Learning Linked Data:
Final Project Report

IMLS Grant no.: LG-51-11-0147-11, October 2011 through September 2012

Project Partners:

University of Washington (lead)
Kent State University
University of North Carolina
JES & Company
3roundstones Inc.

Project Overview

The "Learning Linked Data" project, funded through September 2012 by a one-year IMLS planning grant led by the University of Washington, envisions an online learning environment in support of educating library and museum professionals in the principles and practice of Linked Data -- a "language lab" of software-supported methods for data processing and analysis. A core project group of twenty instructors, students, and technology experts met to develop an Inventory of Learning Topics. The Inventory, outlining a target set of basic analytical and software skills needed across a wide range of pedagogical contexts, was posted on a blog for input from a larger circle of colleagues. The group concluded that a follow-on project should aim at engaging instructors in a dialogue with software developers to match analytical tasks with specific software tools in order to provide instructors and students with a well-organized and carefully documented collection of video microtutorials.

Part 1: Description of Project Activities

Linked Data is data that can fit into a "cloud" of interconnected data sources -- whether those sources are published world-readably on the Web (Linked Open Data) or behind corporate or institutional firewalls (Linked Enterprise Data). For our purposes, Linked Data is data published in a form compatible with the Resource Description Framework (RDF) model of the World Wide Web Consortium (W3C). Compatibility with RDF and practices common to Linked Data has become an important objective of most current initiatives for standardization and service development in the library and museum worlds.

"Learning Linked Data" is a project funded under the IMLS program National Leadership Grants for Libraries from October 2011 through September 2012 and aimed at planning the development of a software-supported environment for learning the principles and practice of Linked Data. The project sought to establish guidelines for a "language lab" for the trainers and university faculty who are teaching current and future library and museum professionals. In order to elicit the requirements for this environment, the project recruited a community of twenty active project participants from a wide range of relevant backgrounds, encompassing university LIS faculty and professional trainers, graduate students, software developers, instructional technology practitioners, and Semantic Web experts.

The planning project was structured around a two-day workshop held at the University of Washington in February 2012. In preparation of the workshop, several concept papers were discussed on a mailing list. The workshop then focused on refining and restructuring the project's key text, an Inventory of Learning Topics. In the Inventory, each learning topic is associated with examples of software tools needed for specific analytical or data-processing tasks. After the workshop, the Inventory was posted on a blog for comment by colleagues. A final version of the Inventory, appended to this report, is intended to serve as the starting point for a larger follow-on project to be proposed to IMLS in early 2013.

February 2012 Workshop

The project convened a meeting of its core community (see List of Participants in the Appendix) on 2-3 February 2012, at the Information School on the University of Washington campus in Seattle. Prior to this meeting, participants collaborated by email to draft lists of learning topics relevant to Linked Data along with tools available for use in instruction about those topics, and to consider ways of documenting and packaging the tools to support effective teaching and self-instruction.

Participants came to the workshop discussions with backgrounds in academia and university-level teaching, information technology development and deployment, software design, library and information science, and metadata applications design. A first-draft inventory of Learning Topics was discussed point-by-point, refined into a second draft, then associated with supporting software tools.

Scoping the project to "basic" learning topics

Much of the discussion at the Seattle workshop revolved around the scope of the learning platform the group had in mind and, more concretely, of a follow-on implementation project. Mindful of the risks of taking on too many topics, at too superficial a level, the participants felt that the follow-on project should focus, at least initially, on the fundamentals of understanding and interpreting Linked Data. This discussion of scope reached the following conclusions.

Specific pedagogical approaches and learning outcomes. Each participant came to the workshop with particular pedagogical scenarios in mind -- courses for particular learners, from particular backgrounds and experience, and with correspondingly specific expected learning outcomes. An early draft of the Inventory of Learning Topics tried to classify the topics into Beginning, Intermediate, and Advanced. However, it emerged that topics considered at a beginning level for library-science students may be considered advanced for a computer science student, and vice versa. Narrowing the focus to a specific audience, it was felt, risked favoring one audience at the expense of others. Specific pedagogical approaches, along with the audiences for which they were tailored, were therefore considered to be out of scope for the Inventory of Learning Topics. Moreover, it was felt that any particular syllabus offered by the project would inevitably need to be customized for a diverse range of instructional goals and instructor preferences and would, in complex ways, go quickly out of date.

As examples, however, specific scenarios did inform much of the discussion. Among software tools, in particular, the group acknowledged rough categories of complexity, inasmuch some are usable by anyone with a mouse while others require familiarity with a Unix command line and scripting languages. Participants agreed that the software-supported learning environment envisioned for the follow-on project would need, at a minimum, to differentiate by the level of IT competence required.

Support for the design of Linked Data. Sound modeling is an essential foundation for creating high-quality Linked Data. However, the group recognized that methodologies for designing data models from scratch are very diverse and selection among them depends on target audience.

Students may come at the problem with backgrounds in databases, UML modeling, formal ontologies, or traditional knowledge organization systems. One approach favored by several participants, the [http://dublincore.org/documents/singapore-framework/ Singapore Framework for Dublin Core Application Profiles], offers a workflow for metadata design well-suited for intermediate-level students. However, the Singapore Framework was considered too specific to be a priority for envisioned learning environment, at any rate in its initial phase. In general, the group felt that the relative lack of mature software tools for supporting the process of designing Linked Data implied that data design should be a second-order priority in the follow-on project.

Support for the practicalities of implementation. Much of the discussion examined the fuzzy line between supporting instruction about the conceptual underpinnings of Linked data versus imparting the skills required to implement Linked Data in practice. For example:

- **Publishing linked data "as Linked Data".** Simply publishing a dataset on the Web as a ZIP file does not make the contents of that dataset available for linking. The publication of Linked Data involves exposing one or more RDF-compatible representations of the contents of a dataset, possibly via content negotiation based on browser settings or user preferences. Publication is thus both a practical implementation skill and an academic learning topic (e.g., the principles behind content negotiation).
- **Visualizing Linked Data.** Visualization of Linked Data relationships was seen not just as a software function, but as a Learning Topic in itself. At the most basic level, node-and-arc diagrams are used to visualize RDF graphs -- webs of RDF statements. However, diagrams can be generated at higher levels of granularity (as in "Linked Data Cloud" diagrams), and sophisticated statistical techniques can be used to depict clusters of related resources or generate "cloud" diagrams of high-level relationships between datasets.
- **Storing Linked Data.** Linked Data can be exposed on the Web as plain-text files holding RDF triples "serialized" in one of several interchangeable syntaxes; in the Inventory, these fall under the learning topic "Creating and manipulating RDF data." RDF triples can be indexed for retrieval by storing them in specifically optimized databases ("triple stores"). Setting up a triple store is relatively straightforwardly an implementation-related skill. However, the participants recognized that the rapid evolution of approaches and software for storage could potentially be seen in the follow-on project as a learning topic in its own right.

"Language lab," or "kitchen"?

The metaphor guiding development of this project was that of a "language lab" for learning Linked Data, and the workshop's goal was to specify how the language lab should be equipped. Participants acknowledged the "language" metaphor as particularly apt for instruction about RDF because it usefully emphasizes the nature of RDF as a conceptual model rather than just a specific data format or concrete syntax. It was however felt that the language metaphor should not be pressed too far (e.g., with headings about the "grammar," "composition," and "translation" of RDF, as in the initial concept papers) because the linguistic terminology is not reflected in mainstream writing about RDF. Participants agreed, therefore, that topic headings should use the native jargon of RDF.

The native jargon of RDF makes few distinctions between types of data; for RDF, "everything is just data." Distinctions common in other fields between, say, "element sets," "value vocabularies," and "datasets" have no exact equivalents in RDF. Ontologies, RDF vocabularies, and SKOS concept schemes – the conceptual structures of Linked Data – are themselves considered as "just data" and expressed with the same formatting as the instance metadata using those structures.

Workshop participants therefore decided that the Inventory of Learning Topics, and any narrative about the future tool collection, should take care to label topics primarily with terminology native to RDF, drawing analogies to terms from other fields only as needed. Cloaking the principles underlying RDF in the terminology of, say, library science, may help students in that discipline to grasp concepts in the short term but does not prepare them well for working with RDF outside of the library-science context. The participants felt it to be the job of instructors, rather than the Learning Linked Data project itself, to shape the material into a form intellectually accessible to specific groups of students. As one participant noted, "Grounding [the project] in RDF allows instructors to use whatever metaphor or mechanism makes sense to them and their students -- whether linguistics, math, or programming logic."

The collection of software-supported methods to be provided by the project was seen as providing a "palette" of functionality from which instructors could draw in realizing particular course concepts. An alternative metaphor of a "kitchen" was proposed, describing the challenge in terms of equipping a workspace with utensils that would allow cooks to prepare a wide range of "courses."

An Inventory of Learning Topics

The resulting list of topics was organized under five categories:

- Understanding Linked Data
- Searching and querying datasets
- Creating and manipulating RDF data
- Visualizing webs of data
- Implementing a Linked Data application

For each Learning Topic, the workshop participants characterized what type of software tools instructors and learners would need to use in the course of learning, illustrating each type of software, where possible, with a known exemplar.

Following the workshop, the Inventory was [http://lld.ischool.uw.edu/wp/ posted on a University of Washington blog][1] for public comment. Workshop participants helped disseminate this invitation to specific colleagues as well as to mailing lists and communities where they were active (see the Outreach list in the Appendix). The resulting feedback greatly enriched the Inventory with pointers to additional tools and with potential user scenarios. Much of this input was documented as blog comments, and much arrived via email and other private channels. All comments were compiled into a document that will inform the development of a proposal for a follow-on project.

Users of tutorials about "Learning Linked Data"

Numerous usage scenarios for tutorials on learning Linked Data emerged at the workshop and as feedback to the blog posts. These fall into two broad categories:

Working professionals seeking to solve problems on-the-job.

- One university library wants to map MARC, EAD, and Dublin Core records to into a master record structure. The process involves extracting parts of several different record formats into RDF triples, merging those into a graph, and turning the result into a single record.
- Another small academic library wants to document faculty research as part of an institutional repository. Metadata would describe faculty expertise, scholarly research profiles, grant applications, publications and related metrics.
- A scientific association manages information in multiple genres -- journals, conference proceedings, committee bulletins, and the like -- that need to be both interlinked among themselves and better integrated with topic-relevant information from external sources.
- A large private-sector organization needs to ensure efficient access to key data across a broad range of non-interoperable repositories -- the solution to which requires data managers to channel efforts away from the construction of separate data silos towards their common expression as Linked Data.

In each scenario, working professionals need to learn to think differently about familiar problems and find ways to instruct themselves about solutions while remaining on the job. This implies either the availability of resources for self-instruction, support for professional training seminars, or a combination of the two.

Students seeking formal qualification.

- In today's environment, people move into the library and museum profession from educational backgrounds ranging from the social sciences and humanities to computer science, both academic and applied. The availability of generic instructional materials -- usable in the context of widely differing pedagogical approaches -- would support formal university courses at all levels.
- Crucially, the faculty members teaching such courses themselves need ways to bring themselves up-to-date on Linked Data principles and practices. The "Learning Linked Data" project itself was initiated in large part by faculty members and professional trainers who themselves feel challenged by the rapid pace of technological evolution and see a vital need to keep ahead of their students by establishing a source and forum for learning about new software and pooling expertise and experience among themselves about its use in university coursework.

Planning the Follow-on Project

Workshop participants agreed that the learning materials provided by the follow-on project should not just support theoretical understanding but also provide students with very practical

skills. Where precisely to draw that line, it was felt, should be the first task of the follow-on project. This project should document how to use software in support of learning the topics outlined in the Inventory, potentially with supplementary information on how to integrate those tools into specific pedagogical approaches.

While early discussion considered that the follow-on project might aim at packaging a set of useful utilities as a "snapshot" distribution, or even take development one step further by integrating the tools into a common interface with "orchestrator" functionality, it was soon recognized that this goal would be expensive to realize and unsustainably difficult to maintain. Rather, taking the Inventory of Learning Topics as a point of departure, the project would identify, under each topic, specific analytical tasks -- tasks, the accomplishment of which could be taught, using existing software, in the form of brief, screencast-based microtutorials. The process of identifying such specific tasks, it was felt, should be undertaken in an open community of instructors and interested learners, anchored perhaps in an existing metadata community platform such as the Dublin Core Metadata Initiative.

Well-specified tasks would provide concrete reference points for recruiting software developers to provide microtutorials targeted on the use of their own tools. The end result of the project would be a set of microtutorials, each focused on a specific task, organized under an evolving set of Learning Topics and Sub-Topics. Instructors and learners would draw on this "palette" of tasks in designing specific courses, selecting and sequencing the possibilities differently according to the background and needs of their students. By developing a publication channel for microtutorials with well-practiced workflows, and anchored in an ongoing metadata community, it was felt that this effort could be undertaken in a way that could prove to be sustainable beyond the end of the project.

The project team completed a draft proposal for a two-year follow-on project for a final review by workshop participants and other stakeholders in September 2012. It is anticipated that the proposal will be submitted to IMLS in January 2013, with implementation anticipated to begin the following October.

References

1. http://lld.ischool.uw.edu/wp/

Part 2: Quantitative Report

Institution Name: The Information School, University of Washington
Grant #: JG - 51 - 11 - 0147 - 11

A. SITE SPECIFIC PROJECT ACTIVITY: Workshop on resources for Linked Data instruction, 2-3 Feb. 2012 at UW iSchool, Seattle, Wash.

1. _____ Total # of collection items conserved, relocated to protective storage, rehoused, or for which other preservation-appropriate physical action was taken.

2. _____ Total # of collection items digitized, scanned, reformatted, or for which other electronic or digital preservation action was taken.

3. _____ Total # of collection items with new or enhanced accessibility (include items that were cataloged or for which finding aids or other records were created or computerized) [includes ____ items made accessible to users other than grantee staff for the first time, ____ items with new or enhanced access for staff only].

4. _____ Total # of lectures, symposia, demonstrations, exhibits, readings, performances, concerts, broadcasts, Webcasts, workshops, multi-media packages, or other learning opportunities provided for the public (do not include PSAs or other promotional activities) [includes _____ out-of-school or after-school programs, _____ exhibits].

5. _____ Total # of tools created, improved, or produced for searching, information management, or information analysis by users other than or in addition to grantee staff.

6. ___1____ Total # of conferences, programs, workshops, training sessions, institutes, classes, courses, or other structured educational events provided.

7._____ Total # of internships, apprenticeships, mentoring opportunities, or other extended educational opportunities provided.

8._____Total # of degrees/certificates earned as a result of the grant [includes _____ Master's, ____ Ph.D. degrees, _____ other (specify): __________].

9._____Total # technology upgrades or improvements (specify): _____

10. Other activities not covered by the categories above:
16 participants -- 4 from University of Washington iSchool, 3 from the Dublin Core Metadata Initiative, 2 from Kent State University, 1 each from University of North Carolina - Chapel Hill, University of Illinois iSchool, New York University Libraries, Library of Congress , 1 each from three private consulting practices in information systems serving libraries, not-for-profit organizations, and corporate enterprise

B. PORTABLE PRODUCTS (relating to the activity named in section A.)

11. _____Total # of research reports, papers, books, reprints, or other publications generated.

12. ____1____Total # of Web sites developed or improved [include URLs/addresses:

___http://lld.ischool.uw.edu/wp/__].

13. ____2____Total # of learning resources produced [includes _____ oral histories, _____ curriculum resources, _____ curriculums, _____ Web-based learning tools, or _____ other (specify): _Inventory of Learning Linked Data Topics, Glossary_].

14.____4____Total # of key management documents created

[includes _____ emergency plans, _____ conservation surveys, _____ strategic plans,

_____ other (specify): __3 blog posts, 1 implementation project narrative__].

15. If your grant created one or more quantifiable products not covered by the categories above, please briefly identify and quantify them here. Attach another sheet if necessary.

1 Inventory of Learning Topics (B.13): 6 main categories, each with 8-13 subcategories
1 panel summarizing the project activity (proposed) for the ALISE 2013 conference

C. PARTICIPANTS/VISITORS/USERS/AUDIENCE (relating to the activity named in section A.)

16. ___>45___Total # of **community organization partners** [includes _>30_ informal partners, _15_ formal partners].

17. _____Total # of **schools** (pre-K through grade 12) that used services provided by your grant (include only schools that actively participated, not those to which material was simply distributed or made available) [includes _____ students participating in field trips].

18._____Total # of **teachers** supported, trained, or otherwise provided with resources to strengthen classroom teaching or learning.

19._____Total # of **pre-K through grade-12 students** served [includes ____ youth 9-19 who used, participated, visited, or otherwise interacted with activities, experiences, resources, or products offered by your grant].

20. _____Total # of **viewers and listeners** for radio, television, and cable broadcasts (for series, include total actual audience for all broadcasts; do not include audience for PSAs or other promotional activities or Webcasts; do not report potential audience).

21.___>200__Total # of **users of Web-based resources** provided by your grant (include all individuals the project served). Choose the measure that best represents your use rate (choose only one): ____ visits (hits), ____ unique visitors, ____ registered users, ____ other measure (specify): __53 page comments form 24 individuals, plus >12 comments by email from 8 individuals___.

22. __>300__Total # of **individuals** benefiting from your grant (include all those from questions 18-21 plus others the project served, including staff or others in your field). Only include those who actually participated or used your project services in some way.

23. This number includes: ___>300__**professionals**, _____ **non-professionals or pre-professionals**, _____ **docents or interpreters**, _____**volunteers**, _____**staff** that received services provided by your grant.

24. If your grant served one or more quantifiable audiences not covered by the categories above, please briefly identify and quantify them here. Attach another sheet if necessary.

6 individuals representing 4 organizations provided Linked Data use cases
>20 software tools and instructional resources identified

Appendix

Workshop Attendees

- Mike Crandall, Joseph Tennis, Randy Orwin, and David Talley from the University of Washington iSchool
- Tom Baker, Diane Hillmann, and Stuart Sutton from the Dublin Core Metadata Initiative
- Riley Stormer and Marcia Zeng from the Kent State University School of Library and Information Science
- Craig Willis (representing Jane Greenberg) from the School of Information and Library Science at the University of North Carolina at Chapel Hill
- Karen Wickett from the University of Illinois iSchool
- Independent consultants Joseph Busch, Karen Coyle, and Marjorie Hlava
- Library software developers Corey Harper and Ed Summers

Outreach Channels

- ALISE - ALA's Association for Library and Information Science Education
- DCMI
- DERI
- Talis
- UKOLN
- Individual iSchools
- ASIS Information Architecture list is active (with alot of professors)
- LODLAM list
- W3C Library Linked Data list
- ISKO, especially the European chapters