Linking Library Data: A Linked Data Based Approach

Kumar Sharma *Ujjal Marjit* *Utpal Biswas*

Abstract

In order to bring library resources into the web the data must be rendered into the underlying representation format provided by Semantic Web technology. In doing so library data have become a part of Linking Open Data (LOD) project, participating into a globally interlinked data hub. In this article, we present how Linked Data approach has been used to represent and publish library data into the Web of Data. Moreover, the library community can thus be influenced by the benefits provided by Linked Data. It avails linking and sharing information with external sources, data enrichment, and reusing the resources among libraries.

Keywords: Semantic Web, Linked Data, Open Data, Metadata, Ontology

1 Introduction

There exist numerous digital libraries to share resources among library domains. However, in the present day, they suffer mainly from disclosing information on the web. By which it is difficult to share library resources with non-library domains. In the past many other domains endeavoured to achieve this using mechanisms such as File Transfer Protocol (FTP) and document based approach. Now-a-days the document based approach has been used by many organizations to visualize their data. But they could not advance using traditional approaches. Because these approaches lack very basic feature of sharing such as interlinking, data reusing and use of unified data modelling framework. Present libraries are able to process their data using various machine readable cataloguing, such as MARC 21. In addition to that, they are able to exchange their data among libraries at a minimum rate. But they are still far away from making data as a part of the traditional web. They should be able to make data more transparent, structured, and easily understandable by the machine as well as humans. Currently, this can only be achieved by using Semantic Web technologies.

Semantic Web [1] technology has been evolved to bring the concept of Web of Data. Semantic Web only deals with data in lieu of document. It suggests data to be an integral part of the web, by which, outside sources interact directly with data rather than the document that holds it. We can say that such kind of data has knowledge bundled with it, which can tell about itself, its location and much other associated information. Thus, Semantic Web, is approaching towards resolving the issues faced by traditional document based approaches. It enables library data to be discovered through the web. It also helps to achieve fast navigation among the library data and improves their visibility, reusability, heterogeneity, and interoperability. Immense numbers of data from government bodies have been published using Linked Data approach. This includes European Union Open Data[1], Japanese Government Open Data[2], UK Government Open Data

[2], and many others. Library resources are also being participated into the Web of Data from many organizations such as British National Library (BNB)3, Europeana Linked Open Data [3], Cambridge University Library dataset4, Hungarian National Library5, and Library of Congress Subject Headings6.

This article presents our previous experiences in making Linked Data service for library resources. Additionally, it is also shown how Linked Data can be used to publish library resources as Web of Data.

2 The Link Open Data Cloud

Linked Data, an extension to the Semantic Web technology, is a semantic approach to publish resources on the web by enabling sharing and interlinking [4]. For a resource to be a part of the Linked Open Data Cloud, one must follow Linked Data principles proposed by Sir Tim Berners Lee [5, 6], which are as follows:

1. Use Uniform Resource Identifier (URI) for naming resources.
2. Use HTTP URIs, so that people can browse those names.
3. When someone looks up a URI, provide useful information, using the standards RDF and SPARQL.
4. Include links to other URIs, so that they can discover more things.

Once the RDF dataset is published using Linked Data basic rules, it becomes a part of linked open data cloud. The Linked Data plays a pivotal role to advertise the data semantically in library domain. Most importantly it helps in representing data using common and unified data modelling framework. Data represented by such models can be accessed easily on the web. These data are so called self-described data, where data itself is sufficient to express its identity. It itself tells about its nature, the source and other related information. Thus, data by becoming a part of the Linked Data web is accessible by many other resources on the web. In Linked Data paradigm, HTTP URIs and ontologies play a crucial role in enabling self-described data. HTTP URI makes resource dereference able as well as provides unique identifier. Ontologies provide main building block of the Semantic Web, by allowing the definition of resource properties and their relationship.

3 Linking Library Data

Library data is stored mostly in inherent format such as MARC 21 and other relational databases. These are called legacy library data. The legacy library data needs to be converted into RDF before publishing them into the web. In a previous work [7], conversion of MARC 21 Format for Bibliographic Data into RDF retaining their links and provenance have been presented. Library data, however, can also be created from scratch, for newly setup libraries. In this case, the data entered by users are directly modelled using RDF. The overall architecture is shown in Figure 1. For newly created library data, we are directly inputting user entered data into RDF generation component and create graph for each resource. Legacy library data is handled by separate component, by which data from legacy data storage are read and processed by the data conversion component. This component converts legacy data into RDF. While generating RDF resources, some important tasks need to be taken care. These include choosing right URIs, ontology selection, RDF link generation, and publication. We will discuss each

of the steps in detail in the later sections. After generating RDF resources they are interlinked with outside sources and stored into triple-store using Jena TDB[1]. TDB is a component of Jena[2], an open source Semantic Web framework in Java. TDB is used for storing and querying RDF triples.

The provenance information has also been tracked for each library resources represented by RDF. Provenance provides metadata such as the source of origin, agent, activity, and process related information. Finally, provenance information is also modelled using RDF and stored using Jena TDB. Discerning provenance information is a separate component which exposes provenance related metadata of the dataset and linked library resources.

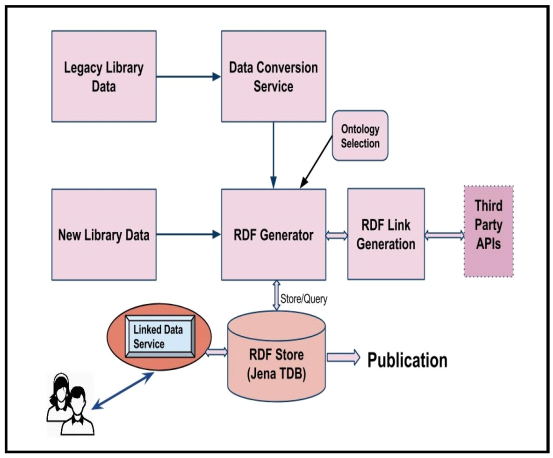

Figure 1: Architecture of Linking Library Data

3.1 Linked Data Service

Resources stored into triple-store can be browsable, using its URI. For this, a Linked Data service is required to process the incoming requests. As RDF data are understandable by machines, machine can perform reasoning over RDF data and can analyze them. Machine always performs semantic search on RDF data. Hence, the Linked Data service must be able to process the request and give response back with resource description in RDF format. However, format of RDF is not readable or clearly understandable for human users, for which, the Linked Data service has to return the resource description in human readable representation, such as HTML. Currently, Java Servlet in server side and JSP pages in client side have been used to implement Linked Data service. The resource URIs when typed in browser it gets result back in HTML representation. There exists certain Semantic web search engines, such as Swoogle[1], which process resources represented in RDF. These applications require resource description to be replied in RDF representation. For a Linked Data service to work correctly, it has to use Content Negotiation mechanism, defined in HTTP specification[2], to serve different representation of a resource. This is handled by Content:<Accept-Type> HTTP header field. Content negotiation has been implemented in such way that on providing the link of particular library resource the Linked Data service determines the request's Content:<Accept-Type> header field. If the Content:<Accept-Type> is text/html or text/plain then it returns resource description in HTML representation. If the Content:<Accept-Type> is rdf+xml then the resource description is returned in RDF representation. Figure 2 illustrates the architecture of Linked Data Service.

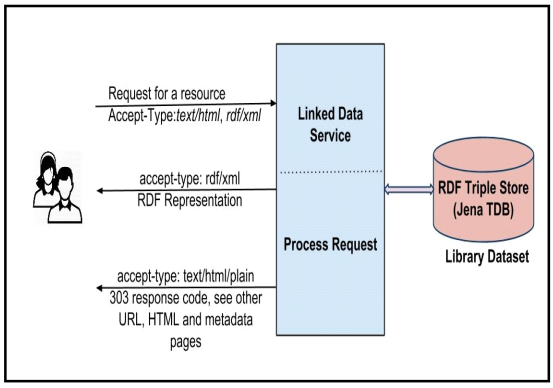

Figure 2: Architecture of Linked Data Service

3.2 Data Modelling and Ontology Selection

Data modelling is the foremost task in Linked Data generation. In Semantic Web RDF is the sole data modelling framework for modelling resources. RDF models a resource in triple representation such that, a triple denoted by T is defined as:

T = {S, P, O}

where,

S = Subject,

P = Predicate or property, and

O = Object

Subject denotes a resource which is being described (for e.g., a Book and its author), Predicate represents resource's property or relationship name (for e.g., bookName, hasAuthor), and the Object represents property value or another resource linked using RDF link attributes. Each of these triples, except for literal values, are identified using HTTP URIs that uniquely identifies a resource on the web. Such data representation forms a directed or labelled graph where subjects and objects (which are nodes in graph) are joined by predicate (the arc). Hence, resources in RDF are always stored in triple form. Each resources has their property and value which resembles the description of metadata. Ontology is the main building block of Semantic Web which provides concepts in describing a knowledge base. Without using ontology it is not possible to model resources in RDF. It provides terms and properties to describe resources. Hence, selection of a correct ontology in a particular domain is important task. In library domain there exist a number of ontologies, taxonomies, and controlled vocabularies. In this work Marc Ontology [8], RDA Group 1 Element Vocabulary[1], DCTerms[2], FOAF[3], BIBO[4] and VoID[5] ontologies have been used.

3.3 Choosing Right URIs

The first and second Linked Data principles suggest that we should use HTTP URIs for naming resources. URIs for resources needs to be chosen carefully. In a domain each and every URIs should be unique so that they do not lead to duplicity. In this work, the format of URI is <domain-name>/<service-name>/BibResources/<resource-id> where the resource-id is the unique resource identifier maintained in the local database. For resource-id, we can provide some unique random characters as well as some name for the resource. Choosing unique random characters is very straightforward which is easy to generate but the URI does not look prominent. Furthermore, choosing some appropriate name of the resource can describe its nature, but it is not easy to maintain the uniqueness. Since similar name may exist by which it leads to duplicity. What type of resource-id is required is depends on the data provider. However, it is better to follow best practices that have been implemented. Christian Bizer et al. [9] have presented the concepts, and best practices regarding Linked Data implementation. They have discussed how to publish Linked Data, the design architecture, approaches on choosing right URIs, and setting RDF links to other data sources.

3.4 RDF Link Generation

The fourth linked data principle suggests to add RDF links to the resources. There are three types of RDF links: relationship link, identity link, and vocabulary link. Relationship link points to the resource defined in outside linked data sources. For e.g., rdfs:seeAlso and foaf:knows are relationship links which relate to the resources from outside

sources. Identity links point to similar resources to get more detail description. For e.g., the owl:sameAs is an identity link. The vocabulary links point to the description of a term defined by a vocabulary, in our work we have used the relationship links fetching from outside sources such as DBpedia[6] and VIAF[7]. These sources provide some APIs by which the related resource URIs are fetched. Currently, this component is at the very basic stage, and it needs to be improved for enriching external links. This can be achieved either by enhancing the current approach or by using well known link discovery frameworks, such as Silk [10].

3.5 Linked Data Publication

One of the critical tasks in linked data generation is publishing final version of linked dataset into the Web of Data. As soon as the dataset is published it will be a part of the globally interlinked data hub. Currently, this component is in active development, which needs live library data from a well known institution. For a dataset to be a part of the data hub, some constraints needs to be fulfilled which are: the data must be usable in some ways to the end users, each and every resources must be linked to two or more external linked data sources, the linked data sources should be well known and must be member of the same data hub, and they also have to provide valuable information. As a whole we can summarize Linked Data publishing steps as: preparation of data, data modelling using RDF (includes appropriate ontology selection and choosing right URIs), setting up the infrastructure such as Linked Data services, interlinking data with external sources, and the publication. There exists several linked data publication tools which help in deploying linked dataset into the web of data. Few of them are Fedora[8], Virtuoso Universal Server[9], Talis platform[10], and Drupal 7[11].

4. Experiments

The approach has been experimented with several MARC 21 legacy data. For experimental purpose openly available MARC 21 data from different sources, such as Harvard Library Bibliographic Dataset[12], have been used. The legacy library data have been converted into RDF, and the data modelling, interlinking resources with external linked data sources such as DBpedia, and VIAF have been achieved successfully. The work is still in progress and requires more refinement mostly in the area of link generation, assigning right attributes to the resources by selecting appropriate vocabularies, and optimization of the conversion process. To model resources using RDF we have been using Java version 1.6 along with Apache Jena version 2.7.4.

5. Future Work

The work is solely based in Linked Data generation from the legacy library resources. Till now only the MACR 21 Format for Bibliographic Data have been used. There are other format of MARC 21 so further research will be carried out on the formats such as MARC 21 Format for Holdings Data, Authority Data, Classification Data, and Community Information. We aim to convert all these formats and make the framework sole implementation for the legacy resources while tracking provenance information, reducing the size for provenance storage and keeping versioning of the resources.

6. Conclusion

In this article, we have presented a linked data based approach for linking library data and make them

available on the web. Linked data provides real benefits of sharing, interlinking, and reusing of the library resources. Data sharing enables cooperation of library resources with non-library domains on the web. Reusing feature tends to resolve the issues of heterogeneity and interoperability. We believe this article assists many librarians to follow the linked data approach in designing Linked Data service and helps in understanding the technical terms and their implementation.

References

1. BERNERS-LEE, T., HENDLER, J. & LASSILA, O. The semantic web. Scientific American, 2001, 284 (5), pp 28-37.
2. SHERIDAN, J., AND TENNISON, J. Linking UK Government Data. In LDOW. 2010.
3. HASLHOFER, B., AND ISAAC, A. data. europeana. eu: The europeana linked open data pilot. In International Conference on Dublin Core and Metadata Applications, 2011, pp. 94-104.
4. BIZER, C., HEATH T., & BERNERS-LEE, T. Linked data-the story so far. International journal on semantic web and information systems 2009, 5 (3), pp 1-22.
5. HEATH, T., and BIZER, C. Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 2011, 1, (1), pp 1-136.
6. BERNERS-LEE, T. Linked data-design issues (2006). Available at http://www.w3.org/DesignIssues/LinkedData.html (Accessed on 15/07/2014).
7. KUMAR, S., UJJAL, M., & UTPAL, B. Exposing MARC 21 Format for Bibliographic Data As Linked Data With Provenance. Journal of Library Metadata 2013, 13, 2-3, pp 212-229.
8. KRUK, SR., SYNAK, M., & ZIMMERMANN, K. MarcOnt—Integration ontology for bibliographic description formats. In International Conference on Dublin Core and Metadata Applications, 2005, pp-231.
9. BIZER, C., CYGANIAK, R. & HEATH, T. How to publish linked data on the web. 2007.
10. VOLZ, J., BIZER, C., GAEDKE, M., & KOBILAROV, G. Silk-A Link Discovery Framework for the Web of Data. LDOW, 2009, 538.

About Authors

Mr. Kumar Sharma, Research Scholar, Department of Computer Science & Engineering, University of Kalyani, India.

Mr. Ujjal Marjit, System-in-Charge, C.I.R.M.(Centre for Information Resource Management), University of Kalyani.

Dr. Utpal Biswas, Associate Professor, Department of Computer Science and Engineering, University of Kalyani, West Bengal.