Linked Data as an Element to Support Resource Discovery: The Need for Harmonization of Metadata Standards

Praveenkumar Vaidya **N S Harinarayana**

Abstract

With the vision of mounting global interoperability of library data on the Web, bringing together people involved in Semantic Web activities to focus on Linked data in the library community, building on existing initiatives, and identifying collaboration tracks for the future, has tempted library professionals to think way beyond traditional methods of resource discovery and user accessibility. The automation of information systems with the advent of digital libraries has proliferated plethora of metadata standards with dissimilar attributes. A common standard ensures to reach the broadest community of information workers which avoids metadata to be repetitious, time consuming and tedious. Hence, the necessity is harmonization and interoperability of metadata standards to ensure consistency and crosswalk with other standards. Resource Description Framework (RDF) does provide a framework well founded in Web architecture and a formal semantics to achieve harmonization to reach final destination of Linked Data.

Keywords: Linked Data, Resource Discovery, Metadata, Semantic Web

1. Introduction

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data (Berners-Lee, 2006). Over the years the bibliographic and subject standards used in library world are translated into Language of Semantic Web and Linked Data. Linked Data is growing body of datasets on the World Wide Web that are interconnected by means of the Resource description Framework (RDF) (W3C, 2004) a language for specifying relationships between things –using web-based Uniform Resource identifiers (URIs , or Web Addresses) and terms used to describe them (Berners-Lee, 2006). Semantic Web designates the technologies underlying Linked Data.

RDF is fundamentally a grammar for a language of data. It is a language designed by humans to express human thoughts in a form amenable to processing by machines. RDF statements follow a simple three-part sentence structure, the *triple*. (Baker, 2011)-each consisting of a subject, a predicate and an object to form RDF graphs. The assertion of an RDF graph amounts to asserting all the triples in it, so the meaning of an RDF graph is the conjunction (logical AND) of the statements corresponding to all the triples it contains (W3C, 2004).

In 2009 the Library of Congress Subject Headings published library standards in RDF to mark the new beginning towards the implementation of new W3C standard, Simple Knowledge Organisation System (SKOS)- the first step towards a comprehensive service for authorities and vocabularies, id.loc.gov (Library of Congress, 2011). In 2011, a Library Linked Data Incubator Group of the World Wide Web consortium (W3C) acknowledged the idea of Linked Data expressed using standards such as RDF (W3C, 2011). Hence RDF has become a common tool to enable the integration of library data on the web for more expansive user access to information.

2. Objective of the study

In view of 'data deluge' on internet the library world is looking at Linked Data with the help of standard like Resource Description Framework (RDF) to connect all available electronic data of the world to enhance resource discovery. This article deals with the qualitative analysis of issues around the Linked Data laced with metadata interoperability and harmonization, in order to ensure that library data meets traditional standards for quality and consistency and also explains the theoretical conclusions of metadata initiatives.

3. Methodology and Scope

This paper describes at the broader shift of the library community towards an RDF data model with emphasis on Library Linked Data. The methodology presented in this paper is explicitly directed to study current metadata standardization practices that helps solve the issues in metadata harmonization.

This article draws huge inspiration from Library Linked Data incubator group final report (W3C, 2011) and specifically to give emphasis on the recommendations suggested in this report and also goes on to explain how harmonization and interoperability of metadata will achieve Library Linked Data environment for future generations.

- The first section will outline the concept of Linked Data and some of the benefits that linked data could have for libraries, librarians and researchers. The emphasis is to enable Library Data to Library Linked Data which has the semantic value for resource discovery.

- Next part willdiscuss on metadata harmonization and interoperability. It presents an analysis of current issues and challenges and progress of metadata harmonization to achieve Linked Data environment and future of Linked Data.

- Lastly, article concludes with findings about RDF, Linked Data and metadata harmonization and how it has helpedlibrariesto move ahead with next generation of standards to achieve Semantic Web technology.

4. Libraries and Linked Data

Tim Berners-Lee's vision of interlinking whole data of the world shifted the thought process of current generation to create highly structured metadata that allow computers to understand the relationships between objects which yield to precise web search, authority control, classification, data portability and disambiguation which were considered to be domain of library professionals. Since 1999 the W3C has been working on a set of Semantic Web standards also known as Linked Data. By marking up information in standardized and highly structured formats like Resource Description Framework (RDF), we can allow computers to better understand the meaning of content, rather than simply matching on strings of text, which allow web search engine to work like relational databases throwing more accurate search results.

As Linked Data initiatives are found more in numbers, obviously there has been increased debate about exactly what we mean when we refer to Linked Data and the Semantic Web. Are the phrases interchangeable? Do they refer to a specific set of standards including RDF, SPARQL query language, and OWL web ontology language? (W3C, 2010) As for "Linked Data" we will accept the two part definition offered by the research team at FreieUniversitat Berlin, "The Web of Data is built upon two simple ideas: First, to employ the RDF data model to publish structured data of the Web. Second, to (use http URIs) to set explicit RDF links between data items within different data sources" (Isele, et al., 2009).This definition gives two distinct aspects of Linked Data: exposing data as RDF and linking RDF entities together.

4.1 Benefits derived from Linked Data Approach

Librarians with their expertise in search, metadata generation and ontology development are in natural position to understand and implement Linked Data and they have explicit mandate to organize information derived from many sources and to make it broadly accessible (Byrne & Goddard, 2010). Linked Data is sharable, extensible and easily re-usable. With the concepts like language-agnostic URIs, it supports multilingual functionality for data and user services.

Linked Data allows anyone to contribute their unique expertise in form which can be reused and recombined with the expertise of the others. By employing globally unique identifiers like works, places, people, events, subjects and other objects of interest, libraries allow the resources to be cited across broad range of resources (W3C, 2011). Hence librarians are also in a unique position to provide trusted metadata services for long term as a data on the web.

In Linked Data ecosystem any attribute that makes it possible for any connection from any unknown resource makes it useful link, which is unique character for resource discovery. A query can draw related information from any link that is available on the world wide network of data, to get useful results from the web. Hence navigation across the web will be more sophisticated and precise in nature. Links between libraries and non library services such as Wikipedia, Geo names, MusicBrainz, the BBC and The New York Times will connect into larger universe of the information on the Web which seamlessly flow to local user.

In Linked Data, the structured data using such as RDF in attitude (RDFa) and microdata plays a vital role in the crawling and relevancy algorithms of search engines. Hence Linked Data will favour interdisciplinary research by enriching knowledge through multiple domain specific knowledge bases. In fact, to have a common format for all data would be huge relief for interoperability and integration of all kinds of system. Libraries working with vendors to collaboratively develop a large shared knowledge base that could act as a library 'linking Hub'. The linking Hub would expose a network of tightly linked information from publishers, aggregators, book and journal vendors, subject authorities, name authorities and other libraries. (Byrne & Goddard, 2010).Hence large quantities of Linked data can handle all functions of selection, ordering, cataloguing, authority control, taxonomy development and search. The resource discovery will be supported by excellent granularity and capability to handle intelligent queries.

4.2 Current status of Libraries: Issues with Library Data

Despite some movements within the library world which has given birth to publication of key element sets such as Dublin Core Metadata Initiative (DCMI) metadata terms and Reference Framework Bibliographic Records (FRBR) in Linked Data compatible formats, majorityof the library data available today that we have is not integrated with web resources still resides in databases. The huge amount of library bibliographic data is not connected to geographic information, persons and organisations available on web which makes libraries redundant and inaccessible.

The existing library standards such as MARC format or information retrieval protocol Z39.50 are developed for library specific context and these standards should be broadened and standardized to Linked Data format for the benefit of global accessibility. The library data which isin natural language text format and also the library metadata available now do not support the standard structure formats, hence prevents libraries from implementing new technology changes. There is also considerable disparity in concepts and terminologies used between libraries and Semantic Web communities and it is essential for both communities to foster mutual understanding to bring their respective expertise.

Even though, there are some hiccups like privacy (Singer, 2009) trust, rights management (Hellmann, 2009) and collaboration with multiple users exist in full fledge adoption of Linked Data, much effort is required by the communities to address these challenges.

4.3 Towards standardization, Interoperability and Harmonization:

Eric Miller noted in a (2004) talk that libraries have four major roles in the Semantic Web or Linked Data

1. Exposing collections- use Semantic Web technologies to make content available;
2. Webifying thesaurus/mapping /services;
3. Sharing lessons learned;
4. Persistence (Miller,2004)

Looking at these four points there are opportunities for individual institutions and librarians to push Linked data work forward.

W3C Library Linked Data incubator group (W3C, 2011) insists on Semantic Web standardization by using available standards such as Simple Knowledge Organisation System (SKOS), Web Ontology Language (OWL) and RDF.Currently digital data in libraries has been managed predominantly in the form of ‘records’ that are bounded sets of information stored in files of a precisely specified structure. The Linked Data, in contrast, structure data as graphs-constructs which, in principle may be boundless. The difference between these two approaches means that the process of translating library standards and datasets into Linked Data must be undertaken to achieve the objective. The official owners of resource data and standards should assign Uniform Resource Identifiers (URI) to make library data to be in compatible with Linked Data format.

In order to maximize linkability with other datasets, library metadata must be expressed in Linked Data terms. The library data has to be mapped or aligned to existing Linked Data vocabularies. 'Alignments' are links between semantically equivalent, similar or related entities across different value vocabularies, metadata element sets or datasets. (W3C, 2011)

But, however lack of institutional support for metadata can threaten the long term persistence of their shared meanings. In case of Functional Requirements of Bibliographic Records (FRBR), which have been expressed in a number of different ontologies, are not aligned, this limits the semantic interoperability of metadata in which their RDF vocabularies are used. The Library Linked Data community should encourage aligning already existing datasets to re-use. Hence, metadata harmonization with other standards becomes essential to achieve maximum benefits of Linked Data and the viability of Linked Data in the long term will depend on the preservation of vocabularies across generations.

5. Metadata, Interoperability, Crosswalks and Harmonization

5.1 Metadata

Metadata has been with us since so many years, it is just "cataloguing'' by other name. This has provided the information professionals an important approach for organizing, managing digital material for the resource discovery.

Today, the term "metadata" usually refers to information with one fundamentally different characteristic as compared tothese more historic notions: it is machine-*processable*, i.e. it is expressed in a way that allows computers to search, sort and present metadata without human intervention. That is, the "data" in metadata refers specifically to information that is readily accessible to computers. Metadata in this modern sense has been part of computer systems since their early days, for example in file systems where file names and file permissions constitute metadata about the file content. It was in this context the term "metadata" became widely used, in the sense of data *about data* or more explicitly (National Information Standards Organization, 2004)

5.2 Metadata Standard and Interoperability

Hence, the metadata is represented in the form of catalogue elements and the organisation of these elements in a systematic format is called metadata standard. The metadata standard is the set of metadata elements and rules for their use that have been defined for a particular purpose (Hirwade, 2011). Very commonly the terms scheme, schema and standards are used interchangeably.

A metadata scheme is the set of descriptor types available to be applied to information. There are numerous standards available to address particular information use and management requirement. Such standards are emerged from the needs of specific interest groups to standardize how they classify information. Many different metadata schemes are emerged and number, size, and complexity of content metadata standards continue to grow in a different user environments and disciplines. Hence, there should be a mechanism to interact with all these metadata standards to share content metadata among the demand to access broad range of information available. In this context, interoperability of standards gains importance to connect metadata schemas. Interoperability is the ability of systems, services, components organizations to work together and exchange information without special effort on either system.

As there is rise in number of metadata standards, supplying the metadata for each standard becomes more repetitive, time consuming and tedious. In order to minimize the amount of time required to create and maintain the metadata and maximize its usefulness to the broader user community, there is a necessity of *one* metadata standard for metadata created and maintained which can be made accessible through related content metadata standards.

5.3 Metadata Crosswalks and Harmonization

A crosswalk is a specification for mapping one metadata standard to another. Crosswalks provide the ability to make the metadata elements defined in one metadata standard available to communities using related metadata standards. But obtaining to develop crosswalk is problematic and also maintaining the crosswalk as metadata standards change becomes even more problematic due to the need to sustain a historical perspective and on-going expertise in the associated standards. (Pierre &LaPlant, 1998) Hence there should be consistency across the metadata standards and it is enabled by data harmonization and this is essential to successful development of crosswalk of metadata standards. The use of harmonization creates only one set of metadata and to map any number of related metadata standards and adequate use of harmonization simplifies the development, implementation and deployment of related metadata standards through the use of common terminology, method and processes.

Various studies are conducted to reveal that many metadata standards have been designed to serve different purposes like describing text, image, manuscripts, video, etc. They contain different elements in each metadata standard.

A study was conducted (Hirwade, 2011) with different metadata standards across disciplines were compared to find out the usability of particular standard. Generally, a single standard fails to fulfil the entire metadata requirement; hence a combination of two or more standards is made to get better results. Sometimes, if a standard is chosen and if it does not contain some necessary elements, such elements can be placed in an optional group to meet the necessity for the element. Using defined metadata schemes, shared transfer protocols and crosswalks between schemes, resources across the network can be searched more seamlessly.

Duval, Hodgins, Sutton and Weibel (2002) set forth four fundamental principles for such harmonization to provide guiding framework for the development of practical solutions for semantic and machine interoperability in any domain using any set of metadata standards or simply refers to the ability to use several different metadata standards in combination of single software system.

- **Modularity:** Metadata modularity has the ability to combine metadata elements from different schemas with syntactically and semantically interoperable way without causing ambiguities or incompatibilities.
- **Extensibility:** Metadata systems must allow to create structural extensions to a metadata standard for application specific or community specific needs in context of diversity of resources and information available.
- **Refinement:** Metadata refinement is essential to create semantic extensions and improve the precision for descriptions to improve the subject access to resources for more coherent search and browsing facilities.
- **Multilingualism:** Metadata system must have the ability to express process and display metadata in a number of different linguistic and cultural circumstances.
- Nilsson et al., (2010) suggested a fifth principle, namely
- **Machine processability:** the ability to automate processing of different aspects of the metadata specifications, so that machines can handle extensions manages modules, understand refinements and provide support for multilingualism.

This principle suggests that harmonization may be realized in an automated fashion, with no need for translations, mappings or manual intervention.

5.4 RDFization of Metadata Standards

A detailed study done by Nilsson (2010) shows that there are challenges and obstacle to achieve metadata harmonization in three broad categories.

- **Conventions:** The different metadata specifications use different methods for identifying and describing metadata elements and terms from value vocabularies.
- **Models:** The specifications differ substantially in how they define metadata records, and in how metadata is structured and processed. A mapping solution is therefore destined to be incomplete and suffer from not being general to extensions.
- **Combinations:** Combining element to form application profiles and encoding them in syntaxes are both processes that rely heavily on models are harmonized, application profiles and syntaxes will become more easily addressable harmonization issues.

However, there has been clear movement towards conventions based on Web architecture and recommendation of identification on URIs and also strong orientation towards describing element and value vocabularies in Web architecture friendly way, using RDF schema for metadata vocabularies and

SKOS (Simple Knowledge Organization Systems) for describing value vocabularies such as controlled vocabularies, taxonomies and classification schemes.

Predominantly, world over the library records are stored in MARC format. Westrum et.al (2012), in their Pode project has applied a method of automated FRBRizing, based on the information contained in MARC records. The project has also experimented with RDF representation and has concluded that a conversion of existing traditions to new standards can be challenging to attain harmonization and also found that precision of resource discovery was high.

Hence, in today's environment the recipe for harmonization is to adopt a common model based on formal semantics, which is RDF model to achieve linked data environment.

5.5 Future of Harmonization and Linked Data

The coherence of Linked Data will increasingly depend on aligning with major vocabularies such as IFLA's ISBD and FRBR review groups which are seeking to align their RDF vocabularies with each other, and also with Resource Discovery and Access(RDA) with Dublin Core Metadata Initiative (DCMI) Metadata Terms. This work will be undertaken in part by new DCMI Bibliographic Metadata Task Group (2011) with support from JISC, the Functional Requirements for Bibliographic Records (FRBR) Review Group, and the International Standard Bibliographic Documentation (ISBD) Review Group. A new Schema.org Alignment Task Group (2011) is developing alignments with new and rapidly evolving vocabularies of Schema.org, an initiative which aims to helping Web developers to embed structured data in Web pages. Also in 2011, DCMI and Friend of a Friend vocabulary (FOAF) project reached an agreement in order to reinforce long term viability for RDF vocabularies in all niches of the Semantic Web ecosystem.

6. Contribution of this Article and Conclusion

This article describes qualitatively in detail how the Linked Data will try to interconnect the whole world of information with metadata standards possible tool such as RDF and challenges involved to attain the idea of harmonization and interoperability. This also emphasises the necessity of all available metadata standards to come together and integrate with different domains. The potential benefit of metadata harmonization is highlighted in context of Linked Data or Semantic Web to reach out to broader user community of the world.

Libraries should embrace the web of information, both by making their data available for use as Linked Data and by using the web of data in library services. Ideally, library data should integrate fully with other resources on the Web, creating greater visibility for libraries and bringing library services to information seekers. In engaging with the web of Linked Data, libraries can take on a leadership role grounded in their traditional activities: management of resources for current use and long term preservation; description of resources on the basis of agreed rules; and responding to the needs of information seekers. (W3C, 2011)

The Linked Data offers a pragmatic, data oriented environment that showcases the true value of harmonized metadata using hundreds of vocabularies in combination (Nilsson, 2010). With great vision of interlinking the world of information, burdened with humongous data and silos, coupled with challenges of multilingual semantics and inherited structural issues from various quarters has made Linked Data as the greatest challenge for communities of this era and apparently, interesting times are ahead for the World in resource discovery.

References

1. Baker, Thomas. (2012) "Libraries, languages of description, and linked data: a Dublin Core perspective", Library Hi Tech, Vol. 30 Iss: 1, pp.116 - 133 available at: 10.1108/07378831211213256
2. Berners-Lee, T. (2006), "Linked Data–design issues", available at:http://www.w3.org/DesignIssues/LinkedData.html
3. Brickley, D., Miller, L. and Baker, T. (2011), "Agreement between DCMI and the FOAF Project", available at: http://dublincore.org/documents/2011/05/02/dcmi-foaf/
4. Byrne, Gillian and Goddard, Lisa (2010). The Strongest Link: Libraries and Linked Data, available at: http://www.dlib.org/dlib/november10/byrne/11byrne.html
5. Duval, E., Hodgins, W. Sutton, S. and Weibel S. (2002), "Metadata Principles and Practicalities", D-Lib Magazine, Vol.8 No. 4, April 2002, available at: http://www.dlib.org/dlib/april02/weibel/04weibel.html
6. DCMI Bibliographic Metadata Task Group (2011), available at http://wiki.dublincore.org/index.php/Bibliographic_Metadata_Task_Group
7. Hellman, Eric (2009). Can Librarians Be Put Directly Onto the Semantic Web?
8. http://go-to-hellman.blogspot.com/2009/08/can-librarians-be-put-directly-onto.html
9. Hirawade, Mangala Anil (2011),"A study of metadata standards", Library Hi Tech News, Vol. 28
10. Iss: 7, pp.18 - 25, available at 10.1108/07419051111184052
11. Isele, Robert, et.al. (2010). Silk - A Link Discovery Framework for the Web of Data, available at:
12. http://www4.wiwiss.fu-berlin.de/bizer/silk
13. LaPlant, W and St. Pierre, M (1998) Issues in cross walking content and metadata standard, available at: http://www.niso.org/publications/white_papers/crosswalk/
14. Library of Congress (2011), "A bibliographic framework for the digital age", available at:
15. http://www.loc.gov/marc/transition/news/framework-103111.html
16. Library of Congress (2011), "Authorities and Vocabularies", available at: http://id.loc.gov/

17. National Information Standards Organization (NISO) (2004), Understanding Metadata, Bethesda,MD, NISO Press, available at http://www.niso.org/publications/press/UnderstandingMetadata.pdf

18. Nilsson,M and Naeve, A. (2010) Metadata harmonization:a roadmap for standardization available at Nilsson, M. (2010), "From interoperability to harmonization in metadata standardization:Designing an evolvable framework for metadata harmonization," Kungliga TekniskaHögskolan, Stockholm, available at:http://kmr.nada.kth.se/papers/SemanticWeb/FromInteropToHarm-MikaelsThesis.pdf

19. Schema.org Alignment Task Group (2011) available at http://www.schema.org/

20. Singer, Ross. (2009). Linked Library Data Now! Journal of Electronic Resources Librarianship 21 no.2: 114-126. http://dx.doi.org/10.1080/19411260903035809

21. Westrum, Anne-Lena Et. al (2012), Improving the presentation of library data using FRBR and Linked data, available at: http://journal.code4lib.org/articles/6424

22. W3C (2004) Resource Description Framework (RDF): Concepts and Abstract SyntaxW3C Recommendation 10 February 2004 available at: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

23. W3C (2010) Semantic Web, available at http://www.w3.org/standards/semanticweb/

24. W3C (2011) Library Linked Data Incubator Group Final Report, W3C Incubator Group Report 25 October 2011, available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

About Authors

Mr. Praveenkumar Vaidya, Librarian, Tolani Maritime Institute. Pune.
E-mail: vaidyapraveen@gmail.com

Dr. N S Harinarayana
Associate Professor, University of Mysore. Mysore
E-mail: ns.harinarayana@gmail.com