LTS and Linked Data: a position paper

Prepared for Cornell University Library Technical Services (LTS)
by Chew Chiat Naun, Jason Kovari, and Steven Folsom
December 16, 2015

Introduction

This document outlines motivations for adopting linked data techniques for describing and managing our collections, and seeks to articulate a specific role for Library Technical Services (LTS) within this enterprise. Engaging in linked data experimentation, data modeling and tool development offers LTS an opportunity to augment and, over time, improve upon existing methods to meet Cornell University Library's (CUL) strategic priorities and major initiative areas.

Linked data is a paradigm for making machine-actionable data and not just human-readable documents fully accessible and repurposable anywhere on the Internet. It uses common Web communications protocols as used by ordinary browser software and search engines to connect machine-readable, structured data across distributed computers. But it also connects the data *meaningfully*. Semantic modeling of linked data allows formal definitions to be made not only within but also between ontologies used in different domains. For example, two terms from different vocabularies that have related meanings can be asserted to be equivalent, or to be hierarchically related; or a relationship can be stated to hold between an item described using one vocabulary with an item described in another. Systems can then be designed or adapted to take advantage of these assertions.

By facilitating interoperability, linked data offers potential solutions for a range of problems that libraries have until now assumed that they and their users would simply have to accept. The following examples are very typical.

- A user finds the works of an author in the library catalog, but has to start over to identify the same author in an article index, because the catalog and the index do not refer to the author by the same name.
- A library wants to offer its users unified discovery of dissertations in different formats from different sources. The systems developer finds a way to provide retrieval of

dissertations from a database, but has to research and then write additional code to account for the very different conventions used to describe similar materials in the MARC catalog.

- In order to support a faculty member's research interests, a selector wants to assess the strength of the library's collection in publications from a particular region within a certain country. A collection analyst tries to find everything published there but first has to research the names of all the distinct places that are located within that region.
- The library expands its acquisitions from a foreign country, and discoverability of these materials is greatly enhanced by the availability of catalog records from their national library. However, this national library has its own authority file and the data it contains, which includes valuable information such as vernacular names, cannot be integrated into our discovery system due to existing conventions that allow only Library of Congress Name Authority File (LCNAF) data.

Linked data makes it feasible to solve problems such as these because it is based on the idea of entity references and descriptions that are shared across disparate systems. In the above examples, linked data solutions have the potential to benefit end users and library staff alike - the former through improved discovery and patron services, the latter through greater consistency and more efficient management of data.

At this stage of its development, there is no guarantee that linked data will live up to the most optimistic claims made for it. As with all advances, a great deal depends on the institutions, practices, and specific implementations that will evolve around the technology. It is also worth emphasizing that some of these needs may be met by solutions that are not based on linked data technologies. However, it is clear that deployment of linked data has the potential to offer significant improvements in services for library users.

In pursuing our research agenda we propose to test the following hypotheses:

- Linked data can improve discovery and provide greater context to library collections
- Linked data and related technologies can achieve targeted enhancements of workflows within LTS
- Linked data methods can be applied successfully to a wide range of collection types, e.g. special and general collections, owned and licensed content

- Creating linked data natively can have a stronger impact on users and production processes than relying on transformation of legacy data

In summary, linked data implementation will provide an opportunity to reassess the costs and benefits associated with describing our collections.

Production Incentives

LTS Priorities, 2013-2015[1] and the Cornell University Library Strategic Plan *Toward 2015*[2] both call for CUL to develop creative methods for how the Cornell community finds and uses the full scholarly record available in our library collections and more broadly on the web. Library Technical Services are increasingly asked to focus on descriptive activities in areas outside the traditional catalog, e.g., institutional repositories and digital collections for archival materials. At the same time, we need to make our traditional collections more discoverable locally and more visible on the web. Linked data, which is built from the outset to be extensible and interoperable, provides one of the keys to meeting both objectives. Adopting linked data techniques can better position LTS to meet CUL's evolving mission.

The search engine community and others are employing linked data to address large-scale aggregation and analysis of data. By leveraging established identities, linkages can go further, whether through more richly linked displays or active suggestion. Projects like Google's and Yahoo's Knowledge Graphs, Bing's Snapshot and Facebook's semantically-aware search offer users enhanced discovery through semantic data. There will also be lessons to learn from other quarters, such as the Linked Open Data in Libraries, Archives and Museums (LODLAM)[3] community. The British Broadcasting Corporation's (BBC) online content and Europeana's site are two examples of mature implementations that provide added value through the use of linked data[4][5][6][7][8]. For example, the BBC's page for the band "The Specials"[9] includes links to related

[1] https://lts.library.cornell.edu/lts/who/prorities13-15
[2] https://www.library.cornell.edu/about/inside/strategic-plan
[3] http://lodlam.net/
[4] http://www.bbc.co.uk/blogs/internet/entries/af6b613e-6935-3165-93ca-9319e1887858
[5] http://www.bbc.co.uk/blogs/internet/entries/78d4a720-8796-30bd-830d-648de6fc9508
[6] http://www.infoq.com/presentations/bbc-data-platform-api
[7] http://www.europeana.eu/portal/
[8] http://goo.gl/RcfFiq
[9] http://www.bbc.co.uk/music/artists/07eb40a2-2914-439c-a01d-15a685b84ddf

news articles, similar bands and Wikipedia data describing the band. Europeana's search tool presents the user with auto-suggestions based on underlying Resource Description Framework (RDF) data, to help disambiguate works of art and other concepts.

In experimenting with linked data, we also have an opportunity to reconsider how we describe, publish, and share our data so that its value extends beyond the library. When significant numbers of our own Cornell community, and certainly a high proportion of people outside Cornell, start almost all their information gathering through a search engine rather than navigating to an institutional home page, we have to realize that the visibility of our information is highly dependent on the efficacy of how we expose our information to search engines. Libraries stand to increase their share in the information economy and their visibility on the web when external entities link to our data and we link to theirs. In a very real sense, linked data embodies the notion of information as a public good.

From Cornell LTS's experience with the Linked Data for Libraries (LD4L) project we know we face challenges with converting our legacy data to meaningful RDF linked data, especially when reconciling entities produced by these conversions with entities on the broader web. Experience shows that the more we can prepare our legacy data through preprocessing, the more semantically rich our resulting linked data will be. We have hypothesized that creating data natively in RDF will allow us to better evaluate the overall effect linked data may have on production methods and the impact of richer descriptions on the user experience.

Linked data does not by itself solve the production pressures that all technical services departments face. That said, an important advantage linked data offers LTS is the opportunity to accept and reuse data produced outside the library domain, potentially allowing LTS to meet its production obligations while achieving greater descriptive depth about our collections than is currently possible. As in all areas, LTS makes judgements concerning where to invest resources. LTS's investment in metadata production can have a much greater impact if it takes a form suitable for consumption outside existing library systems. Linked data affords this opportunity in that the standards are developed, documented and used extensively outside the library domain.

Benefits

1. Sharing data for resource discovery has been a long-standing effort for the library community. Moving natively to web data standards opens up possibilities for more flexible frameworks and tools to achieve these ends.
2. We can more easily integrate library data using data models that are community-defined. This will enable us to create and reuse data using models that better suit the resource, e.g. using the Music Ontology to describe music recordings and performances. We will also be able to take advantage of a wider range of vocabularies (e.g. Faceted Application of Subject Terminology (FAST), LinkedBrainz, and the Getty vocabularies) - including those produced and used by communities outside the library world.
3. We can enhance discovery environments using the wealth of data from non-library sources that describe works, persons, and other entities represented in our catalogs.
4. Conversely, we can more effectively expose our collections and data in non-library discovery environments and increase our visibility on the web.
5. We will be much less obliged to create data redundantly for entities already described on the web as linked data. For example, linking to Wikidata may be a viable alternative to replicating a more limited form of the same data in our authority file. We can instead explore ways of extending the creation and sharing of authorities that can be reused by others, and make the processes more efficient and scalable.
6. Because of the formal logic of RDF, RDF Schema (RDFS), and the Web Ontology Language (OWL), we can make implicit knowledge explicit and through inferencing support queries on data not necessarily supplied by the cataloger[10], e.g. using transitive properties[11]. For example, all works said to be published in the Central New York Region would include those published in the Finger Lakes Region.

Community Engagement

The potential of linked data will not be fulfilled by individual institutions acting in isolation. The library ecosystem is complex and libraries engaging in linked data objectives have a vested interest in assuring that their findings are impactful and transferable to the broader library community. Linked data in libraries is in some ways still in its formative stages, and this

[10] http://www.vivoweb.org/files/presentations/12ws7/Workshop_part_2_slides.pdf
[11] http://www.w3.org/TR/owl-ref/#TransitiveProperty-def

presents opportunities for a library like CUL that has a history of pushing boundaries for the benefit of libraries and their users.

Several factors come together to make this an ideal time for LTS to play a leadership role in library linked data: CUL's participation in the LD4L project and (if awarded) its successors, the nascent partnership between CUL-IT and LTS on linked data work, LTS's innovative and flexible organizational culture, its excellent working relationship with service providers like OCLC, and its strong ties to professional communities such as VIVO, the Program for Cooperative Cataloging (PCC), the Association of College and Research Libraries' (ACRL) Rare Books and Manuscripts Section (RBMS), the Music Library Association (MLA), the Association for Library Collections and Technical Services (ALCTS), and now Kuali OLE. Indeed, there is a strong case for viewing OLE not as a competing priority to linked data, but as an opportunity to play a leadership role in integrating linked data solutions into mainstream library technology operations. And by engaging with groups like RBMS and MLA, with their strong roots in specific user communities, in data modeling for specific content types, we can create data that better serves the research needs of those communities.

LTS is consciously pursuing a strategy of working with partners within the library community and funding agencies to engage with in this effort. Exploratory work of this nature falls into the realm of research and development, which LTS has always considered to be part of its mission. Our participation in major collaborative projects, with the support of Mellon, has given us the opportunity to make a tangible impact on community understanding and practice, as discussed below.

Benefits

1. As an early experimenter, LTS will have input in developing and addressing foundational use cases that can properly motivate suitable data models and the building of cataloging tools to support them. By participating actively in the linked data community, we stand a greater chance of being influential in how new technologies are applied before standards are fixed. A good example of this is the ongoing collaboration between the Linked Data for Libraries project and the Library of Congress to evolve the BIBFRAME ontology.
2. We can begin the process of understanding how linked data will better the library user experience. While first generation linked data based discovery environments may not yet

be as polished as the best of existing library discovery layers, our understanding of the underlying data positions us to contribute to efforts by CUL and others to build these environments.

3. We can influence the migration path from current practices to ones that more fully exploit linked data techniques.
4. By building on commonly used technologies, linked data can make libraries less reliant on our traditional systems vendors and content providers, and create opportunities for productive collaboration with a wider range of partners. For example, the use of identifiers offers an attractive alternative to traditional ILS functionality for authority maintenance.

How We Move Forward

In the foregoing sections of this paper we considered the reasons for investigating linked data and the partnerships that can help us to do so. In this final section we consider how we can best pursue these efforts, and offer some thoughts on an appropriate linked data agenda for LTS at this juncture, given our particular combination of strengths, capacity, and responsibilities. We should expect this agenda to evolve further in the light of experience.

Groups such as the World Wide Web Consortium's (W3C) Library Linked Data Incubator Group[12] have been considering the effect of linked data methods on library processes for some time now. Many of the projected benefits of linked data echo the seven operational definitions of value provided by bibliographic description in the Final Report of the Task Force on Cost/Value Assessment of Bibliographic Control, 2010[13] [14].

We believe that the prima facie case for exploring the practical applications of linked data to technical services work is strong, and that pursuing these investigations is in keeping with CUL's institutional culture - the "spirit of adventure" we proclaim to our peers. But this boldness must also be tempered by a spirit of critical inquiry. We can certainly see the theoretical benefits of linked data - just as we understand the actual costs and diminishing returns of remaining with

[12] http://www.w3.org/2005/Incubator/lld/wiki/Use_Cases#Use_Cases_.2F_Case_studies
[13] https://journals.ala.org/lrts/article/view/5501/6757
[14] http://connect.ala.org/files/7981/costvaluetaskforcereport2010_06_18_pdf_77542.pdf

the status quo. That, however, is not the same as understanding what specific investments in the linked data sphere are strategic for us. While watching larger developments with interest, we believe our own interests are best served by concentrating our efforts on specific areas where our actions can both provide tangible benefits in the short to medium term and provide lessons to inform our future choices. We believe our efforts will have the most value if we give primary focus to examining how linked data techniques can solve known problems in representing our collections and to marshaling community practices that will best support our emerging needs.

The strategy we propose is to undertake projects that involve an iterative process of assessment on two fronts. On the demand side, we should examine the impact of linked data on matters directly affecting the user experience, such as exposure of collections via Resource Description Framework in Attributes (RDFa) for search engine optimization and the development of prototype user interfaces that support identified user tasks by leveraging additional relevant data not available through traditional library sources. The experiences of libraries who have experimented with publishing their collections as linked data suggest that it is a productive strategy: the Bibliothèque nationale de France (BNF), for example, reported a large spike in traffic following implementation,[15] as did our Columbia University colleagues in a project for their institutional repository.[16] These results also suggest that a good understanding of search metrics will be an important tool in the deployment of linked data. The LD4L use cases provide a starting point for development and evaluation of interfaces supporting a range of user tasks.[17] The Discovery and Access project has a strong practice of making incremental enhancements to our discovery environments based on user feedback and usability studies; in some cases linked data strategies may be appropriate solutions for identified areas for improvement. Pursuing this work will involve close collaboration with other CUL stakeholders, including usability experts and collection curators.

On the supply side, we should explore the preconditions and implications of moving to linked data workflows - what tools and services can best support them, how cataloging practices can evolve, where efficiencies can be traded off against production investments, and what we can and should expect from industry partners such as service providers, integrated library systems

[15] https://wiki.duraspace.org/download/attachments/68060801/LD4L-bnf.pdf?version=1&modificationDate=1425313952249&api=v2
[16] http://academiccommons.columbia.edu/item/ac:167931
[17] https://wiki.duraspace.org/display/ld4l/LD4L+Use+Cases

(ILS) vendors, and standards organizations. Some of these themes are being explored in larger collaborative projects in which CUL is involved, such as the Linked Data for Production (LD4P) proposal to Mellon. However, we think it is also important to keep in view our own institutional interests and agenda. A large part of our emphasis here will be on transitional strategies that can pay off in the near future even as the environment continues to evolve. Examples of such strategies include moving toward the use of URIs to manage authorities, continuing to work on transformation tools, and exploiting third party identity aggregation services. We also see value in carefully scoped early experimentation with native linked data production, particularly where we can see immediate advantages over existing methods. Some projects - such as efforts to leverage VIVO - build on technical foundations already laid at Cornell and elsewhere, and are novel only in the sense that we have found new uses for tools that were already at our disposal. Other major undertakings, notably the Kuali OLE implementation, present an opportunity to exploit the new methods in our next generation of systems.

LTS has become a recognized player in linked data efforts not only within CUL, but also in the wider library community. Our catalogers and metadata staff have made significant contributions in community modeling decisions and also in reimagining authority workflows in pilot projects with service providers and open source communities. LTS's batch processing unit has been an early implementer of using identifiers to maintain data in MARC.

The following are areas where investment of effort can further the objectives we have outlined and provide test cases for our hypotheses.

- Linked data standards development
    - Contributing to ontology development and debate to ensure best practices and alignment, e.g. among BIBFRAME, Visual Resources Association (VRA) RDF, VIVO, the Virtual International Authority File (VIAF), Friend of a Friend (FOAF), the PROV-O provenance ontology, and schema.org.
    - Providing feedback on improved modeling and serialization of vocabularies currently available in RDF.

- Exploitation of linked data in discovery environments
    - Assessing usability of discovery environments based on linked data.

    - Identifying opportunities for full exploitation of ontologies and inferred data.
    - Evaluating available linked data, both within the library domain and beyond, as sources for scalable solutions to meet existing and new needs.
        - Measuring the effectiveness of using OCLC WorkIDs to support collocation of various formats and editions of works.
        - Assessing the potential of VIAF data to support preferred language searching.
    - Leveraging VIVO efforts to make actionable connections between the research community, faculty networks, and library content. Much of the data in VIVO can be exploited further to support discovery, e.g. LTS's current project to include thesis advisors in our Voyager thesis records.

- Outreach and advocacy to communities of metadata practice
    - Contributing to the modeling of existing library controlled vocabularies as they transition to linked data, e.g. RBMS vocabularies.
    - Establishing best practices for storing Uniform Resource Identifiers (URIs) in MARC and non-MARC records, with the goals of easier management of data and preprocessing for linked data.
        - Library of Congress (LC) Authority URIs in CUL bib records.
        - Faceted Application of Subject Terms (FAST) Vocabulary URIs in CUL bib records.
        - PCC Task Force on URIs in MARC. [18] [19]
        - VRA Core 4 Oversight Committee best practices for work URIs in VRA Core 4 XML.

- Tool development
    - Shaping hybrid environments that make connections between current production systems that use MARC and non-MARC with linked data platforms and services.
    - Contributing to the development of authority control functionality in Kuali OLE to ensure that it both supports and takes advantage of linked data methods.

[18] http://www.loc.gov/aba/pcc/documents/URIs-MARC-taskgroup.docx

[19] https://goo.gl/GkbUFH

- Leveraging Vitro software and infrastructure to create native RDF instance data for Cornell collections.
    - Linked data cataloging pilot.
    - Creation and maintenance of local authorities/entity information in Simple Knowledge Organization Scheme (SKOS) and other ontologies
- Participating in OCLC entity lookup pilot.
- Contributing to the continuing development of transformation tools.