Why Linked Data?

Go to LD4L Wiki Gateway

The Semantic Web is a term coined by Sir Tim Berners Lee in a seminal article[1] published in Scientific American in 2001. Berners Lee articulated a vision of a World Wide Web of data that machines could process independently of humans, enabling a host of new services transforming our everyday lives. While the paper's vision of most web pages containing structured data that could be analyzed and acted upon by software agents has not materialized, the Semantic Web has emerged as a platform of increasing importance for data interchange and integration through the growing community implementing data sharing using international semantic web standards[2], called Linked Data[3].

There are many current examples of the use of semantic web technologies and Linked Data to share valuable structured information in a flexible and extensible manner across the web. Semantic Web technologies are used extensively in the life sciences to facilitate drug discovery by finding paths across multiple datasets showing associations between drugs and side effects via genes linked to each.[4] The New York Times has published its vocabulary of approximately 10,000 subject headings developed over 150 years as Linked Data and will expand coverage to approximately 30,000 topic tags; they encourage the development of services consuming these vocabularies and linking them with other online resources.[5] The British Broadcasting Corporation uses Linked Data to make content more findable by search engines and more linkable through social media; to add additional context from supplemental resources in domains like music or sports; and to propagate linkages and editorial annotations beyond their original entry target to bring relevant information forward in additional contexts.[6] The home page of the United States data.gov site states, "As the web of linked documents evolves to include the Web of linked data, we're working to maximize the potential of Semantic Web technologies to realize the promise of Linked Open Government Data."[7]

Linked Data is a publishing paradigm for making data and not just human-readable documents fully accessible and inter-linkable anywhere on the Internet. Linked Data uses the same common Web communications protocols as ordinary browser software to connect machine-readable data across distributed computers. In 2006, and updated in 2010, Berners Lee described a 5-star rating system for published data to be considered Linked Data[8]:

- Available on the web
- Available as structured data readable by a machine
- Available in a non-proprietary format
- Expressed using open World Wide Web Consortium (W3C)[9] standards
- Linked to other data on the web

While Linked Data can be used internally within an institution or across a collaborative group, it becomes much more valuable when it is Linked ***Open*** Data, and is publicly shared using an open license such as the Creative Commons CC-BY[10] or CC0[11] licenses, or the United Kingdom's Open Government License[12]. For our Linked Data for Libraries project, our intention is that all SRSIS instances will share Linked Open Data with the world.

Linked Data conforms to a common data format known as the Resource Description Framework, or RDF[13]. Each unit of Linked Data expressed in RDF has a subject, predicate, and object comparable to the simple sentence structure of human language. All subjects, predicates, and objects (other than simple data values) are encoded or represented as uniform resource identifiers, or URIs[14], intended to be resolvable as uniform resource locators (URLs)[15] on the Internet. Just as typing the URL of an ordinary web page in a browser should produce an HTML (Hyper-Text Markup Language) document in response, a Linked Data request should trigger a response in the form of one or many RDF statements, known informally as triples. Triples may be served from static data files or generated on the fly from data stored as XML or in a relational database; a native semantic web application persists its data in a type of database optimized for storing and querying semantic triples, known as a triplestore.[16]

Understanding what these data refer to requires another key component of the Semantic Web – a way to encode meaning. An ontology declares a set of defined types and relationships that are referenced within Linked Data to express what is being referred to and the nature of the relationships involved. Ontologies are shared as broadly as possible to reduce the friction and overhead of interpretation. The Friend of a Friend ontology[17], for example, declares a type *foaf:Person* that is universally recognized in Linked Data as a reference to a human being. Ontologies are not limited to the simple hierarchical classification structures of a taxonomy or partonomy or to the small set of broader/narrower /related relationships typical among terms in a thesaurus; we can express, for example, that one person *knows* another person or a work of fiction *has as primary source* a historical document.

Software applications consume Linked Data from multiple sources – where that Linked Data has been structured to describe known people, places, organizations, events, and associated concepts and relationships – as the foundation for new, dynamic, information-rich services. These services can request Linked Data via the Hypertext Transfer Protocol (HTTP)[18] without knowing anything about the internal storage schema or application programming interface (API)[19] of a remote data source, and the response arrives via HTTP and conforms to a standard data format, RDF – a stark contrast to data silos accessible only to those with the keys and a detailed map to their individualized contents or APIs.

In addition to breaking down silos, Linked Data has also, through its fundamental dependence on ontologies, charted new ground in practices for data description, or metadata. Changes to metadata practice driven by the adoption of Linked Data can best be summarized as making once implicit statements explicit. Declaring the subject of every metadata statement with a URI as its identifier and using defined types and properties (also specified by URIs) for expressing the content of metadata in RDF eliminates much of the ambiguity in what is being referred to, in where the intended meaning has been defined, and in how the information referenced can be directly accessed. The ability to support explicit references also encourages the practice of converting “strings to things.” A string is just a sequence of characters to interpret or match as best one can to other strings; a Linked Data URI is a reference to any amount of structured information with the potential to guide interpretation or feed automated processes. For example, compare the information content of “Twain, M.” to the structured information in the VIAF Linked Data record for Mark Twain associated with the URI http://viaf.org/viaf/50566653/.

Of course, simply being able to link, request, and assemble data more easily does not by itself produce new insights or guarantee utility. Many challenges remain in discovering the existence of data sources; in interpreting the meanings encoded in the ontologies used for expressing the data and creating mappings among ontologies; and in resolving multiple identifiers that may or may not refer to the same person, place, organization, or thing. Tools are maturing,[20] scalability is improving,[21] and services are developing to resolve co-references to different datasets,[22] but the scale of the Internet is vast.

For this project we will begin with a constrained set of institutions and highly structured library catalog records linked to authority records and Library of Congress Subject Headings (LCSH)[23]. As other sources are brought into the mix, we can expect higher rates of uncertainty in identifying authors, keywords, place names, and terms from multiple vocabularies. Our premise is not that we will achieve perfect alignment; the goal is to make large bodies of information more discoverable and interoperable than they have been, as a significant step forward benefitting users well beyond the bounds of our three institutions. We expect problems as well as successes, and will work throughout to clarify the remaining challenges as the blueprint for a research agenda looking farther into the future.

Libraries are a natural home for serving and consuming Linked Data and building innovative new services. This proposal envisions a set of software tools, ontologies, and user-facing services capable of representing, discovering, and integrating human knowledge currently outside the confines of traditional library catalogs, web pages, and online information services.

---

[1] Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web". *Scientific American Magazine*. Retrieved August 21, 2013.
[2] http://www.w3.org/standards/semanticweb/
[3] http://linkeddata.org
[4] D.J. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019
[5] http://data.nytimes.com
[6] http://www.cmswire.com/cms/information-management/bbcs-adoption-of-semantic-web-technologies-an-interview-017981.php?pageNum=2
[7] http://www.data.gov
[8] http://www.w3.org/DesignIssues/LinkedData.html
[9] http://www.w3.org
[10] http://creativecommons.org/licenses/by/3.0/
[11] http://creativecommons.org/publicdomain/zero/1.0/
[12] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[13] http://www.w3.org/RDF/
[14] http://tools.ietf.org/html/rfc3986
[15] http://tools.ietf.org/html/rfc3986#page-7
[16] http://en.wikipedia.org/wiki/Triplestore, http://semanticweb.com/introduction-to-triplestores_b34996
[17] http://www.foaf-project.org
[18] http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol
[19] http://en.wikipedia.org/wiki/Application_programming_interface
[20] Janev, V. and Sanja Vraneš, Maturity and Applicability Assessment of Semantic Web Technologies, *Proceedings of I-KNOW ’09 and I-SEMANTICS ’09*, 2-4 September 2009, Graz, Austria
[21] http://www.w3.org/wiki/LargeTripleStores and http://www.w3.org/wiki/TripleStoreScalability
[22] http://sameas.org
[23] http://id.loc.gov/authorities/subjects.html