Linked Data in Library Workflows

Anne Washington
Presenter

Jeff Mixter
Presenter

Abstract

As the library community continues to explore linked data for cataloging and data management, there are a variety of new opportunities as well as challenges to ensure that data is interoperable and manageable across bibliographic datasets. Data model interoperability is critical to ensure we can continue to manage bibliographic materials at scale across the wide range of material types and metadata providers and consumers. Additionally, the interoperability of the linked data that connects our bibliographic materials together—traditionally seen as authorities—must be evaluated and managed to improve the way users find and interact with library materials. To meet the challenges, OCLC, in partnership with others, is building a robust and interoperable set of linked data that can be used in future-facing data management workflows and discovery interfaces.

Keywords: cataloging, metadata, linked data, authorities, continuing resources, discovery

Background

Library professionals have been experimenting with linked data for over two decades. The initial motivation for migrating to Resource

Description Framework (RDF) linked data was to help facilitate the transition from the Machine-Readable Cataloging Record (MARC) standard to a standard that could be more easily used and understood on the web. Consequently, early library linked data modeling efforts focused on replacing MARC. Work that followed—and that which continues—was the creation of library models and formats, then the translation and transformation needed to move to and between these new models. This work has had its biggest impact on library cataloging workflows.

Cataloging

When evaluating cataloging from a library workflow perspective, consider its three primary purposes: inventory control, registration, and knowledge work. Inventory control involves traditional supply chain activities, but also curating the data in a way that is relevant to the library's infrastructure and the communities served. Registration, similar to inventory control, is focused on traditional supply chain management and involves a library registering one or more inventory items, such as books, to an inventory control record that describes them. These purposes alone do not warrant investment in linked data and transitioning away from MARC. For these two activities, the value of linked data is efficiency.

Linked data is inherently multi-lingual, which helps resolve the data management concern around language when cataloging a MARC record. Linked data entities can also connect to a wide variety of additional library-controlled vocabulary identifiers, which provide the potential to access vocabulary information more dynamically, like the most current preferred label for a subject heading. From an efficiency standpoint, this lessens the burden of cataloger localization of MARC metadata records.

Knowledge Work

The third purpose, knowledge work, is the creative process library staff engage in to create and curate data that is critical for library workflows.

It is the practice of describing library materials and their relationships to other entities—such as people, organizations, concepts, places, and events—to facilitate end-user discovery and, increasingly, to enable library workflows other than those related to cataloging. The latter use of knowledge work is of particular importance because it bridges currently disparate but interconnected library workflows and therefore provides an opportunity for streamlining work for library staff. The result of linked data providing efficiency gains to inventory control and registration is that catalogers can dedicate more time to knowledge work; providing better metadata for discovery services and curating data that bridges library workflows.

Broaden Impact

Linked data broadens the impact of knowledge work beyond cataloging. Catalogers and other metadata professionals contribute not only bibliographic data, but data about other things—or entities—related to those resources: people, organizations, places, subjects, and events. By contributing to a growing, shared graph of data, knowledge workers collectively build re-usable connections between resources and to other entities.

New workflows such as researcher information management (RIM) and scholarly output data management are becoming central to libraries and their stakeholders. Linked data provides library staff with new opportunities to both expand what they can say about information materials while also aggregating data across the library data ecosystem. Shared models across workflows and material types bridge existing workflows—such as bibliographic, and scholarly communication workflows—and enable new ones as well.

Better Connect Existing Data

Library data has expanded dramatically since the advent of modern cataloging. Library materials are broadly categorized into print,

electronic, and digitized formats. Library administration and organizational structures often reflect these delineations, and these items have traditionally been cataloged and managed as siloed datasets. Although they might be pulled together in a unified discovery layer, the disparate nature of the metadata management causes these resources to not share strong or meaningful cross-silos relationships.

Using linked data, making more connections between resources and across silos is possible. Figure 1 illustrates an example of this. In the center is a print book, *You Should Have Been Here an Hour Ago* by surfer Phil Edwards. Phil Edwards is a linked author heading in the metadata. A general search on Phil Edwards also returns the "Surfin' Shorts" collection of short films, where his name appears in the description but not as a linked access point. Similarly, the search returns an image from Pepperdine University Libraries digital collections without a linked access point to Phil Edwards. The surfboard image metadata also includes the company, Hobie, that created the surfboard. Searches on Hobie return an article on the founder, Hobie Alter, and the Hobie company's business forecast. Discovery systems pull these results across formats, but their connections are tenuous, based on strings.

Using unambiguous identifiers for Phil Edwards, Hobie Alter, and the Hobie surfboard company in the metadata better connects these

Figure 1. Linked data can help connect library materials across print, electronic, and digital silos

resources across data silos and improves the ability to return relevant and related search results across materials. It also provides a way to incorporate information about the person and business entities—for example, birth dates, places of residence, founding dates—into the search and discovery experience. Building identifiers into data and workflows across publishers, content providers, aggregators, and library departments is critical to building a graph of these connections.

Say What You Need to Say

Linked data provides an underlying structure that expands what can be said about bibliographic materials as well as their related entities: people, organizations, places, subject, and more. The MARC cataloging format outlines what can be said about a particular resource in a single record. A cataloger may have more information that is important to capture, but they are limited by the scope of the MARC record and often use authoritative strings of text over identifiers. Catalogers may be able to capture this additional information in an authorities or scholarly communication workflow, but it is not always possible to bridge these workflows or to bring that data together in ways that are needed for cataloging or discovery. By using flexible data models, permanent entity identifiers, and interoperable services, catalogers and metadata professionals can say what they need to say about the bibliographic resources, organizations, people, chronology, and more in a seamless way that enables connections across workflows.

Continuing resources catalogers often have additional context to add to their MARC records to capture changes to the title, issuing bodies, and topics. Capturing this information in ways that can truly link resources, people, and organizations and be used across different workflows is possible through linked data.

Figure 2 is an example of these connections in the context of serial publications. Title A is published by Issuing Body A with a focus on Topic A. Later, the title changes to Title B, with no change to the issuing

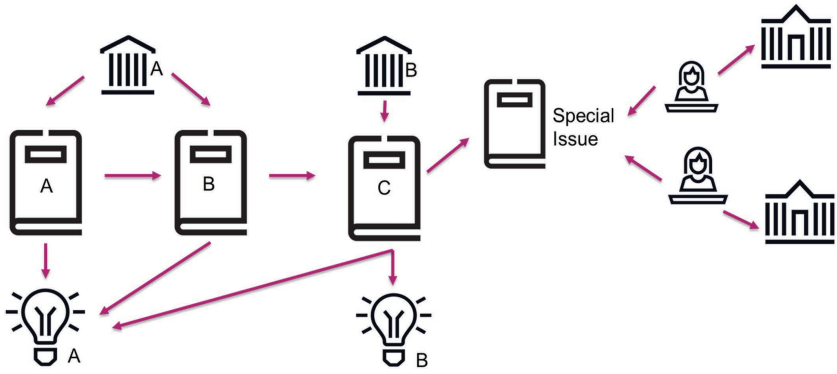

Figure 2. Linked data graphs connect issuing bodies, serial titles, topics, special issues, and editors

body or topic. Next, the title changes to Title C, which is now issued by Issuing body B, and an additional Topic B is added. Title C has a special issue with two co-editors. Each editor is affiliated with a research institution. Building and exposing this information as a graph of data allows for connections made between the special issue authors and Topics A and B, the research institutions and the serial titles, and makes it possible to find articles by searching across all of the serial titles.

Make Data Work Harder

RDF natively supports multilingual metadata, which makes it possible to more easily search and view data in a user's preferred language. Because a representation of a given entity can contain multiple languages, it is not necessary to create and maintain authority records for each language. A single entity for a subject or organization, for example, can have a preferred or alternative label in any language. As Figure 3 illustrates, instead of maintaining three separate authorities for the World Health Organization—one each for English, Spanish, and Chinese—a single linked data entity includes a preferred label and description in each language. An application can query for and return the label in the requested language.

Figure 3. One representation of a given entity can contain labels and other string data in multiple languages

Shi and Donathan's analysis of issues in CrossRef data highlights the challenges of organization names in scholarly article data.[1] Differences in spelling and acronyms for the same entity in multiple languages may make it difficult to connect an author with their affiliated organization and make it practically impossible to bring these works together. Using an identifier for the organization with labels in different languages better connects that author to an organization and makes it possible to find works related to a given institution. Working with publishers, library staff, and aggregators to incorporate identifiers for researchers and their affiliations will enable accurate attribution and broader reach for publications.

Provide More Meaningful Discovery

When one looks at the various parts of knowledge work, graph-based data can provide a more meaningful discovery experience for end users. Continuing resources library staff provide rich, complex data about a publication's relationships to earlier and later titles, and to issuing bodies and topics. Although this information is captured in MARC records, these connections have not been used to their full potential. Linked data models unlock this data to make it truly linked and actionable. In a graph-based discovery system, users could search across all titles of a given serial, instead of searching each title

separately, saving them time and increasing the chances they will find what they are looking for.

The contextual entities used to connect resources in support of the cataloging workflow and other workflows, such as a researcher information management workflow, can be used to inform search algorithms and provide users with related resources. Importantly, the incorporation of the connected contextual entities can be used to not only provide end users with related materials, but also provide context or reasoning behind why they were suggested to the user.

Contextual entities such as subjects, and academic departments could be used to provide a library user with related results based on a found article. Current discovery systems already have features to provide related items for library users. But in most cases, the results are based on data features that are opaque to the library users (and sometimes even the people building the discovery application).

In a graph-based system, the application could use the related contextual entities to not only provide related items to library users, but also an explanation of how they are related. For example, the application could present items related to the article the user found and provide context, for those related items: they were published by an author who worked for an academic institution that participated in the same grant as the author; they share a common subject; or were published near the time period in which the grant was given. This type of relationship tracking is exactly what graphs are designed to do. New, expansive data models that can represent relationships across domains and material types allow for the full benefit of connected library graph data.

OCLC's Current and Future Work

OCLC is working to build a linked data ecosystem that can optimize current library workflows, bridge disparate workflows, and enable new workflows, all with a focus on knowledge work. Linked data is not an

end unto itself. Instead, it is a means to an end, to realize the full potential of the rich resource descriptions and connections that library staff create in order to support library workflows and fulfil users needs. Keeping the ends in mind helps us better understand how to produce models and data that help library staff where and when they are doing their work.

Acknowledgments

We thank our OCLC colleagues Charlene Morrison, Laura Ramsey, and Hanning Chen, for their helpful insight and support as we developed the NASIG presentation and paper.

Contributor Notes

Anne Washington is a Product Analyst at OCLC.

Jeff Mixter is a Senior Product Manager at OCLC.

Note

1 Julie Shi and Dennis Donathan II, "Metadata for Everyone: Identifying Metadata Quality Issues Across Cultures," (presentation, NASIG Annual Conference, Pittsburgh, PA, May 24, 2023).