Submitted on: 16.07.2018

Proposal for Implementing Linked Open Data on Libraries Catalogue

Esraa Elsayed Abdelaziz
Computer Science, Arab Academy for Science and Technology, Alexandria, Egypt.
E-mail address: esraa_el_sayed@hotmail.com

Saleh Mesbah Kaffas
Computer Science, Arab Academy for Science and Technology, Alexandria, Egypt.
E-mail address: saleh.mesbah@gmail.com

Abstract:

The semantic web vision was created in 1999 by Tim Berners Lee. In 2001, the W3C declared the semantic web as the web of shared data. Hence, the realization of this vision requires that current data adopts the new technology in order to be part of the global shift of the current web technology. Library data often encodes some of the most important, unique and authoritative information in the world. Hence very essential that library data speaks the same language as the language spoken by the current web. Bibliographic data stored in traditional record formats has reached its limits of efficiency and utility. Semantic web technology enables users to move from viewing, and managing bibliographic data as records towards managing data as entities (works, people, places, etc.). Moreover, the library data should be treated as part of the web, in other words woven into the web and integrated into the sites and services that library users visit daily, and be part of the linked data vision of the Semantic Web.
This research aims to develop a new method for implementing the conversion of bibliographic data to semantic web. It will provide semantic methodologies to link cataloguing records to external Linked Data resources, as Wikipedia and DBpedia. The purpose is to extend the knowledge provided in the cataloguing information and map it to the largest online encyclopaedia the DBpedia. The proposed attempt succeeded in connecting the British Library with the DBpedia through the SPARQL "SPARQL Protocol and RDF Query Language" queries and succeeded in facilitating the searching, retrieving and connecting data all over the web.

Keywords: Semantic web, Dbpedia, Resource Description Framework (RDF), SPARQL.

1. Background

Semantic web is the era of web technology known as web 3.0. Web 1.0 is when the information age was born. It was the web of static documents; in fact, it was called "The read only web." In other words, the early web allowed for users to search for information and read it, and there was very little in the way of user interaction or content contribution. Web 2.0 focused on linking pages, sharing data among websites and applications, the use of social networking, and community-based knowledge management. Web 2.0 is characterized as the web of user interaction. Web 3.0 is the "Semantic Web." [1] The semantic web "provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries." Hence the semantic web is the web of shared data. There is a lot of information on the web, but it is not part of the shared data. The semantic web would represent all kinds of content that can be easily understood by machines, independent of the underlining system, or programing technology. The ultimate goal is to satisfy the sophisticated needs of internet users. This is done via communicating with common semantic meanings using semantic web languages. With the release of the semantic web, the linked data appears. The linked data was defined as "The semantic web is a web of data, however to make the web of data a reality, it is important to have the huge amount of data on the web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the semantic web need access to data, but relationships among data should be made available, too, to create a web of Data. This collection of interrelated datasets on the Web can also be referred to as Linked Data" [2] (W3C definition of Linked Data). Linked data is explained in details in next sections in this proposal. Along with the shift of library material from print material to electronic data, and the revolution of cyber knowledge; came the demand of the library users who expect to find and access materials online. It's not enough to have pages "on the web"; library data must be part of the web, – in other words woven into the web- and integrated into the sites and services that library users visit daily – Google, Wikipedia, social networks. Digital libraries hidden behind websites such as OPAC, are completely isolated from search engines, and web services especially in the semantic era of the web. Information seekers need to be connected to their local library resources from wherever they are on the web. Library data often encodes some of the most important, unique and authoritative information in the world. When this information can be referenced as entities, it enables data relationships to be rendered in many more contexts, and to be linked and related to many other linked data resources, which in turn will increase the relevance of libraries within the wider information ecosystem. Semantic web relies on basic standards and definitions that will be presented in the following sections; Metadata, RDF (Resource Description Framework, Ontology, and SPARQL)

1.1 Metadata

Meta data is one of the main concepts and techniques used in databases needed in applying the sematic web. A digital library is a library in which collections are stored in digital formats, and accessed by computers. Metadata is structured information that describes, explains or locates information resources. The main purpose of using metadata is to make it easier to retrieve, use, or manage information. Metadata is often called data about data, or information about information. In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital. Traditional library cataloging is a form of metadata. Now cataloging is moving towards converting the cataloged metadata to linked data. Thus the information provided by the bibliographic records, will be extended to link to other information resources in an interoperable fashion. [3]

1.2 RDF (Resource Description Framework)

RDF is the core syntax of Semantic Web Languages; it addresses the issue of managing distributed data. The fundamental data structure to represent statements in RDF syntax is named triple. The triple representation is a statement broken into three parts: Subject, Predicate, and object. For example, if we say “The sky has blue color” this means the subject is “The sky” and the predicate is “has” and the object is “blue color” [4]. The subject, the predicate, and the object in the triples are all resources. In the Semantic Web, there has to be a unique identifier for each resource. Subject, predicate and object are identified by the URI (Unified Resource Identifier) The same word might have different senses (ambiguity) or different resources can reference the same concept. In addition, in some cases, a node from one graph is merged with node from another graph, in other words they reference the same resource. All data in semantic web uses RDF as the primary representation language. RDF is based on XML specifications. RDF is used in the implementation of the proposal.

1.3 Ontology

Ontologies are the main model for knowledge representation in the Semantic Web. Ontologies describe the entities that consist of a set of types, properties, and relationships. Ontologies are expressed in the Semantic Web language defined by the W3C committee, which consists of the Resource description framework (RDF), Resource Description Framework schema (RDFS), and Web Ontology Language (OWL). They are all used to define ontologies with different level of semantics. Researchers in [5] have found that ontologies are the ideal knowledge model to formally describe web resources and their vocabulary, and hence, to make the underlying meaning of terms included in web pages explicit. The focus here is not on the specification and the syntax of each language, but on the differences of the semantic capabilities of each one relative to others, and its role in creating a semantic model for a given domain. In the Semantic Web literature, the languages and the reasoning technologies have always been arranged as a stack called the layered cake.

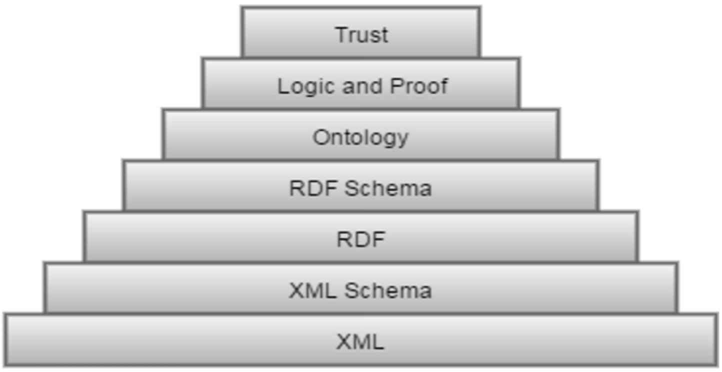

Figure 1 Semantic languages layered cake

Figure 1 illustrates the hierarchy of the representation languages, where each layer exploits and uses the semantic capabilities of the one below, and each language is compatible with all the languages on top of it in the stack. Languages in the semantic layered cake are organized so that each language’s expressive capabilities build on those of the ones below, i.e. as we go up in the layered cake, the level of expressivity increases. [6] The first layer, Uniform Resource Identifier (URI) and Unicode, follows the important features of the existing of semantic web. Unicode is a standard of encoding international character sets and it allows that all human languages can be used on the web using one standard form. URI is a string of a standardized form that allows to uniquely identify resources. A subset of URI is Uniform Resources Locator (URL), which contains access mechanism and network location. Extensible Markup Language (XML) layer, together with XML namespace, and XML scheme definitions, represent the common syntax used in the semantic Web. XML is a general purpose markup languages for documents containing structured information. RDF is the core syntax of Semantic Web languages. It addresses the issue of managing distributed data. All other Semantic Web standards build on the RDF specifications. RDF reuses many of the fundamental features of the web while extending them to suite the definition of distributed network of resources in ontologies. Its syntax is based on the Unicode XML syntax for exchanging data. All the data in the Semantic Web uses RDF as the primary representation language. The normative syntax for serializing RDF is XML, in the RDF/XML form. To allow standardized description of taxonomies and other ontological constructs, an RDF Schema (RDFS) was created by extending the formal semantics within RDF. RDFS is a language with the expressivity to describe the basic elements of the ontologies. RDFS can be used to describe taxonomies of classes and properties and use them to create lightweight ontologies, which are ontologies which include concepts connected by general associations, rather than strict formal connections.

1.4 SPARQL

SPARQL is a semantic query language for triple stores able to retrieve and manipulate data stored in RDF format. SPARQL stands for SPARQL Protocol and RDF Query Language and it is used to fetch, store, and update queries over triple stores.

2. Linked data and Bibliographic data

The goal of the linked data is to transform the web into a global knowledge base. The term linked data is mostly referred to as the set of best practices for publishing and connecting structured data on the web. In the past four years, data providers have been adopting these practices. Linked Data is the language machines can read and understand in order to seamlessly fetch results. And not only results in the form of a list, but anything that might be important and relevant to the information quest (resources, concepts, topics, ideas). Linked Data technologies are unique for they allow software agents to find, share and integrate information across diverse resources easily and effectively. Libraries have started rethinking their catalogues and reshaping them along the lines that have been set by popular search engines and online retailers [9]. Efforts like Bibframe are a step to implement linked data based cataloguing. On the other hand efforts for creating linked data involve collaborations that focus on creating links between concepts and their definition in global data sets, to form the open linked data cloud like Wikipedia, DBpedia, WordNet…etc. The cataloging based on linked data technology will be able to extend the bibliographic cataloging records to include properties and information from corresponding nodes in the linked data cloud. It will allow library data to interact with larger web information, consistently, and in an interoperable fashion. This means that web pages and information resources such as Wikipedia can have direct links to library data and users will be able to move easily between the web and library resources. Linked data can collect bibliographic data about one topic from multiple resources as library of congress data, British library, and many national libraries and connect with other web organizations. [7] The concept of linked data for the library bibliographic data implemented on the basic of concept URI's. URI is used to describe the resource on the internet for example its type, its name and its description on the web.

3. Proposed method to link bibliographic data to open linked data.

The purpose of this work is to publish bibliographic records as linked data. Which in turn will allow to link entities in the library catalogues, to their related entities on the open linked data cloud. We will use bibliographic records of the British Library and link it to the DBpedia entities. This will allow extending the bibliographic information to include more data that we can rarely find in a library database. Searching in a library database can provide information about a resource (book, article, author, list of publications, etc). Yet we are able to query the origin of the author of a book or get detailed information about his life. The goal of linking bibliographic data to linked data is to be able to extend the library records to provide the user with more detailed information to related entities. As a start we will provide some background information on the technical tools and concepts and dataset will be used. The British library was one of the major libraries that published their full database in RDF format and public for legal use.

The DBpedia is a community project that extracts structured, multilingual knowledge from Wikipedia and makes it freely available using the semantic and linked data standards. This project started in 2006. The goal of the DBpedia was to parse the information in different Wikipedia infobox templates, into RDF. However, DBpedia ontology schema was not developed until late 2010. The ontology organizes the knowledge on Wikipedia in 320 classes, which form a sub assumption hierarchy, and are described by 1650 different properties. The DBpedia includes 1.8 billion facts, linked to more than 30 other data sets in the linked open data cloud, which makes it the most important data set in the linked data[6]. Hence, we rely on the DBpedia for defining the creating a link between bibliographic records and their corresponding entities in the DBpedia schema. The proposal target is to establish relationship between British library subjects and corresponding terms in external data which is in our case we used the DBpedia. These relationships enable linking between data resources and expose the British library content to wider audience. We aim to set out if we could populate a field for each term in the British library with a link to an external resource. If we could provide links between British Library data and external resource and if we import definitions to the British Library from external matching external resources like the DBpedia we will be applying the semantic web knowledge. Adding linked data URI's to the subject areas would facilitate making the British Library Data available as part of the semantic web of linked vocabularies. DBpedia is used for this trail for two reasons. Firstly, as stated by the DBpedia "The DBpedia knowledge base is served as linked data on the web. As DBpedia defines linked data URIs for millions of concepts, various data providers have started to set RDF links from their data sets to DBpedia, making DBpedia one of the central interlinking-hubs of emerging web of Data" [8]

The DBpedia is updated based on the frequently-used Wikipedia pages, so has a method to stay current, and a way to add content to DBpedia pages, providing inbounds links so people can link to British Library subject area landing pages via DBpedia.

4. Proposed implementation

The purpose behind this proposal is to connect the library cataloguing RDF data with the DBpedia data, we developed a prototype semantic search engine that extracts selected information from the RDF databases of the British Library and use the results to linked it to the DBpedia databases to be shown as joining the libraries data with external databases. Figure 2 illustrates the work flow for exploring Dbpedia to extend bibliographic data from British Library.

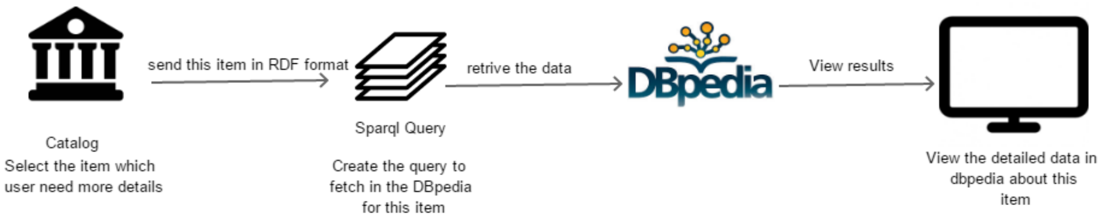

Figure 2 Work Flow for Implementation

In the practical work we implemented the process on 4 types of information from the British Library, we searched with the ISBN, the author, the Title and the subject of a book, we searched for them in the DBpedia database and we preview the output in our webpage. We created a a prototype interface the search process using PHP programming language, we entered the title of a book and the search engine retrieved the author biography from the DBpedia.

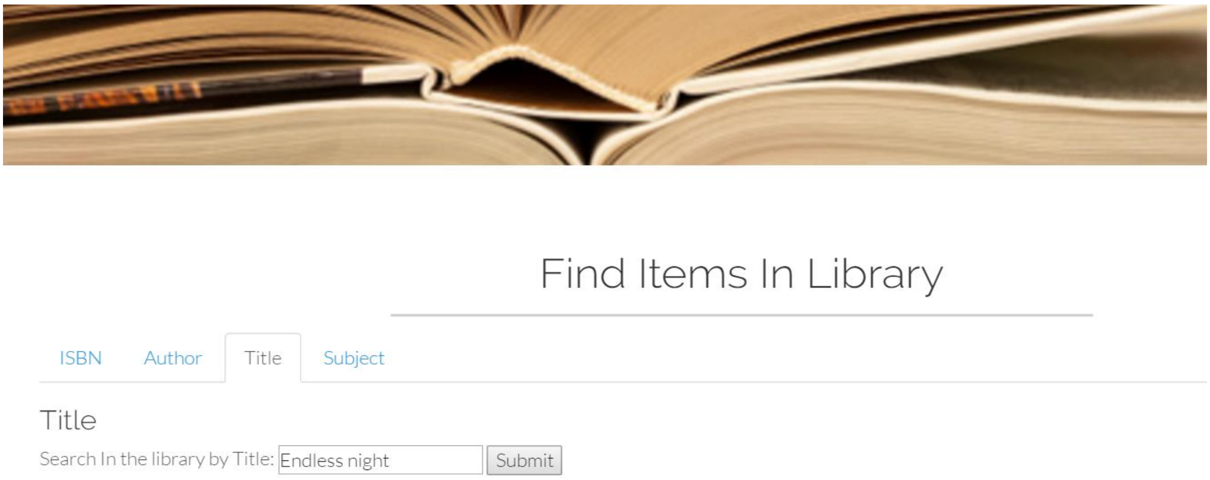

Figure 3 Input field

First we tested the proposal with the Title of the book, we used “Endless night” book as shown in Figure 3. Next, the SPARQL query takes the Title of the book and retrieved the authors of the book then extracts the authors biography from the DBpedia as shown in figure 4.

```
$query='
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX bio: <http://purl.org/vocab/bio/0.1/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX event: <http://purl.org/NET/c4dm/event.owl#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX isbd: <http://iflastandards.info/ns/isbd/elements/>
PREFIX madsrdf: <http://www.loc.gov/mads/rdf/v1#>
PREFIX org: <http://www.w3.org/ns/org#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdau: <http://rdaregistry.info/Elements/u/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX event: <http://purl.org/NET/c4dm/event.owl#>

SELECT DISTINCT ?authorname  WHERE {
    ?book dct:creator ?author ;
    dct:title "' . $searchTitle . '".
            ?author foaf:name ?authorname. }';
```

Figure 4 SPARQL Query to retrieve author details from DBpedia

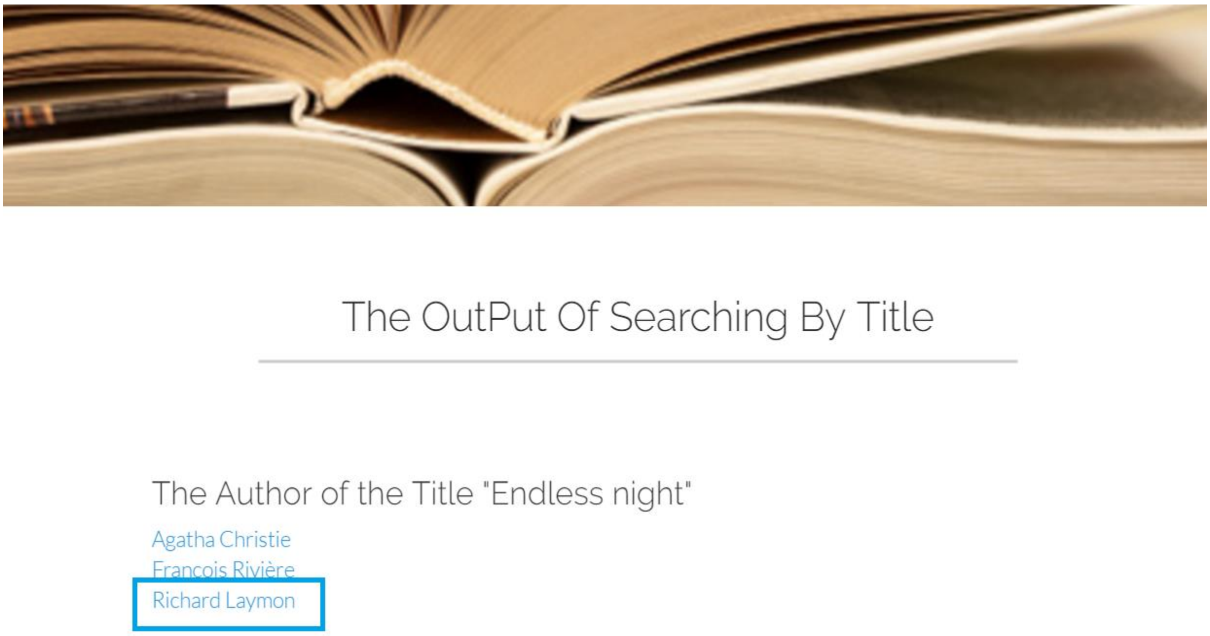

Figure 5 Select Author Name to get his biography

By selecting the author name as shown in figure 5 we get the author biography from the DBpedia in figure 6.

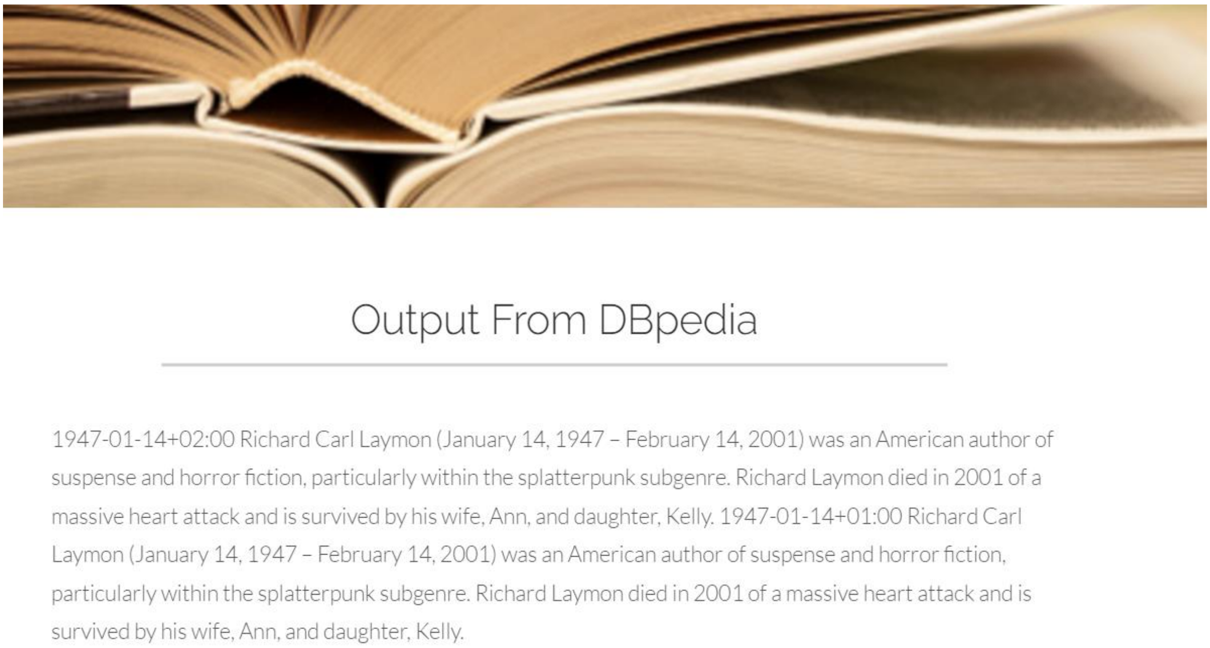

Figure 6 Author biography from DBpedia

5. What comes next

The proposal shows the connection of the British Library with the DBpedia through extracting information and getting output with brief details for the item. We are aiming in the future to apply more connections and more Sparql queries between the British Library and the DBpedia. We target in the future to extract more data from DBpedia like label, the publisher, years of publication and, description. Semantic digital libraries facilitate for the user finding more meaningful results. We are moving towards the artificial intelligence to be a massive web for highly intelligent interaction. The library and the linked data in the web are two concepts devoted to increase the information access and knowledge discovery. The research shows the applicability of the library resources with linked data.

References

1. Koivunen M-R, Miller E. W3c semantic web activity. Semantic Web Kick-Off in Finland. 2001:27-44.
2. "Data - W3C". *W3.org*. N.p., 2016. Web. 13 May 2016.
3. Retrieved January 25, 2016, from https://en.wikipedia.org/wiki/Cataloging
4. Retrieved January 25, 2016, from https://en.wikipedia.org/wiki/Resource_Description_Framework
5. R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data & Knowledge Engineering,* vol. 25, no. 1, pp. 161-197, Mar. 1998.

6. D. Allemang, and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, 2nd ed., Waltham, MA, USA: Morgan Kaufmann, 2011.
7. Linked Data for Libraries. Retrieved January 25, 2016, from https://www.youtube.com/watch?v=fWfEYcnk8Z8
8. "PLOS and DBpedia – an Experiment towards Linked Data | PLOS Tech." *PLOS Tech*. Web. 24 Feb. 2016.
9. Christensen, Anne. "Next-generation catalogues: what do users think." *Catalogue* 2 (2013): 1-15.