Ellen Phillips

Info 241-12

Technology topic report: Linked open data and institutional repositories

November 23, 2015

In a collection development environment characterized by standardized database offerings, the unique resources produced by researchers and faculty members at colleges and universities are likely to gain in prominence. However, success hinges on how easily items in institutional repositories (IR) can be found by users and on how closely associated those items can be to the institution, funding bodies, and other stakeholders that produced them. This is particularly important for securing grants, as many countries and non-profit foundations now require free dissemination of data and knowledge gleaned through publically funded research. While search engine discovery is an important part of IR strategy, information is not automatically integrated across any systems within the larger university landscape and can therefore be difficult to quantify.

The major IR solutions such as open source dSpace and bepress' Digital Commons are designed around flat bibliographic standards and the information is stored in relational databases (RDB), (Latif, Borst & Tochtermann, 2014). Although both contain elements from the Dublin Core, due to their respective extensions, they have disparate metadata fields. Furthermore record quality is degraded through the practice of self-archiving in the IR and the use of author-based article posting systems such as those hosted by the Social Sciences Research Network (SSRN) and ResearchGate (RN). Private systems like these rely on various controlled vocabularies, or none at all, to organize their contents. Differences in record quality were noted in two separate studies; one a large review of records from different types of repositories (Palavitsinis, Manouselis & Sanchez-Alonso, 2014) and the other a smaller analysis of sixty records in three IR, (Kurtz, 2010). Both found inconsistent metadata, incorrect use of fields, duplication and ambiguous date data in records that were uploaded by the authors themselves rather than librarians.

There are multiple compelling reasons to collect, analyze, and make available the research data produced by an institution of higher learning. There is an identified need for researchers and faculty members to publicize and measure the impact of their scholarly work in order to advance in their field. Herreid, Prud'homme-Généreux, Schiller, Herreid, and Wright (2015) reported on survey data that showed that 80% of faculty members at four-year colleges and universities felt that producing original research was important for promotion and tenure. Many scholars also feel that knowledge is a public good and that sharing it moves society forward. Other needs involve proposed legislation such as the Fair Access to Science and Technology Research Act (Electronic Freedom Foundation, 2015) as well as numerous policies from individual institutions that mandate open access reporting of research findings (Welcome to ROARMAP, 2015).

The convergence of big data with the notion that scholarly activities should be graphed using a computational business model has created a market for software solutions that can do this using linked open data (LOD). Research Information Management (RIM) systems, also sometimes called Current Research Information Systems (CRIS), have been developed to fit these diverse needs, aiming to coordinate data while reducing workload with the intended outcome of presenting a clear snapshot of the institution's scholarly activities as a whole (Dempsey, 2014). While raising the research profile can translate into more grant awards, it is also clear from the marketing collateral from the various vendors that these systems have administrative functions related to Human Resources and accreditation reporting ("Streamline internal processes," 2015; "Sail through accreditation," 2015).

Some universities such as Columbia, SUNY Stony Brook, Stanford University, and Cornell, have created their own systems. With the exception Stony Brook, all make use of LOD. Most of the vendor solutions also run on LOD. The major ones include Sympletic's Elements, Elsevier's Pure, and Thomson Reuter's Converis. Pure and Converis are part of a larger suite of products that are well integrated with the database offerings of the major publishers that own them. Symplectic is an incubator project of Macmillan Publishers, a collection of companies under the name Digital Science (Dempsey, 2014).

Mitchell (2013) states that LOD data is often modeled in resource description framework (RDF), (p.13). This means it will be necessary to either migrate or map data stored in RDB to one based on RDF, also called a triple-store. Triples form the basis of linked open data, as information is stored in uniform resource identifiers (URI) as object-predicate-subject statements. This is an important goal to pursue not only in light of the capacity of enterprise-wide big data, but also to contribute to the 4.7 billion RDF triples that were on the web of data as of 2009 (Bizer, Heath, Berners-Lee, n.d.).

Konstantinou et al., (2014), identified several methods to use data from an RDB as RDF triples (p.836). One solution is to extract the data migrating it completely to a new format without maintaining any links (p.836). The other enables the two to be aligned and links between them are maintained (p.836). OpenLink Virtuoso RDF Views is middleware that enables the instantaneous “mapping [of] arbitrary collections of relational tables into SPARQL accessible RDF” (Erling, n.d.). Konstaninou et al. state that open source solution D2RQ makes use of a “table-to-class and column-to-predicate approach” (p.836) to automate the creation of read only mapping files. Triplify is open source, runs on top of an RDB and according to their organization’s web site, is a “small plugin for Web applications,” consisting of several files with less than 500 lines each (“Overview, introduction and news,” 2015). The second approach maintains the two sources side by side and Konstaninou et al. make the point that “ontology mapping and alignment is an ever-changing domain” (p.840).

By using systems that leverage big data against flexible and open architecture based on the principles of LOD, it will be possible to disseminate and track the impact of research as never before. While these activities may look different across the breadth of disciplines, metrics from these scholarly undertakings are critical in order to produce quantified analyses of scholarly work in ways that are meaningful to different stakeholders.

References

Bizer, C., Heath, T., Berners-Lee, T. (n.d.) Linked data – the story so far. (preprint version) Retrieved from http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Electronic Freedom Foundation, (October 20 2015). FASTR Ensures that Publicly Funded Research Belongs to the Public. Retrieved from https://www.eff.org/deeplinks/2015/10/fastr-ensures-publicly-funded-research-belongs-public

Dempsey, L. (2014, October 26). Research information management systems - a new service category? [Web log post]. Retrieved from http://orweblog.oclc.org/archives/002218.html

Erling, O. (n.d.), Declaring RDF Views of SQL Data, retrieved from http://www.w3.org/2007/03/RdfRDB/papers/erling.html

Herreid, C. F., Prud'homme-Généreux, A., Schiller, N. A., Herreid, K. F., & Wright, C. (2015). A peek behind the curtain of tenure and promotion. *Journal Of College Science Teaching*, *45*(1), 61-65.

Konstaninou, N., Spanos, D., Houssis, N. & Mitrou, N. (2014). Exposing scholarly information as linked open data: RDFizing DSpace contents. *The Electronic Library, 32,* 834-851. DOI: dx.doi.org/10.1108/EL-12-2012-0156

Kurtz, M. (2010). Dublin Core, DSpace, and a Brief Analysis of Three University Repositories. *Information Technology and Libraries 29*, 40-46. DOI: http://dx.doi.org/10.6017/ital.v29i1.3157

Latif, A., Borst, T., & Tochtermann, K., (2014). Exposing Data From an Open Access Repository for Economics As Linked Data. *D-Lib Magazine, 20*(9/10). doi:10.1045/september2014-latif

Mitchell, E. (2013). Building blocks of linked open data. *Library Technology Reports, 49*, 11-25.

Overview, introduction and news. (2015). Retrieved November 21, 2015 from http://triplify.org/Overview

Palavitsinis, N., Manouselis, N., Sanchez-Alonso, S. (2014), Metadata Quality in Digital Repositories: Empirical Results from the Cross-Domain Transfer of a Quality Assurance Process. *Journal Of The Association For Information Science And Technology, 65,* 1202–1216.

Sail through accreditation. (2015). Retrieved November 23, 2015 from http://www.digitalmeasures.com/activity-insight/benefits/accreditation.html

Streamline internal processes. (2015). Retrieved November 22, 2015 from http://www.digitalmeasures.com/activity-insight/benefits/personnel-reviews.html

Welcome to ROARMAP. (2015). Retrieved November 21, 2015 from http://roarmap.eprints.org/

Table of Research Information (RIM) Systems

These systems were custom built to meet the needs of the university that created them.	
Columbia University Scholarly Profiles (CUSP)	http://irvinginstitute.columbia.edu/cusp/cgi-bin/ww2ui.cgi/splash

Cornell University (VIVO)	http://vivo.cornell.edu/, http://www.vivoweb.org/about
Stanford University Community Academic Profiles (CAP)	https://cap.stanford.edu/
SUNY Stony Brook (Faculty Profiles)	https://it.stonybrook.edu/services/faculty-profiles
University of Texas at Arlington (Collaborative Profile Partnership)	http://www.uta.edu/research/collaborate/restricted/onestep.php
These systems were built by for-profit companies:	
Digital Measures	http://www.digitalmeasures.com/
Elsevier Pure	https://www.elsevier.com/solutions/pure
Symplectic Elements	http://symplectic.co.uk/products/elements/
Thomson Reuters Converis	http://converis.thomsonreuters.com