Databib: Cataloging the World's Data Repositories

Michael Witt

Abstract

Databib (http://databib.org) is a curated, global, online catalog of research data repositories. Librarians and other information professionals have identified and cataloged over 500 data repositories that can be easily browsed and searched by users or integrated with other platforms or cyberinfrastructure. Databib can help researchers find appropriate repositories to deposit their data, and it gives consumers of data a tool to discover repositories of datasets that meet their research or learning needs. Users can submit new repositories to Databib, which are reviewed and curated by an international board of editors. All information from Databib has been contributed to the public domain using the Creative Commons Zero protocol. Supported machine interfaces and formats include RSS, OpenSearch, RDF/XML, Linked Data (RDFa), and social networks such as Twitter, Facebook, and Google+.

Introduction

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 the United States 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 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 (http://databib.org) begins to address these needs for an audience of librarians, data users, data producers, publishers, and research funding agencies.

Databib: A Tool and Resource for Locating Repositories of Research Data

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.

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

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.

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.

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.

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[12] 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 has been recruited to ensure the coverage and currency of content and metadata in Databib. The Editorial Board solicits and reviews submissions and expands 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 the tool as well. As outreach is conducted to new regions and user communities, Databib seeks to achieve its goal of providing a comprehensive catalog of the world's data repositories that is fully integrated into the research data lifecycle.

Acknowledgements

Initial support for Databib was provided by a National Leadership Sparks Innovation Grant (LG-46-11-0091-11) from the Institute of Museum and Library Services (IMLS) and was developed in collaboration with Purdue University Libraries and Penn State University. This paper was adapted from an unpublished white paper submitted to the IMLS in July 2012. Additional support for travel to India and accommodations was provided by the Purdue University Office of Engagement in collaboration with Jawaharlal Nehru University and INFLIBNET.

References

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.

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

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

7. http://twitter.com/databib

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

9.http://www.w3.org/TR/xhtml-rdfa-primer

10.http://dublincore.org

11.http://www.foaf-project.org

12.https://dmp.cdlib.org

About Author

Mr. Michael Witt, Assistant Professor of Library Science, Purdue University
West Lafayette, Indiana, USA
E-mail:mwitt@purdue.edu