantiated in

[1] http://research.microsoft.com/en-us/collaboration/fourthparadigm

[2] http://www.nsf.gov/pubs/2005/nsb0540

[3] Association of Research Libraries, To Stand The Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering: ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. 2006.

August 2010 by a survey of 57 ARL libraries, 21 of which reported they were providing infrastructure or support services for e-Science, with an additional 23 libraries reporting they were in planning stages.[4] A number of academic and research libraries are beginning to take a more active role in data management on their campuses, applying library science principles to help address the data deluge. This includes a wide range of activities such as helping researchers formulate funder-required data management plans, adapting library practice to help organize and describe research datasets, developing data collections and data repositories, taking responsibility for digital preservation, and encouraging data literacy.

Librarians are in a good position to provide these services; unfortunately, there is currently no framework in place to support the organization and discovery of data repositories. For example, many funding agencies are requiring their sponsored researchers to submit their data to repositories without giving further instructions to them, raising a host of questions, such as:

- Which repositories are appropriate for a researcher to submit his or her data to?
- How do potential users find relevant data repositories and discover datasets that meet their needs?
- How can librarians help patrons who are looking for data find and integrate these data into research, learning, or teaching?

Databib begins to address these needs for an audience of librarians, data users, data producers, publishers, and research funding agencies.

In addition to being an important reference resource for these user groups, the Databib platform goes beyond the traditional bibliography to serve and integrate bibliographic content using new technologies. One technology in particular, Linked Data[5], shows a great deal of promise for delivering a "web of data" (i.e., the Semantic Web) and giving librarians a new toolkit for describing and classifying data in a relational manner that spans institutions and industries and aids in resource discovery.

About Databib

Using the Databib website, users can search for data repositories using a basic keyword or advanced search. Searchable metadata fields include:

- Title of the data repository
- URL
- Who maintains the repository
- Brief description of the contents of the repository and its intended audience
- Who can access the repository
- Who can deposit datasets
- Licenses and how downloaded data may be reused
- Library of Congress Subject Headings
- Annotations from other users

[4] C. Soehner, C. Steeves, and J. Ward, E-science and Data Support Services: A Study of ARL Member Institutions. Association of Research Libraries, 2010.
[5] http://linkeddata.org

Users can browse data repositories alphabetically or by subject. Subject headings for each record are linked so that users can see other repositories in the same subject areas.

The system supports three classes of users: anonymous, user, and editor. *Anonymous* users can access as well as contribute, edit, and annotate Databib records. Creating an account enables a *user* to log in, track, and get credit for his or her contributions. All contributions are queued for review and approval by an *editor* before they are posted. The Databib software provides its own authentication and interfaces to support this workflow.

Databib is built on the widely used, free, and open-source "LAMP" software stack: Linux as the operating system, Apache as the web server, MySQL as the relational database, and PHP as the programming language. In addition to PHP as the server-side programming language, Databib also makes use of the jQuery Javascript library for dynamic client-side functionality, such as the integration of Library of Congress Subject Headings into automatically completed form elements.

The Databib application uses a three-tier architecture to separate sections of the application for maintainability. These sections are the user interface, the business logic, and the data access layer. For every user interface page there is a corresponding business logic module, which lives in its own area of the code repository. The business logic module makes calls to the underlying database via the data access layer, which executes database queries and then returns the data back to the user interface, which renders the data for the end-user.

Integrating Databib

The purpose of Databib is to maximize the connections that can be made between researchers and data repositories in a bibliographic context. Open data encourage sharing and making these kinds of connections. For this reason, all data associated with Databib are made available to the public domain using the Creative Commons Zero protocol.[6]

A dynamic feed of records is available from Databib by subscribing to its RSS feed. Announcements about new data repositories in Databib are also made via the @databib Twitter account.[7] Users may recommend or share a particular repository with social networks; over 300 are supported including Facebook, Twitter, Google+, and FriendFeed. Records in Databib can be dynamically generated and downloaded in RDF/XML[8] individually as well as in batch mode. Databib also supports OpenSearch, which allows users to save a query and receive results on demand.

An important goal of the Databib project was to connect research data repository records into the rich and growing ecosystem of structured semantic Linked Data available on the web. To that end, each Databib metadata element was mapped to an appropriate ontology or vocabulary term and, where feasible, values for the element were selected from an appropriate thesaurus.

[6] http://creativecommons.org/publicdomain/zero/1.0

[7] http://twitter.com/databib

[8] http://www.w3.org/TR/REC-rdf-syntax

Every approved record in Databib exposes semantic data via the RDFa[9] format embedded within the hypertext markup, a common technique for publishing Linked Data on the open web. The embedded RDFa metadata may be utilized by crawlers to improve discoverability and provide richer "snippets" in search results, or they may be harvested by aggregation services seeking to wrap Databib records in another context.

All of Databib's metadata elements are mainly using these vocabularies: Dublin Core[10], which is widely used in the Linked Data world both within and outside of the cultural heritage domain; Friend of a Friend (FOAF[11]), which is also widely used; and Databib Terms. A handful of the Databib metadata elements did not have obvious corollaries in the Linked Data world (for example, "repository type," "deposit policy," and "access status"), so we created a small vocabulary to ensure that these terms were included in the RDFa expressions of a record's metadata. The Databib Terms vocabulary is now published on the web for similar projects to use, should they have need of such terms.

We planned to link out to as many vocabularies as feasible for metadata element values once we had accumulated enough records to determine which vocabularies fit the data. Mapping subject values to Library of Congress Subject Headings was an obvious fit; however, few of the other elements seemed to fit widely used thesauri. We have recently added support to include geographic values for the "location" element. Linking out to more thesauri is an area of improvement for future development.

3. Process

As a small, short-term project, Databib loosely followed agile methodologies so as not to delay the project with formalism and undue process. Aspects of our project management methodology were brief, weekly check-ins; the production of relatively few management artifacts such as plans, charters, roadmaps, and timelines; the usage of a shared wiki for knowledge management; and frequent commits of code to the shared repository.

Development check-ins were held every Friday. While an hour was reserved on our calendars, we typically needed no more than 20-30 minutes to review progress from the past week, set goals for the next week, identify and remove blocking issues, and assess the project's overall progress. Both investigators participated in these calls with the graduate student programmer, and on occasion a member of the advisory board would participate as well. Agendas were sent out in advance of every meeting to give participants an opportunity to shape them, and minutes were generally distributed soon after each meeting to all project personnel.

Major decisions were recorded on the wiki, as was all of the project's documentation, including information on assessment and evaluation, sustainability ideas, links to related work, the project's rubric for deciding how to choose appropriate repositories for Databib, technical specifications, metadata mappings and RDFa information, and a to-do list.

[9] http://www.w3.org/TR/xhtml-rdfa-primer
[10] http://dublincore.org
[11] http://www.foaf-project.org

Funding from the Institute of Museum and Library Services (IMLS) supported principal investigator Michael Witt (5%) as project director and Mike Giarlo (5%) as technical architect for nine months as well as a graduate student programmer (50%) for four months. Cost savings were reallocated to provide stipends for the interns. Purdue University contributed the efforts of Gretchen Stephens (5%) as community bibliographer through cost-share and provided website hosting.

4. Project Results

A "soft launch" of Databib and public beta test began with a tweet on March 15, 2012, on the Twitter social network.[12] The announcement was subsequently re-tweeted and forwarded with no other publicity. The beta test concluded in June 2012.

Databib consists of 17,789 lines of code (primarily in PHP) and has been released as open source via Google Code[13] under the GNU General Public License.[14] In addition to the code, the entire content of Databib is available to the public domain in RDF/XML format.[15] Over 200 data repositories were cataloged during the project. A rubric, "Databib: Guidelines for Bibliographers," was written to help give direction for the identification and consistent description of data repositories.[16]

Evaluation and Preliminary Assessment

The proposal defined five outcomes for the grant: 1) a functional and useful Databib platform as described in the project design; 2) the original description and annotation of primary repositories of research data represented by records in Databib; 3) a rubric for evaluating new repositories for inclusion in Databib; 4) documentation and supporting activities to catalyze a community of bibliographers; and 5) this white paper. These outcomes were successfully accomplished. Progress was guided by a small panel of advisors who were invited to our monthly calls, given access to our project wiki and code, and who reviewed this white paper. Our advisors were Jenn Riley, Gail Steinhart, Ed Summers, Mary Vardigan, and Todd Vision.

For the period of the beta test between March 15 and April 30, 2012, the following usage of Databib was recorded and reported using awstats (does not include robot/spider traffic):

Unique visitors	1,746
Number of visits	3,404
Page views	34,372
Hits	56,183

[12] https://twitter.com/#!/mwittin/status/180356801374597120

[13] http://code.google.com/p/databib

[14] http://www.gnu.org/licenses/gpl.html

[15] http://databib.org/include/serializeRDFXMLAll.php

[16] http://databib.org/Databib_Rubric_Draft.pdf

Different countries	38
Average duration of visit	5 minutes, 6 seconds

At the time of this report, 1,523 Library of Congress Subject Headings had been assigned to 204 records in Databib. One hundred users had created accounts and registered as bibliographers. The Databib account on Twitter had 102 followers receiving notifications when new repositories were added. Many users reported bugs and gave feedback using a link on the Databib website.

Dissemination and Outreach

Although travel was not supported in this grant, project members took every opportunity to share Databib and foster its adoption. Notable activities include:

- 9/22/11: Purdue press release[17]; picked up by ACRL, Microformats and Semantic Web, SemanticWeb.com, Web Newswire, and Highbeam Research.
- 11/11/11: Presentation by Michael Witt to the Science Journalism Laureates[18] about Databib and the scholarship of data that included reporters from Nature, Wired, Science, NPR, and IEEE Spectrum. West Lafayette, IN.
- 3/22/12: Poster presented by Mike Giarlo and Witt at the American Society for Information Science and Technology (ASIST) Research Data Access and Preservation Summit.[19] New Orleans, LA.
- 3/23/12: Databib internships reported by the Indiana University-Indianapolis School of Library and Information Science.[20]
- 4/3/12: Poster presented for the ARL Diversity Scholars[21] visit to Purdue University. West Lafayette, IN.
- 4/18/12: Brown bag presentation by Michael Witt and library science interns to Purdue librarians and IU SLIS students. West Lafayette, IN.
- 4/24/12: Databib project re-tweeted by Todd Park, Chief Technology Officer of the President of the United States, during his #bigdata online chat session on Twitter.[22]
- 4/25/12: Databib linked from the DMPTool.[23] Databib is also linked from DMP Online, a similar tool maintained by the Digital Curation Centre in the UK.
- 6/23/12: Invitation accepted by Witt to present Databib at the American Library Association (ALA) Annual Conference by the Data Curation Interest Group of the Association of College and Research Libraries (ACRL). Anaheim, CA.
- 6/24/12: Poster accepted by Witt and interns for the ALA General Poster Session. Anaheim, CA.

[17] http://www.purdue.edu/newsroom/general/2011/110922WittDatabib.html
[18] http://www.purdue.edu/sjl/laureates/index.html
[19] http://www.slideshare.net/asist_org/databib-michael-witt-rdap12-poster
[20] http://www.slis.indiana.edu/news/story.php?story_id=2389
[21] http://www.arl.org/diversity/init
[22] https://twitter.com/todd_park
[23] http://blogs.library.ucla.edu/dmptool/2012/04/24/databib

- 7/10/12: Poster accepted by Witt and Giarlo for the 7th International Conference on Open Repositories. Edinburgh, UK.
- 10/6/12: Poster accepted by Witt for the Library Information Technology Association (LITA) National Forum. Columbus, OH.

Challenges and Opportunities

The main challenge was working within resource and time constraints to complete the project. To accomplish the proposed work in nine months with 15% total FTE and a half-time graduate student for a semester was very ambitious. We ran out of time and did not generate MARC records for Databib, for example. One lesson that we learned is that it is difficult to perform quality work on any particular task related to a project with less than two to four hours per week to dedicate to it. Some work was done quickly, more in the mode of prototyping than production. There are opportunities to improve the Databib website, rubric, cataloging, and software that we did not have time to realize. The project would also benefit greatly from support for travel to present Databib and to engage a wider audience of users and potential collaborators at conferences and professional meetings.

While it was not part of our original proposal, establishing two internships for library science graduate students at IU-Indianapolis SLIS provided the project with necessary labor, energy, and fresh perspectives. Interns cataloged the majority of the data repositories in Databib and provided valuable feedback and input into our iterative development. We also worked with Dr. Erik Mitchell at the University of Maryland's School of Information to develop an assignment for his cataloging class that yielded another 14 records for Databib. It was rewarding to give future librarians the experience of working on Databib to augment their professional training and resumes.

Another challenge was to identify appropriate vocabularies and to construct useful Linked Data from metadata in Databib. There is no clear answer to the question, "What is a data repository?", and it is difficult to define and differentiate data repositories, data archives, data portals, websites that host collections of data, web-accessible databases, data coupled with analytical tools, etc. The need for creating such an ontology is identified as future work that we intend to pursue. It was also a challenge to transform data entered by users to create linkages to other concepts and content on the web. We addressed this challenge, in one case, by importing all of the Library of Congress Subject Headings and their associated URIs into Databib and developing code that auto-completes the subject heading and stores its associated URI in the local database.

One audience of users surprised us with their enthusiasm and engagement: the managers of data repositories. Many repository managers re-tweeted the addition of their repositories to Databib, which in turn were retweeted by the users of the repository. Many also issued releases to their blogs or news outlets reporting their addition to Databib. These repositories have some of the most complete and correct records in Databib, because the records have been edited and enhanced by their respective communities. To further develop this engagement, our interns created an email template to begin proactively identifying and contacting repository managers.

There are many opportunities for future development. One key development is to connect resources related to data management planning for researchers. For example, the DMPTool[24] has been used to create more than 1,000 data plans for grant proposals. This represents a point of need for a researcher who may be wondering which data repositories may be appropriate for him or her to deposit research data. Ideally, metadata and integration could automate this process so that a researcher creating a data plan would have appropriate repositories automatically recommended based on the funder, keywords in the plan, or other contextual information.

There are also many organizations and potential collaborators in the data curation arena. In order to steer the future direction of Databib and identify opportunities for collaboration and resourcing, we have assembled an international Advisory Board. Experts from the Digital Curation Centre, DataONE, National Academy of Sciences, Dryad, Jawaharlal Nehru University, California Digital Library, SPARC, DataCite, DMPTool, re3data, Chinese Academy of Sciences, and Australian National Data Service have volunteered to serve as advisors for three-year terms. A complementary Editorial Board will be recruited to ensure the coverage and currency of content and metadata in Databib. The Editorial Board will solicit and review submissions and expand coverage of under-represented repositories (e.g., by subjects or country) to realize a global scope and impact for Databib.

In conclusion, Databib has been received enthusiastically by a broad community of people interested in research data curation. The need for the resource has been clear. During the beta test, many librarians began including Databib in their library resource guides, instruction, and outreach. Researchers have followed the @databib Twitter account from a wide variety of disciplines. The managers and user communities of some data repositories have embraced Databib as well. A full assessment and evaluation of Databib will need to be conducted after its launch and promotion in order to conclusively measure its adoption and impact.

Resources

The following resources were produced by the project:

1. Databib project website, http://databib.org (CC0)
2. Databib software source code, http://code.google.com/p/databib (GNU)
3. Databib bibliographic records, http://databib.org/include/serializeRDFXMLAll.php (CC0)
4. Databib: Guidelines for Bibliographers, http://databib.org/Databib_Rubric_Draft.pdf (CC0)
5. Databib on Twitter, http://twitter.com/databib
6. Databib Terms vocabulary, http://databib.org/ns#

[24] https://dmp.cdlib.org