3 | Ensuring Discoverability of IR Content

Kenning Arlitsch, Patrick OBrien, Jeffrey K. Mixter, Jason A. Clark, and Leila Sterman

Discoverability of content through Internet search engines is paramount to the success and impact of institutional repositories (IRs). Overwhelming evidence suggests that library and IR Web sites attract relatively little direct traffic, and instead the vast majority of users begin their research with search engines (DeRosa et al., 2010) and land at local Web sites only through referrals. Americans conduct 18 billion searches per month in Internet search engines (comScore, Inc., 2014), so the potential market for visitors is deep, but library Web sites and repositories typically see only a minuscule fraction of that traffic. Libraries find themselves struggling to become effective in a discovery environment that "means syndication to search engines, to disciplinary resources, or to other specialist network-level resources" (Dempsey, Malpas, & Lavoie, 2014). This directive speaks to making IR content available and usable to a variety of user agents on the Web through data interchange standards that are widely accepted and supported.

Search engines must be able to access IR metadata and make sense of its structure. Even the best repository software will fail if it offers metadata that is incomplete, lacks context, or is not understood by machines. The user experience is also a significant factor for search engines. Google is very concerned with delivering a superior experience to its customers and makes it clear that sites can improve ranking in search results by addressing the user experience (Google Inc., 2015b). This includes providing high-quality

content with rich descriptive text that is useful, presented in a logical linking structure, and easily accessed by both users and Web crawlers (Google Inc., 2015a).

The extent to which IR content draws attention from search engines and ranks in search results is contingent on the search engine optimization (SEO) practices that are built into the repository. While SEO itself has been described in great detail elsewhere, this chapter discusses SEO issues unique to IR as well as several newer Semantic Web techniques that can help improve the discoverability and relevance ranking of IR content, including structured metadata, Semantic Web Identity, PDF cover sheets, and semantic description of content through Linked Data.

STRUCTURED METADATA

The Metadata Problem

Structured metadata is a fundamental underpinning of digital library work, and it can help address the lack of search engine attention to IR content. Metadata must be accessible and organized for machines as well as humans. Several types of user agents must be considered in the formula for discovering metadata in IR:

1. Commercial search engine crawlers (Google, Bing)
2. Specialized search engines (Google Scholar)
3. Intelligent software agents (Semantic Web bots)
4. Human users

Search engine crawlers don't actually crawl through repository databases. Instead, they systematically trigger the display of Web pages by following links, and when an HTML page is generated they harvest its contents. It is at the crucial point of page display that all the metadata necessary to represent the content must be simultaneously visible to the human and comprehensible to the crawler. Other potential obstacles to crawlers may include IR websites that don't provide clear and quick paths to content; overuse of graphics that crawlers can't decipher; conflicting sitemaps and robots.txt files; slow server response; and content that is moved without

appropriate messaging to inform crawlers of the changes, whether temporary or permanent (Arlitsch & OBrien, 2013).

In 2011 Google Scholar announced that institutional repositories should "use Dublin Core tags as a last resort" because the schema isn't appropriate for describing scholarly works (Google Scholar, n.d.a). Dublin Core doesn't include unambiguous fields for each part of a bibliographic citation: volume, issue number, first page, last page, or a field for the PDF URL. Nor are there appropriate fields that distinguish a published article from a preprint, a dissertation from a thesis, or a book chapter from a book. In short, Dublin Core cannot provide the parsed bibliographic information that Google Scholar gets from publishers who use other schemas such as Highwire Press, PRISM, EPrints, and bepress. Google Scholar's dismissal of Dublin Core has been a major factor in the poor visibility of open access IR content (Arlitsch & O'Brien, 2012).

Beyond the specific requirements that enable discovery in Google Scholar, there are broader possibilities in the areas of semantic markup and Linked Data that help to establish higher engagement and use of IR content. The content of an IR must be classified so that machines may understand the site in broad context. Schema.org, a collaborative project between Google, Bing, Yahoo, and Yandex, is a vocabulary for defining things on the Web. The vocabulary of Schema.org tends to skew toward description for e-commerce settings, but classes and properties are being actively defined and are increasingly applicable to scholarship and academe. Active W3C Working Groups (WG), such as the Schema BibExtend WG (http://goo.gl/ZKbE4J), are open for participation in these defining activities. This growth in the vocabulary is key for accurate description in IR settings. Several Schema.org types help guide the semantic markup for IR content, including:

- schema.org/Article
- schema.org/Dataset
- schema.org/ScholarlyArticle

The work needed to establish Semantic Web Identity and convert legacy IR metadata into Linked Data is described in more detail below.

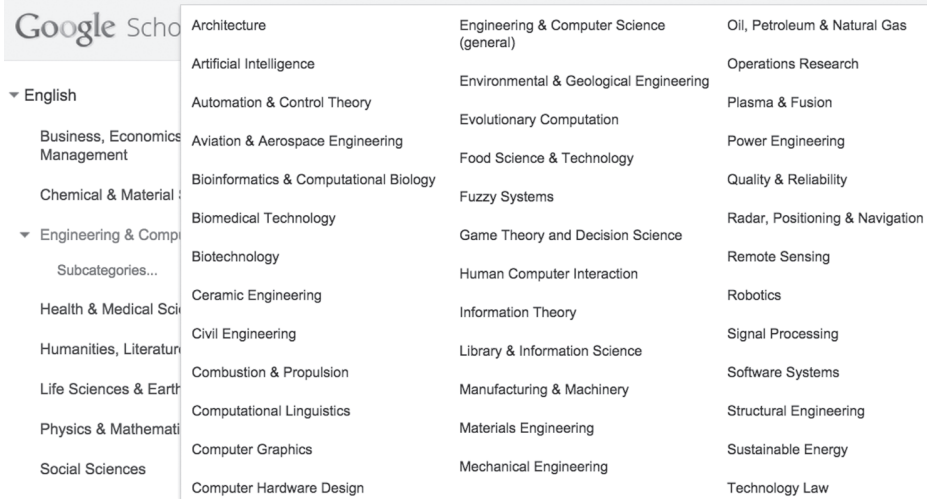

Figure 3.1. Google Scholar Metrics "Engineering & Computer Science" category and its subcategory taxonomy.

Consistency of Metadata

Much of the work of ensuring discovery of IR content has focused on machine-readable markup and semantic modeling practices, but providing consistent metadata for IR items is a core requirement. IRs are often part of the library ecosystem, and practices like applying Library of Congress Subject Headings may already be a part of the ingest process. It is important for both humans and machines that the application of terms is consistent. It may be obvious for items that have specific names (departments, colleges), but it is similarly important to apply consistent metadata in all fields. A machine may not know that "biology," "Biology," and "Biological sciences" could be synonymous in the organizational structure. There are a large number of other controlled vocabularies that IR managers can choose from, and most pertain to specific fields or domains. One possibility for assigning "Web-friendly" vocabularies are the facets that Google applies in its own systems. For example, Google Scholar citations (http://goo.gl/TejdTK) uses an academic taxonomy consisting of 8 broad categories and 253 subcategories that could provide a useful framework for organizing IR content (Figure 3.1).

DISCOVERY IN GOOGLE SCHOLAR AND OTHER SEARCH ENGINES

The ubiquity of Google and Google Scholar has established them as the paradigms of commercial search engines. Google's mission is to "organize the world's information and make it universally accessible and useful" (Google Inc., 1999). Google Scholar (GS) is a specialized search engine designed to find and index scholarly literature; it is a separate part of the Google organization and uses different algorithms and methods to analyze Web content. The different approaches of these two related search engines underscores the challenge to IRs trying for a presence in both: they must present content on a single Web page for various audiences. Below is an example of Modern Language Association, Seventh Edition (MLA) citation information presented for human readability:

> **Human-Readable MLA Citation Format**
> Arlitsch, Kenning, and Patrick S. O'Brien. "Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar." *Library Hi Tech* 30.1 (2012): 60–81.

Humans benefit from their ability to grasp context and parse a citation into its individual elements. We can determine the difference between title, journal, volume, issue, and page numbers, regardless of the various formats and styles that are available. But machines see only strings of characters and need help identifying the string of text as a bibliographic citation, parsing the citation's elements, and establishing relationships between fields.

The crawlers that gather information for search engines prefer each of these elements to be provided in defined fields. Figures 3.2 and 3.3 are respective examples of structures that help general search engines like Google and academic search engines like Google Scholar understand a bibliographic citation. They show the same citation with each element in specific Schema.org and Highwire Press tags.

Key information provided to general search engines via Schema.org:

- Lines 3 and 4 indicate this is a scholarly article as defined by Schema.org (i.e., http://schema.org/ScholarlyArticle).
- Lines 9–11 indicate the exact "Kenning Arlitsch" we are referring to per

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar",
  "author": [
    {"@type": "Person",
        "name": "Kenning Arlitsch",
        "sameAs":[ "http://viaf.org/viaf/294187294",
        "https://scholar.google.com/citations?user=KWrhbCMAAAAJ&hl",
        "http://www.lib.montana.edu/people/about.php?id=31" ] },
    {"@type": "Person",
        "name": "Patrick S. O'Brien",
        "sameAs": ["http://viaf.org/viaf/306101244",
        "https://scholar.google.com/citations?user=tWV-IE4AAAAJ&hl",
        "http://www.lib.montana.edu/people/about.php?id=21" ] }
    ],
  "isPartOf": [
    {"@type": "PublicationVolume",
        "volumeNumber": "30" },
    {"@type": "PublicationIssue",
        "issueNumber": "1" },
    {"@type": "Periodical",
        "datePublished": "2012",
        "name": "Library Hi Tech",
        "issn": "0737--8831",
        "publisher": "Emerald" } ],
  "pageStart": "60",
  "pageEnd": "81",
  "associatedMedia":
    {"@type": "MediaObject",
        "encodingFormat": "PDF",
        "contentUrl":
        "http://scholarworks.montana.edu/xmlui/bitstream/handle/1/3193/Arlitsch-Obrien-LHT-GS-final-revised_2012-02-18.pdf"
        },
  "sameAs": [
    "http://dx.doi.org/10.1108/07378831211213210",
    "hhttp://scholarworks.montana.edu/xmlui/handle/1/3193" ]
  }
</script>
```

Figure 3.2. General search engine markup applying Schema.org.

VIAF, Google Scholar, and Montana State University's URI Linked Data. This becomes very important when an author has a common name, such as "John Smith."

- Lines 18–27 indicate this scholarly article is part of the Library Hi Tech journal, Volume 30, Issue 1, published by Emerald.
- Lines 30–35 indicate that a PDF of the scholarly article is available via the MSU Scholarworks IR URL provided.
- Lines 36–39 indicate that the Web page containing the code above is about the same "thing" (i.e., schema.org/ScholarlyArticle) as the HTML page in the MSU Scholarworks IR and the doi.org URI.

```
<!--  Title & Author -->
<meta name="citation title"
    content="Invisible Institutional Repositories: Addressing the Low Index Ratios of IRs in Google Scholar" />
<meta name="citation_authors" content="Arlitsch, Kenning" />
<meta name="citation_authors" content="O'Brien, Patrick" />
<!-- Publisher & Journal -->
<meta name="citation_journal_title" content="Library Hi Tech" />
<meta name="citation_publisher" content="Emerald Insight" />
<meta name="citation_date" content="2012" />
<meta name="citation_volume" content="30" />
<meta name="citation_issue" content="1" />
<!-- Article Location -->
<meta name="citation_firstpage" content="60" />
<meta name="citation_lastpage" content="81" />
<meta name="citation_pdf_url"
    content="http://scholarworks.montana.edu/xmlui/bitstream/handle/1/3193/Arlitsch-Obrien-LHT-GS-final-revised_2012-02-18.pdf" />

```

Figure 3.3. Highwire Press tags for academic search engines like Google Scholar.

While these figures may look complicated, the markup is designed for machines to parse the information and provides a method, format, and syntax that both Google and Google Scholar understand.

IR SITE STRUCTURE

Content is more easily found by both humans and machines if there is a short and efficient pathway from the home page to item-level content (Google Inc., 2015a). IRs also benefit from providing a clear sitemap directing search engines to the most important content, such as item pages. In addition, libraries can structure the human-readable links on the IR entry Web site to match the organization of the institution, thereby ensuring consistent and clearly defined content. Matching the hierarchical structure of the institution (College > Department > Item) or providing a similar logical structure can assist human navigation.

Ranking algorithms are enormously important in the search engine business. One method of ranking "objectively and mechanically" (Page, Brin, Motwani, & Winograd, 1999), called "PageRank," was Google's first algorithm and still plays into the many factors that help Google give order to the vast World Wide Web. PageRank is largely based on the number of inbound links a site has from other Web sites, as they are interpreted by search engines as a vote of confidence. IRs can improve their rank in search results by encouraging organizations or centers on campus to link back to relevant sections of the IR from their own Web sites and social media

profiles. Although many of Google's current 200+ "signals" (Dean, 2014) that rank search results are secret, they are largely based on the standards of SEO best practices and machine-readable markup, which are outlined in webmaster guidelines and tools that some search engines provide.

PDF Files and Cover Sheets

One goal of IRs is to ensure that the public has easy access to the content. The portable document format (PDF) is currently the most common way to deliver scholarly articles. Google Scholar recommends maximum PDF file sizes of 5 MB (Google Scholar, n.d.b), and the filename should be the article title, with words separated by hyphens.

A standardized PDF cover sheet may also be helpful to humans as it identifies the source of a downloaded file, and it is useful for machines because it provides another standard method of communicating citation information. Google Scholar makes recommendations for optimized IR PDF cover pages (Google Scholar, n.d.c). Some software generates cover sheets automatically, though it may be prudent to check the created page against Google Scholar's recommendations.

BEST PRACTICES FOR THE FUTURE

Establishing Semantic Web Identity

Although humans are good at inferring meaning from words and context, machines are not. Homonyms, or more specifically in this case, homographs, are a challenge to machines trying to discern varying definitions from the same string of characters and can cause them to deliver inaccurate search results. Does that "jaguar" on a Web site refer to the animal, car, sports team, supercomputer, or an old Macintosh operating system?

Things or concepts can be established as "entities," which helps search engines understand and trust them, and that in turn may help increase visitation and use. Google's Knowledge Graph is an effort to build a knowledge base of semantically related and vetted information about established entities. Using data collected through its Knowledge Graph, Google has thus far rolled out three enhancements to search results: Knowledge Card, Carousel, and Answer Box.

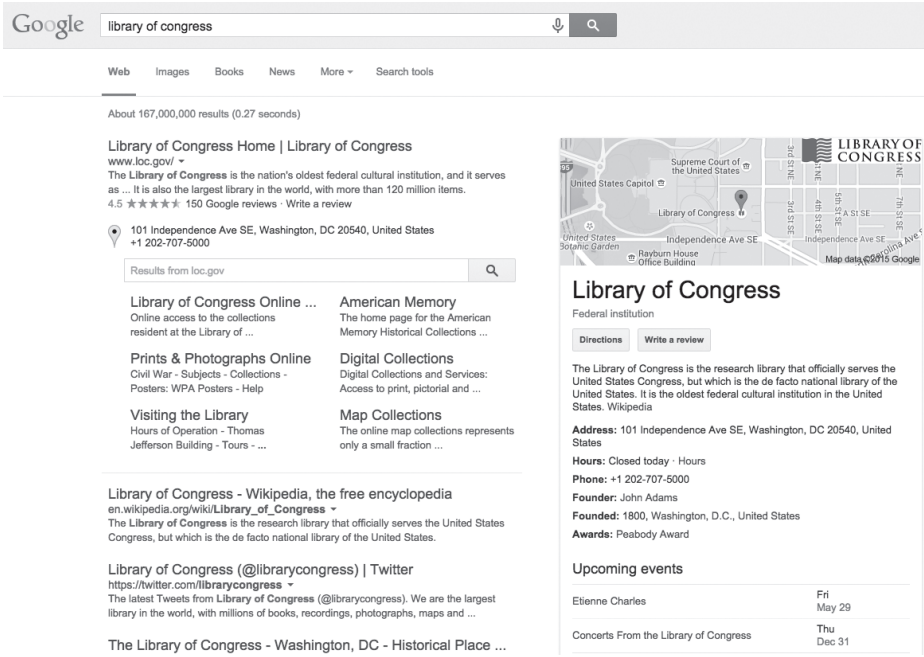

Figure 3.4. A Google search for "Library of Congress" displays a Knowledge Card for the organization.

The Knowledge Card (see Figure 3.4) is a panel that now often appears to the right of Google search results and displays information about specific entities (e.g., people and organizations). The Carousel (see Figure 3.5) is a group of instances that comprise a concept and appears across the top of the search results screen (e.g., sports teams, universities in a given state). The Answer Box (see Figure 3.6) provides facts about concepts or things that haven't necessarily been established as entities and is embedded at the top of traditional search results.

Each of these enhancements is populated with information that the Knowledge Graph compiles from certain sources on the Web that are trusted to establish entities. Chief among these sources is structured data generated from Wikipedia entries. Other sources may include Google My Places, Google+, Wikidata, and Schema.org markup consistent with the human-readable content in Web sites. Ensuring that these sources are populated with accurate information helps create Semantic Web Identity.

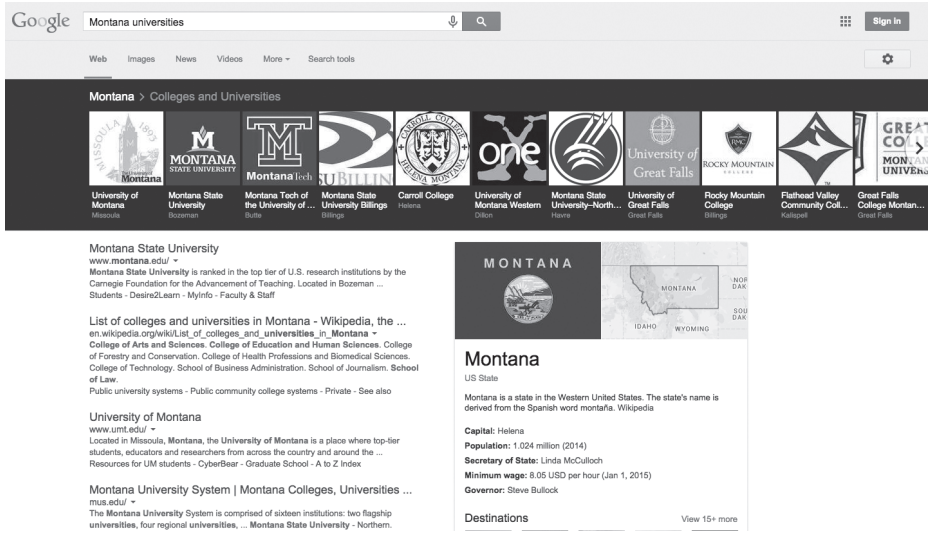

Figure 3.5. A Google search for "Montana universities" displays a Carousel with logos from each of the schools.

A Google search for "Montana State University Library" in 2013 demonstrated what happens when a thing (an organization in this case) doesn't have an established Semantic Web Identity. Instead of displaying the flagship library of the Montana State University (MSU) system, located in Bozeman, the Knowledge Card display showed another MSU campus in Billings, Montana (see Figure 3.7). The phrase "Montana State University Library" was simply a text string to Google, and it interpreted the organization incorrectly because the data sources contained erroneous information about the MSU Library. As a result, Google incorrectly identified the MSU Library as a building in Billings, Montana. A screenshot from 2015 demonstrates that the authors have successfully corrected the problem (see Figure 3.8).

There were several reasons why the MSU Library in Bozeman was misidentified in Google's Knowledge Card: (1) no one had claimed the property or verified facts about the library in the trusted data feeds to Google's Knowledge Graph; and (2) no article about the MSU Library had been created in Wikipedia.

The example of the Semantic Web Identity problem of the MSU Library can be extended to IRs as well. The concept of an institutional repository is currently not well understood by Google because it hasn't

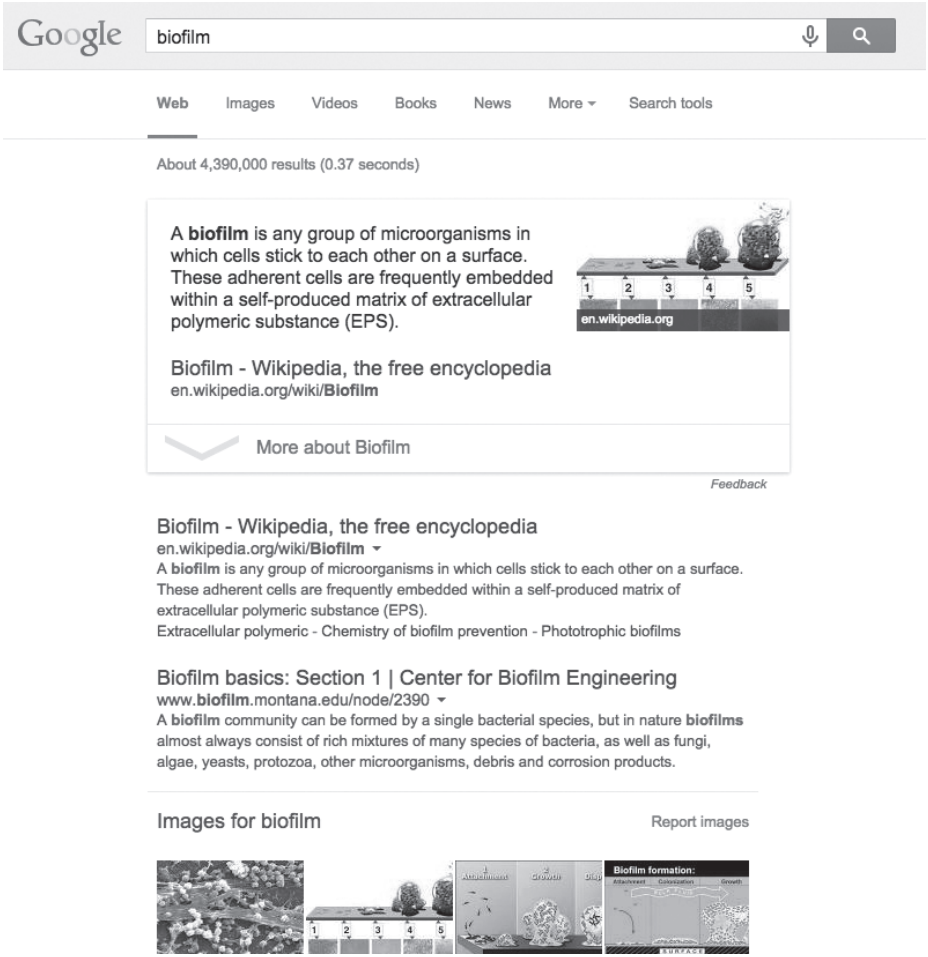

Figure 3.6. A Google search for "biofilm" displays an Answer Box containing a definition from Wikipedia.

been carefully defined for machines by librarians in Google's trusted data sources. Currently, searching for "institutional repository" in Google brings an "Answer Box" based on a Wikipedia entry. The Wikipedia entry contains descriptive text, but it has no machine-understandable properties (i.e., parent institution, topics represented, languages, etc.). Moreover, there are zero *instances* of the "concept" of an institutional repository. In other words, the IR is a *described* concept only, and machines would be hard pressed to provide a list of IRs, let alone point to one. Wikipedia

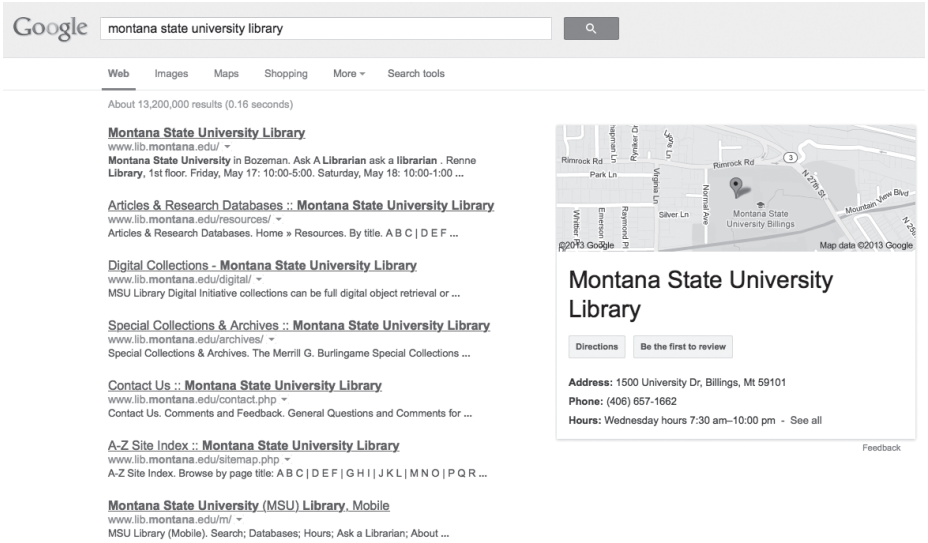

Figure 3.7. A Google search for "Montana State University Library" in 2013 displayed a Knowledge Card for a branch campus in Billings, Montana.

has a loosely related "List of Repositories" (http://en.wikipedia.org/wiki/List_of_repositories) containing fewer than 20 repositories, and none are from the United States.

Contrast that situation with a Google search for "Montana universities," where a rich Carousel display appears that includes a list (instances) of all the universities in Montana with their logos, as well as a robust Knowledge Card display about the state in which they are located. This kind of display makes it clear that Google has verified each of those organizations as "university" entities located in the entity of "Montana" and is anticipating that the searcher will have questions about the state of Montana. Currently, the Semantic Web lacks similarly structured data about individual IRs from trusted sources.

DESCRIBING ITEMS ON THE SEMANTIC WEB

An adequate description of a library organization on the Semantic Web must be followed by descriptions of the items held by the library. The process of describing library items in a way that is helpful to search engines is no trivial task, and given the current infrastructure used by most libraries (i.e., OPAC and content management systems), syndication of library data

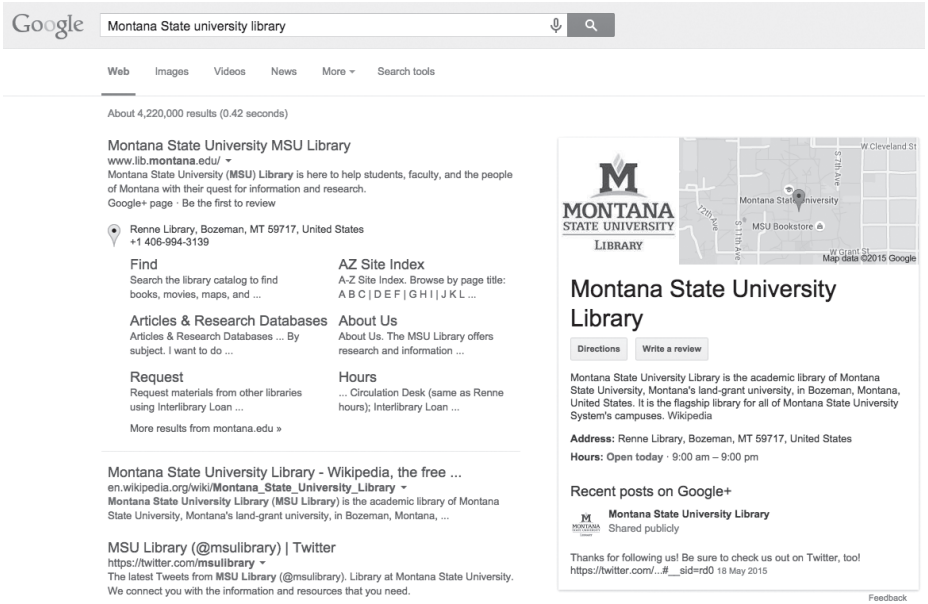

Figure 3.8. A Google search for "Montana State University Library" in 2015 displays a Knowledge Card with correct information about the organization.

can prove to be a difficult challenge. Libraries cannot just describe items on their Web sites using basic HTML because it is a markup language that is neither intended nor useful for semantic description. RDF (resource description framework) is a W3C standard designed to describe things on the Web in a way that allows machines to consume and understand the item. The model structures data in a simple sentence-like syntax (Mixter, 2014):

Subject => Predicate => Object

This framework allows for the structured description of things on the Web using domain-specific or general-purpose vocabularies. Domain-specific vocabularies tend to narrowly focus on a particular area of interest, such as bibliographic material, and have few ways of describing things outside of that domain. Domain-specific vocabularies are not always understood and consumed by search engines. General-purpose vocabularies, like Schema.org, were developed and published by search engines (Google, Yahoo!, Bing, and Yandex), so they were designed to describe a wide

variety of things on the Web and to be understood by those machines. Since its release in 2011, Schema.org has become the lingua franca for describing things on the Web. Using RDF as the basic framework and Schema.org as the vocabulary, libraries can describe their items on the Web in a format that allows search engines to understand, consume, and index the data.

Data Cleanup

With a basic understanding of Semantic Web infrastructure for syndicating data, IRs can begin to clean up existing metadata. For the purposes of this discussion, data cleanup refers to the process of turning string values into URIs (uniform resource identifiers) that can be dereferenced online. For example, a URI for Aldous Huxley, the author, is http://dbpedia.org/resource/Aldous_Huxley. Machines that follow the URI link will be presented with more structured data about the thing, such as a class (e.g., person, book, place) and its properties (e.g., name, birthdate, birthplace, occupation, etc.). Some of these properties themselves will be URIs that machines can follow to learn even more. This chain reaction allows search engines to place the initial thing, in this case the author Aldous Huxley, into a much broader context and understand how he connects to other entities on the Semantic Web.

The following list presents a basic library use case:

- A search engine crawls a library Web page (with structured metadata) for the book Brave New World. That Web page describes Aldous Huxley as the author of the book.
- The search engine follows the URI for Aldous Huxley and learns that he was born in http://dbpedia.org/resource/Godalming (Godalming, United Kingdom). The DBpedia link provides the search engine with additional information about Godalming.
- The search engine can also learn that Aldous Huxley wrote http://dbpedia.org/resource/The_Doors_of_Perception (The Doors of Perception). This type of information is used by search engines to help users discover other relevant items.

Semantic Web graph theory is explained well in a blog post published by Google that describes the Google Knowledge Graph and how it is different from traditional search engines (Singhal, 2012).

Existing metadata in an IR can contain errors and inconsistencies, and improving the quality of that metadata is a prerequisite to giving it the structure that is appropriate for search engines. Data cleanup can be done a variety of ways, but open source tools will be sufficient for most IRs, given the limited number of metadata records that IRs typically contain. OpenRefine is a tool that can import a variety of data formats such as Excel spreadsheets, TSV (tab-separated value) or CSV (comma-separated value) documents, and JSON (JavaScript object notation). Most repository software allows for export of data into these common data formats, so there should not be any need for costly or difficult initial format conversion. Once the dataset is loaded into OpenRefine, built-in tools can be used to clean up the data (Verborgh, De Wilde, & Sawant, 2013). Of particular importance is the reconciliation tool, which can be used to query string labels in metadata fields, such as "Aldous Huxley" against trusted entity datasets such as DBpedia.org. The services will automatically match and pull over the entity URI or if there are multiple matches, the user will be prompted to select the correct one. Figure 3.9 illustrates the high-level theory behind the reconciliation process that turns text strings into defined entities understood by search engines.

One of the most difficult tasks in converting legacy metadata into RDF data is converting the strings that do not reconcile into unique entities. In instances where strings do not match existing entities, libraries may need to create their own entity descriptions (Mixter, OBrien, & Arlitsch, 2014a). Once the dataset is cleaned up, it is ready for conversion into RDF.

Data Conversion

An RDF vocabulary must be applied before data can be converted to RDF. As previously mentioned, the RDF framework can be broken down into three basic parts: Subject; Predicate; Object. When this syntax is applied to data, the result is a triple in which two entities are connected by a property:

Machine-Readable Serialization:
<http://www.worldcat.org/oclc/2457589>
<http://schema.org/author>
<http://viaf.org/viaf/71392434>

Human-Readable Serialization:
"Brave New World" => authored by => "Aldous Huxley"

In the example above, the two entities are the book "Brave New World" and the person, "Aldous Huxley." They are connected by a property that indicates that the book was authored by the person. At a very basic level, an RDF vocabulary is used to describe things and the relationships between them. Figure 3.10 is a diagram of how an RDF vocabulary can be used to describe theses and dissertations.

An RDF extension (http://refine.deri.ie/) for OpenRefine can be used to apply an RDF vocabulary to an existing dataset (http://refine.deri.ie/rdfExport). After the mapping is complete, the dataset can be exported as RDF, at which point it is almost ready for syndication on the Web.

Data Syndication

After the dataset has been cleaned up and converted into RDF, there is still a need to serialize it on the Web so that search engines can consume it. This can be somewhat difficult because RDF has a variety of serializations that are geared toward different audiences, such as databases, humans, or machines. RDF is the underlying framework for all of the serializations, and conversion between them is seamless/lossless. However, search engines do not consume all serializations. Search engines prefer RDFa and JSON-LD serializations of RDF, and consequently, it is important for libraries to use one of these two serializations when they syndicate their RDF data on Web

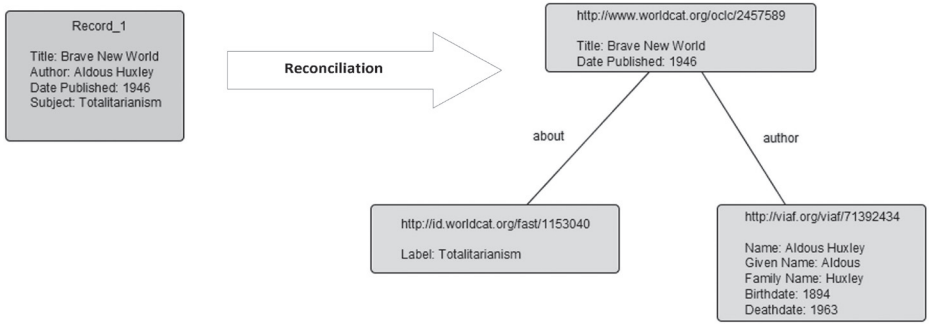

Figure 3.9. Converting records to entities.

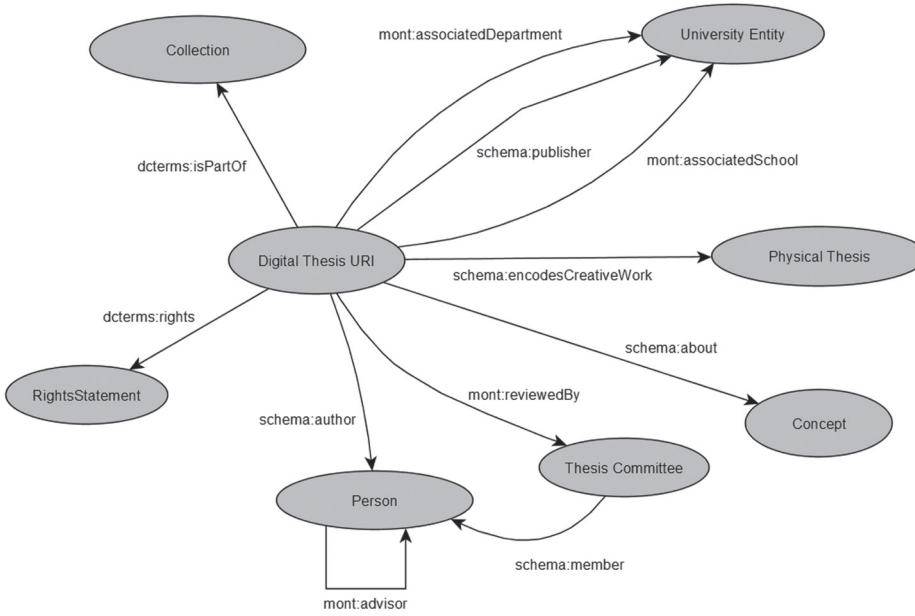

Figure 3.10. Concept map for theses and dissertations (Mixter, OBrien, & Arlitsch, 2014b).

pages (Google Developers, 2015). RDFa is a W3C recommendation serialization that uses HTML tags and attributes to encode RDF data. Since RDFa uses HTML, it is a natural choice for syndicating RDF on Web pages, but RDFa can be difficult to construct and debug. JSON-LD is a W3C recommended serialization and can be embedded directly into Web pages the same way that JavaScript is embedded. Although JSON-LD is easier to embed on Web pages than RDFa, there is a concern that search engines will not trust all JSON-LD markup, since the semantic data are not visible to human users of the Web page. Google recommends using JSON-LD for specific types of entities (e.g., Events) but otherwise recommends using RDFa for semantic markup (Google Developers, 2015). Regardless of which serialization is chosen for syndication, libraries will need to make sure that there is a mechanism in their content management systems for displaying serialized data on Web pages. In addition to syndicating the RDF data about the bibliographic items, there is also a need to store and syndicate the data about entities that were locally created, such as students, faculty (that

do not exist in VIAF, ORCID, or ISNI), local subject headings, and so on. These entities can be stored in a local triple store and syndicated using open source software such as Pubby (http://wifo5-03.informatik.uni-mannheim.de/pubby/). Once the RDF is syndicated, it is prudent to check that items are described and displayed well and that the syndication is recognized and consumed by search engines.

SUMMARY

Although IRs preserve a wealth of knowledge, much of the content remains hidden to Internet users because of poor or inconsistent discovery by external search engines. This chapter has focused on some SEO techniques that can help improve discovery of IR content by search engines, and these include structured metadata applied consistently and accurately for a variety of user agents, user experiences, cover sheets, and accessible site structures. It also described some techniques IR managers can employ to participate in entity-based search on the Semantic Web. Librarians would do well to become familiar with Semantic Web Identity and be more active in helping to develop robust entity definitions of IR and related library concepts in data sources trusted by search engines. IR should add a layer of Linked Data, which will help improve comprehension for humans and machines.

Linked Data entities will grow organically as items in repositories are explicitly defined and linked to other data sets. As the ecosystem evolves, machines will more clearly understand what an IR is, what it contains, and the value in directing users to trusted information sources. Publishing IR content as Linked Data will increase the number of connected entities on the Semantic Web, increasing the value and meaning of each data point as it is connected to other entities on the Web. Consistent application and practice of these SEO and Semantic Web techniques will help ensure that IR content is discoverable on the Web.

REFERENCES

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech, 30*(1), 60–81. http://dx.doi.org/10.1108/07378831211213210

Arlitsch, K., & OBrien, P. S. (2013). *Improving the visibility and use of digital repositories through SEO*. Chicago, IL: ALA TechSource, an imprint of the

American Library Association. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=578551

comScore, Inc. (2014, April 15). comScore releases March 2014 U.S. search engine rankings. Retrieved from http://www.comscore.com/Insights/Press-Releases/2014/4/comScore-Releases-March-2014-U.S.-Search-Engine-Rankings

Dean, B. (2014, August 8). Google's 200 ranking factors: The complete list. Retrieved from http://backlinko.com/google-ranking-factors

Dempsey, L., Malpas, C., & Lavoie, B. (2014). Collection directions: The evolution of library collections and collecting. *Portal: Libraries and the Academy, 14*(3), 393–423. http://dx.doi.org/10.1353/pla.2014.0013

DeRosa, C., Cantrell, J., Carlson, M., Gallagher, P., Hawk, J., & Sturtz, C. (2010). *Perceptions of libraries, 2010: Context and community* (p. 108). OCLC, Inc. Retrieved from http://www.oclc.org/reports/2010perceptions.htm

Google Developers. (2015). About Schema.org. Retrieved from https://developers.google.com/structured-data/schema-org

Google Inc. (1999, November 5). Google: Company info. Retrieved from https://web.archive.org/web/19991105194818/http://www.google.com/company.html

Google Inc. (2015a). Steps to a Google-friendly site. Retrieved in 2014 from https://support.google.com/webmasters/answer/40349?hl=en

Google Inc. (2015b). Webmaster guidelines: Quality guidelines. Retrieved in 2014 from https://support.google.com/webmasters/answer/35769#quality_guidelines

Google Scholar. (n.d.a). Inclusion guidelines for webmasters. Retrieved in 2011 from http://scholar.google.com/intl/en/scholar/inclusion.html

Google Scholar. (n.d.b). Inclusion guidelines for webmasters: Content guidelines. Retrieved in 2011 from https://scholar.google.com/intl/en-US/scholar/inclusion.html#content

Google Scholar. (n.d.c). Inclusion guidelines for webmasters: Troubleshooting. Retrieved in 2014 from http://scholar.google.com/intl/en-US/scholar/inclusion.html#troubleshooting

Mixter, J. (2014). Using a common model: Mapping VRA core 4.0 into an RDF ontology. *Journal of Library Metadata, 14*(1), 1–23. http://dx.doi.org/10.1080/19386389.2014.891890

Mixter, J., OBrien, P., & Arlitsch, K. (2014a). Describing theses and dissertations using Schema.org. In *Proceedings of the International Conference on Dublin*

Core and Metadata Applications. Austin, TX. Retrieved from http://dcevents.dublincore.org/public/dc-docs/2014-Master.pdf

Mixter, J., OBrien, P., & Arlitsch, K. (2014b, October). Describing theses and disserations using Schema.org. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Austin, TX. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/269/313

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the Web* (No. SIDL-WP-1999-0120). Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

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

Verborgh, R., De Wilde, M., & Sawant, A. (2013). *Using OpenRefine*. Birmingham, UK: Packt Publishing.