Libraries and Linked Open Data

Paul J. Albert[1], Jon Corson-Rikert[2], Melissa A. Haendel[3], and Kristi L. Holmes[4]

[1]Weill Cornell Medical College, Samuel J. Wood Library; [2]Cornell University, Albert R. Mann Library; [3]Oregon Health & Science University, Library; [4]Washington University in St. Louis, Bernard Becker Medical Library

Libraries play increasingly important roles in many of the Linked Open Data and other Semantic Web projects, many of which serve libraries' core mission and service areas. These activities leverage library strengths and expertise in areas such as: data structure and sources, ontology and controlled vocabulary expertise, education and outreach, and programming and technical support.

With the recent emphasis on interdisciplinary science, software applications for research networking and discovery have also become more important. These tools use web-based standards for representing metadata about scholars' activities and resources[i]. Applications that use open, machine-readable descriptions of data improve the efficiency of this process and are instrumental in supporting the development of Open Science.

A linked data infrastructure allows for: enhanced discovery, visualization, integration between research and clinical data, and deep faceted searching across multiple types and sources of data. Robust data and smart web applications can support inter-institutional partnerships and enhance opportunities for data reuse. By breaking data out of single-purpose data silos, research networking platforms can promote a network effect: the value of the network increases with the amount of linked data available and applications available to consume that data.

To better understand this emerging discipline, we highlight below the work at three library-based programs from Weill Cornell Medical College, Cornell University, and Oregon Health Sciences University. Topics discussed include a description of the product or service and how librarians get involved, a description of library-based outreach activities, and a discussion of experiences or skills necessary for libraries to support these efforts at their institutions.

Paul Albert, Assistant Director for Research and Digital Services at Weill Cornell Medical College Library

1. Describe the product or service that you provide in the role that you play at your library?

VIVO[ii] is an institutionally-driven semantic application for representing authoritative data about researchers. In my capacity as project manager, I work 50% of my time on VIVO. This role includes outreach, defining and negotiating requirements, meeting with stakeholders, running meetings, and modeling data. At Weill Cornell Medical College, VIVO is a partnership between the library and

Information Technologies and Services. This collaboration has produced a productive and sincere dialogue about the competing needs of user versus what is (and ought to be) possible with the technology.

2. How can other librarians use this product or service?

Developers can download a virtual machine to play around with the application[iii]. You can also search any of the individual VIVO sites. Or, any of the growing number of sites such as VIVOsearch[iv] that aggregate VIVO data across sites.

3. How has your library reached out to your institutional community and how have you earned support for this particular service?

VIVO is not yet in production beyond the Weill Cornell campus. Our roll-out is scheduled for January 1, 2014. Thus far, our outreach has consisted of a small pilot of 12 faculty and one-on-one interviews with faculty. We plan to have a road show about VIVO in the fall.

For us, the way to earn support from our institutional community has less to do with presentations and pamphlets, and more to do with providing services and tools of unquestionable value. So which fundamental questions can we solve for our faculty? Our short list includes: an accurate and complete list of publications, a polished web presence, inferences regarding expertise, and an ability to easily make changes to data from authoritative sources. Other stakeholders include administrators and research support. For them, we can provide reports such that five people across campus are not, to varying degrees of success, looking up the publication history for a given person.

4. What skills or experience do you think librarians need to acquire to meet the needs of escience and data management and can you provide examples of the skills and services that you or your other staffs have in this particular area?

For me, this question speaks to how libraries and librarians remain valuable partners in the business of getting science done. My idealized vision for librarians is that they are proactive, not reactive. They realize their field is in flux, but they remain a bridge between humans and metadata. As a sort of technological intermediary, they have well-informed opinions – taste, if you want to call it that – about metadata. They know where it lives, how to access it, how to structure, store it, annotate it, manage it, and derive meaning from it.

Which grant opportunities are relevant to which researchers? Who has or should work with whom? Which languages do your faculty speak? What is their global health expertise? Many of these are questions that can be answered by asking faculty directly. But what if they don't answer or their answers are wrong or incomplete? With a modicum of resourcefulness and occasionally a bit of programmer time, librarians should be well-equipped to take authoritative data and make inferences about people.

A key component of reporting is data visualization. With a modicum of HTML experience, you can produce really nice visualizations using d3[v]. The Science of Science (Sci2) Tool Sci2[vi]doesn't require any HTML background and can be used for network visualizations and burst analyses.

Finally, the recent wave in transparency, accountability, and demonstrating value makes maintaining a well-functioning institutional repository eminent sense.

Jon Corson-Rikert, Head of Information Technology Services at Mann Library, Cornell University

1. Describe the product or service that you provide in the role that you play at your library?

I work with VIVO at Cornell University[vii], where we support VIVO as a local tool as well as contribute to the national/international community through development, implementation support, ontology extensions, and community activities such as weekly calls. As the head of an IT group within the Cornell University Library I see this as part of our library's connection with our faculty and researchers as well as a way to connect our researchers with new developments in their disciplines at institutions beyond Cornell.

2. How can other librarians use this product or service?

VIVO is an open-source ontology, software suite, and community with many opportunities for engagement through weekly calls, webinars, regional meetings, an annual implementation fest, and an annual Conference. Further information is available at http://vivoweb.org.

3. How has your library reached out to your institutional community and how have you earned support for this particular service?

We have worked very closely with schools and colleges at Cornell to improve the quality of data they gather about their faculty, especially their publications and other research activities, and to integrate information about their academic and research activities into VIVO as a university-wide, cross-disciplinary platform.

We are also connecting VIVO to our library online discovery environment to make faculty and researcher expertise more visible to students and other faculty looking to expand their knowledge through online searches. By converting our library catalog's MARC metadata to RDF we can make explicit connections between the works of current researchers and their other academic and research activities, as well as highlight Cornell authors in our library discovery system. We believe this enhanced catalog will help expose the vitality of the Cornell community and bring more awareness of current research to the forefront.

4. What skills or experience do you think librarians need to acquire to meet the needs of escience and data management and can you provide examples of the skills and services that you or your other staffs have in this particular area?

At Cornell the best example of new library activities in eScience and data management is the Research Data Management Service Group (RDMSG). The RDMSG[viii] is a collaborative, campus-wide organization linking faculty, staff, and students with data management services to meet their research needs, from specialized data management services to the development of data management plans. Services may be requests at any stage of the research process, from initial exploration to grant preparation, data gathering, analysis, and longer-term preservation and access.

The RDMSG is jointly sponsored by the Senior Vice Provost for Research and the University Librarian, and operates with representatives from the library, the Center for Advanced Computing, the Cornell Institute for Social and Economic Research, Cornell Information Technologies, and the Weill Cornell Medical College.

Melissa Haendel, Lead Ontologist, Ontology Development Group at OHSU Library; Assistant Professor, Department of Medical Informatics & Clinical Epidemiology, OHSU

1. Describe the product or service that you provide in the role that you play at your library?

My department (Ontology Development Group) generally speaking, provides data classification and management strategies. We also develop ontologies for the eagle-i application[ix] (and now VIVO), and support data collection and development of this software tool for the inventorying of research resources. Finally, we work with numerous other community members to support interoperability and data standards amongst different applications and data sources, with a particular aim towards making data available on the Semantic Web as Linked Data.

2. How can other librarians use this product or service?

eagle-i is an open-source application and is usually installed on a per-institution basis. Most of the ontologies that we work on are indexed by the OBO Foundry[x]. Data standards are available at BioSharing[xi]. Both of these sources can be used to help researchers provide quality metadata on their data for publication, either in conjunction with a journal, in a new data journal (such as the new Nature journal), in institutional repositories, or community repositories such as Dryad[xii].

3. How has your library reached out to your institutional community and how have you earned support for this particular service?

This is an area with which we struggle. Whilst we are funded by national grants and have many collaborators around the world, we do less well locally. We are occasionally approached by local researchers and administrative staff to help with their data classification needs (such as the building of a new Clinical and Translational Activity Repository for quality research profiling) or to aid in search for biospecimens in the local biobank. That said, we don’t currently have data management, data publication, or data classification services listed on the library website yet, but we are working on this. We do participate in local information science training venues to help promote adequate instruction in this area.

We recently won the 1K challenge at the Beyond-the-PDF Conference[xiii] on a project that will aim to identify researcher needs with respect to data handling. We anticipate the results of this project to better inform both our library as well as others in defining adequate service and training for libraries to help with these efforts[xiv].

4. What skills or experience do you think librarians need to acquire to meet the needs of escience and data management and can you provide examples of the skills and services that you or your other staffs have in this particular area?

My team is very diverse, and I believe that dealing with data requires a certain degree of diversity. We have had PhD computer scientist, a Masters in Engineering, three PhD scientists, and a Masters in Library Science on our team. Skills that are needed and that I don't often hear about much from traditional librarian training are:

- *General software development practices, such as user-centered design, version control, and agile development*. These concepts are just as important when working with a client to help them with their data, even if you aren't building any software.
- *Semantic Logic*. Librarians are traditionally trained to use vocabularies to tag documents. If this capability could be expanded to learn more about the use of logic to build and leverage more sophisticated vocabularies and inferencing, we could help support getting more out of our data.
- *Scientific background*. A lot of time is spent understanding scientific researchers' needs (we are a health university), and having a firm grasp on the scientific method is important to build systems or manage data that meets their need.
- *Communication skills*. Data management and classification services are iterative and require drawing the requirements out of the end users. Good communication and documentation strategies are critical.
- *Scripting*. A big part of data management is turning the data from one form into another, lexical matching, automatic annotation, etc. Basic scripting skills can be a great help here.

Sources cited:

- https://www.ctsacentral.org/best%20practices/research%20networking
- http://vivoweb.org/
- https://wiki.duraspace.org/display/VIVO/Virtual+Appliances
- http://beta.vivosearch.org/
- http://d3js.org/
- https://sci2.cns.iu.edu/user/index.php
- http://vivo.cornell.edu/
- http://data.research.cornell.edu
- https://www.eagle-i.net/
- http://obofoundry.org
- http://biosharing.org/
- http://datadryad.org/
- http://www.force11.org/beyondthepdf2
- http://www.force11.org/1Kchallenge#1k2