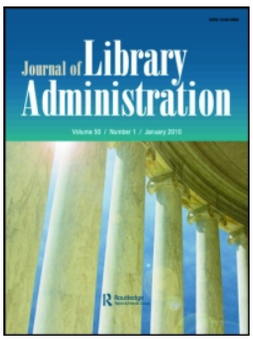

Journal of Library Administration

ISSN: 0193-0826 (Print) 1540-3564 (Online) Journal homepage: http://www.tandfonline.com/loi/wjla20

Semantic Web Identity of Academic Libraries

Kenning Arlitsch

To cite this article: Kenning Arlitsch (2017) Semantic Web Identity of Academic Libraries, Journal of Library Administration, 57:3, 346-358, DOI: 10.1080/01930826.2017.1288970

To link to this article: http://dx.doi.org/10.1080/01930826.2017.1288970

Journal of Library Administration, 57:346–358, 2017
Published with license by Taylor & Francis
ISSN: 0193-0826 print / 1540-3564 online
DOI: 10.1080/01930826.2017.1288970

Routledge
Taylor & Francis Group

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.*

SEMANTIC WEB IDENTITY OF ACADEMIC LIBRARIES

Abstract. *Semantic Web Identity (SWI) is proposed as the condition in which search engines recognize the existence and nature of entities. The display of a Knowledge Graph Card in Google search results is an indicator of SWI, as it demonstrates that Google has gathered verifiable facts about the entity. Such recognition is likely to improve the accuracy and relevancy of Google's referrals to that entity. This article summarizes part of the research conducted for a recent doctoral dissertation, showing that SWI is poor for ARL libraries. The study hypothesizes that the failure to populate records in appropriate Linked Open Data and proprietary Semantic Web knowledge bases contributes to poor SWI.*

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.

KEYWORDS Semantic Web Identity, search engines, Google Knowledge Graph, Knowledge Graph Cards, Knowledge Cards, academic libraries, Association of Research Libraries, ARL.

INTRODUCTION

The term "Semantic Web Identity" (SWI) is proposed as the condition in which Internet search engines recognize the existence and nature of entities. An entity can be a person, place of interest, concept, or an organization, but for the purposes of this article "entity" refers to a member organization of the Association of Research Libraries (ARL). The display of a Knowledge Graph Card (KC) (Google, Inc., 2016) in Google search engine results pages (SERP) is an indicator of SWI, as it implies that the search engine has gathered enough verifiable facts to make a formal presentation of the entity that offers convenience and immediate answers to users. A robust KC contains information elements about the organization that are descriptive, provide contact information, and may include appearance elements such as logos or photographs (see Figure 1).

The implication of SWI goes far beyond the simple display of a KC in search results. SWI implies machine comprehension of entities, which allows them to infer relationships with other entities and contextualize queries as intended on the Semantic Web. In the academic setting, SWI may help search engines more accurately match academic organizations with searchers' interests, which in turn may positively impact the award of research funding, student enrollment, faculty recruitment, and even university rankings. A research center that has not established its SWI, to take the first example, may not appear to grant proposal reviewers to be engaged in the type of research it is proposing, thereby weakening its chances for funding.

The study of SWI is related to Search Engine Optimization (SEO), and it includes business elements of marketing and branding. This article summarizes part of the research for a recent doctoral dissertation (Arlitsch, 2017) that examined the SWI of ARL libraries and other academic organizations, and tested the hypothesis that active engagement with proprietary and Linked Open Data (LOD) knowledge bases facilitates SWI.

BACKGROUND

The first iteration of the World-Wide-Web that was introduced by Tim Berners-Lee and his colleagues in the early 1990s (Berners-Lee, Cailliau, Luotonen, Nielsen, & Secret, 1994) revolutionized the Internet and ushered in an era of tremendous technological growth, forever "changing the way we work, socialize, create, and share information" (Manyika & Roxburgh, 2011).

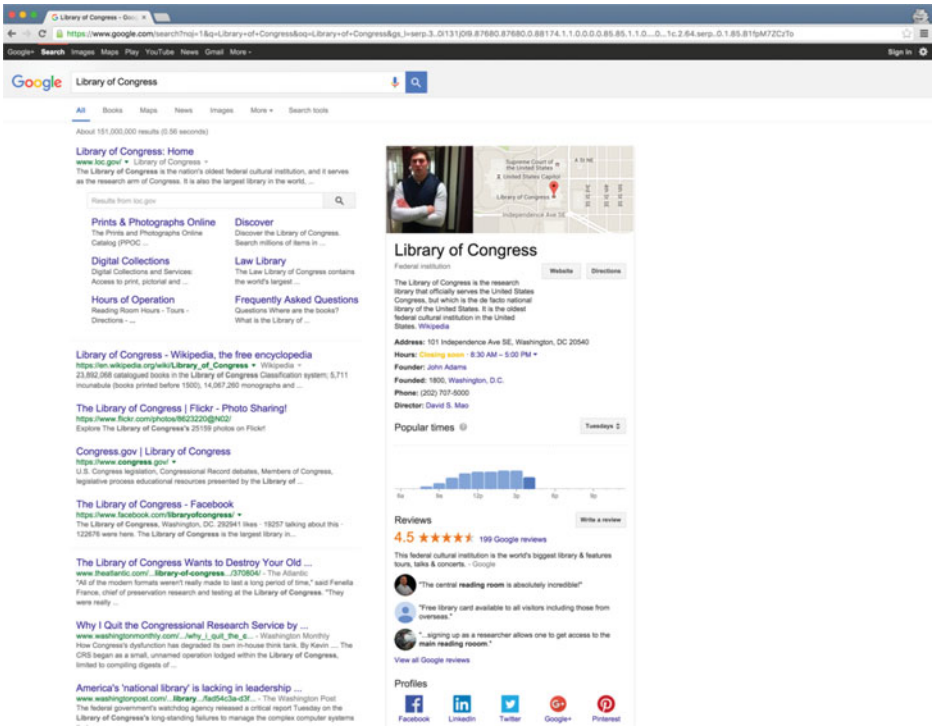

FIGURE 1 Screen capture showing a robust Knowledge Graph Card for the Library of Congress.

Propelled by the hypertext transfer protocol (HTTP) and the hypertext markup language (HTML) that allowed the unprecedented linking of documents around the world, the Web was immediately recognized for its transcendent ability to connect people to information. But explosive growth of documents posted on thousands, then millions of websites quickly led to a disorganized state in which it was difficult to find anything. The initial solutions to the information organization problem included manually-created subject directories, following a path employed by librarians for generations. Examples of such directories included the *Yahoo Directory; DMOZ; Scout Report; and the Librarians' Index to the Internet*. Nearly all of these, including the *Yahoo Directory*, which was once "the most common way people found websites" (Sullivan, 2014) eventually became irrelevant because manual organization by humans simply isn't feasible in the dynamic environment of the Web. Even before the subject directories reached their peak, Berners-Lee and colleagues realized that the deluge of information on the Web would require the processing power of machines; computers had to play a greater role than simply storing and serving documents on the Web (Berners-Lee, 1996).

Greater success in finding documents came with search engines whose algorithms weighed numerous "signals" in websites to return appropriate responses to search queries. The search engine landscape was crowded in the 1990s and early 2000s ("Timeline of web search engines," 2016), but the "PageRank" ranking algorithm developed by Sergey Brin and Larry Page (Brin & Page, 1998) proved superior to other models. The Google search engine that they developed has since has become the dominant public method of search on the Internet, consistently controlling nearly two-thirds of the United States search engine market. Microsoft's Bing search engine, which also provides organic search results for Yahoo! is the only other significant contender in the United States (comScore, Inc., 2016). Google's market share in the countries of the European Union has been estimated to exceed 90% (Meyer, 2015).

Long before the size of the "indexed" World-Wide-Web rose to nearly 5 billion pages (de Kunder, 2016), search engine and Web developers began to realize that the "strings-based" search environment of the Web, where algorithms matched strings of text to search queries, was a limited solution. The Web had to evolve into a new environment where the subtleties of context and meaning could be gleaned by machines, and where search engines were more able to provide answers to queries rather than simple referrals to documents where the answers might reside. In 2001, Berners-Lee and his colleagues formally introduced the concept of the Semantic Web, an evolved version of the World-Wide-Web where data and information could be processed automatically by computers that have access to structured collections of information (Berners-Lee, Hendler, & Lassila, 2001). Although the growth of the Semantic Web was slow in the early years, recent developments in both search engines and the data sources available to them have dramatically improved the ability of machine comprehension and communication. In 2012, Google introduced its Knowledge Graph (Singhal, 2012), a graph database that operates behind the scenes to gather verified facts and produce enhancements to search results, such as Knowledge Graph Cards, sometimes referred to as Knowledge Panels, Information Cards or Knowledge Cards (Goel et al., 2013).

The Semantic Web promises an improved user experience through more intelligent use of data and integration with semantic technologies. Wide-ranging effects are possible when machines can tap structured data sets to understand entities and their relationships. Search engines that have been informed by Semantic Web knowledge bases can provide accurate answers in SERP instead of lists of websites. Information can be handed off to semantic technologies, such as mapping or voice recognition applications. In short, the transition from the strings-based environment of the Web to the entity-based environment of the Semantic Web helps search engines offer more accurate and relevant search results and answers. But providing search

engines with accurate and verifiable structured data records is a crucial part of this equation.

RESEARCH HYPOTHESIS

The portion of the dissertation research that is reported in this article evaluated the SWI of the 125 members (decreased to 124 members during this study) of the Association of Research Libraries (ARL). The study tested the hypothesis that certain proprietary and open Semantic Web knowledge bases must be actively engaged so that an organization can be optimally recognized and understood as an entity by Google.

The problem of organization names looms large in this research. Each member library of the ARL has a primary (official) name that it voluntarily submits to the ARL membership directory,[1] and 94 of the 125 members also often use an alternate (unofficial) name when they communicate about their organizations. For example, "Yale University Library" is the official name listed in the ARL directory, while "Sterling Memorial Library" is the alternate name for the main library at Yale University. The research questions are structured to consider both primary and alternate names.

Research Questions

RQ 1: What is the current state of Semantic Web Identity of ARL libraries, as indicated by the presence of accurate Knowledge Graph Cards in Google search results when the primary and alternate names of those libraries are searched?

RQ 2: Are records or profiles present for ARL primary and alternate library names in the following knowledge bases: Google My Business, Google+, Wikipedia, DBpedia and Wikidata?

RESEARCH METHODS

A data set was collected by conducting Google searches for the member organizations, observing the presence or lack of KC in the SERP, and capturing evidence of those results with screen capture software. Five proprietary and LOD knowledge bases were also searched to determine whether records, profiles, or articles existed for the library organizations. Results were recorded in binary code in a spreadsheet: "1" indicated the presence of a KC or a knowledge base record and "0" indicated the lack of a KC or record. The data set comprises over 1400 screen capture files, two spreadsheets,

[1] ARL membership directory - http://www.arl.org/membership/list-of-arl-members

and R statistical equations that demonstrate the state of SWI for ARL libraries during the data collection period from December 2015 through April, 2016 (Arlitsch, 2016).

RQ1: Google Searches

The total of 219 primary and alternate names was searched in Google to determine whether accurate KC appeared for each library name in SERP. In cases where a KC displayed for both the primary and alternate name of the library, notice was taken as to whether those KC were identical, indicating that the search engine has correctly determined that both names refer to the same organization. This "same as" semantic designation can be explicitly stated in certain knowledge bases, such as Google My Business and Wikidata.

RQ2: Knowledge Base Searches

Google's Knowledge Graph is known to gather information about entities from a variety of sources, and when it has enough verified facts it generates a KC for the entity. The scholarly and nonscholarly literature (including pronouncements from Google itself) gives some indication of the sources of this information. For this study, five knowledge bases were hypothesized to influence the generation and population of KC. These five sources included two proprietary knowledge bases (Google My Business and Google+), and three knowledge bases that are considered significant data sources in the Linked Open Data cloud (Wikipedia, DBpedia, and Wikidata). The 219 primary and alternate names of ARL member libraries were searched in each of the five knowledge bases and records or profiles that appeared for each name were recorded. The presence or lack of records in these knowledge bases was then compared with the presence or lack of accurate KC.

The five knowledge bases were selected for the following reasons:

1. **Google My Business (GMB)** is recommended by Google for registration of an organization (Google, Inc., 2017), implying that GMB may be a significant source of information for its Knowledge Graph. Creating a record in GMB requires a formal "claim and verify" process that involves a two-step communication response. In this way, Google verifies that the location of the claimant is legitimate.
2. **Google+** contains profiles for many organizations and is tightly integrated with other data sources in the Google empire. Verified Google+ profiles are automatically generated when an organization has been claimed and verified in GMB. However, it is also possible to create a profile directly within Google+, which usually results in an unverified profile.

3. Apart from its size and documented influence on the Semantic Web, **Wikipedia** was chosen because the textual descriptions evident on most KC are indicated by Google as having been drawn from Wikipedia.
4. **DBpedia** records are automatically generated from Wikipedia articles and are offered as linked data, which led to the hypothesis that Google might take advantage of these highly structured data records.
5. **Wikidata** is a relatively new creation of the Wikimedia Foundation but it has recently been significantly augmented by the migration of Freebase data. The now defunct Freebase was formerly acknowledged as a major source of information for Google's Knowledge Graph, and it seemed reasonable to assume that Google might now turn to Wikidata for structured data records.

FINDINGS

The main findings from the data collection and analysis are described in this article with the help of visual displays called table plots, which are generated from the R statistical software (Gentleman & Ihaka, 2016). Table plots compare two (in this case) columns from the spreadsheets, and each row in the table plot represents a primary or alternate name of an organization. Colors in the table plots represent the presence (1) or lack (0) of the item being recorded. For instance, the blue rows in column 1 of Figure 2 represent the primary names of the 125 members of the ARL (score = 1), while the orange rows represent the 94 alternate names of the libraries (score = 0).

RQ1: Google Search Results

Google SERP reveal that accurate KC appeared for 46% of primary name searches and 79% of alternate name searches for ARL member libraries. Combined, 132/219 (60%) primary and alternate names displayed accurate KC, leaving 87 (40%) of primary and alternate names that either displayed no KC at all or displayed a KC that was inaccurate for the library name being searched. (see Figure 2).

Semantic "Same As" relationships

The number of accurate KC that displayed for the 125 ARL member organizations for either the primary or alternate name was 102/125 (82%). However, in many cases the KC that displayed for the alternate name was not the same as the KC that displayed for the primary name. In Figure 3, for instance, a Google search for the primary name "University of Rochester Libraries" displayed an incomplete KC for the "Rush Rhees Library," the alternate name of

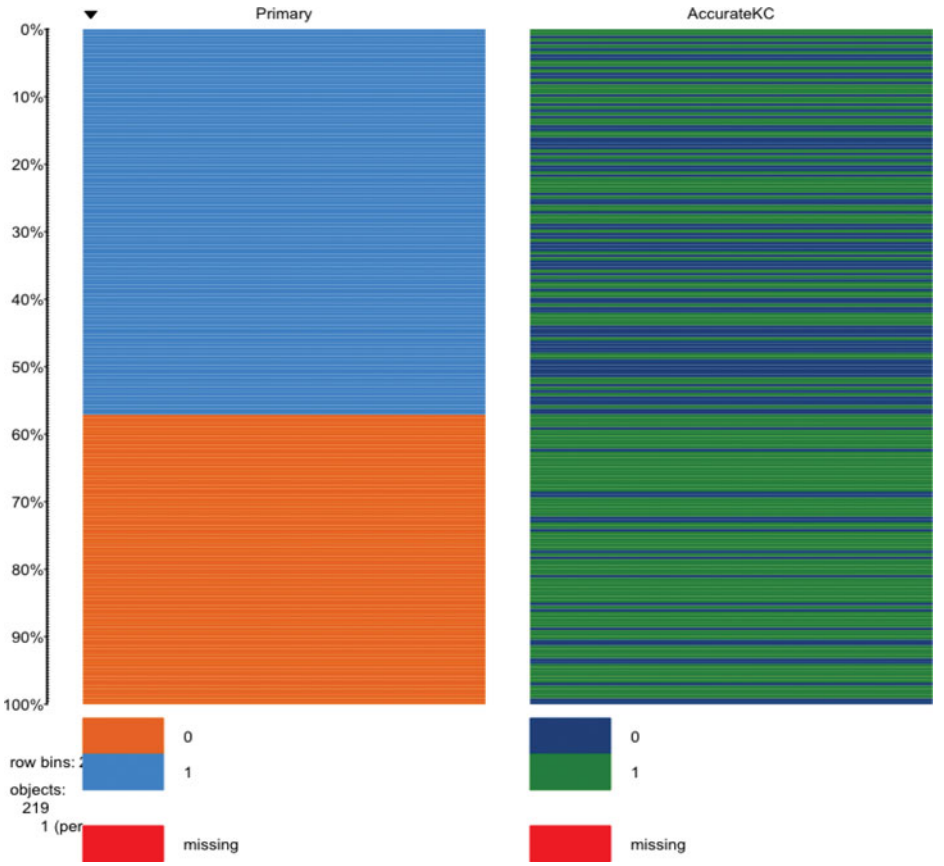

FIGURE 2 Table plot showing that the 94 alternate library names for ARL members (column 1, orange rows) display more accurate KC (column 2, green rows) in Google SERP.

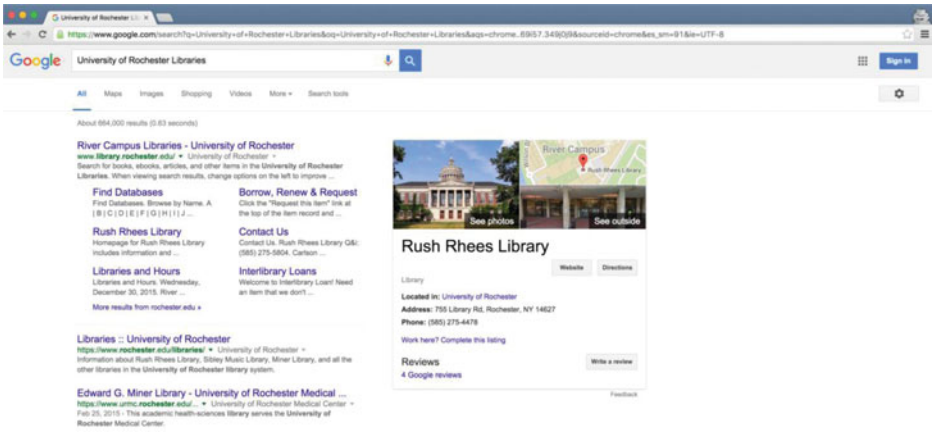

FIGURE 3 A search for "University of Rochester Libraries" displayed an incomplete KC for Rush Rhees Library.

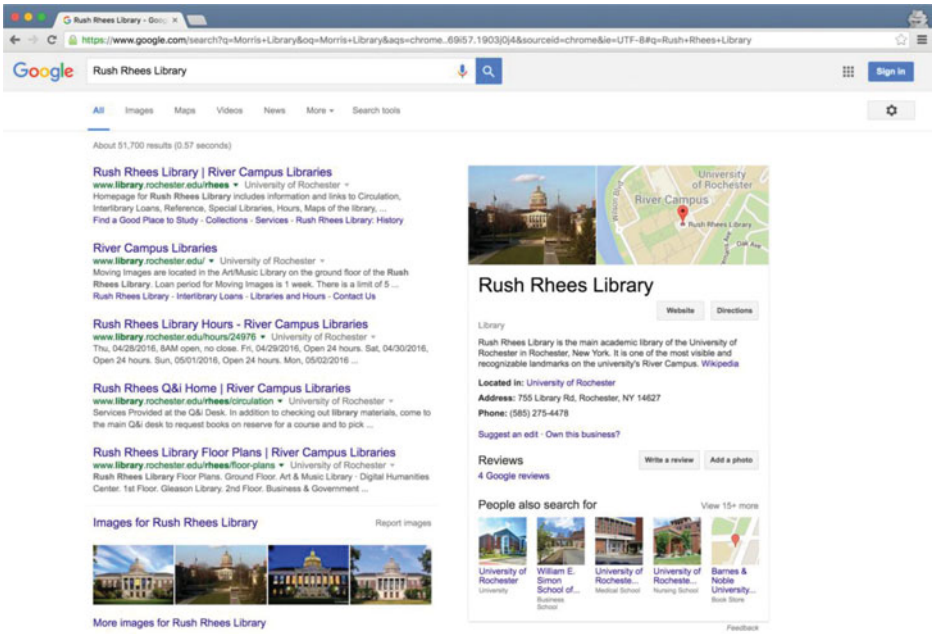

FIGURE 4 The KC that displayed for a search of the "Rush Rhees Library" was different than the KC in Figure 4.

the main library on campus. However, a Google search for the "Rush Rhees Library" displayed a different KC for the Rush Rhees Library than the one that appeared in the first search (see Figure 4).

This important nuance raises the issue of semantic "same as" relationships. An established "same as" relationship would indicate that the search engine recognizes the primary and alternate names as belonging to the same library organization, causing it to display the same KC regardless of which name is searched. Figure 5 provides a visual demonstration that the "same as" relationship of primary and alternate library names is lacking in most cases. Identical KC for both primary and alternate names displayed for only 37% (46/125) of the member libraries.

RQ2: Records and Profiles in Knowledge Bases

Table 1 summarizes the number and percent of records or profiles that appeared for primary and alternate library names in each of the five knowledge bases. Only 22% of the primary library names and 43% of the alternate names showed a "claimed and verified" business in Google My Business (GMB). Verified Google+ profiles existed for 41 (19%) of primary or alternate names, but another 42 unverified Google+ profiles were found, often displaying information that differed from the verified profiles.

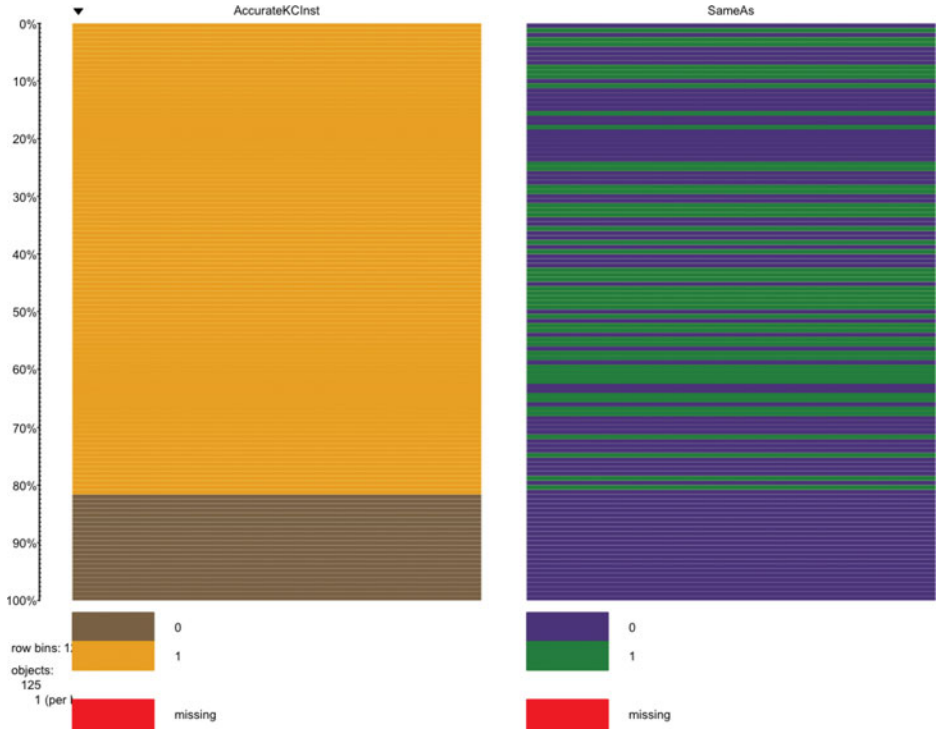

FIGURE 5 Table plot showing that 82% of 125 ARL member libraries displayed an accurate KC when either the primary or alternate name was searched (column 1, yellow rows). However, KC that appeared for both the primary and alternate name of the library were often different, indicating a lack of "same as" understanding by the search engine (column 2, purple rows).

Wikipedia articles were found for only 82/219 (37%) of primary or alternate names, but 26 of those lacked an infobox, which supplies the structured data that is crucial for generation of DBpedia records. DBpedia records were found for only 32% of primary or alternate names, and Wikidata records were found for 29% of primary or alternate names. Wikidata records were only counted if more than one field was populated, as this helped eliminate

TABLE 1 Number and percent of records or profiles in knowledge bases for primary and alternate library names.

Knowledge Base	Primary (% of 125)	Alternate (% of 94)	Total (% of 219)
Google My Business	28 (22)	40 (43)	68 (31)
Google Plus (unverified)	25 (20)	17 (18)	42 (19)
Google Plus (verified)	22 (18)	19 (20)	41 (19)
Wikipedia (w/o infobox)	10 (8)	16 (17)	26 (12)
Wikipedia (w/infobox)	30 (24)	26 (28)	56 (26)
DBpedia	30 (24)	39 (41)	69 (32)
Wikidata	26 (21)	37 (39)	63 (29)

shell records that had been automatically generated from Wikipedia articles. However, most of the Wikidata records that were counted were only minimally populated.

DISCUSSION AND CONCLUSION

The research described in this article demonstrates that ARL libraries have significant room for improvement of their Semantic Web Identity (SWI). The member libraries are poorly represented in the knowledge bases that influence Google's Knowledge Graph and help generate Knowledge Graph Cards (KC) in search engine results pages (SERP). As a result, only 60% of the combined primary and alternate names searched in Google displayed accurate KC.

A significant finding of the research is that ARL libraries are inconsistent in the communication of their names, a practice that creates more difficulties on the Semantic Web than in the analog world. While all ARL libraries submit their primary names to the ARL membership directory, most have alternate names and tend to use them more frequently than primary names on their websites and in knowledge bases. This results in more KC appearing for more alternate names than for primary or official names of the organization. Furthermore, librarians seem to make little effort to reconcile those names by establishing "same as" relationships in machine readable knowledge bases that would help search engines better understand that both names refer to the same organization.

The findings also imply that in the absence of a clear organizational marketing strategy for the Semantic Web, individuals in libraries sometimes take it upon themselves to engage with the knowledge bases. Google+ is the easiest of the examined knowledge bases with which to do this, and the data set shows many unverified Google+ profiles, including some that exist for units within libraries when the parent library organization has no profile at all. When multiple profiles exist, they often contain multiple addresses and other contact information. Combined with a lack of formal engagement with other knowledge bases, these unsanctioned efforts undermine consistency and create confusion for search engines trying to gather authoritative facts.

It is important to understand that the display of a KC is merely an indicator of SWI, and that the larger meaning behind the lack of a KC is that the search engine doesn't understand the existence and nature of the organization. While it is interesting to apply technical practices to compel the generation and population of KC, there is a larger implication about librarians' willingness to step into the arena of the Semantic Web. As a profession, we have generally been slow to engage with Semantic Web data sources and technologies, to the detriment of our organizations and

our work. Beyond SWI for library organizations, many common concepts in the field of library and information science are also poorly defined on the Semantic Web due to this lack of engagement. Even now, the DBpedia record for "Library" displays subject headings of "Book promotion; Library; and Library science," and the RDF types include terms like "Artifact; Object; Physical entity; and Structure." These surely are not the best descriptions of modern libraries, but they paint the picture of the structured data that we make available to the machines that have an increasing impact on our relationships with users.

The sobering fact is that search engines do not have a good understanding of academic libraries, and this may hinder referrals. While the member libraries of the ARL served as a convenient and discrete data set, the state of their SWI is no different than other areas of academia. Other parts of the dissertation research not described in this summary article demonstrate that the problem extends well into all academic organizations.

While the current SWI situation is unimpressive in libraries and other academic organizations, it represents an opportunity. Librarians can improve the SWI of their own organizations and offer that capability as a service to other academic organizations. In late 2015, the Montana State University (MSU) Library launched a new SWI service, which is currently being used to improve the SWI of other MSU academic organizations. A description of that SWI service and its impact will be the subject of a future article.

ORCID

Kenning Arlitsch http://orcid.org/0000-0002-5919-735X

REFERENCES

Arlitsch, K. (2016). *Data set supporting the dissertation "Semantic Web Identity in Academic Organizations: Search engine entity recognition and the sources that influence Knowledge Graph Cards in search results."* Montana State University Scholar Works. Retrieved from https://doi.org/10.15788/M2F590

Arlitsch, K. (2017). *Semantic Web Identity of Academic Organizations: Search engine entity recognition and the sources that influence Knowledge Graph Cards in search results*. (Dissertation). Humboldt Universitätzu Berlin, Berlin, Germany. Retrieved from https://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=43177

Berners-Lee, T. (1996). WWW: past, present, and future. *Computer*, *29*(10), 69–77. doi:10.1109/2.539724

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., & Secret, A. (1994). The World-Wide Web. *Communications of the ACM*, *37*(8), 76–82. doi:10.1145/179606.179671

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34–43. doi:10.1038/scientificamerican0501-34

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. doi:10.1016/S0169-7552(98)00110-X

comScore, Inc. (2016). comScore releases February 2016 U.S. desktop search engine rankings [Commercial]. Retrieved from https://www.comscore.com/Insights/Rankings/comScore-Releases-February-2016-US-Desktop-Search-Engine-Rankings

de Kunder, M. (2016). The size of the World Wide Web (The Internet). Retrieved from http://www.worldwidewebsize.com

Gentleman, R., & Ihaka, R. (2016). R (Version 3.2.4) [X86_64-apple-darwin13.4.0 (64-bit)].R Foundation. Retrieved from https://www.r-project.org

Goel, K. J., Thakur, S. A., Levy, J. L., Dhanaraj, C. R., Carmi, E., Provine, J. R., & Moxley, E. K. (2013). Knowledge Panel. Retrieved from https://www.google.com/patents/US20130311458

Google, Inc. (2016). Request a change to a Knowledge Graph card in search results. Retrieved from https://support.google.com/websearch/answer/6325583?p=kg_edit&rd=1

Google, Inc. (2017). Get your free business listing on Google [Commercial]. Retrieved from https://www.google.com/business/

Manyika, J., & Roxburgh, C. (2011). *The great transformer: the impact of the Internet on economic growth and prosperity*. (pp. 1–10). New York, NY: McKinsey Global Institute. Retrieved from http://www.mckinsey.com/industries/high-tech/our-insights/the-great-transformer

Meyer, R. (2015). Europeans use Google Way, Way More than Americans Do: Google's huge market share is part of what strengthens the EU's antitrust case. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2015/04/europeans-use-google-way-way-more-than-americans-do/390612/

Singhal, A. (2012). Introducing the knowledge graph: things, not strings. Retrieved from http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Sullivan, D. (2014). The Yahoo Directory—once the Internet's most important search engine—is to close. [Commercial]. Retrieved from http://searchengineland.com/yahoo-directory-close-204370

Timeline of web search engines. (2016). In *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation, Inc. Retrieved from https://en.wikipedia.org/wiki/Timeline_of_web_search_engines