AN OVERVIEW OF LINKED DATA AND ITS APPLICATION IN LIBRARIES

-Sanat Kumar Behera and Bulu Maharana

INTRODUCTION

Library as the knowledge centre organizes different types of documents physically and electronically. People's access to information has changed; they mainly access it through digital gadgets. As a service organization library always works on how to provide all possible relevant information. It has to focuses on the semantic aspect of its product. It uses different Standards such as MARC, Dublin core etc. and tries to build a contextual relationship among its products.

The web has a lot of data, including pictures, documents, links etc., incorporated into its pages. When looking at these pictures or data, we understand what it is all about or whether it is a cow or buffalo. Computer does understand the images, links, documents, but they don't know its relationship with the pages. Linked data make the computer understand the meaning of the data and its relationship with the pages. HTML& HTTP only provide a link to the pages. If we evaluate the search engines only depends on the word frequency. They don't know the structure or contextual relations available in the pages. The introduction of the semantic web concept by Tim Berners Lee has paved the way for different technology which facilitates a meaningful web. The components of Linked data URI, RDF and other languages like OWL SPARQL is making the web understandable to both human and machine.

WHAT IS LINKED DATA?

Linked data is a set of principles and standards which provide a contextual meaning and create a structural relationship between data available on the web. It is not referring to simple hypertext but focuses on the contextual concept of the data and documents identified by URIS and processed with RDF (Resource Description Framework) and other structure data tools. In the words of Tim Berners Lee, it's not about putting data on the web instead proving links and relationships to each data available data on the web. He has mentioned four rules in executing the linked data into reality. The four rule focuses on using URI to name things, using HTTP URI for standard exchange of information, & using RDF, XML, SPARQL etc., to describe information and connect with other data linked to them. (Berners-Lee, 2006).

TECHNOLOGICAL LANDSCAPE OF LINKED DATA

Uniform Resource Identifier (URI)

Uniform Resource Identifier (URI) is the unique identifier which identifies any objects or things over the web. It identifies a resource with name, location or both. HTTP URI which is important element in the linked data system is the URI using HTTP (Hypertext Transfer Protocol) of W3C standards. For standard exchange of data, the HTTP URI is used.

RDF (Resource Description Framework)

RDF (Resource Description Framework) is a data model published as a W3C recommendation in 1999. It's a language to describe the structure of the web resource. A resource may be anything data, documents, abstract etc. It provides a common framework in expressing and interlinking web resources over the web. The general structure or the RDF statement always follows Triple, subject, predicate, and Object. The predicate is called a property in the case of machine or programming language. Its model which is used to identify the resources, things any other entity

RDF Data model

RDF Data model is the process using which The RDF statements are presented over the web. There are various data model available which can be use in expressing the RDF statements.

RDF Triple

Let's take an example of a statement: Wings of Fire Published in 1999. In Fig.1, the sentence classifies into three primary facets subject, predicate and object. We as humans understand that it is a sentence, and these are its facet. But machines don't recognize its syntax to be machine-understandable. We have to describe each entity and its relationship with the other

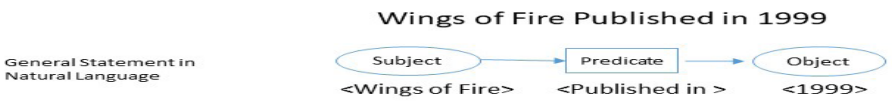

Figure 1: Natural Statement

The structure in which RDF presented is called as RDF statement. RDF statement always express in RDF triple that is subject, property and object. Subject always have a URI and Predicate is described with the RDF vocabulary which also have an URI. Object or value may be an URI or a Literal. Literal is a character string, a data value which sometimes express with XML Schema data types recommended by W3C.

.**Figure 2: Examples of an RDF Triple**

RDF Graph Model is the most common form model which RDF statements are presented. The RDF graph is a set of objects called "nodes" connected by lines or arrows, which comprises the relationship between the edges or nodes. Below is the Graph model of the fig-2

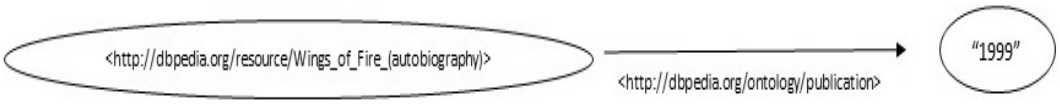

Figure 3: Graph Model

RDF Serializations

RDF Serialization is process or format in which the RDF statements are described. The most common formats are N-Triple, RDF Turtle (Terse RDF Triple Language), RDF/XML, RDFa (RDF embedded in HTML) etc.

RDF Graph

RDF Graph data model is the most common form model that presents RDF statements. The RDF graph is a set of objects called "nodes" connected by lines or arrows, which comprises the relationship between the edges or nodes. In the below figure 2, the book "Wings of Fire" resource is connected with other nodes with ray lines and presents a relation between the nodes with ontology terms author and publications. Each of the nodes in RDF is closed in a circle and connected through a relationship.

<http://dbpedia.org/resource/Wings_of_Fire_(autobiography)> <http://dbpedia.org/ontology/publication> "1999"
<http://dbpedia.org/resource/Wings_of_Fire_(autobiography)> <https://dbpedia.org/ontology/author>
<http://dbpedia.org/resource/A._P._J._Abdul_Kalam>

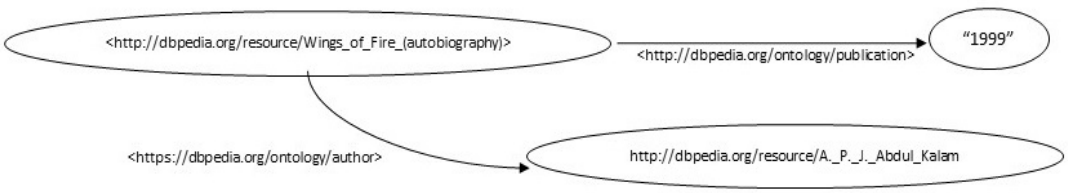

Figure 4: RDF Graph

RDF N-Triple

The most common and basic format used in the RDF sterilization is RDF N-Triple. It follows the principles of RDF triple. The subject and property identified with URI are located in a web base, and the object may be URI or literal closed in an inverted comma.

<http://dbpedia.org/resource/Wings_of_Fire_(autobiography)><http://dbpedia.org/ontology/publication> "1999"

<http://dbpedia.org/resource/Wings_of_Fire_(autobiography)><https://dbpedia.org/ontology/author>

<http://dbpedia.org/resource/A._P._J._Abdul_Kalam>

In the above statement, The Book «Wings of Fire» is identified with resource URI and has a relation «published in» in 1999. The source of the URI is DBpedia, located in a web base. «Published in» is identified with ontology URI «publication,» and 1999 is a literal closed in an inverted comma as per the data types rule of W3C. The URI of the subject and property always expresses the angle bracket. Here, the object is a literal(year), so described within an inverted comma. In the second line, the Book has an author, & the author is Dr. A.P.J Abdul Kalam, and both have a relationship identified using ontology author.

RDF turtle

The turtle notation is the extension to the N-triple. It presents the information concisely using different abbreviations and provides various ways to write URI and literal. @prexies directives shorts the long URI and closed within the angle bracket in @base directives for relative URIs..

@prefix dbo: <http://dbpedia.org/ontology/>

@base <http://dbpedia.org/resource/>

<Wings_of_Fire_(autobiography)> dbo: publication "1999"

<Wings_of_Fire_(autobiography)> dbo: author <A._P._J._Abdul_Kalam>

In the above statements, the @prefix is used to identify and abbreviate the vocabulary ontology URI instead of long URI and @base is for the resource URI. In the beginning the directives are declared that is http://dbpedia.org/ontology/ to dbo: and @base to only subject within angle brackets.

Besides the above sterilization process, RDF Graphs also presented RDFa, RDF/XML, JSON-LD etc. RDFa uses the HTML, XHTML and HTML5 attributes and some additional attributes to present the data. RDFa data model is used to build semantically enabled websites. XML is the syntax to describe the properties attributes of the documents. RDF/XML is a process in which XML attributes and namespace are used to describe the RDF graph in the form of statements. The nodes and predicates are uses while presenting data in RDF/XML. JSON-LD (JavaScript Object Notation(JSON) for Linked Data) required javascript in its process. It is easy to write and read; however, it is complicated and costly to parse.

RDF Schema(RDFS), called RDF vocabulary Description Language, is a semantic extension to the RDF. RDF describes the syntactic structure, and RDFs describe the semantic relationship of the term. It describes classes, properties etc. which further assist in expressing the relation between the properties and resources. In RDFS, the "classes" are defined as rdfs: class and each sub-classes or member is called as an instance. For example, rdf: property is an instance of rdfs:class and rdfs:range is an instance of rdf:property that describes a property's value

OWL (Web ontology Language) an extension to the existing standards and languages like XML, RDFS etc. It provides additional vocabulary to the existing RDFS vocabulary. It designs to express semantic relationship between things which machine can interpret and understand. OWL has three different sub-languages OWL lite, OWL DL and OWL FULL. OWL define classes, properties and instances. owl: Things is the root of all classes and owl: Nothing is the empty classes.

- OWL Lite is meant for the user requiring a classification hierarch. It provides a basic sub class hierarchical classification. It is less expressive and computational.
- OWL DL is mean for the user who need more expressiveness with loosing the computational value. It has all the features of OWL Lite.
- OWL Full is very expressive and compatible with RDF and RDFS. The computational logic is very less.

SPARQL

SPARQL is a query language of RDF like SQL for different Databases. It can be used as a query language in any database through middleware to preview or store RDF data. The query in SPARQL follows a triple pattern called graph pattern, which comprises Subject, Predicate, and Object. It has four types of queries ASK, SELECT, CONSTRUCT and DESCRIBE. The SELECT query returns variable bindings and CONSTRUCT query form an RDF graph. ASK finds out whether any match is available or not, and DESCRIBE returns the RDF graph.

Use of linked Data Globally

There are various project going on worldwide in providing link data service or doing research on its applications. Below are some of the important initiates.

Schema.Org is a collaborative effort of Bing, Yahoo, Google, and Yandex to create, maintain, promote the structure data on the web. It is a first shared vocabulary that webmaster can use. It follows two hierarchies, one is for textual value, and another is for things. The Data model used in Schema.org is RDF Schema. RDF Schema is derived from the CyCL ontology Language.

Google Knowledge Graph is an initiative of Google to improve search accuracy and relevance. It enables the users to "search for things, people or places that Google knows about—landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art" and other relevant results. It has facts that solve instant queries of the users. The Fact data compiled from public sources, license content and owners contributed data.

Open Graph protocol is the creation of Facebook to track people's relationships and content available on Facebook. Google, WordPress uses this protocol. It has used the RDFa and Dublin core metadata elements. https://ogp.me/

Wikidata is an open, multilingual and collaborated project by the Wikimedia Foundation. It is central storage for the structured data available on Wikipedia, Wikimedia, Wikisource, Wiktionary, etc. Each of the resources and entities of the Wikidata is identified by URIs. The URI always follows a pattern of " http://www.wikidata.org/entity/QX" wherein "QX" X always changes and is an integer that identifies the source.

DBpediais a community project launched in 2007 and available in the free license. It extracts the structural content and infobox of Wikipedia using different software and identifies them with HTTP URI. It served as linked data on the web. SPARQL Endpoint, DBOntology, DBpediaArchivo are some of the valuable services of DBPedia, which facilitates linked data service over the globe.

WHY LINKED DATA FOR LIBRARIES?

Library Users are getting smarter, and their information needs and consumption mode also changed. The users need relevant information and want similar works in their searched topic. The library has huge collections of hard copies and digital resources that need visibility. Further, to satisfy the information need library have to connect fellow libraries for more documents/information. Library data need to be structured and compatible with the web standards for better retrieval. The accessibility of the resources and how much relevant query our system is solving is also pushing us to follow the semantic technology. Using the semantic, we can build a relationship within our existing resources and connect resources of other libraries. Linked data will facilitate the library in linking resources, mining the available records, and providing a contextual relationship in Digital Resources. LD is not new for librarians in cataloguing and classification; we are already classifying describing the different entities & properties of the documents through metadata and other tools. See Also, Tracing, Subjects, added entry some of these are there for the beginning, which assists the library in providing relevant information to the user.

Standard Practices and Initiatives of linked data in library

Many libraries and groups worldwide are taking initiatives and applying linked data in library collections & services and contributing to the linked data over the web. They are developing different standards though which other libraries can participate and contribute in future.

IFLA has developed two conceptual models: FRBR (Functional Requirements of Bibliographic Records) and FRAD (Functional Requirement of Authority Data). One is for bibliographic data, and the other is for Authority Data.FRBR uses the bibliographic record's conceptual entity-relationship model, which satisfies the user need. The main Fundamental feature of the FRBR is Find, Identify, Select, Acquire. FRBR has three groups of entities. The entities are the foundation of the FRBR principles.FRAD has used the principles of FRBR; however, additionally, it has identified the creation of other entities related to the authority data. It has added additional attributes like a controlled access point, rules, agency, and identifier to support the authority data.FRAD model identifies the four tasks of users of authority data Find, Identify, Contextualize and Justify.

Library of Congress has converted LCSH and authority files to RDF statements. It has developed RDA based on the principles of FRBR and FRAD. LC developed a bibliographic description data model named BIBFRAME (Bibliographic Framework) using the linked data principles. It was designed to replace the MARC standards. BIBFRAME 2.0 was released in 2016 and focused on three core categories of abstraction work, instance and item. BIBFRAME 2.0 includes two vocabularies, BIBFRAME and MADSRDF (Metadata Authority Description Schema), and other ontologies like OWL, RDFS, SKOS etc.

Online Computer Library Centre (OCLC) the owner of the worldcat which is the biggest catalogue of the world using its huge record in linked data service. It has integrated schema.org, FRBR principles with the bibliographic record of the worldcat and developed different new linked data schema.

FAST (Faceted Application of Subject Terminology) In collaboration with the LC, OCLC converts the LCSH to RDF statements, and the purpose is to provide subject heading schema in simple and

navigation friendly. It is available under the open license of Creative Commons. The schema is accessible open, and anyone can export MARC XML, RDF, ISO MARC format.

The VIAF (Virtual International Authority File) consists of the authority file of more than 40 organizations over the globe maintained by OCLC. The contributed authority file matches with VIAF. The given description of each entity merges and converts into a single record.

CONTENTdm is a digital collection management system designed to build, preserve and customize the digital collection. CONTENTdm linked data project evaluates the descriptive data practices given on the digital resource metadata and converts them to the entity-based data model. It "Focused on the specific needs around migrating descriptive practices to an entity-based model, with the end goal of increasing discoverability and sensemaking." (Proffitt, 2021).

Share-VDE (https://www.share-vde.org/sharevde/clusters?l=en) is the project of a group of libraries in which they bring all their bibliographic records and authority files together to create a linked data environment. It is an advanced virtual discovery platform that connects 22 libraries. Using BIBFRAME and other ontologies, library catalogues are transforming into RDF. Any user can access the MARC21, BIBFRAME and RDF versions of the record. The record is available for export in PDF, Excel, XML, HTML, RDF etc.(Share-VDE Project, 2022)

COMET (Cambridge OPenMETadata) is a joint initiatives of Cambridge University and OCLC which converted the catalogue bibliographic records to linked data. The data is accessible in the form of RDF/XML and other notation.

USING LINKED DATA IN LIBRARY

Library manage its resources using different standards like MARC, Dublin core etc. and uses different software & tools to increase the visibility of its in-house and purchased collections. Converting Library data to Linked *Data* is not easy task as MARC is not fully compatible with web standards. There are some MARC converters like marc2rdf-modeler, marcmods2rdf etc. tools to convert MARC to RDF. Each of the MARC data has to be described as entities following the principles of FRBR and FRAD or based on DRA. Each of the field or tags has to be mapped with the data models.

The library can integrate linked data to the library websites based on Drupal 7 and WordPress. Word press has some plugins like wp-linked-data, wp-LDP and WordLift, which assist in publishing linked data. Wolrd-lift converts each content category and publishes them into open linked data cloud. Drupal7 with RDF extension module, RDF mapping API enriches the content's metadata. Drupal 8 has an RDF UI extension, which integrates site builder with schema.org in mapping the content and fields.

Eprints and DSpace repository software has also added semantic feature into their system. Both the software uses persistence URI and HTTP URI. EPrints has included a feature to export the repository content to RDF XML, RDF N-Triples, RDFa etc. DSpace is using org.dspace.rdf.conversion package which include org.dspace.rdf.conversion.ConverterPlugin to convert the repository content to RDF. The SPARQL version of the repository can be created and published using Apache Fuseki.

CONCLUSION

The web needs to be structured and follow uniformity so that machines can understand things available on the web. Global initiatives on linked data like schema.org, DBpedia, and WikiData initiatives are making the way of linked data applications. The ongoing projects and adaptation of linked data may be slow among libraries. However, it has witnessed a remarkable development in providing linked data services. FRBR, FRAD, RDA, BIBFRAME, DCMI-LD etc., standards in the library fields giving birth to many initiatives worldwide in providing linked data service. Linked data features of EPrints, DSpace, Drupal WordPress etc., also boosting the linked data service in libraries.

REFERENCES

- Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012). Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World*, 113(11), 549–570.
- Breeding, Marshall. (2020). Smarter Libraries through Technology: Linked Data Brings Challenges and Opportunities to Libraries. *Smart Libraries Newsletter* 40 (4), 1-2.
- Byrne, G., & Goddard, L. (2010). The strongest link: Libraries and linked data. *D-Lib Magazine*, 16(11–12), 1–8.
- Godby, J. et al.(2019). Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage. Retrieved from: https://www.oclc.org/content/dam/research/publications/2019/oclcresearch-creating-library-linked-data-with-wikibase-project-passage-a4.pdf
- Hallo, M., Luján-Mora, S., & Trujillo, J. (2014). Transforming library catalogs into linked data. ICERI2014 Proceedings, 1845–1853.
- Mitchell 2, E. T. (2016). Library Linked Data: Early Activity and Development. In Library Technology Reports 52(1).
- Mixter, J., & Washburn, B. (2021). On the way to Library Linked Data : The view from OCLC Research. September.
- Neish, P. (2015). Linked data: What is it and why should you care? *Australian Library Journal*, 64(1), 3–10.
- OCLC Research Archives and Special Collections Linked Data Review Group. (2020). Archives and Special Collections Linked Data : Navigating between Notes and Nodes. Oclc Research Position Paper, 14. Retrieved from: https://doi.org/10.25333/4gtz-zd88
- Pace, A. K., & Chapman, J. (2020). OCLC Linked Data: Research, experimental applications, and shared infrastructure. July.
- Proffitt, M. (2021). Wikimedia and Libraries : From Vision to Practice. Retrieved from: https://doi.org/10.25333/hy6t-sg17
- Rasmussen, K. B. (2018). Digital curation after digital extraction for data sharing. *IASSIST Quarterly*, 42(3), 1–2.
- Wang, Y., & Yang, S. (2018). Linked Data Technologies and What Libraries Have Accomplished So Far. *International Journal of Librarianship*, 3(1), 3.

NOTES

1. http://viaf.org/
2. https://blog.google/products/search/introducing-knowledge-graph-things-not/
3. https://dbpedia.org/
4. https://ogp.me/
5. https://schema.org/docs/about.html
6. https://www.oclc.org/en/contentdm.html
7. https://www.oclc.org/en/viaf.html
8. https://www.oclc.org/research/publications/2021/oclcresearch-transforming-metadata-into-linked-data.html
9. https://www.share-vde.org/sharevde/clusters?l=en
10. https://www.w3.org/2001/sw/RDFCore/Schema/200212bwm/#ch_introduction
11. https://www.wikidata.org/wiki/Wikidata:Main_Page
12. http://cul-comet.blogspot.com/p/about.html
13. https://www.w3.org/TR/2012/NOTE-rdf-interfaces-20120705/
14. https://www.youtube.com/watch?v=FaeOtrRsZt4
15. https://wiki.lyrasis.org/display/DSDOC5x/Linked+%28Open%29+Data