Connecting libraries to the unfamiliar data and metadata resources in the Linked Open Data (LOD) Universe -- The Metadata Vocabulary Junction (MV-Junction) Project

1. Assessment of Need

1.1 Background

This research responds to the question of how libraries can benefit from the data and metadata resources that are now available to use in the wake of the Linked Open Data movement. Within the last three years, the world has welcomed the meteoric rise of linked data, as Sir Tim Berners-Lee highlighted in his talk, "The Year Open Data Went Worldwide," at the TED conference in February 2010[1]. Linked Data is about using the Web to connect related data formerly isolated in small or large repositories (often called "silos") and not previously linked. It is about using the Web to lower the barriers to link data that have previously only been brought together using other more cumbersome methods.[2]

Growth of linked datasets has been very quick; for example, a Linked Open Data (LOD) cloud that started with only twelve datasets in May 2007 had already exceeded 200 sets by September 2010. The LOD cloud graph categorizes datasets into seven types: Media, Geographic, Publication (including bibliographic data and authority data from library communities), User-generated content, Government, Cross-domain, and Life Sciences.[3] Joining 13 other nations in establishing open data, the United States government also launched an official service, Data.gov, in May 2009, where thousands of datasets and tools can now be found. The Data.gov service hosts one of the largest open collections of RDF datasets in the world[4] and is presented as one dataset in the LOD cloud.

1.2 Libraries as Contributors and Users of the Linked Data

The Linked Data wave has already pushed libraries to the forefront of innovative services, as the term "Linked Data" is being echoed at many library and information professional conferences, reports, and Webinars during the last two years. One after another, the national libraries and a number of major archives in the world have published their bibliographic data, authority data, and controlled vocabularies as Linked Data and formed an impressive list of datasets in the LOD cloud.[5] In summer of 2010, a W3C Linked Library Data Incubator Group (LLD XG) was established to bring all these activities together in pursuit of common goals. The LLD XG intends to "explore how existing building blocks of librarianship—such as metadata models, metadata schemas, standards and protocols for building interoperability and library systems and networked environments—encourage libraries to bring their content, and generally re-orient their approaches to data interoperability towards the Web, also reaching to other communities."[6] Libraries have heard the call to open the silos. Soon, best practices guides and use cases will be available to the library community because of LLD XG's efforts and examples set by pioneering libraries. Libraries are being sought to contribute their data and metadata to the Linked Open Data universe.

In addition to their roles as contributors, there is another role for libraries: as users of Linked Data. With Linked Data technologies, libraries can efficiently reach a much wider range of data content, including geographical, scientific, social science, biographic, and historical information, as well as a much richer and more diverse data universe, including archival collections, statistics, moving images and sound materials, multimedia, real-time data, and so on. Linked Data can help libraries more effectively provide services to their users through their already existing digital collections, web-based directories, and catalogs, without significantly increasing the library's workload or doing large reengineering projects for their existing bibliographic databases and Websites. Further more, with Linked Data technologies, libraries can aggregate

---

[1] *Tim Berners-Lee: The year open data went worldwide.* TED. (2010 February). Retrieved from http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html

[2] Linking Open Data (LOD) Project Webpage. Retrieved from http://linkeddata.org/

[3] The Linking Open Data cloud diagram. (2010). Retrieved from http://richard.cyganiak.de/2007/10/lod/

[4] Data.gov. (2011). Retrieved from http://www.data.gov

[5] CKAN (Comprehensive Knowledge Archive Network). *Library Linked Data.* Retrieved from http://ckan.net/group/lld

[6] W3C. (2010). Library Linked Data Incubator Group. Retrieved from http://www.w3.org/2005/Incubator/lld/

data based on the pieces/chunks of information they need from a dataset without integrating a whole database or converting full metadata records. They can mash up metadata statements (not whole records) from different namespaces or aggregate data from a variety of resources, based solely on what is needed.

Many datasets available as Linked Data have previously been unknown or unfamiliar to many libraries, however. Examples of such unfamiliar datasets include: DBpedia – the RDF form of Wikipedia; DBLP Bibliography, which includes more than 1.5 million publications and 400,000 authors;[7] GeoNames, which encompasses 10 million geographical names and consists of 7.5 million unique features;[8] The Friend of a Friend (FOAF) dataset, which is a Web of machine-readable pages describing people;[9] and, over 100 specialized government and scientific datasets. In order to mine these datasets and obtain what might be useful, one has to understand the data structures and metadata terms (also known as "properties" in LOD terminology) used by those datasets.

1.3 Sorting out diverse metadata vocabularies

Dozens of metadata standards were developed by various user communities in the 1990s in response to the rise of the World Wide Web and the invention of digital libraries,. Examples of popular standards used in memory institutions include the Dublin Core Element Set (DC), the Metadata Object Description Schema (MODS), Categories for the Description of Works of Art (CDWA), Encoded Archival Description (EAD), and the Core Categories for Visual Resources (VRA Core).[10] In the 2000s, more specialized metadata standards have joined the crowd through two main avenues: (a) independently developed schemas such as PBCore for public broadcasting, the IPTC photo metadata standard that comes with Photoshop software, and XMP metadata for all Adobe documents; (b) application profiles of an existing metadata standard, such as the Dublin Core-based Scholarly Works Application Profile and the National Library of Medicine Metadata Schema. Both the independently developed schemas and the application profiles reflected a natural and initial reaction to the need of "structured data about data" for describing and discovering the attributes of various resources.

The diversity of metadata tools also embraces vocabularies that focus on the "things," rather than the media carrying them. In the library and archive world, these tools include the MARC Authority format and the Encoded Archival Context Initiative (EAC), which provide encoding descriptions of persons, corporate bodies, and families related to works and archival records. Beyond these cultural heritage communities, there are many other vocabularies that provide similar functions; for example, the FOAF Vocabulary is for describing people and the things to which they are related. Darwin Core is designed to help describe biodiversity data, thus it includes metadata terms such as "phylum," and "taxonRank," and "nomenclaturalCode".

A related issue is whether the data carried by these vocabularies is interoperable. When anyone needs to aggregate the data values or make links through RDF triples, it is important for them to understand what the relationships are between and among these metadata terms from different namespaces. Are the values encoded with the following metadata terms (i.e., properties) interchangeable: dc:creator, mods:name, foaf:name, and vra:agent? If a person's name was encoded with these metadata terms, could they be linkable: dc:subject, foaf:focus, mods:name, and skos:concept? If a date range is encoded in one system, could it be converted into another: vra:earliestData + vra:latestDate = foaf:age, = madsrdf:DateNameElement? These questions are not merely a matter of properties matching practices; the data models behind these metadata vocabularies need to be compared when aligning the metadata elements.

Adding to the diversity and complexity is the mixing of expression methods used by these metadata vocabularies. In the Linked Data environment, ontologies are developed to establish relationships, properties, and functions between terms or concepts, rather than metadata element sets. An example is the newly developed ontology, Bibliographic Ontology. More conventional metadata vocabularies are utilizing RDF vocabularies,

---

[7] Schloss Dagstuhl. (2011). Open-Source Reference for Scientific Literature in Informatics: DBLP and Schloss Dagstuhl join forces. Retrieved from http://www.dagstuhl.de/no_cache/en/about-dagstuhl/news/detail/meldung/345/

[8] GeoNames (2007-). About GeoNames. Retrieved from http://www.geonames.org/about.html

[9] The Friend of a Friend (FOAF) project. Retrieved from http://www.foaf-project.org/

[10] For details of the many metadata vocabularies mentioned in this proposal, please consult the supplmentary Reference list and an additional list: *Metadata Standards--Metadata schemas, application profiles, and registries.* Zeng, M. L. (Ed.) Available at: http://www.metadataetc.org/book-website/readings/appendixaschemas.htm

such as the new MADS/RDF (Metadata Authority Description Schema in RDF) vocabulary resented as an OWL ontology, and others which are in development, e.g., the Digital Multimedia Repositories Ontology (DMRO).

Without a bird's eye view of what is out there on the ever expanding metadata landscape, without a place that functions as a junction where the many routes, lines, or roads of descriptive practice can meet, link, or cross each other, it will be difficult for libraries to connect to those unfamiliar resources in the Linked Open Data world. This project aims to research and build such a junction through the alignment of metadata terms that are used by different communities and data providers within and beyond the library world.

2. Intended Results and National Impact

2.1 Intended Results

The problem to be addressed in this research proposal is the alignment of metadata terms from diverse namespaces. Specifically, this proposed Metadata Vocabulary Junction (MV-Junction) project aims to become a bridge that helps connect libraries to the unfamiliar data and metadata resources in the LOD universe by analyzing and aligning metadata terms that are used by different communities and data providers. The researchers will store the results of these alignments in a knowledge base, using the tool WebProtégé[11], which can then be output on demand as various deliverables including: web pages, tables, forms, concept maps, comparison matrices, and text through a dedicated website. Aligned metadata terms will be linked to the datasets found in the LOD cloud.

The outcome of this proposed research will differ from products generated from previous efforts in metadata mapping and crosswalking:

1. The resulting MV-Junction will not be another crosswalk or registry, nor will it aim to just attach more schemas to existing crosswalks or register more vocabularies in a central service, although these already-established resources will be consulted as part of the project. Rather, the focus of the MV-Junction research will be to align metadata terms from different namespaces according to their semantic meaning, and to analyze the degree of alignment when similar metadata terms are compared with one another. For example, the term "RightsSummary" as found in the Public Broadcasting Data Dictionary (pdcore) would be compared with the term "rightsHolder" as defined in the Dublin Core Metadata Terms (dcterms) to determine whether the former conveys a broader meaning than the latter.
2. Such semantic relations between metadata terms as found in the above example will be expressed with W3C recommended languages including SKOS, OWL, and RDF Schema. Thus, the two metadata terms will not just be "crosswalked" as:

   pbcore:RightsSummary | dcterms:rightsHolder

   but will be expressed as:

   pbcore:RightsSummary *skos:broadMatch* dcterms:rightsHolder

   Other semantic relationships to be expressed include same-as, close-match, exact-match, broad-match, narrow-match, related-match, among others.
3. Metadata vocabularies to be included in the MV-Junction will be far more numerous than those considered as "core" in the libraries, museums, and archives fields. Take the metadata used in biodiversity research for example, the pertinent vocabularies include not only Dublin Core, but also CiteData Metadata Scheme for the Publication and Citation of Research Data, Darwin Core, and the OBIS (Ocean Biogeographic Information System) Schema.
4. The metadata alignments will reach a number of metadata vocabularies found in other subject domains and communities, with other types of data than that in the bibliographic world. The primary criterion for inclusion in the project will be the potential of a vocabulary to be used to connect libraries to useful outside datasets.
5. The alignment results will not simply align the metadata terms next to each other (as most crosswalks have done) based on the schemas. The proposed research will conduct analysis in the context of datasets, with the understanding of actual data instances attached to each metadata term included in the study. Alignment

[11] Stanford Center for Biomedical Informatics Research. (2010). Protégé. Retrieved from http://protege.stanford.edu/

results will be evaluated through additional data conversion testing, thus the results will be reliable and applicable in the real environment.

Because it is not possible or realistic to include every metadata vocabulary found in every subject domain or for every conceivable format in this research, the MV-Junction project and intended results aims to create a structure for further development and lay the framework for additional research. It is also not reasonable to require a library to become a research lab or production unit when considering datasets to be connected (refer to 3.3 Limitations). The project will select the most closely related areas and well-established metadata vocabularies as the starting point and identify additional potential targets as the project moves on.

2.2 Significance of the Project

Few research projects, if any, have addressed this research question from this perspective (see the section above). The LOD project represents a new generation of the "Web of data" approach that supports the reuse and discovery of data and metadata resources, encourages open source development, and consequently simplifies the process of metadata creation and data aggregation. Any effort to facilitate the discovery and use of the treasures in the LOD datasets would have a national impact, as all members in the communities of libraries, museums, and archives will be able to benefit from the resulting rich metadata resources. Historically, when a new technology becomes available it is the larger, more resource-rich institutions that are well positioned to invest in the experimentation that results in new services. Middle- and small-size libraries, archives, and museums, which tend to focus more on collections and content rather than the new technologies, are naturally the secondary group to benefit from the technologies. The LOD universe presents a great opportunity for any institution to reach and discover the treasure of data resources, regardless of its status. Currently, there are few tools available when a library needs to explore this opportunity. It is also clear that simple, yet powerful tools are needed for a novice creator to generate his/her specialized metadata or mash up the available data that implement RDF principles. The proposed Metadata Vocabulary Junction will be one of the tools available to assist any size or type of library (and other memory institutions) to connect to this treasure.

The intellectual merit of this proposed study is that it will advance our knowledge of how the current data models and metadata vocabularies influence the reuse and advancement of existing services of the library field. The LOD project provides a unique resource for metadata term alignment study. It includes many datasets that contain structured data or semi-structured data supported by heterogeneous data structures. The proposed research will be a pioneer study to address issues of libraries as the users of Linked Data. It will also be a unique study on how to align metadata terms from diverse communities, specifically how to relate metadata vocabularies familiar in the library community to the unfamiliar resources of the LOD universe. In addition to uncovering inter-relations of metadata vocabularies, problems that could arise as the result of inappropriate alignments and aggregation will also be studied and documented.

Already, it is clear that the technological groundwork has been laid for libraries to benefit from Linked Data. What is now needed is the intellectual preparation to support the discovery and reuse of LOD datasets as well as the practical tools to facilitate such activities. The MV-Junction project will fulfill both of those needs. With a "junction" showing the available metadata terms of diverse namespaces and how they can be used correctly and unambiguously, the library community will be able to take full advantage of these valuable data resources.

3. Project Design and Evaluation Plan

3.1 Goals and Objectives of the Project

The overall goal of the project is to provide the libraries with an integrated tool, the Metadata Vocabulary Junction, which will enable them to connect, discover, and use the unfamiliar data and metadata resources in the LOD universe. This tool will pave the way for transparent application of sophisticated diverse metadata vocabularies from different namespaces. Ultimately, the objectives of the study can be summarized as follows: (1) to determine the semantic relationships and align the metadata terms in the context of the metadata vocabularies from different namespaces; (2) to develop an integrated tool to facilitate discovery of matched metadata terms, reorganization of metadata terms, and use of the selected metadata terms to link and aggregate useful data based on the needs.

3.2 Research Methodology and Process

The problem addressed in this research proposal is the alignment of metadata terms from diverse namespaces. It examines semantic interoperability of metadata as well as related intellectual and technological solutions for libraries in the LOD universe. To reach this end, the research will deal with eight interrelated tasks.

1. Selecting datasets to be studied from the LOD cloud. The first group of metadata vocabularies to be included in the research (around twenty) will depend upon the datasets selected. A pre-test of the project found that available datasets may be selected according to different dimensions: datasets for "things" (e.g., places, events, people, species); medium (e.g., news, bibliographic data, research data, census and other statistics), and user community (e.g., biodiversity, education, government). The results of other tasks (see below) may lead to further selection for the second group: either more datasets for the same kind of data but with different data structures (e.g., various geo-name datasets), or new types of datasets for other things, media, and user communities.
2. Examining selected metadata vocabularies or data structures that are used by the selected LOD datasets. The focus of this task is on their patterns at different levels, including but not limited to: structure of the schema, core metadata terms, usage constraints, class ranges, targeted user communities, and existing mapping with other vocabularies.
3. Collecting and examining existing individual metadata "crosswalks," or "maps," their quality, and the consistency among the crosswalks which have covered the same metadata vocabularies.
4. Creating the knowledge base for the MV-Junction for the metadata terms studied and their semantic relationships. An open source tool, WebProtégé, will be used to store classes, properties, instances, and annotations. This key product of the research will provide a semantically controlled structure to individual metadata terms as well as their mapping status, which are expressed with SKOS mapping properties. Aligned metadata terms will also be linked to the datasets found in the LOD cloud.
5. Collecting and examining metadata instance samples from selected datasets and forming a test bed. Fifty to one hundred records will be downloaded from each selected dataset using a random sampling method. The test bed will contain two types of testing data: packed complete metadata records (each containing all statements about one resource) and individual metadata statements (each containing one property and value(s) about one resource).
6. Evaluating the results stored in the knowledge base of MV-Junction using the metadata instances stored in the test bed. These metadata instances provide the context of the meaning and usage of specific metadata terms. The project will test the alignment result using two approaches: 1. Analyzing metadata instances packaged as part of the datasets to verify the alignment decision. 2. Using the alignment decision to go back to datasets and compare the "before" and "after" data conversion results. The results will also be evaluated by peer reviewers and consultants.
7. Modifying the knowledge base according to the testing results and the evaluation comments from peers and consultants.
8. Creating a Web-based tool for browsing and generating various deliverables of the data derived from the MV-Junction's knowledge base on demand. The deliverables of document formats may include web pages, tables, forms, concept maps, comparison matrices, textual explanations, and flowcharts.

3.3 Limitations of the Project

The study has several limitations: (1) The research sample to be used is limited to the metadata vocabularies of selected datasets found in the LOD cloud as of Feb. 1, 2010. It will be beyond the capability of this project to include every metadata vocabulary found in every subject domain or for every conceivable format. (2) Although all datasets could potentially be of value for a library's users, it is also not reasonable to require a library to become a research lab or production unit. Therefore when considering datasets to be connected, the project will select only the most closely related areas. (3) Only well-established metadata vocabularies and data structures of datasets will be selected. (4) The study covers descriptive metadata only, excluding administrative and technical metadata. (5) Although the alignment results from this study will be

grounded in comprehensive intellectual work, the final product may not suitable for implementation in metadata linking or aggregation projects without further testing.

4. Review of Literature and Related Projects

References that are closely related to this project first come from metadata mapping methods and technologies used in the library community prior to the LOD movement. Additionally, "alignment" is a newer terminology about mapping, and "metadata term" and "property" often replace the term "elements." The literature review will use conventional terminology respectively.

4.1 Mapping Methods and Technologies Used in the Library Community

The process of mapping essentially consists of establishing equivalencies between elements in different metadata element sets. In *direct mapping*, one-to-one mapping is usually applied when two (or a limited few) schemes are involved. *Cross-switching* is another kind of model usually applied to reconcile multiple schemes. Instead of mapping between every pair in the group, each scheme is only mapped to a switching scheme, e.g., Getty Research Institute's crosswalk, which allows multiple metadata schemas to all crosswalk to CDWA.[12]

The mapping results usually are presented as metadata *crosswalks* -- often a chart or table that represents the semantic mapping of data elements in one data standard (source) to those in another standard (target) based on similarity of function or meaning of the elements.[13] Common aspects may include a semantic definition of each metadata element and other issues including such pre-conditions as mandatory requirements, cardinality, constrains of relative elements, allowable data range, and flexibility for local terms.[14] Major efforts in metadata mapping have produced a substantial number of crosswalks. Almost all schemas have created crosswalks to widely applied schemas such as Dublin Core and MARC. Metadata specifications may also include crosswalks to a previous version of a schema as well as to other metadata schemas, e.g. the Visual Resource Association Core Categories (VRA Core).[15]

Efforts to establish a crosswalking service at OCLC have indicated the need for robust systems that can handle validation, enhancement, multiple character encodings, and allow human guidance of the translation process.[16] Researchers at the National Science Digital Library (NSDL) have also included a crosswalking service in their sequence of metadata enhancement services.[17] Both element-based and value-based crosswalking services assist in achieving semantic interoperability and improve the reusability of metadata in a variety of knowledge domains.

Two crosswalking approaches have emerged in practice. The *absolute crosswalking* approach requires exact mapping between involved elements (e.g. vra.title → dc.title) of a source schema (e.g., VRA Core) and a target schema (e.g., DC). Where there is no exact equivalence, there is no crosswalking (e.g. vra.technique → [empty space]). Absolute crosswalking ensures the equivalency (or closely-equivalent matches) of elements, but does not work well for data conversion because when there are unmapped elements, data associated with them will not be converted. To overcome this problem, *relative crosswalking,* (which can be considered as an "*alignment*" process), is used to map all elements in a source schema to at least one element of a target schema,

---

[12] Harpring, P., Woodley, M., Gilliland-Swetland, A., & Baca, M. (Eds.). *Metadata Standards Crosswalks*. Retrieved from http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html

[13] Baca, M., Gill, T., Gilliland, A. J., & Woodley, M. S. (2000). *Introduction to Metadata: Pathway to Digital Information*. Online edition, Version 2.1. Glossary. Retrieved from http://www.getty.edu/research/conducting_research/standards/intrometadata/glossary.html

[14] St. Pierre, M., & LaPlant, W. P., Jr. (1998). *Issues in Crosswalking Content Metadata Standards*. Bethesda, MD: NISO Press. Retrieved from http://www.niso.org/publications/white_papers/crosswalk/

[15] *Visual Resource Association Core Categories (VRA Core).* Retrieved from http://www.vraweb.org

[16] Godby, C. J., Young, J. A., & Childress, E. (2004). A repository of metadata crosswalks. *D-Lib Magazine*, 10(12). doi:10.1045/december2004-godby

[17] Phipps, J., Hillmann, D. I., & Paynter, G. (2005). Orchestrating metadata enhancement services: Introducing Lenny. *Proceedings of the International Conference on Dublin Core and Metadata Applications. Madrid, Spain*. 57-66. Retrieved from http://arxiv.org/ftp/cs/papers/0501/0501083.pdf

regardless of whether the two elements are semantically equivalent or not (e.g., vra.technique → dc.format).[18] While many crosswalks exist, the degrees of mapping (exact, close, broader, narrower, and overlapping, etc.) have NOT been implemented in the metadata mapping practice, although they have been used in subject vocabulary mapping. This proposed research will employ the mapping terminology and semantics into metadata terms alignment process when conducting relative crosswalking.

Regrettably, most metadata crosswalks did the mapping based on metadata specifications, and not on real data conversion results. The major challenge in *converting records* prepared according to a particular metadata schema into records based on another schema is how to minimize loss, or distortion, of data. In studies by Zeng and colleagues,[19,20] it was found that when data values were involved, converting may become imprecise and conversion tasks become more complicated. Their study on metadata quality provides strong evidence for the impact of crosswalks on quality when converting large amounts of data. The most serious difficulties include: misrepresented data values, important data values lost, incorrectly mapped elements and data values, and missing elements. It is the concern of the proposed MV-Junction that when connecting to unfamiliar data resources these blind mapping process and the consequential errors may occur. This is also why the project wants to implement the evaluation component, to avoid mismatched elements as well as the values that are encoded with them.

When a metadata record includes values from multiple controlled vocabularies, *co-occurrence* of values allows for an automatic, loose mapping between vocabularies. As a group, these loosely mapped terms can answer a particular search query or a group of questions. Existing metadata standards and best practice guides have provided the opportunity to maximize the co-occurrence mapping method. A good example is the VRA Core Categories, which recommend the use of five controlled vocabularies for culture and subject elements. Additionally, metadata records often include both controlled terms and uncontrolled keywords. With *co-occurrence mapping* method, loosely-mapped values will become very useful in productive searching with highly relevant results.

In the LOD cloud graph, all datasets point towards to DBpedia. DBpedia's pages are automatically generated from RDF triples and each include a great number of duplicated metadata statements. For example, for a city's location (e.g., Cleveland, Ohio), different metadata properties from different namespaces will bring the same or related values.[21] These results give us the opportunity to examine the mapping degree of those properties. A recent study in which Zeng took part[22] looked at 250 concepts in agriculture and twenty concepts relating to oil spills to determine which properties included in the DBPedia pages may provide synonyms/equivalents. The following properties were found to have synonyms or equivalents: skos:preferredLabel, owl:sameAs, rdfs:label, dbprop:name, dbpprop:offcialName, foaf:name, dbpprop:nativeName, dbpprop:nickname, etc. Approaching these findings from the opposite direction, the values that are considered to be synonymous or equivalent can be used to support the goal of aligning properties (i.e., metadata terms) gathered by DBpedia.

Registries for metadata vocabularies, powered by semantic technologies such as RDF, SKOS, OWL, have emerged in recent years. Their primary functions include registering, publishing, and managing diverse vocabularies and schemas, as well as ensuring they are crosslinked, crosswalked, and searchable. The basic components of a metadata registry include identification of data models, elements, element sets, encoding schemes, application profiles, element usage information, and element crosswalks.[23] The most active and well-

---

[18] Zeng, M. L., & Chan, L. M. (2010). Semantic Interoperability. *Encyclopedia of Library and Information Sciences.* 3rd edition. 1:1, 4645-4662. (Online publication date: 09 December 2009) Bates, M.J., & Maack, M.N. (Eds.). New York, NY: Dekker Encyclopedias, Taylor and Francis Group.

[19] Zeng, M. L., & Xiao, L. Mapping metadata elements of different format. *E-Libraries 2001, Proceedings, May 15-17, 2001, New York.* Medford, NJ: Information Today: 91-99.

[20] Zeng, M. L., & Shreve, G. (2007). *Quality Analysis of Metadata Records in the NSDL Metadata Repository*. A research report submitted to the National Science Foundation.

[21] See Cleveland, Ohio page generated by Dbpedia: http://dbpedia.org/page/Cleveland

[22] Morshed, A., Johannsen, G., Keizer J., & Zeng, M. (2010). Bridging End Users' Terms and AGROVOC Concept Server Vocabularies. *Proceedings of the International Conference on Dublin Core and Metadata Applications, Pittsburgh, October 20-22, 2010.* (Poster) http://dcpapers.dublincore.org/ojs/pubs/article/view/1015/981

[23] SCHEMAS Registry, UKLON CORES project. Retrieved from http://cores.dsd.sztaki.hu/

known metadata registry is the Open Metadata Registry[24] (formerly the National Science Digital Library (NSDL) Registry). It is one of the production deployments of SKOS and serves across metadata, terminology, and services.

4.2 Linked Open Data and the Library Community

Evidence found in Linked Data-related literature and in projects now emerging from the library community suggests that libraries are taking a more active role in contributing to the LOD universe. The activities include numerous presentations at conferences, including those devoted to Linked Data and the libraries: Code4Lib 2009 conference on Linked Data, Talis Open Days on Linked Data and Libraries, 2010, annual International Conference on Dublin Core and Metadata Applications since 2008, and the special session in IFLA 2010 on Libraries and the Semantic Web. In the 2008 LOD graph, there was one dataset that was contributed by the library community: the Swedish National Library's Libris catalogue. In just the last two years, many experimental and full launches of Linked Data services have appeared, which use bibliographic data and vocabularies from national libraries and archives. The Library of Congress has been publishing its vocabularies as Linked Data, including full *Library of Congress Subject Headings*, using SKOS as its language.[25] These implementations from the library sector have been occurring so quickly that reports of such tools are usually only found in conference presentations, announcements, and workshops.

Since August 2010, the W3C Library Linked Data Incubator Group (LLD XG) has been meeting each week through Web conferencing and a number of face-to-face meetings, while also conducting a comprehensive environmental scan. Its mission is to help increase global interoperability of library data on the Web, by bringing together people involved in Semantic Web activities—focusing on Linked Data—in the library community and beyond, building on existing initiatives, and identifying collaboration tracks for the future.[26] One of the deliverables that the group plans is "a use-case document that describes a number of real-world use cases, case studies, outreach and dissemination initiatives targeted to the library community and related sectors express requirements to approach library environments to the Semantic Web."[27] Over fifty use cases were submitted to the LLD XG[28] in 2010 and curated by the XG members, resulting in eight clusters of research and implementation interests.

The library community has concentrated on building its own conceptual models over the last decade, with the most significant result being the FRBR[29] family models for bibliographic data, name authority data, and subject authority data. The connection of all of these FRBR-related models with the Linked Data movement has been established through the RDA properties' registration in the Open Metadata Registry, as well as through the adoption of FRBR in domain models of various metadata application profiles (such as the Scholarly Works Application Profile[30]), and the mapping of the FRSAD (Functional Requirements for Subject Authority Data) concept model to OWL and SKOS data models.[31]

SKOS defines classes and properties sufficiently for representing the common features found in a knowledge organization system such as thesaurus, taxonomy, controlled term list, and other KOS structures. Each SKOS concept is defined as an RDF resource and each concept can have RDF properties attached. The

---

[24] Open Metadata Registry. Retrieved from http://metadataregistry.org/

[25] Library of Congress. (2009 --). Authorities and Vocabularies. Retrieved from http://id.loc.gov/

[26] W3C. (2010). Library Linked Data Incubator Group-Charter. Retrieved from http://www.w3.org/2005/Incubator/lld/charter

[27] W3C. (2010). Library Linked Data Incubator Group Wiki. Retrieved from http://www.w3.org/2005/Incubator/lld/wiki/Main_Page#Deliverables

[28] W3C. (2010). Library Linked Data Incubator Group-Use Cases/Case Studies. Retrieved from http://www.w3.org/2005/Incubator/lld/wiki/UseCases

[29] IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR). (1998). *Functional Requirements for Bibliographic Records - Final Report*. Munich: K.G. Saur.

[30] *Scholarly Works Application Profile (SWAP)*. SWAP Working Group, JISC Digital Repositories programme. Coordinated by J. Allinson and A. Powell. Retrieved March 25, 2010, from http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile

[31] IFLA Working Group on the functional Requirements for Subject Authority Records (FRSAR). (2010). *Functional Requirements for Subject Authority Data – A Conceptual Model.* Zeng, M.L., Žumer, M., & Salaba, A. (Eds.). Retrieved from http://www.ifla.org/node/1297

SKOS mapping properties are used to state mapping (alignment) links between SKOS concepts in different concept schemes, where the links are inherent in the meaning of the linked concepts.[32] SKOS has been naturally adopted and used by libraries for publishing controlled vocabularies. However, for SKOS to be used beyond subject vocabularies is an emerging trend. An interesting movement is the alignment of existing data models to SKOS, as demonstrated by the change from the original VIAF ontology and MADS/RDF[33] ontology to the SKOS/RDF patterns recently.[34,35]

The proposed MV-Junction will benefit greatly from the projects that have already fully mapped properties to SKOS (such as MADS and VIAF). Meanwhile, the project will employ the SKOS mapping properties to encode the semantic relations between metadata terms that will be included in the MV-Junction. This kind of application of SKOS beyond controlled vocabularies has not been reported, to the knowledge of this project's researchers. In addition, although projects and services have been reported, they are mostly operational projects rather than research. The proposed MV-Junction project will be among the pioneer efforts that researches LOD for library use through metadata terms alignment.

5. Project Resources: Budget, Personnel, and Management

5.1 Budget

The total project budget is $274,557.72. The total amount requested from IMLS is $203,386.94. The School of Library and Information Science will provide cost sharing of $71,170.78 to partially cover costs of salary and wages, fringe benefits, supplies, computing facilities (i.e., server space), and some indirect costs.

As outlined in the Project Budget, funds are requested from IMLS for salaries and wages ($95,534.22), fringe benefits ($17,096.88), tuition benefits for a graduate assistant ($26,701.72), consultant fees ($1,000), travel ($6,400), and some indirect costs ($56,654.47). This project requires no additional equipment, no start-up costs, no construction, and no services.

5.2 Personnel

The Project Director and Principal Investigator, Dr. Marcia Lei Zeng, is Professor of Library and Information Sciences at Kent State University. Zeng's major research interests include database quality control, knowledge organization and representation with the focuses on metadata, markup languages, and controlled vocabularies. She has over 60 scholarly publications and four books (including a co-authored textbook titled *Metadata* that are widely used). She conducted a bibliographic database quality study for OCLC bilingual MARC records in the 1990s and developed production rules and a quality-checking guide for eliminating the most-frequently-occurred errors. She also led a large metadata quality research for the National Science Foundation (NSF) National Science Digital Library (NSDL) in the 2000s, which involved metadata mapping as well as data conversion from diverse resources. Her working experiences related to the Linked Data include preparing a recommendation to produce LOD-enabled metadata for an EU project, VOA3R, as a consultant of the FAO of the UN. She has chaired IFLA Functional Requirements for Subject Authority Records (FRSAR) Working Group and currently also serves as an Invited Expert of W3C Library Linked Data Incubator Group.

Co-Director and Co-Principal Investigator Dr. Karen F. Gracy is Assistant Professor of Library and Information Science at Kent State University. Dr. Gracy brings expertise in the areas of digital preservation and audiovisual archiving. Her research and teaching interests include digital preservation, audiovisual archiving, preservation education, and the social contexts of information creation and use, with a focus on ethics and values. Her first book, *Film Preservation: Competing Definitions of Value, Use, and Practice*, was published by the Society of American Archivists in 2007. Other recent publications include "Moving Image Preservation

---

[32] Miles, A. & Bechhofer, S. (Eds.) *SKOS Simple Knowledge Organization System Reference* (2009). W3C Candidate Recommendation 18 August 2009. Retrieved from http://www.w3.org/TR/skos-reference/

[33] Note: The MADS/RDF (Metadata Authority Description Schema in RDF) vocabulary is a data model for authority and vocabulary data used within the LIS community and presented as an OWL ontology.

[34] Young, J. (2011). Email to the LLD XG Public list. Retrieved from http://lists.w3.org/Archives/Public/public-xg-lld/2011Jan/0037.html

[35] Library of Congress. (2010). MADS/RDF Vocabulary Description. Retrieved from http://www.loc.gov/standards/mads/rdf/#t1-5

and Cultural Capital," which appeared in the Summer 2007 issue of *Library Trends* (v. 56, no. 1), and "Film and Broadcast Archives," which appeared in the third edition of the *Encyclopedia of Library and Information Sciences* in 2009. Most recently, her research has focused on the impact of mobile technologies on the consumption and use of moving images in library and archival contexts, particularly how users are incorporating moving image material into their own works, circulating these materials in their social networks, and performing information work through appraisal, description, and preservation activities. She is particularly interested in formal and informal metadata structures for moving images online, and is eager to connect these structures via the proposed MV-Junction.

5.3 Management of Project

The project activities will be monitored and managed by Drs. Zeng and Gracy on a regular basis. Activities to be supervised will include data selection, collection, and analysis; the construction of a knowledge base for the alignment results; and testing of results, evaluation, and development of the deliverables. Drs. Zeng and Gracy will also conduct evaluation of the project tasks throughout the grant period, as well as at the completion of the project. SLIS has the resources, facilities, and infrastructure to successfully complete and manage the project. Please consult the attached schedule of completion, which will provide an overview of all aspects of the project, including a timeline and budget distribution of requested funds over the two-year period.

6. Communication Plan and Sustainability

The research findings of this project will be documented and made available on the Website of the *Metadata Vocabulary Junction*. The research methodology, literature review, database structure, and *Metadata Vocabulary Junction* information architecture will be made available through reports, research papers, and presentations, so that they can be reused or implemented in other projects.

The researchers request support from IMLS for this two-year project. At the end of the project, the researchers plan to continue to donate their time and their Graduate Assistants' time toward the maintenance of the Website. With further funding opportunities and collaboration from other interested parties, the knowledge base and test bed will grow and new metadata vocabularies will be included as the LOD cloud grows. The importance of the project is its methodology, which will be tested and refined throughout the two year project, and will be easily built into a workflow for continuing use by the researchers and practitioners in the library and LOD communities.