IMLS Level I Planning Grant

Title: Planning for development of an integrated learning platform to teach the principles and process of metadata design for a Linked Data environment

Assessment of Need

Since Tim-Berners Lee coined the term in his explication of the Semantic Web in 2006[1], Linked Data has become a focal point for publishers of structured data on the World Wide Web. A growing “cloud” of open linked data resources has emerged around DBPedia[2], a database of information extracted from Wikipedia, with structured information ranging from descriptions of people (FOAF[3]) to GeoNames[4] , census data, and bibliographic descriptions (DBLP[5]). Publishers of Linked Data now range from the BBC, NASA, and New York Times to government agencies in numerous countries, notably the US, UK, and the Netherlands. Major national libraries, first and foremost the Library of Congress[6], are leading a trend to publish authority files, catalogs, and datasets as Linked Data. The use of Semantic Web methods to expose information to the Linked Data cloud promises to improve access to the resources of businesses, public agencies, and cultural memory organizations for the benefit of all.

Linked Data is based on standards and practices that facilitate the post-coordinated integration of information published on the Web across a diversity of sources. The move to Linked Data signifies a fundamental shift to new bases for library and information science (LIS) and thus in the skill set required of 21st century librarians and information professionals.

In traditional IT environments, metadata "records" need only make sense locally, within a closed system – even if that system is very large, as with the union catalog of a library federation. Traditionally, there was no requirement – nor was there a technological basis – for designing metadata to be linked or merged with data sources outside of local systems. Conventionally, LIS professionals have worked primarily by starting with tools pre-configured for standardized data formats such as MARC (for library catalogs) or Simple Dublin Core, the pre-Linked-Data format for item-level description required since 2001 by the Open Archives Initiative (OAI) Protocol for Metadata Harvesting promoted by IMLS for use by its digitization projects[7].

Today's Web environment, in contrast, presents both the opportunity and, increasingly, the requirement to create rich linkages from data held locally to other sources of data on the Internet. Linked Data is designed to express such connections inasmuch as it uses URIs ("URLs" used as Identifiers), whenever possible, to identify resources being described, their properties, and descriptive relationships to other resources. In effect, Linked Data uses the Web's Domain Name

[1] Tim Berners-Lee (2006-07-27). "Linked Data - Design Issues". W3C. http://www.w3.org/DesignIssues/LinkedData.html. Retrieved 2010-12-18.

[2] "DBpedia dataset". DBpedia. http://wiki.dbpedia.org/Datasets. Retrieved 2010-12-18.

[3] http://www.foaf-project.org/

[4] http://www.geonames.org/

[5] http://www4.wiwiss.fu-berlin.de/dblp/

[6] http://id.loc.gov/

[7] http://www.imls.gov/applicants/guidelines/pdf/FY11_NLG_Guidelines.pdf, p. 10.

System (DNS) as a distributed dictionary of institutionally managed, globally unique identifiers for its data elements.

If URIs provide the words for a "language" of Linked Data, the grammar for that language is provided by the Resource Description Framework (RDF)[8]. RDF properties function roughly like verbs (e.g., "isReferencedBy") and RDF classes like nouns (e.g., "Person"). As in natural languages, where utterances acquire their meaning by following a sentence grammar, RDF statements ("triples") follow a simple and consistent three-part grammar of subject, predicate, and object. Analogously to higher-order pieces of writing, RDF statements are aggregated into RDF "graphs".

RDF statements support the process of connecting dots between information silos – and therefore of creating "knowledge" – through providing a generalized linguistic basis for expressing the linkages. Just as English as a second language often provides a basis for communication among non-native English speakers, RDF also provides a common second language into which local data formats can be translated and exposed for the purposes of interoperability.

The foundational vocabularies of Linked Data, notably Dublin Core (for describing information resources), FOAF (for describing people and organizations), and SKOS (for expressing thesauri as Linked Data), are all fundamentally simple and, in principled ways, extensible – traits which have made them popular starting points for teaching metadata, whether to university students, mid-career professionals, or application development teams.

The task of teaching the design of metadata for this new Web environment is roughly analogous to that of teaching a natural language[9]. Students need to master the basics of a grammar (RDF); learn a starting vocabulary of several dozen "words" from Dublin Core, FOAF, and SKOS; practice "reading" the meaning of data they may encounter; engage in drills in the use of common idioms for "expressing" information using Semantic Web principles; and absorb best-practice principles for designing metadata applications – linguistically the equivalent of higher-order outputs such as essays and dissertations. Just as mastery of, say, college-level French requires more than the memorization of verb paradigms, learning to produce solidly designed metadata – metadata that follows best-practice design principles and plays well in the Linked Data space – presents students with unique challenges at the level of abstract thinking and modeling.

The expressive flexibility afforded by RDF is becoming increasingly necessary as the Web evolves its explosively diverse ecosystem. It encompasses not just the familiar forms of publishing and scholarly discourse, but new and ever-evolving forms of scientific, commercial, and multimedia information at all levels of granularity, from collections seen as wholes down to individual data points, data tables, illustrations, and named gene sequences. The one-size-fits-all

[8] http://www.w3.org/RDF

[9] The language metaphor has been part of the metadata design discourse for over a decade. See, for example, Thomas Baker. (2000). "A Grammar for Dublin Core". *D-Lib Magazine*. vol 6, no. 10. Available: http://www.dlib.org/dlib/october00/baker/10baker.html and
Elaine Svenonius. (2000) *The Intellectual Foundation of Information Organization*. Cambridge, MA: The MIT Press, p.59.

type of standard which worked reasonably well to describe buildings full of books, recordings, and videos at the item level appears inadequate to the challenge presented by such diversity.

This proposal for an IMLS Planning grant aims at bootstrapping the creation of an integrated learning environment to support the teaching of metadata design for the Linked Data environment in both university and professional training contexts – in effect, to create a "language lab" for learners and designers of Languages of Description.

As a language designed by humans for processing by machines, RDF requires competence in the use of software tools for ingesting, visualizing, transforming, and interpreting its URI-based statements. Happily, the accelerating expansion of the Linked Data cloud since 2006 has pushed the development an increasingly rich offering of open-source software tools for processing that data. The challenge in creating a language lab for Semantic Web data is therefore one of selecting open-source components, integrating them into a unified environment, and providing the documentation necessary for helping teachers and learners to use the platform effectively.

In order to plan such a language lab, the project will convene a workshop of circa 15 to 20 participants to assess the requirements of teaching Linked Data in light of available tools. The workshop will bring together professors and instructors in the area of knowledge organization – people who understand the requirements of learners and of classroom teaching – with software providers and technology developers who have deep understanding of Semantic Web standards and up-to-date knowledge of available open-source tools that can be used in support of those requirements.

Starting from an analysis of pedagogical requirements, the workshop will specify the potential components for a learning platform, elaborate scenarios for use of the platform in classrooms, and specify what sort of documentation and training would be necessary to enable instructors to use the platform effectively (i.e., to "teach the teachers"). We envision the platform as a basis for the development of course modules by people involved in both formal and informal teaching and learning environments, and something that would integrate easily into commonly used interfaces such as browsers. The platform would allow learners to accomplish tasks such as the following:

- To brainstorm user requirements for "agile" application development with the help of RDF-compatible mind-mapping tools.

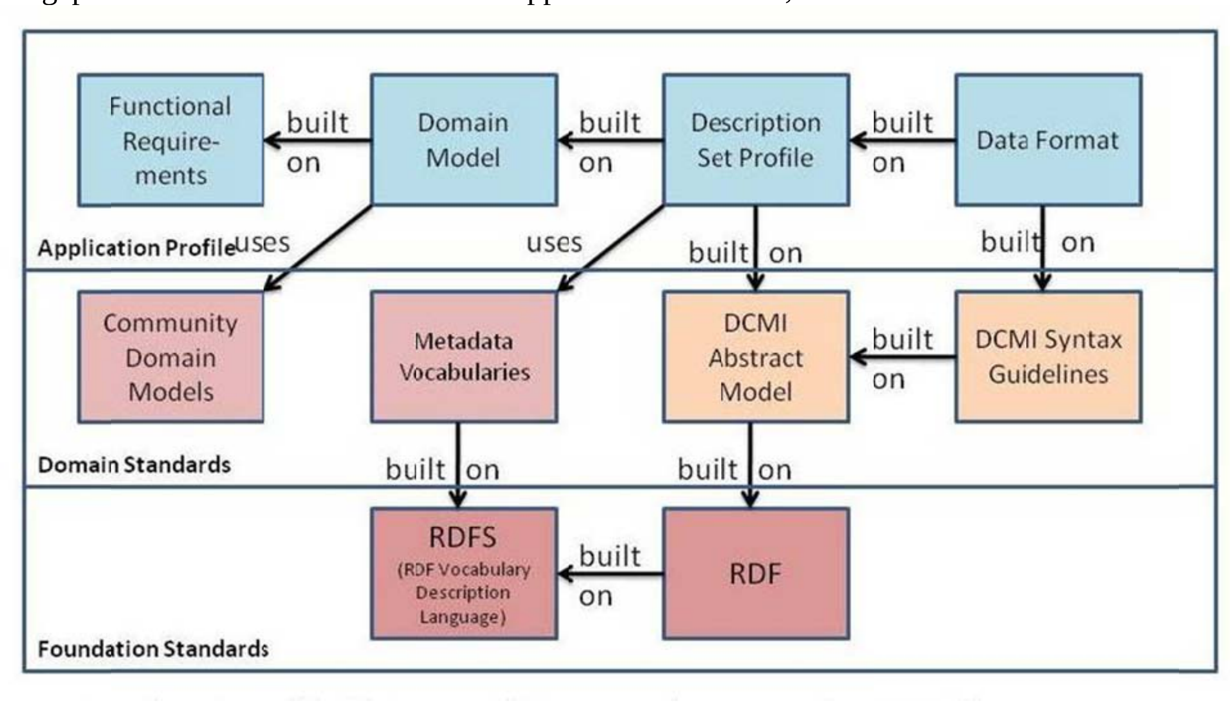
- To create customized Languages of Description – concept schemes, element sets, value vocabularies, ontologies, and application profiles – tailored to target communities yet semantically consistent with other sources in the Linked Data environment.

- To translate legacy knowledge organization systems for use in Linked Data using the W3C standard Simple Knowledge Organization System (SKOS).

- To read, interpret, manipulate, and visualize RDF data, converting between interchangeable syntactic representations and using Semantic Web search engines to discern usage patterns in data encountered "in the wild".

- To scale up and automate metadata creation through effective use of "semantic" wikis and modern, RDF-enabled content management systems such as Drupal[10].

The overarching challenge is one of supporting the full life-cycle of structured data – starting with the development of a metadata design from user requirements, defining a domain model of resources to be described, decorating that domain model with descriptive properties, declaring new properties and classes to fill gaps in available vocabularies, pulling vocabularies and domain models together with application-specific constraints into Application Profiles, and instantiating the resulting Application Profiles in concrete implementation formats. One important conceptualization of the relationships between these descriptive components is offered by the Singapore Framework for Dublin Core Application Profiles[11], illustrated below.



National Impact

The national impact of Linked Data was addressed in the opening section of this narrative. Linked Data represents a crucial new paradigm for publishing information on the Web and a critical stepping-stone to the emerging Semantic Web. By identifying data relationships and categories with URIs, Linked Data effectively provides both footnotes to its own intellectual sources and globally valid hooks for merging data across information silos. Linked Data underpins key initiatives to expose eGovernment data for reuse by citizens and to make traditionally stand-alone scientific datasets available for interdisciplinary exploration.

The challenge facing libraries is how to expose their rich legacy of data – catalog records, authority files (e.g., VIAF[12]), thesauri and classification schemes, and all forms of contextual

[10] http://drupal.org/drupal-7.0

[11] http://dublincore.org/documents/singapore-framework/

[12] http://viaf.org/

resource description – as Linked Data and make it available, and thus discoverable, through cross-referencing and rich interlinking with related sources of information outside the library walls, thereby extending the reach and impact of library data in the wider context of the Web and pulling into the library environment rich supporting resources from elsewhere.

Working groups such as the W3C Library Linked Data Incubator Group[13] are focusing on issues involved with "translating" existing library and standards into the language of linked data. This IMLS proposal focuses on a smaller but crucial aspect of the problem of moving libraries into the Linked Data space: that of providing faculty and professional trainers with a technological platform for teaching the methods and practices of Linked Data to the new generation of librarians and information professionals that will be needed for this approach to have an impact on a broad scale.

In a general sense, this goal aligns with the IMLS 21st Century Skills Initiative, which aims at developing the role of libraries in imparting 21st century knowledge and skills to current and future generations in support of a world-class workforce. We note an interesting parallel to the Gateway to 21st Century Skills supported by the project partner JES & Co[14]. While different in focus, the goals of the two initiatives are clearly complementary. Whereas the IMLS concept emphasizes the role of libraries and museums in educating their patrons about modern information skills and technologies, JES & Co's Gateway concept targets teachers and learners – both students and mid-career professionals – who need to use and share curriculum materials across state and national boundaries.

This proposal for a IMLS Collaborative Planning Grant further narrows the focus on 21st century skills by targeting the creation of a technological platform for teaching Semantic Web design skills to next-generation librarians and information professionals.

Inasmuch as the content accessed by students and learners will be the increasingly vast corpus of vocabularies, authorities, catalogs, and datasets produced and managed by libraries and others as Linked Data, the project will make these resources available for advanced analysis and management through innovative use of technology-based tools. A unified platform providing a learning environment for novice learners will encourage experimentation and innovation by new professionals and thus speed the integration of library resources into the Linked Data cloud.

Intended Results

This one-year planning project will focus on preparing, holding, and following up on a workshop that will bring together software and technology experts with teachers and trainers for the purpose of identifying the needs and requirements of an open-source technology platform usable in classroom or seminar settings – online or face-to-face – to teach students the fundamentals of metadata design using Languages of Description. Definition of critical components of this platform will be driven by requirements deriving from an analysis of essential knowledge and methods students will need to master in order to interpret Linked Data and create effective applications.

[13] http://www.w3.org/2005/Incubator/lld/

[14] http://www.jesandco.org/weblink-cat-ourprojects/web-cat-gateway

The intended results of this IMLS planning grant will therefore be:

1. a set of requirements for a technical (software) platform, based on an analysis of teaching needs in the area of metadata design using Languages of Description;
2. an inventory of existing and emerging open-source tools available for integration into such a platform;
3. an assessment of software development effort required to integrate the components, perhaps using a plug-in architecture with an "orchestrator" function[15] for scripting typical workflows; and
4. the specification of documentation that will be needed to make the package usable by its intended audience.

The project team intends to make this result available for feedback and comment through a website hosted at the University of Washington (see Phase 3 in Project Design below). We will also disseminate on mailing lists of the Dublin Core Metadata Initiative[16] and World Wide Web Consortium, notably the emerging Library Linked Data community[17] and through college and university level LIS partners in the growing iCaucus.[18]

This result, in turn, is intended to inform a proposal for a subsequent NLG project grant that would support the actual development of the technical platform. This follow-up proposal would be submitted to IMLS in January 2013, with an anticipated kickoff in October 2013 and a two-year development cycle. The proposal would be submitted under the category "Advancing Digital Resources" – the NLG program which appears to provide the closest fit to our ideas inasmuch as it includes the use or development of tools and practices "to enhance access, use, and management of digital assets over their entire life cycle" and which have the potential to enhance teaching and learning.

The ultimate goal of both the IMLS Level I Planning Grant and the follow-on IMLS project on Advancing Digital Resources is to package a well-documented set of tools that professors at information schools, trainers in libraries, and self-guided learners at their workstations can use to experiment with the creation and consumption of Linked Data – from the initial brainstorming of requirements through the creation of application profiles to the manipulation and interpretation of the resulting data. Documentation and guidance in the use of the platform will be used to "teach the teachers", enabling them to use the platform as the basis for hands-on curriculum modules for teaching the principles and processes of Semantic-Web-enabled metadata.

Project Design and Evaluation Plan

As indicated in the section on Intended Results, above, the primary activity for this planning grant will be to organize and conduct a two-day workshop in January 2012 to bring together key participants. The workshop will focus on developing a framework for a platform that can be

[15] http://dcpapers.dublincore.org/ojs/pubs/article/view/803/799
[16] http://www.jiscmail.ac.uk/lists/dc-general.html
[17] http://lists.w3.org/Archives/Public/public-lld/
[18] http://www.ischools.org/site/about_icaucus/

used to effectively teach design of metadata in a Linked Data environment. This planning grant will achieve a clear formulation of the goals and objectives for a full-scale project by providing a structured process for bringing together developers who have been working on critical components of the software necessary for an instructional platform and educators who have been involved in teaching elements of metadata design.

The principal investigators will ensure that there is adequate room for discussion and deliberation among this group and others prior to, during and after the workshop, to evaluate the key elements of the proposed learning platform and ensure that it meets the educational community's needs.

To achieve these goals, a phased approach will be used for the project:

- Phase 1 (October 2011-December 2012): The project partners (University of Washington, Kent State University, University of North Carolina, JES& Co and Talis Inc) will work together during the three months prior to the planned workshop to develop a tentative set of requirements for a learning platform and a preliminary list of open source tools that can be used as a starting point for discussions. These requirements and other materials will be sent to the workshop participants prior to the workshop as preparatory reading to ensure that workshop time is used effectively.
- Phase 2 (January 2012): At the workshop itself, the participants will methodically work through these initial requirements, identifying gaps, fleshing out details, and concluding with a set of recommendations for a fully realized framework, with areas needing further work defined and scoped. We will also extend the list of potential open source tools, develop a preliminary gap analysis which will inform areas where further development might be needed, and identify documentation and supporting materials necessary for effective implementation of the platform. In addition, time will be devoted to defining a list of interested parties who should be invited to review and comment on the draft framework, to gather wider input.
- Phase 3 (March 2012-June 2012): To foster discussion and comment on the framework following the workshop, the draft framework and other results of the workshop will be published on a website hosted at the University of Washington, with the extended group of participants identified at the workshop (along with the original workshop participants) specifically invited to contribute comments and suggestions for improvements.
- Phase 4 (July 2012-September 2012): The results of this discussion will be consolidated and integrated with the workshop products into a final deliverable during the summer of 2012, comprising the four elements identified in the Intended Results section.
- Phase 5 (September 2012-January 2013): While not part of this planning grant, an important follow-on activity will occur after the end of the grant, as key participants from the planning grant will work together to construct a full scale proposal to IMLS in 2013 for a National Leadership project grant to build the learning platform and test it in both academic and professional settings.

The ultimate measure of success for this planning grant will be the production of a well-constructed framework and approach for developing an open platform for teaching Linked Data that can be used in both academic and practical settings. Acceptance by the broader community through the comment period embedded in the grant proposal will be the critical test for validation of the result. A formal proposal that is ready for submission to IMLS by the deadline for NLG research project grants for the 2013 cycle is an added outcome that will provide evidence of the success of the work undertaken in the planning grant.

Project Resources: Budget, Personnel, and Management

Project Personnel and Management

The University of Washington Information School will be the lead for this project, with Joseph Tennis and Michael Crandall acting as co-principal investigators; a graduate research assistant from the Information School will be engaged to help synthesize the results of the workshop and produce the final deliverables during the summer.

Dr. Tennis has been active in the knowledge organization and metadata community for many years, through his involvement with the International Society for Knowledge Organization, the Dublin Core Metadata Initiative, the American Society for Information Science and Technology Classification Research Group, and his teaching and research.

Mr. Crandall has extensive industry experience with metadata implementation and software, as well as being a founding member of the Dublin Core Metadata Initiative Oversight Committee and a frequent speaker at conferences on metadata and information organization. He is currently Chair of the Information School's Master of Science in Information Management Program, and has been a principal investigator or consultant on two IMLS grants in recent years.

An external consultant, Thomas Baker, will be engaged as Research Manager and thought partner in the project. Dr. Baker is currently Chief Information Officer of the Dublin Core Metadata Initiative (DCMI) and Chair of the DCMI Usage Board, the committee for maintaining DCMI's standards, such as DCMI Metadata Terms. With an MLS from Rutgers and a PhD from Stanford University, Tom has worked with one foot in the research world and one in libraries. From 2006 to 2009, he co-chaired the W3C Semantic Web Deployment Working Group, which standardized Simple Knowledge Organization System (SKOS), currently used by the Library of Congress, New York Times, and numerous other organizations as the de-facto standard for making existing thesauri and classification schemes usable in a Linked Data context. He is currently Co-Chair of the W3C Library Linked Data Incubator Group, which is bringing together key institutions to formulate a coordinated strategy for implementing Linked Data in libraries, and a member of the W3C Semantic Web Coordination Group. Tom is a co-author of key DCMI technical specifications that were used as a basis for software development by Talis Inc and JES & Co.

These principal personnel have all been involved in Linked Data efforts or organizations that have worked with precursors to Linked Data in educational and other settings for a number of years. More information on their activities and qualifications is available in the attached List of Key Project Staff and Consultants and Resumes.

Partners and Workshop Participants

A number of organizations and individuals have already committed to participating in this project with the University of Washington, including four formal partners. Two educational institutions (Kent State University and the University of North Carolina) have provided partnership commitments. Marcia Zeng and Jane Greenberg, the two leads from these institutions, bring to the project many years of experience teaching metadata design at university schools of information, and have been very active in the broader community engaged in discussions of languages of description for many years. Learnings from the IMLS-funded HIVE project[19] at UNC's Metadata Research Center will be of particular value in this planning grant.

In addition, two software organizations, JES & Co and Talis, Inc. have signed on as formal partners. Individuals who have indicated they will participate from these organizations include Diane Hillmann, Jon Phipps, and Stuart Sutton from JES & Co, who have collaborated on the development of the open metadata registry, GLRC Metadata Commons; and Dave Wood, Vice President for Engineering at Talis Inc, a veteran of standardization and implementation of Semantic Web technologies and member of the W3C Semantic Web Coordination Group, which plans activities of the World Wide Web Consortium in this strategic area.

We have also obtained informal commitments from a number of key participants for the workshops who will bring essential knowledge and contributions to the project. These include trainers in institutional environments such as the independent consultant Karen Coyle and library software developer Corey Harper, each of whom brings deep knowledge of library standards and an up-to-date grasp of technological solutions to courses and seminars they teach to mid-career professionals. Finally, Emma Tonkin, a representative of UK-funded (JISC) projects and member of the DCMI community, will bring a perspective formed from extensive testing of software-supported methods for eliciting metadata requirements from experts and end-users – a crucial phase at the start of the workflow leading from user requirements to concrete metadata formats and software applications.

Starting from this core of participants, the project team will extend the size of the workshop up to a maximum effective size of fifteen, filling out the list so as to ensure a healthy balance between people who bring a "pedagogical" perspective, and those who bring a "technological" (software and standards) perspective, bearing in mind that many of the planned participants are experienced on both sides of the equation.

Budget

The research manager will be budgeted at a constant salary throughout the project to ensure continuity in management and ongoing attention to oversight. The principal investigators will each contribute one full month of salaried time, distributed as needed throughout the project. Other costs are specifically related to the workshop itself, and cover travel and expenses for 15 participants beyond the project principals.

A more detailed discussion of budget items is contained in the Budget Justification attachment.

[19] http://ils.unc.edu/mrc/hive/