Linked Bibliographic Data: An Approach to Integrating Bibliographic Data with External Datasets from the Web

Hendrik
Department of Informatics
Universitas Islam Indonesia, Yogyakarta, Indonesia
Tel: (+62) 274-895-287, Email: hendrik@uii.ac.id

Eko Budi Prasetyo
Master's Program in Informatics
Universitas Islam Indonesia, Yogyakarta, Indonesia
Email: echo.koby@gmail.com

Teduh Dirgahayu
Department of Informatics
Universitas Islam Indonesia, Yogyakarta, Indonesia
Tel: (+62) 274-895-287, Email: teduh.dirgahayu@uii.ac.id

Abstract. Digital libraries share their catalogue services to expand access to their collections. Currently MAchine-Readable Catalogue (MARC) format is much used as a standard in sharing bibliographic records among libraries. However, this standard is limited to bibliographic data interchange among libraries. It cannot exploit data available on the Web. This paper presents an approach to enrich bibliographic data with related information from the Web. The approach develops linked bibliographic data by integrating an existing bibliographic system with external data sources on the Web. The approach consists of three steps, i.e. (i) analyse the bibliographic system's data structure; (ii) develop a linked bibliographic data model; and (iii) incorporate the resulted linked data into the existing system. To illustrate the approach, we conduct a case study in which we integrate an open source library management system, called SLiMS, with datasets from the Web, i.e. DBpedia, VIAF, and WorldCat. In the case study, bibliographic subjects are enriched with abstracts and labels; authors data are enriched with information related to the authors, including their other collections. Information about collections availability is made accessible so it can be linked to other libraries that implement this approach.

Keywords: linked data, library management system, bibliography, data integration, external datasets.

1. INTRODUCTION

Digital library is a new form of library organization that uses information technology (IT) to provide information services. It does not only consist of a collection of digitized books and other collections, but also electronic resources as well as technical capabilities. It allows users to create, discover and use bibliographic information (Sun and Yuan, 2012). The use of IT enables libraries to establish a digital library network. Such network allows the libraries to share their catalogue services to expand access to their collections.

A widely-used standard for interlibrary catalogue search is Z39.50 protocol(NISO, 2003). Using this protocol, databases of library's catalogues can be accessed over a TCP/IP computer network. Modern libraries use MAchine-Readable Catalogue (MARC) [1] standard to share their bibliographic records with other libraries. However, this standard is limited only to interchange bibliographic data among libraries. It cannot be used to exploit data available on the Web to enrich bibliographic data. Enriched bibliographic data would give users more information about the books and other collections.

Linked Data (LD) could be a promising technology to address this issue. LD refers to a set of best practices for

---

[1] https://www.loc.gov/marc/

publishing and connecting structured data on the Web. It is a new way for integrating data from many resources on the Web (Galiotou and Fragkou, 2013). LD is based on four basic principles: (i) things are identified using Unified Resource Identifiers (URIs); (ii) the URIs shall be dereferenceable by using HTTP URI; (iii) information shall be provided using Web standards, e.g. Resource Description Framework (RDF); and (iv) links to other URIs shall be provided to allow related information discovery (Heath and Bizer, 2011).

There are hundreds LD datasets available on the Web, that are grouped into several classifications such as government, media, life sciences, and so on. Among them, DBpedia[2] is one of the most popular LD datasets and becomes the central hub in Linked Data Cloud[3]. It extracts structured information from Wikipedia into RDF format so it is accessible on the Web (Bizer et al., 2009). DBpedia consists of more than 38.3 million things in 125 languages. VIAF[4] (Virtual International Authority File) is a name authority service that combines multiple name authority files from a number of national libraries in more than 50 countries. It is operated by Online Computer Library Centre (OCLC). Worldcat[5] provides a combined catalog from a very large number of libraries in more than 170 countries as its LD dataset.

This paper presents an approach to integrating bibliographic data with related information from external datasets available on the Web. The approach is to enrich bibliographic data so as to users would get more information about the library collections. To illustrate the approach, we conduct a case study in which we integrate an free, open source library management system, called Senayan Library Management System (SLiMS)[6], with DBpedia, VIAF, and WorldCat.

SLiMS is widely used by many libraries in Indonesia and other countries such as Thailand, Malaysia, and Bangladesh. It has been developed since 2006 by several developers; some of them have librarian background. Thus, the system fits with standards in the library domain. SLiMS has been localized into several languages, e.g. German, Thai, Malay, Arabic, Brazilian Portuguese and Spanish.

The remainder of this paper is organized as follows. In section 2, we discuss works that are related to ours, especially the implementation of LD in the library domain. In section 3, we present our approach to integrating bibliographic data with external datasets. It is illustrated with a case study. Then, in section 4, we present the result of this work and discuss relevant findings. Finally, we conclude our work with conclusions in section 5.

2. RELATED WORKS

In the digital era, the need to integrate library data with other data from any domain is a must. Currently, bibliographic collaboration and data sharing occurs among libraries only (Hannemann and Kett, 2010). To deal with this issue, it is suggested that existing library models such as Resource Description Access (RDA) and Functional Requirements for Bibliographic Records (FBBR) are made compatible with LD. By doing so, several benefits can be gained by digital libraries, e.g. promoting openness and share ability of metadata, enriching metadata with links, and so on. (Alemu et al., 2012)

Several works have been done to exploit LD in order to enhance interoperability of library data on the Web. A project called Europeana Linked Open Data[7] is aimed at publishing Europeana, that is a collection of digitized library data of Europe's institutions, into LD. Its dataset consists of a metadata of 2.4 million Europeana's collections, e.g. books, art, artefacts, and audio clips. This also linked to other LD datasets such as Swedish Open Cultural Heritage (SOCH) aggregator, GeoNames, GEMET and DBPedia (Isaaca and Haslhofer, 2012).

LD is also used to enhance a library catalogue model. For example, LIBRIS[8] is the Swedish Union Catalogue that is provided as a LD dataset. By employing LD, it enables users to find related information about a resource in the catalogue (Malmsten, 2008). A more comprehensive work that implements LD in the library domain is conducted by Linked Data for Libraries (LD4L)[9] project. This project covers the production of ontology, architecture and set of tools (Krafft, 2011).

3. APPROACH

We propose an approach to integrating bibliographic data with external datasets from the Web. The approach consists of three streps as follows:

1. analyse the bibliographic system's data structure,

2. develop a linked bibliographic data model, and

3. incorporate the resulted linked data into the existing system. In the following subsections, we describe those steps by applying the approach in the aforementioned case study.

[2] http://wiki.dbpedia.org/

[3] http://lod-cloud.net/

[4] https://viaf.org/

[5] https://www.worldcat.org/

[6] http://slims.web.id/web/

[7] http://labs.europeana.eu/data

[8] http://libris.kb.se/

[9] https://www.ld4l.org/

3.1 Analyse the bibliographic system's data structure

This step consists of two tasks. First, we analysed the database schema of SLiMS to identify which tables are related to bibliographic records. Note that SLiMS is a library management system; it comprises not only a cataloguing system, but also a circulation system. Those tables were candidates to be modeled as RDF resources. As a result, we identified five tables that fit to our objective. The tables are Biblio, Author, Publisher, Topic, and Place. The relationships between those tables are depicted in an RDF model as Figure 1.

Second, we analyse those tables to identify which properties can be enriched with external data. The analysis was done by considering the availability of external data that are related to the bibliographic data. As a result, we identified four bibliographic properties, i.e. author, publisher, topic or subject, and publish place.

3.2 Develop a linked bibliographic data model

This step consists of three tasks as described in the following subsections.

3.2.1 Designing a URI pattern

To implement the basic principles of LD as mentioned earlier, we designed a URI pattern to be used in the identifiers of resources. The patterns is as follows.

```
http://opendata-slims.javanehese.net:2021/resource/
{ClassName}/{ID}
```

ClassName refers to the name of a class in our model. **ID** is a sequence number to uniquely identify each instance of the class.

3.2.2 Designing a class mapping

Since SLiMS uses a relational database model, we had to map it into the RDF model. The followings are guidelines to do the mapping.
1. Each table column's name is mapped into a class property.
2. Each primary key value is mapped into an ID for a class.
3. Each data is mapped into the property's value or object of a resource.

For example, Table 1 shows the mapping of Author table into Author class. **rdf**, **slims**, and **schema** are namespace prefixes to shorten the URIs of the resources.
Table *2* lists the URIs of those namespace prefixes.

Table 1: Class mapping for Author table

Property	Object
rdf:type	schema:Person
slims:authorId	mst_author.author_id
schema:name	mst_author.author_name
schema:birthDate	mst_author.author_year
slims:authorityType	mst_author.authority_type
schema:seeAlso	mst_author.seeAlso

Table 2: Namespace prefixes used in the model

Prefix	URI
rdf	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`
slims	`http://opendata-slims.javanehese.net:2021/resource/`
schema	`http://schema.org/`

We used four columns of Author table as the properties of Author class. We added two properties, i.e: **rdf:type** and **schema:seeAlso**. rdf:type property is used to define the class type, whose value is schema:Person. schema:seeAlso property is used to link to other information related to an author. This property is to implement the fourth basic principle of LD. Example of an instance of this class is depicted in Figure 2.

3.2.3 Linking to other datasets

We examined several LD datasets available on the Web and discovered that three datasets can be used to enrich our bibliographic properties, i.e. DBpedia, VIAF and WorldCat.

DBpedia dataset covers information related to persons, creative works, places, organizations, etc. We used this dataset to enrich information related to author, topic, publisher and publishing place. VIAF dataset provides information on name authority. We linked this dataset with author property. Since WorldCat is the world's largest library catalog, thus we used this dataset to give a more detailed information of a bibliographic record.
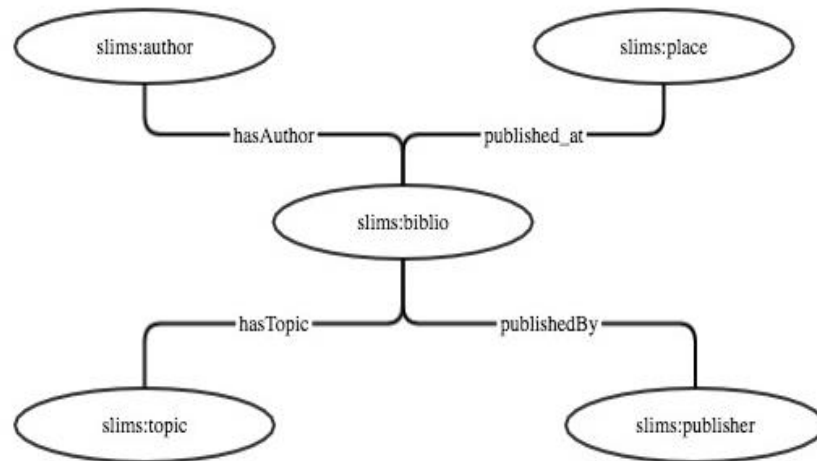
Figure 2: SLiMS Bibliographic RDF class model

3.3 Incorporate the resulted linked data into existing system

To incorporate the resulted linked bibliographic data (LDB) into SLiMS, we designed an integration architecture as depicted in Figure 3. There are two data sources for this system, i.e. internal and external data sources. The internal data source is the SLiMS database, which is a relational data model. The external data sources are from DBpedia, VIAF, and WorldCat.

To integrate data from both data sources, we mapped the internal data source into RDF data model using D2RQ framework[10]. We created a mapping file that contains mapping rules following the guidelines described in Section 3.2.2. This mapping file is saved in form of N-triple file (.n3 extension). The mapping file can be created manually or generated using a command in D2RQ framework. In the case study, we generated the file then modified it manually to insert additional data. We added a vocabulary to describe classes and their properties. We also added URIs of external data sources that are linked with our internal data.

The data conversion from SLiMS relational model to RDF model is done on the fly through that mapping mechanism. The RDF data are not stored into a physical storage. The RDF data can be accessed via browser either in SPARQL endpoint interface or SLiMS' interface.

4. RESULT AND DISCUSSION

To realise the integration of LBD into SLiMS, we customised SLiMS by modifying an end-user module and adding a module for administrative task. In Figure 3, the first

module is represented by Biblio Linked Data block. While the later is represented by Interlinking block. In the case study, we used SLiMS version 8.0 (Akasia). The modification and addition of these two modules are elaborated in the followings subsections.

4.1 End-user module

The end-user module is a landing page for SLiMS application. It is also called as Online Public Access Catalog (OPAC) module. It provides users a simple interface to search any library collections either in a simple or advanced mode. The searching process will result in a list of library collections containing several fields, i.e. the title, author and cover of the collections. Users can click on the title of a collection to get more detailed information about the collection such as the author, publisher, published date, etc.

Figure 4(a) depicts the original version of the detailed information page of a collection titled "The Definitive Guide to MySQL 5" authored by Michael Kofler and David Kramer. Figure 4(b) shows the modified version of the detailed information page of the same collection after integrating LBD. The modification are indicated by box A, B, C and D in Figure 4(b).

The modified version of the detailed information page contains enriching features as follows.

a. More works created by the authors. This information is retrieved from WorldCat. It can be obtained by clicking on WorldCat icons next to the author names (see Figure 4(b) box A). The information will appear in a pop-up window containing a list of the author's works.
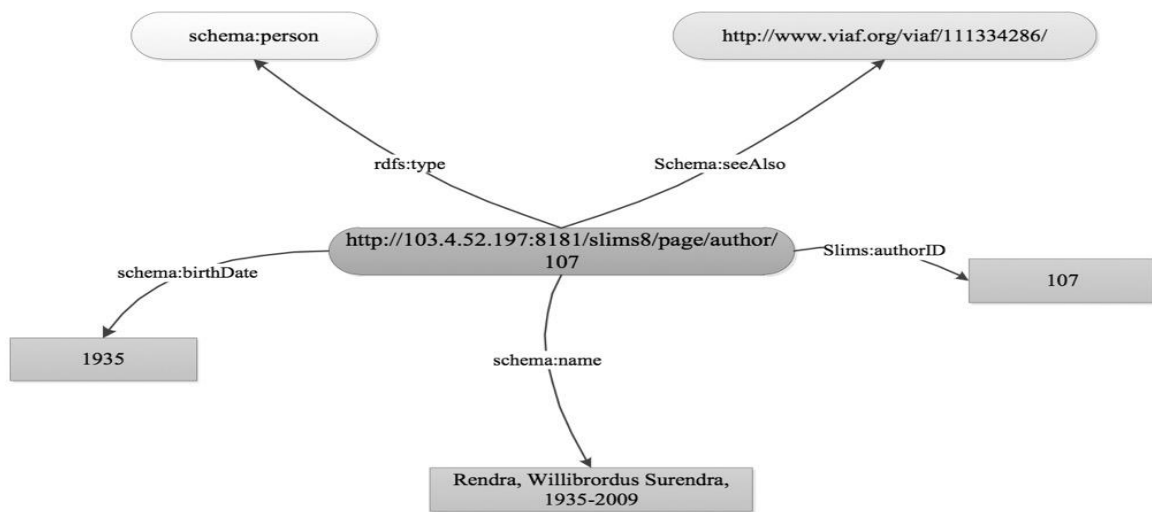

Figure 3: An instance of Author Class

b. Description about the subjects or topics covered by the collection. The description is retrieved from DBpedia and displayed in an infobox with header Subject Information (see Figure 4(b) box B).

In the figure, the subjects of the collection are Database and RDBMS. The infobox displays the description about RDBMS in Bahasa Indonesia because the librarian sets the data source to id.dbpedia.org (a localization of DBpedia in Bahasa Indonesia). This setting is done via the administration module.

c. Links to Wikipedia and DBpedia about the subjects (see Figure 4 box C). These links enable users to visit the original data source pages directly to obtain related information about the subjects.

d. Collections availability in other libraries. This information is retrieved from other SLiMS that implements LBD. It can be accessed by clicking on button "Other Libraries" (see Figure 4(b) box D). The information will appear in a pop-up window displaying a list of libraries have the collection as shown in 錯誤! 找不到參照來源。. This feature enables libraries to share their bibliographic data in ther digital library networks.
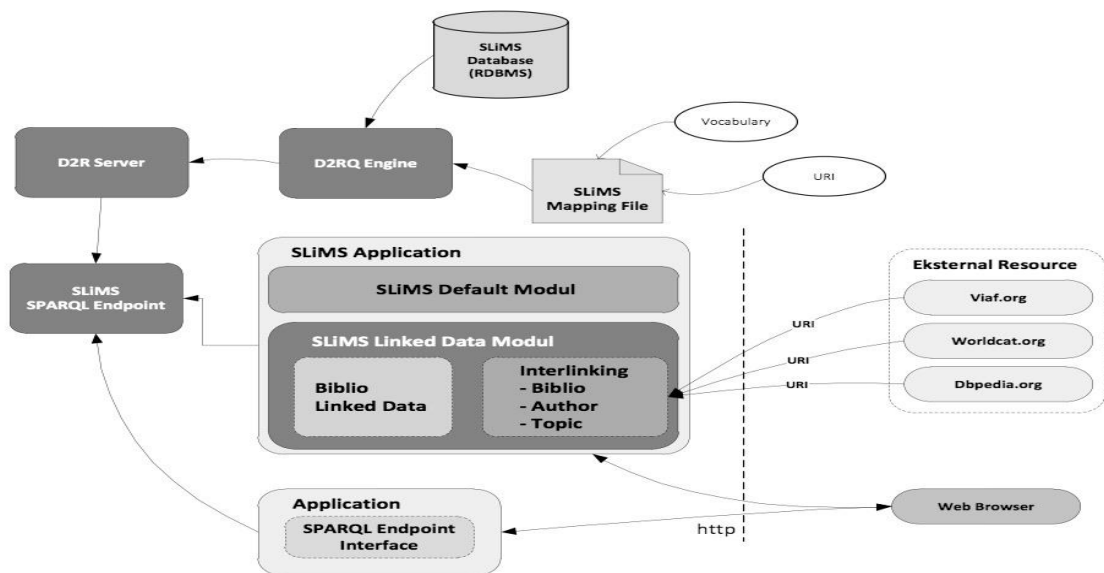
Figure 4: Integration architecture

(a) Original version   (b) Modified version

Figure 5: Detailed information page

4.2 Linked data administration module

This module allows librarians to manage the configuration of LBD, i.e. the linking to other data sources and the SPARQL endpoint. The SPARQL endpoint is to allow other libraries to interact with this SLiMS. The following subsections describes those two configurations.

4.2.1 Linking to other data sources

This configuration is to manage links or URLs of related data sources about authors, subjects, publishers, or places of the publication of collections. The links will be used as objects/values of seeAlso property. Two modes of adding links are provided, i.e. automatic and manual.

In the automatic mode (see Figure 6) , the system checks a pre-defined external data source whether it contains resources related to terms that already added in the bibliographic system. , e.q. a publication place named "America". If the term is founded, the system will return one or more URLs that point to that term/resource. Related to the instance, the system will return a number of URLs; some of them are:

- http://id.dbpedia.org/resource/Afrika-Amerika

- http://dbpedia.org/resource/United_States

- http://id.dbpedia.org/resource/Bank_of_America

- http://id.dbpedia.org/resource/Voice_of_America

In this case, the librarians shall choose the appropriate URLs that will be exposed to users. For example, the librarian chooses the second item in the above lists, the system will then add the URLs to seeAlso property. This URL will be displayed in "See Also" column.

In the manual mode, librarians must add new resources manually with some support from the system. As depicted in Figure 7, a librarian chooses an external data source for looking up the resource and type a term (a place name) to check, i.e "Austria". When, the librarian hit "Get URI" button, the system will search the data source to find information about the term. It then displays the searching result and the librarian can choose one or more URIs that are appropriate for the term. In this example, the librarian choose http://id.dbpedia.org/resource/Austria.

4.2.2 SPARQL Endpoint

This configuration is to register other libraries that implement modified SLiMS with LBD. These libraries will be looked up when users check collections availability as described in Section 4.1.

Librarians can register a library by providing relevant information, i.e. the library's name, location, URL, and SPARQL endpoint URL. Librarians can also edit and delete registered libraries. List of available SPARQL endpoints can be seen in Figure 8.

5. CONCLUSION

In this paper, we have presented an apporach to integrating bibliographic data with external datasets from the Web. The integration approach makes use of Linked Data technology. It consists of three steps, i.e. (i) analyse the bibliographic system's data structure; (ii) develop a linked bibliographic data model; and (iii) incorporate the resulted linked data into the existing system.

We have applied the approach in a case study, in which we modified SLiMS so as to be able to retrieve data from DBpedia, VIAF and WorldCat. By integrating bibliographic data with external datasets, users and librarians can get richer information. In addition, Linked Data allow SLiMS in different libraries to check collections availability with each other.

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. *World Library and Information Congress 78th IFLA General Conference and Assembly*.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*,

7(3), 154–165. Retrieved from http://www.sciencedirect.com/science/article/B758F-4WS9BS0-1/2/83cd58f9b584b76ccaa85cda59cca3a2

Galiotou, E., and Fragkou, P. (2013). Applying Linked Data Technologies to Greek Open Government Data: A Case Study. *Procedia - Social and Behavioral Sciences*, *73*(December 2015), 479–486. doi:10.1016/j.sbspro.2013.02.080

Hannemann, J., and Kett, J. (2010). Linked Data for Libraries. In *World Library and Information Congress: 76th IFLA General Conference and Assembly*.

Heath, T., and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. *Topology*. Morgan & Claypool Publishers.

Isaaca, A., and Haslhofer, B. (2012). Europeana Linked Open Data – data.europeana.eu. *Semantic Web Journal*.

Krafft, D. B. (2011). Linked data for libraries: A project update. *CEUR Workshop Proceedings*, *1486*.

Malmsten, M. (2008). Making a library catalogue part of the semantic web. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications* (pp. 146–152). Dublin Core Metadata Initiative. Retrieved from http://dl.acm.org/citation.cfm?id=1503418.1503434

NISO. (2003). Information Retrieval (Z39.50): Application Service Definition and Protocol Specification.

Sun, J., and Yuan, B.-Z. (2012). Development and Characteristic of Digital Library as a Library Branch. *IERI Procedia, 2, 12–17. doi:10.1016/j.ieri.2012.06.044*

Figure 6: automatic mode for linking to other data sources

Figure 7: manual mode for linking to other datasources

Figure 8: List of available SPARQL endpoint