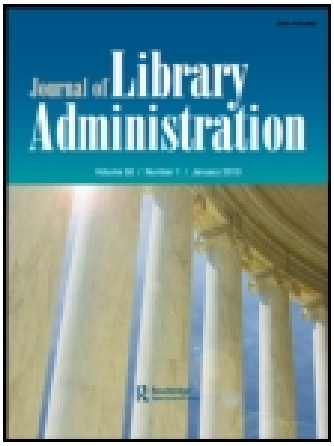

Journal of Library Administration
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/wjla20

Demonstrating Library Value at Network Scale: Leveraging the Semantic Web With New Knowledge Work

Kenning Arlitsch[a], Patrick Obrien[b], Jason A. Clark[c], Scott W. H. Young[d] & Doralyn Rossmann[e]

[a] Dean of the Library, Montana State University, Bozeman, MT, USA

[b] Semantic Web Research Director, Montana State University, Bozeman, MT, USA

[c] Head of Library Informatics and Computing, Montana State University, Bozeman, MT, USA

[d] Digital Initiatives Librarian, Montana State University, Bozeman, MT, USA

[e] Head of Collection Development, Montana State University, Bozeman, MT, USA

Published online: 30 Aug 2014.

To cite this article: Kenning Arlitsch, Patrick Obrien, Jason A. Clark, Scott W. H. Young & Doralyn Rossmann (2014) Demonstrating Library Value at Network Scale: Leveraging the Semantic Web With New Knowledge Work, Journal of Library Administration, 54:5, 413-425

To link to this article: http://dx.doi.org/10.1080/01930826.2014.946778

Journal of Library Administration, 54:413–425, 2014
Published with license by Taylor & Francis
ISSN: 0193-0826 print / 1540-3564 online
DOI: 10.1080/01930826.2014.946778

posIT

KENNING ARLITSCH, Column Editor
Dean of the Library, Montana State University, Bozeman, MT, USA

Column Editor's Note. *This JLA column posits that academic libraries and their services are dominated by information technologies, and that the success of librarians and professional staff is contingent on their ability to thrive in this technology-rich environment. The column will appear in odd-numbered issues of the journal, and will delve into all aspects of library-related information technologies and knowledge management used to connect users to information resources, including data preparation, discovery, delivery and preservation. Prospective authors are invited to submit articles for this column to the editor at kenning.arlitsch@montana.edu.*

DEMONSTRATING LIBRARY VALUE AT NETWORK SCALE: LEVERAGING THE SEMANTIC WEB WITH NEW KNOWLEDGE WORK

KENNING ARLITSCH
Dean of the Library, Montana State University, Bozeman, MT, USA

PATRICK OBRIEN
Semantic Web Research Director, Montana State University, Bozeman, MT, USA

JASON A. CLARK
Head of Library Informatics and Computing, Montana State University, Bozeman, MT, USA

SCOTT W. H. YOUNG
Digital Initiatives Librarian, Montana State University, Bozeman, MT, USA

DORALYN ROSSMANN
Head of Collection Development, Montana State University, Bozeman, MT, USA

Address correspondence to Kenning Arlitsch, Dean of the Library, Montana State University, P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: kenning.arlitsch@montana.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/wjla.

INTRODUCTION

The Internet has been redefining library work for two decades, but the development of the Semantic Web is driving new requirements for data preparation. Metadata for human consumption is becoming less relevant as the Internet of Things allows machines to converse with each other and to consume and comprehend data directly. Americans submit nearly 20 billion queries each month to commercial Internet search engines (comScore, 2014), but few libraries optimize their data for the machines that have become the center of the discovery process. Libraries have also been slow to engage in new methods that authoritatively establish their semantic identities for search engines. Library content and services are more effectively discovered and used after these identities are established and optimized according to evolving Semantic Web protocols and services.

It's not only libraries that face this challenge. Every organization in an academic institution benefits from visibility that drives the use of its products and services, and most of these organizations struggle with piecemeal efforts that are often ineffective in the long term and don't benefit the institution at large. Central IT organizations rarely have the resources or knowledge to successfully and consistently deploy optimization services to campus organizations to ensure their visibility on the Web, which in turn leads some organizations to contract with students or other Web site service providers. Only sometimes do those consultants understand the complex array of elements that comprise a successful Web presence; in many cases the efforts prove unsustainable when the consultant moves on or the funding expires.

Opportunity accompanies change. Librarians may enjoy new roles as trusted facilitators who can develop effective and replicable optimization services by delivering measurable value based on metrics that matter to each organization's leadership. The Montana State University (MSU) Library is engaged in Semantic Web research on several fronts, which we will describe in this article. Our concept of "new knowledge work" encompasses the discoverability, accessibility, and usability of content and services in the Semantic Web. In this article, we survey the following new services that libraries can offer their users and campus partners to aid discovery and understanding of resources at the network scale:

1. Establishing semantic identity for content and entities.
2. Structuring metadata for machine ingest and leveraging external search mechanisms.
3. Centralizing management of faculty activity data for efficient population of Institutional Repository (IR) and other reporting outlets.
4. Developing programmatic social media strategies to connect communities and content.

5. Advancing the role of the library as publisher to include the creation of open extensible book software and reading interfaces that participate in the web of linked data.

ESTABLISHING SEMANTIC IDENTITY

Search engines face the nearly impossible task of fulfilling search requests, a task that may be likened to a stranger approaching another person, uttering one or two words and then waiting for a response. The difficulty is not only the dearth of information in the query, but also the ambiguity of language. Does the search for "mustang," for instance, refer to a car, a horse, a sports team, a brand of shoes, or a pair of jeans? It is difficult for search engines to deliver meaningful results from indexes comprising the "flat" or one-dimensional metadata that libraries have traditionally produced.

The Semantic Web represents the nexus of technological development and structured data, offering added dimensionality to Web content for improved search capability. It aims to reduce the ambiguity of language and provides the structure for delivering richer, more accurate, and more relevant information to the user in search results. Context can come in the form of metadata "triples" of subject, predicate, and object. Metadata presented this way establishes relationships and helps search engines respond more accurately and precisely to queries. Anything that can be assigned a Uniform Resource Identifier (URI) can be described and "understood" by the machines that are designed to improve our discovery and access to the information we value.

Search engines rely heavily on the Linked Open Data (LOD) cloud to seed their understanding of knowledge and concepts. DBpedia, considered the center of the LOD universe, is a trusted knowledge base extracted from structured data in Wikipedia. Before a search engine can refer users to the entity we know as a "library," the search engine must understand the concepts that define a library. This requires considerable computational power to identify the different words that may refer to the same concept or "Thing." For example, the Renne Library is an alternate name for the Montana State University Library in Bozeman, Montana. A human can easily interpret that the two are the same, however, accurate interpretation by search engines is much more difficult and requires a great deal of data. Search engines populate their Semantic Web indexes by relying on a variety of trusted linked data sources, such as Wikipedia, DBpedia, Google MyBusiness, Google+, Freebase, and webmaster-provided Schema.org markup. Additional work to disambiguate people for machine understanding is another facet of this activity and initiatives such as VIAF, ORCID, and ISNI are striving to solve these semantic problems in LOD settings. Ensuring these LOD sources are accurate, robust, consistent and connected is critical for search engines to understand

that libraries and faculty researchers possess qualities and connections that are valuable and relevant to human users.

The library community has neglected the act of defining the library as a Semantic Web entity in Wikipedia and DBpedia. Within the Semantic Web and machine-learning context, existing library concepts are limited and poorly describe the services and resources a library provides. Neither DBpedia nor Schema.org make it easy to describe library concepts such as an IR, data management, education services, or special collections to a machine. This limited conceptual framework in turn affects search engines' abilities to direct users and has wide implications for how libraries are used, promoted, and valued. Improving this framework requires a concerted effort by the library community to improve library-related semantic concepts, as well as librarians rethinking their relationships with Wikipedia.

Google introduced its Knowledge Graph in 2012, which leverages the Semantic Web to change the way search is conducted (Singhal, 2012). Its aim is to shift from the traditional practice of matching character strings to queries, and instead to match defined entities and facts about them. The Knowledge Graph display is a newer element that is currently positioned to the right of, or above, Google search results. It does not displace the traditional list of search results but rather supplements them with additional graphical and textual information relevant to the search. The type of information presented in the Knowledge Graph display varies and can be based on popular searches of that item. For restaurants, an address, telephone number, link to a menu, and diner reviews may be deemed the most useful information to display. For historical figures it may be biographical information and accomplishments. Profiles of athletes may include statistics about their play, career highlights, images, and teams they played for.

In 2012, a Google search for "Montana State University Library" revealed a surprising entry in Google's Knowledge Graph display. Instead of displaying the flagship library of the Montana State University (MSU) system, located in Bozeman, the Knowledge Graph display showed another MSU campus in Billings, MT (see Figure 1). From the perspective of the Google search engine, the MSU library in Bozeman simply was a text string and did not exist as a "Thing." The information being ingested by the Knowledge Graph led Google to incorrectly conclude the MSU library was a building in Billings, Montana.

There were several reasons why the MSU Library in Bozeman was misidentified in Google's Knowledge Graph display. We believe the two most important reasons were: (a) no one had claimed, verified, or corrected facts about the MSU Library in Google's Knowledge Graph data feeds; and (b) the MSU Library did not exist as a Thing of interest because no Wikipedia Article about the MSU Library had been written.

Librarians have spent years dismissing Wikipedia as an unreliable source of information that students shouldn't trust. A Google search for the phrase

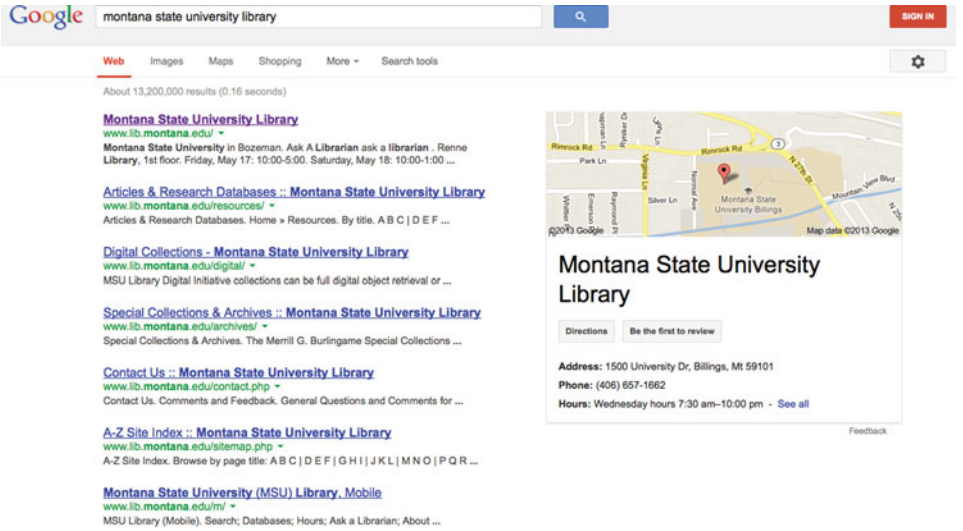

FIGURE 1 Google search results for "montana state university library" in 2012.

"Wikipedia authoritative source" reveals dozens of Web sites and guides created by librarians advocating avoidance of the online encyclopedia. But while the library community focused on the limits of Wikipedia as an information source for research, it became a trusted resource for providing structured data to search engines. The lack of a formal Wikipedia Article was a strong signal to Google and other search engines that the string of text "Montana State University Library" was not important as a Thing.

The fallout from this lack of engagement extends further. Currently "institutional repositories" do not exist as a Thing in Google's Knowledge Graph. While sites like the Registry of Open Access Repositories (ROAR) have manually cataloged a list of IR, and Wikipedia structured data has a "topical concept" description of IR, these are just text strings to search engines. A robust machine-understandable IR concept does not exist because the library community has not defined the structured data necessary within Wikipedia, Freebase, or Schema.org.

Establishing semantic identity for concepts and entities matters because it helps search engines understand and trust them, which in turn will increase traffic to those sites. The strategy of establishing and maintaining semantic identity (see Figure 2) is an example of operating at network scale.

SEMANTIC STRUCTURED DATA FIRST

New knowledge work in the Semantic Web also affects the traditional development of library digital collections, and here again, operating at network

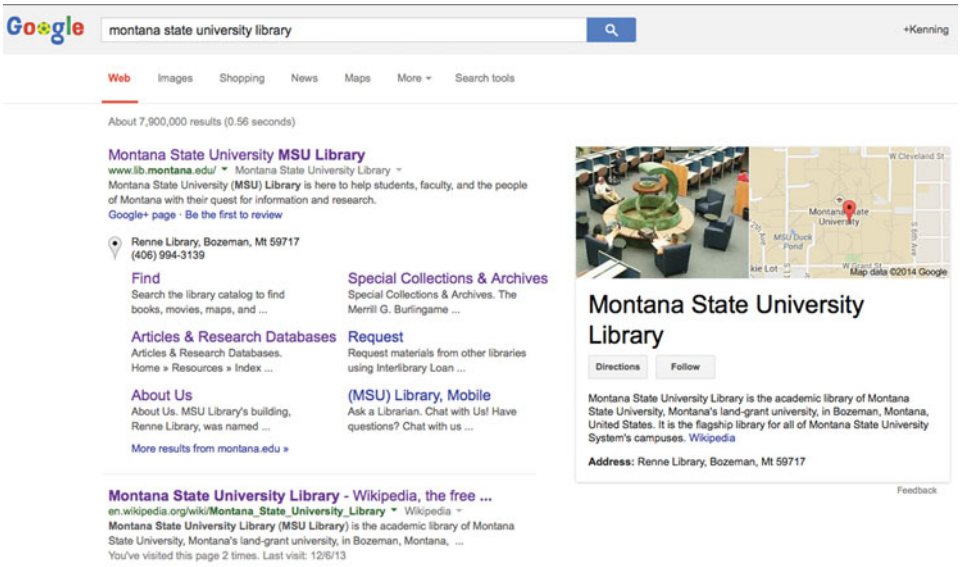

FIGURE 2 Google search results for "montana state university library" in 2013 after semantic identity was established.

scale can have a dramatic impact. Processes and routines for building the digital library have typically centered around tools like CONTENTdm and Solr/Blacklight to build local search indexes and user interfaces for digital content. Commercial search engines, however, have developed the technological standards that users depend on to discover and access information. Spending time and resources developing a competing discovery layer may be a poor use of library resources when patrons clearly prefer to access powerful, freely available commercial search engines such as Google, Bing, and Yahoo. Mapping broad machine-readable concepts, creating semantic HTML markup using the Schema.org vocabulary, and encoding resources in a version of the Resource Description Framwework (RDF) are the primary work of the semantic digital library development model. This work emphasizes supporting or building a digital library architecture whose primary function is to create indexable content and improve its discoverability, access, and visibility. Under this model, the digital library application has the following primary architecture:

1. An "about" Web page (i.e., schema.org/CollectionPage) with rich, structured metadata identifying the title, the major linked data topics, and primary concepts of the collection.
2. A database that can produce item level views (i.e., schema.org/ItemPage) of each record as Web pages with rich, structured metadata.
3. A sitemap XML file listing all files to index and providing indexing directives for search engines.

This architecture does not include searching and browsing as primary components. The point of this omission is to focus digital library development on improving indexing by the major search engines through quality metadata that describes the content and provides context. This metadata and indexing work is the means by which libraries can create network scale around the discovery of our content and materials. Among the benefits of this "semantic structured data first" methodology for digital library applications are:

- An efficient alignment of digital library personnel around the work of building interoperable and indexable structured data markup with Schema.org and RDFa
- A lowering of a significant barrier to creating a searchable collection index and the efficiencies gained in optimizing digital collections data for commercial search engines
- An investment in clean metadata that is optimized to work across digital library platforms and can move as systems change over time.

Within the "semantic structured data first" digital library model, searching and browsing are still a user interface (UI) feature. The difference is that these functions are externalized as a low cost, or free, service provided by search engines. In our application of this model, we are using the Google Custom Search Application Programming Interface (API) and Web service to power our local collections search (arc.lib.montana.edu/digital-collections/). This new model introduces efficiencies in our practices around local search development and refocuses digital library work on creating indexable content. It also exposes our content as open, semantic data that can be accessed and used by machines to enhance knowledge discovery and presentation in ways that are still evolving. Most importantly, the practice of "semantic structured data first" focuses resources on a librarian core competency: improving data quality to enable the discovery and use of information resources.

CENTRALIZING FACULTY ACTIVITY DATA

Conversations around IR use and usefulness often focus on filling the repository with faculty publications and related material. Much of this material already exists, yet current processes for ingesting it into the IR require the time and effort of a busy faculty researcher who is already being asked to submit similar information for grant submission and administration, faculty activity reporting, tenure and promotion processes, etc. Library approaches to collecting IR submissions usually involve an inefficient and time-consuming process of pursuing faculty for their curriculum vitae and PDF versions

of pre- or post-print articles, clearing copyright, and manually re-keying metadata.

The MSU Library is implementing a new strategy that focuses on the library as an information hub that facilitates automated data input to populate network scale services that are used to discover academic organizations, faculty researchers, and research output. Importing data from Google Scholar, faculty activity reports, and other commercial and LOD sources is key to an efficient strategy that also minimizes errors introduced by re-keying.

Our team took advantage of the IR's potential strategic value being recognized outside the library to collaborate with university administration efforts toward collecting intellectual output data in the form of publication citations. During the time our library staff collected a few hundred publications from 20 to 30 faculty researchers through traditional methods the university administration provided us with over 11,000 publication records from more than 600 faculty exported from the faculty activity database. Another benefit of working with the university administration was that our library-maintained data set includes the inherent relationships between authors, their organizational affiliations, and research activities that make a few innovations possible using traditional search engine optimization (SEO) and the Semantic Web.

Most university organizations have well-established Web sites that search engines trust and rank highly. However, these Web sites typically lack the organizations' most valuable content—full text intellectual output that search engines can easily index. As the university's research support hub, the library is uniquely suited to improving the university's network value through search engine positioning by using the IR to connect each university Web site to its intellectual output. This effort will produce network effects that benefit the entire university. We believe organizing IR content around subject areas relevant to college deans, department heads, faculty and researchers makes implementing traditional SEO much easier. For example, asking the university's computer science department to provide a link from their Web site to the section of the IR containing their "computer science" publications is easy and has benefits. The link-back provides a very strong signal to search engines that (a) the content is about "computer science" and (b) a trusted authority, the computer science department, thinks this IR Web page is a very good resource for anyone seeking information about computer science related topics.

On the Semantic Web side our research team is taking advantage of the library-maintained data set to prototype new knowledge work that extends beyond the traditional boundaries of the library and provides the following benefits:

1. Analysis to help university administrators, college deans, and department heads with hiring, budgetary, and space allocation decisions based upon

the scope and depth of content subject areas of their organization's intellectual output.
2. Help existing, and potential, faculty or graduate students illustrate patterns that facilitate cross-domain research collaboration valued by administrators, but are very difficult to foster without face to face contact or in depth knowledge of research practices.
3. Systematically improve library staff efficiency and semantic data quality while populating and maintaining the scholarly output published in the university's IR.
4. Evaluate IR value based on real data about research activities.

SOCIAL MEDIA COORDINATION & AMPLIFICATION

Social media (SM) provides new opportunities for libraries to connect members of the research community, to be involved in conversations among researchers, to assist faculty with establishing authorship identity and surfacing of the university's intellectual output, and to make valuable contributions to scholarship by engaging users at the network scale. While researchers may have interacted in the past with librarians and libraries in the form of face-to-face interactions or downloading subscribed articles, SM allows libraries to engage users in new points of conversation that include disciplinary peers. The SM discussion with researchers may involve promoting connections with other researchers, educating about how to optimize various SM for discoverability and shareability, explaining the benefits of establishing a single semantic identity for authorship attribution, or introducing resources and services that may be outside of what others might expect of libraries.

Much of the library literature surrounding SM suggests that they are venues for marketing the library in order to prove the library's value. Libraries should consider reframing this conversation to a broader concept of SM as one of many places to be active members of a research community; increased use of library resources follows as a result of libraries being actively engaged in the community.

Through social media optimization (SMO) and implementation of concerted and consistent metadata, libraries offer new opportunities to researchers by surfacing and bringing together information via social networking systems (SNS). To start the conversation, librarians can educate researchers that the SM they choose may influence further discoverability. Some SM such as Twitter, Google+, and Tumblr allow their content to be indexed by search engines while Facebook is more restrictive and each reaches a different audience. SMO "refers to optimizing a website and its content in terms of sharing across social media and networking sites" (Wikipedia, 2014). Among SMO techniques are Twitter Cards and Facebook OpenGraph HTML tags which can be used in Web pages to increase click-through-rates.

Another SMO practice, targeted hashtag campaigns with consistent terms used through SM such as Twitter, Instagram, and Facebook, will pull together information within the SM's native search or via search engines in ways not otherwise connected. Libraries engaging in SM should first establish a strategic program with consistent profile information across sites such as Google+, Facebook, and Twitter.

These efforts can be applied in conjunction with basic SEO techniques to enhance the discoverability of Web content through search engines such as Google, Bing, and Yahoo. Such efforts include implementing Google Webmaster Tools and ensuring the correct use of server robots.txt files, using sitemap XML files, having well-written titles and descriptive Web copy, and applying structured data using the Schema.org vocabulary. When information is shared by a researcher or library through SM and subsequently linked to by members of that community, then SEO can enhance the discoverability and SMO can enhance the shareability of these linked resources.

To explore the benefits of having a strategically connected SMO and SEO plan, we hypothesize that library-led SM campaigns will generate positive effects for discovery and sharing, thus adding value to research networks, libraries, and ultimately academic institutions. We expect that social media activity focused on IR content will lead to greater numbers of page views, visitors, downloads, and inbound link references. As researchers see value in increased exposure, faculty will increase their rate of deposit in the IR. Our research will allow us to build an understanding of researcher communities across SNS and to determine how IR content can most effectively become a part of the conversation within those communities. From a broad viewpoint, SM is a powerful tool for engaging library users and faculty researchers. With strategic coordination, programmatic campaigns, and promoting established authorship identity, libraries can facilitate conversation that amplifies the access of and visibility around library services and resources.

RETHINKING THE BOOK CONTAINER

Reading technologies have long been bound to the codex format and the fixity of the printed page, and the book as a container presents an interesting challenge to the emerging needs for data that can be interpreted by machines. Books have evolved quickly in recent years, with various eBook technologies vying to become the next dominant book container. In the transition to the Semantic Web there is an opportunity to revise the publisher role of the library and update the book as a container for knowledge.

We are developing new possibilities for publishing book content through today's primary and most powerful information technology, the Web. In our research, we have created a "web book" framework that treats

content as open, structured data and decouples it from any proprietary interfaces such as Amazon's Kindle or Apple's iBook readers. While one could easily convert the content data into a proprietary format, the real opportunity is formatting that content using standard HTML, CSS, and JavaScript for reading in any Web browser used by any device. Our first prototype was an MSU History class project to publish a book in the browser. The working linked data model and publishing software prototype using recipe and essay content is available at http://arc.lib.montana.edu/book/home-cooking/. Our "web book" mode of publishing represents an innovative step forward by drawing together emerging Web technologies to create a book that is more accessible, discoverable, shareable, and analyzable. Because the framework is "of the Web," we can get metrics from Google Analytics, such as "Time on Page," and optimize each page to be findable within search engines, which is a level of indexing that is hard to achieve using current eBook models. Additionally, the framework employs a new method for publishing within a Web browser using HTML5, structured data with RDFa Lite and Schema.org markup, linked data components using JSON-LD, and an API-driven data model that unlocks the book by transforming its content into a semantic, machine-readable and extensible platform. In keeping with our structured data advocacy, we apply the open data techniques above to ensure that the book content can be crosswalked into newer formats and UIs as technology changes.

The semantic aspect of applying linked data and RDFa markup to the book is a worthwhile contribution to a wider area of research related to open publishing models and possibilities for the eBook. Our research is driven by two fundamental questions:

1. How do we build ontological models to describe and publish book content?
2. Can libraries working as publishers create viable, open frameworks for long-form reading?

We have grounded our research in the concept of the evolving book—what it means for the book as a medium to be hyperlinked, marked up, styled, and analyzed as a full participant in the web of data. Our work benefits librarians, publishers, and researchers looking to understand new ways of publishing and linking traditionally book-based data for machine interpretation and discovery. More importantly, the work around this prototype offers a rich expression of the new knowledge work and possibilities for libraries as we take on new roles in software development and publishing.

CONCLUSION

In this article we have discussed several areas of new knowledge work where libraries can begin to establish services at the network scale that will increase our value to users and other organizations on campus. These services include:

1. Establishing and maintaining semantic identities for organizations and people so that they are recognized by search engines as individual entities.
2. Structuring metadata for machine ingest and leveraging external search mechanisms that improve access and visibility while reducing technology investments.
3. Centralizing faculty activity data management to:
   a. better inform administration decisions related to intellectual output.
   b. identify existing and potential cross-domain collaboration.
   c. improve efficient population of IR and other reporting outlets.
4. Connecting communities by developing programmatic social media strategies that value the university's research activities and intellectual output.
5. Rethinking the book container to include the creation of open extensible software and reading interfaces that advance the role of the library as publisher.

Establishing semantic identities, structuring metadata, centralizing faculty activity data, implementing social media optimization, and developing new publishing models represents network-enhancing activities that connect a university's Web properties to its intellectual output and leads to improved discoverability and access. Funding agencies and university administrators interpret reach and visibility as an expression of organizational value and libraries would do well to recognize and organize around this understanding. As we have developed expertise and skills in practicing Semantic Web development we find ourselves well positioned to offer these services to other organizations within the university. Campus organizations are investing time and resources in their Web presence, however, they lack the expertise and knowledge that forward-thinking librarians can provide. The library accelerates "network effects" as it increases the number of linked university nodes and improves the quality of semantic information about organizations, people, events, resources, and research activities. This new knowledge work provides the semantic information that search engines need to improve discoverability and access to the information that users value.

REFERENCES

comScore. (2014, June 11). *comScore releases May 2014 U.S. search engine rankings* [Press release]. Retrieved from the comScore, Inc. website: https://www.comscore.com/Insights/Market_Rankings/comScore_Releases_May_2014_US_Search_Engine_Rankings

Singhal, A. (2012, May 16). Introducing the Knowledge Graph: Things, not strings [Web log post]. *Google: Official Blog*. Retrieved from http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Wikipedia. (2014, June 10). Social Media Optimization (SMO). Retrieved from http://en.wikipedia.org/wiki/Social_media_optimization