Reframing Description

A Practical Introduction to Linked Data and the Bibliographic Universe

Brinna Michael, Emory University

ABSTRACT Linked data has been on the tongues and minds of librarians for years, but the concept still manages to remain somewhat mysterious and unreachable to the broader technical services community. Recognizing the broad divide between the complex conceptual debates and the reality of practical application, this session seeks to bridge this gap. In providing an entry-level introduction to the practice of using linked data to represent bibliographic descriptions, this session also seeks to spark interest in further participation in the wider linked data movement.

Linked data as the practice of encoding information so that it can be contextually connected to other information by defining relationships in a way that is machine readable/actionable. By defining and identifying data in this way, it is possible for machines to begin processing that information in a way that mimics the manner in which human brains draw connections between concepts. But why is this important? Traditional methods of encoding and recording bibliographic data store concepts as strings and do not provide a means for the machine to independently extract meaning from those strings. As illustrated in figure 1, the human brain can see a set of four completely different strings and 1) recognize that they are all actually representing the same idea, and 2) recall other concepts related to the strings. Given that same list, a machine algorithm will only recognize the strings as completely separate entities unless it is provided with an encoding framework that defines them otherwise.

It is this divide between human cognitive and machine data processing that linked data practices strive to bridge. From the beginning, this goal has been intrinsically tied to the concept of the Semantic Web, defined by Oxford English Dictionary (3rd ed. 2014) as "a proposed development of the World Wide Web in which data in web pages is structured and tagged in such a way that it can be read directly by computers . . ." However, Tim Berners-Lee and his

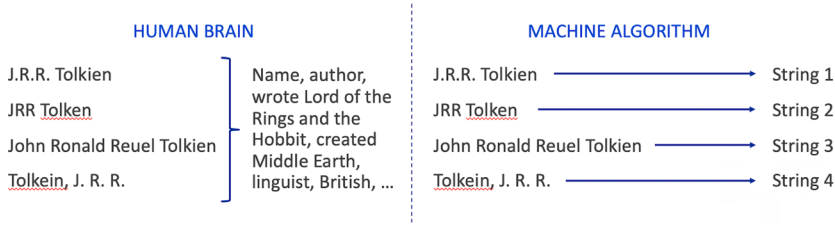

FIGURE 1: Comparison of human cognitive and machine data processing of strings.

colleagues (2001) took this a step further, presenting the Semantic Web as "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

If the Semantic Web is the end goal, linked data practices are the means to attain it. In his work to realize a functional Semantic Web, Berners-Lee (2009) defined the following core requirements for linked data:

1) Use URIs as names for things.
2) Use HTTP URIs so that people can look up those names.
3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL, etc.).
4) Include links to other URIs so that they can discover more things.

The first requirement, to use URIs (Universal Resource Identifiers) as names for "things," including concepts, ideas, people, places, objects, and more, is the core of functional linked data. URIs can be any string of characters and/or numbers, but the most important aspect is that the URI must be unique to the thing it is naming. Requirement two further specifies that URIs should be HTTP compliant, making them searchable on the internet. This compliance lays the groundwork for linking these URIs in a way that is machine-actionable. The third and fourth requirements give linked data its purpose and function. Simply assigning URIs to things is important but, without additional context, they lose functionality. To make that contextualization usable by machines, and further the goal of improving human/machine cooperation, using encoding standards like RDF or SPARQL is critical. Finally, connections and links must be made between URIs to build a system of pathways for machines to follow to better facilitate discovery. Without creating these initial links

and thereby defining these relationships, the data remains static and unusable.

LINKED DATA AS STRUCTURE

As previously mentioned, traditional methods of encoding bibliographic data have taught us to think of that data in terms of records—flat, isolated, and often strictly string-based. Each record is separate from each other and does not offer a machine-readable connection to other records. Linked data, however, operates as a network structure that is, by nature, dynamic and interconnected, utilizing the machine actionability of HTTP-compliant URIs. A network structure allows for easier use and reuse of data in reference to each other.

To illustrate this, consider the following scenario: below are two partial MARC records for books that share a specific set of characteristics, including subjects, content type, mediation type, and carrier type.

Table 1: Comparison of Related MARC Records

BOOK 1		BOOK 2	
FIELD	CONTENT	FIELD	CONTENT
100	$a Tolkien, J. R. R. $q (John Ronald Reuel), $d 1892-1973.	100	Wood, Ralph C.
245	$a The lord of the rings / $c J.R.R. Tolkien.	245	$a The gospel according to Tolkien : $b visions of the Kingdom in Middle-earth / $c Ralph C. Wood.
250	$a Seven volume edition; Millennium edition.		
264	$a London : $b HarperCollins, $c 1999.	264	$a Louisville, Ky. : $b Westminster John Knox PRess, $c [2003].
264	$c ©1966	264	$c ©2003
336	$a text $b txt $2 rdacontent	336	$a text $b txt $2 rdacontent

Table 1: Comparison of Related MARC Records			
BOOK 1		BOOK 2	
FIELD	CONTENT	FIELD	CONTENT
337	$a unmediated $b n $2 rdamedia	337	$a unmediated $b n $2 rdamedia
338	$a volume $b nc $2 rdacarrier	338	$a volume $b nc $2 rdacarrier
600	$a Baggins, Frodo $v Fiction.	600	$a Tolkien, J. R. R. $q(John Ronald Reuel), $d1892-1973. $t Lord of the rings.
		600	$a Tolkien, J. R. R. $q (John Ronald Reuel), $d 1892-1973 $x Religion.
650	$a Middle Earth (Imaginary place) $v Fiction.	650	$a Christianity and literature $z England $x History $y 20th century.
		650	$a Fantasy fiction, English $x History and criticism.
		650	$a Christian ethics in literature.
		650	$a Middle Earth (Imaginary place).

As MARC records, these similarities can only be connected by use of string-matching methods, which rely on flawless data entry to create consistent matches. Such a need is partially responsible for the robust and strict formatting and data entry requirements set forth by the MARC standard and RDA (Resource Description and Access).

The first step towards reframing these two records as a network of interconnected data points is to identify the common concepts between the two. In table 2, we can see the concept or value and its relationship to the book it is describing. Note that in the case of "Tolkien, J. R. R. (John Ronald Reuel), 1892-1973," the relationships between the name and the books it is describing are different, yet it still constitutes a point of connection between the two books.

Table 2: String and URI descriptive values in relation to two books

STRING VALUE	URI	BOOK 1 Relationship	BOOK 2 Relationship
Tolkien, J. R. R. (John Ronald Reuel), 1892-1973	http://id.loc.gov/authorities/names/n79005673	author	subject
text	http://id.loc.gov/vocabulary/contentTypes/txt	content	content
unmediated	http://id.loc.gov/vocabulary/mediaTypes/n	mediation	mediation
volume	http://id.loc.gov/vocabulary/carriers/nc	carrier	carrier
Middle Earth (Imaginary place)	http://id.loc.gov/authorities/subjects/sh85085022	subject	subject

With these values identified, the next step is to locate or create a unique URI to represent the concept. In the case of this example, each of these values has already emerged from existing controlled vocabularies—courtesy of the Library of Congress (LC)—which have been fully converted to linked data in an effort to expand linked data practices within the wider cultural heritage community. In the URI column of table 2, you can see the URIs that represent each of the chosen values.

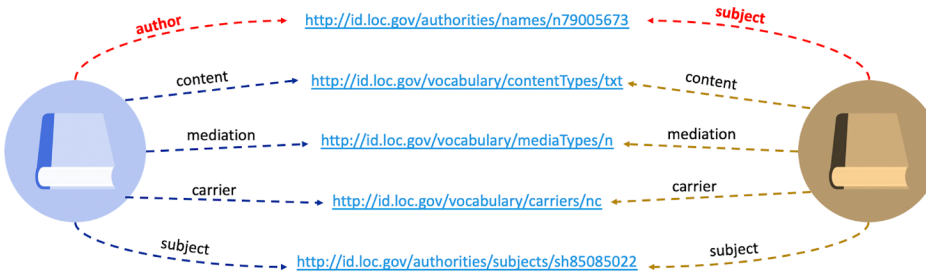

FIGURE 1: Network of shared relationships between two books and five descriptive values. School textbook, textbook icon by Boca Tutor, found at www.iconfinder.com/icons/1741323/school_textbook_textbook_icon, used under a CC BY-SA 3.0 license.

The next step is to create links between the two books being described by defining their relationships to the URIs that represent the identified descriptive values. Figure 1 provides a visual representation of the network structure of these relationships. In the diagram, the book icons represent the core URIs that would represent the physical books being described. The labeled arrows denote URIs for the predefined relationships as described in common schemas such as Dublin Core or Schema.org. Finally, the central column lists the URIs for the previously identified descriptive values.

These structures, known as triples, are the foundation of linked data, and can be expanded almost indefinitely through the use of URIs. For example, figure 1 represents a fragment of the wider network to which these two metadata sets belong. If we expand our view of the network describing these two books (figure 2), we see additional descriptive URIs that are not shared with one another but may link each book to even more resources. Each of these relationships could start its own networked pathway of connections to a wide web of other ideas and resources. However, those relationships must first be defined and the links created, so that a machine can be guided through the logical thought process that human brains follow naturally.

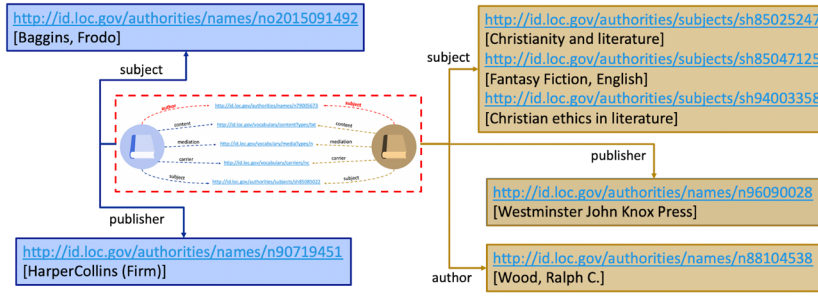

FIGURE 2: Extended network of relationships between two books and descriptive values. School textbook, textbook icon by Boca Tutor, found at www.iconfinder.com/icons/1741323/school_textbook_textbook_icon, used under a CC BY-SA 3.0 license.

LINKED DATA AS CONTENT

Having laid out the way in which structuring the relationships between data creates a wider network, it is important to consider how the actual content of the data works to further the goal of human/machine cooperability. Using one of the descriptive values

from before, it is important to remember that, in assigning names, strings, or URIs to things in the real world, a placeholder is being created for that thing in the digital environment.

For example, J. R. R. Tolkien was a real person who lived, worked, and wrote. He, as a specific and unique person, is represented by LC with the string "Tolkien, J. R. R. (John Ronald Reuel), 1892-1973" and the URI, http://id.loc.gov/authorities/names/n79005673. But how are these two connected? The string is human readable, yet a machine cannot parse that it is made up of a surname, initials, a first and middle names, and his dates of birth and death. On the other hand, the URI is machine readable and actionable, yet there is no way a human would know it was a place holder for J. R. R. Tolkien if presented only with the URI.

This is where the contextualization of data, those third and fourth core requirements laid out by Berners-Lee, becomes important. Table 3 presents an excerpt from the RDF (Resource Description Framework) document underlying the human display version of the entry for Tolkien in the LC's linked data service:

Table 3: RDF serialization of LC term "Tolkien, J. R. R. (John Ronald Reuel), 1892-1973

	SCRIPT[a]
1	<rdf:RDF xmlns:skos="http://www.w3.org/2004/02/skos/core#"
2	xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3	xmlns:rdfs="http://www.w3.org/1999/02/22-rdf-schema#"
4	xmlns:cs="http://purl.org/vocab/changeset/schema#"
5	xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
6	<rdf:Description rdf:about="http://id.loc.gov/authorities/names/n79005673">
7	<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
8	<skos:prefLabel>Tolkien, J. R. R. (John Ronald Reuel), 1892-1973</skos:prefLabel>
9	...

a Taken from the "SKOS – RDF/XML" format available for download at id.loc.gov/authorities/names/n79005673.

Lines 1–5 consist of the declaration of the encoding schemes which give definition to the relationships being drawn—in this case, RDF, RDFS (RDF Schema), SKOS (Simple Knowledge Organization System), SKOS-XL (an extension of SKOS), and Changeset. Declaring namespaces in this way defines a short prefix to represent the base URI for the identified schema throughout the remainder of the document. Line 6 declares what concept will be described by the document, in this case the URI http://id.loc.gov/authorities/names/n79005673. Finally, in line 8, the "skos:prefLabel," a human-readable label, is assigned to the URI as "Tolkien, J. R. R. (John Ronald Reuel), 1892-1973." These three aspects work together to embody a human and machine cooperative representation of Tolkien in a digital environment, which can be seen when following the active URI link.

LINKED DATA AND BIBLIOGRAPHIC DESCRIPTION

Although linked data may seem far away as a daily practice, there are already aspects of that practice ingrained into common cataloging and descriptive metadata procedures. Use of controlled vocabularies has been common and required for decades in most standard descriptive best practices. In fact, it would only take a small leap from using the string values found in controlled vocabularies to using URIs from those same vocabularies to begin satisfying Berners-Lee's first core requirement. A large number of vocabularies encoded as linked data already exist, including the majority of LC's vocabularies, the Getty vocabularies (Art and Architecture Thesaurus, Union List of Artist Names, etc.), FAST (Faceted Application of Subject Terminology), VIAF (Virtual Internet Authority File), and many more domain-specific vocabularies. In recent years, catalogers also began experimentally use $0 and $u in MARC records to integrate actionable URIs into records, beginning the process of shifting descriptive norms away from simply providing the string placeholder to providing both a string and URI. Finally, the very practice of copy cataloging reflects aspects of linked data practice, namely the improvement of efficiency in descriptive practices as well as the reduction of unnecessary duplication of work and data.

With this in mind, what changes might we expect to see as the world of bibliographic description shifts further towards a linked data baseline? We can certainly expect changes within cataloging software, including back-end integration with controlled vocabu-

laries encoded as linked data, potentially in the form of dropdowns or search suggestion features in the cataloging interface. Cooperative cataloging platforms, like OCLC Connexion, may look slightly different, perhaps with an interface that more closely resembles a webform, rather than the familiar MARC record input screen. Alongside the changes to cataloging interfaces, discovery layers may begin to include more options for customization to enable users to customize their search experience in ways that were previously impossible, including improved search filters and, potentially, the ability to select the desired display language. Additionally, with bibliographic data encoded as linked data, bibliographic information will be able to be exposed via discovery layers to large search engine algorithms, making our holdings more openly accessible to potential users on a broad scale. Finally, changes to our conceptual cataloging and metadata frameworks have already been seen and will continue to be seen, including changes from a strictly FRBR (Functional Requirements for Bibliographic Description)-based RDA to an IFLA-LRM (Library Reference Model)-based RDA. Alongside these changes, there will be an increased need to work towards assigning URIs to some of the more complex conceptual aspects of bibliographic entities, like Work and Instance.

Such changes require a wide variety of technical services roles to support them. Software developers and programmers will be called on to develop this new front- and back-end software to support cataloging and discovery. As that software is developed, systems managers and systems librarians will need to keep pace with maintaining and supporting the functionality of these systems. Catalogers and metadata librarians will have to adapt to new input mechanisms and accept a level of disruption in their workflows and consistency, which accompanies the shifting of technologies and conceptual models. Luckily, many technical services individuals have experienced shifts from one ILS (Integrated Library System) to another—an experience which is likely to feel quite similar to a transition into a linked data system. Most importantly, it is important to recognize that such a transition is an extended process, requiring interdepartmental cooperation and a willingness to contribute and step outside our comfort zones.

CONCLUSION

We can already see aspects of a linked-data, bibliographic future coming to life. ExLibris has begun experimenting with automatically generating BIBFRAME (Bibliographic Framework) records from existing MARC records in the Alma system (ExLibris, "BIBFRAME"). These generated records include a number of automatically encoded URIs for descriptive metadata, including the resource language, content type, and mediation, which are drawn from LC's linked data service. Even some validated headings, such as names and subjects, have URIs pulled from VIAF and the LC Subject Headings and Name Authority Files.

On the search and discovery front, linked data will connect libraries with a wider audience through major search engines. For example, when searching for a book on Google, a helpful sidebar is included on every results page. All the information contained in that sidebar is enabled through some form of linked data, drawing on sources such as Wikipedia, Google Books, Goodreads, and more. Recently, a search for *Lord of the Rings* showed a new section in this sidebar entitled, "Borrow." Under this section, information on e-book holdings from local libraries was displayed with a link to the library's catalog record.

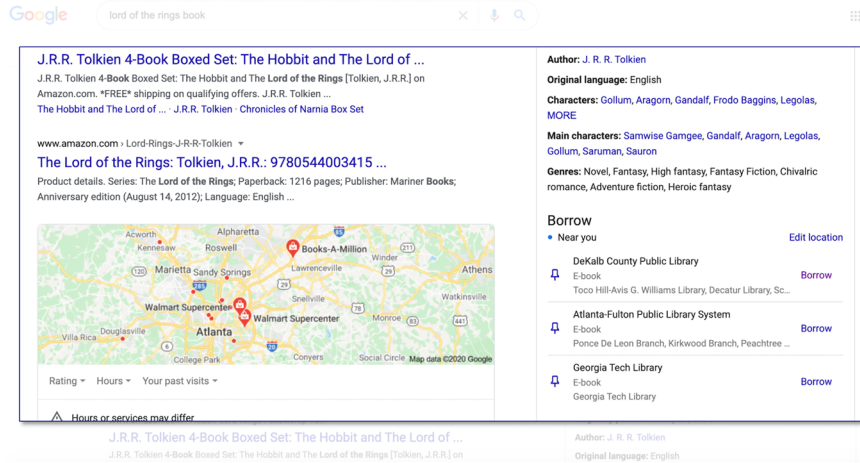

FIGURE 3: Extended network of relationships between two books and descriptive values.

For now, it seems that third-party providers, such as Hoopla and Overdrive, are responsible for enabling this exposure of bibliographic and holdings description to search engines. Nevertheless, it shows a very real opportunity for connecting with users who may not be comfortable searching a library catalog.

While a fully linked-data bibliographic world is still years away, the process of shifting away from the static, records-based processes that have been the staple of descriptive procedures for decades has already begun in full force. In practice, the creation of linked-data networks and the use of linked-data behaviors are not so far from what has been the current norms, but we must be willing to embrace these changes fully as we move forward.

WORKS CITED

Berners-Lee, Tim. "Linked Data." Updated June 18, 2009. www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities." *Scientific American*, May.

ExLibris. n.d. "BIBFRAME." *ExLibris Developer Network*. Accessed June 30, 2020. developers.exlibrisgroup.com/alma/integrations/linked_data/bibframe.

Library of Congress. n.d. "Tolkien, J. R. R. (John Ronald Reuel), 1892-1973." Updated August 9, 2019. id.loc.gov/authorities/names/n79005673.