JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

Invited Paper
J. of infosci. theory and practice 3(2): 06-15, 2015
http://dx.doi.org/10.1633/JISTaP.2015.3.2.1

Deploying Linked Open Vocabulary (LOV) to Enhance Library Linked Data

Sam Gyun Oh
iSchool, Library & Information Science, Sungkyunkwan University
Republic of Korea
E-mail: samoh@skku.edu

Myongho Yi *
Department of Library and Information Science, Sangmyung University, Republic of Korea
E-mail: josephlee@smu.ac.kr

Wonghong Jang
iSchool, Library & Information Science, Sungkyunkwan University
Republic of Korea
E-mail: jangwonhong@gmail.com

ABSTRACT
Since the advent of Linked Data (LD) as a method for building webs of data, there have been many attempts to apply and implement LD in various settings. Efforts have been made to convert bibliographic data in libraries into Linked Data, thereby generating Library Linked Data (LLD). However, when memory institutions have tried to link their data with external sources based on principles suggested by Tim Berners-Lee, identifying appropriate vocabularies for use in describing their bibliographic data has proved challenging. The objective of this paper is to discuss the potential role of Linked Open Vocabularies (LOV) in providing better access to various open datasets and facilitating effective linking. The paper will also examine the ways in which memory institutions can utilize LOV to enhance the quality of LLD and LLD-based ontology design.

Keywords: Linked Data, Library Linked Data, Linked Open Vocabularies

1. INTRODUCTION

Tim Berners-Lee proposes Linked Data (LD) as the crucial building block for the Semantic Web as a means through which humans and machines could efficiently collaborate. Unlike the hyperlinks characteristic of the Web of Documents, LD consists of raw data published according to the four principles proposed

Open Access

Accepted date: June 9, 2015
Received date: November 19, 2014

***Corresponding Author:** Myongho Yi
Assistant Professor
Department of Library and Information Science
Sangmyung University, Republic of Korea
E-mail: josephlee@smu.ac.kr

by Berners-Lee for realizing a clearly identifiable and richly interrelated Web of Data. LD is based on the RDF (Resource Description Framework) data model and can be represented in multiple, interchangeable serialization formats such as RDF/XML or N-Triples (Oh, Kim, & Jang, 2011). The library community has made consistent attempts to publish library data as LD; major institutions such as the Library of Congress and OCLC have either already published their data as LD or made plans to do so in the near future. While the amount of LOD (Linked Open Data) is steadily increasing due to various LD-related projects across an equally wide variety of institutions and industries, much improvement is required in terms of LD quality, the proper expansion of LD, and appropriate ontology designs in support of LD. The purpose of this research is to analyze LOV systems and use its findings in the expansion of LD, the improvement of LD quality, and the enhancement of LD-based ontologies for memory institutions. Therefore, the study will 1) evaluate the structures of LOV systems, 2) survey the current status of LOV data and its main features, and 3) compare the results of the LOV analyses with those of major LD and open data repositories in order to identify ways in which the overall quality of LD may be improved.

2. BACKGROUND OF LINKED DATA

2.1. Linked Data

Tim Berners-Lee initially proposed the principle of Linked Data in 2006 as a practical approach to the implementation of what the Semantic Web had envisioned as a collaborative effort between human beings and machines in the exploration and discovery of useful data and new connections (Berners-Lee, 2006). Technically, LD is not only clearly defined and machine-readable but also incorporative of mechanisms oriented toward linkage to external data. LD is also defined as openly published data that can be referenced by external datasets (Bizer, Heath, & Berners-Lee, 2009). Whereas the traditional Web of Documents simply connected HTML documents via hyperlinks, the new Web of Data provides rich connections between raw data using RDF, resulting in information that is richer in meaning and more open to the creative data processing of human and non-human users than ever before. With such advantages of Linked Data in mind, Berners-Lee recommended that LD be published according to the following four principles: 1) Use URIs as names for things; 2) Use HTTP URIs so that people can look up those names; 3) When someone looks up a URI, provide useful information using standards (RDF*, SPARQL); and 4) Include links to other URIs so that they can discover more things (Berners-Lee, 2006).

The main characteristic of LD lies in its potential to create valuable knowledge based on diverse data by sharing identifiers. The LD tends to be generated in various sizes and domains by individuals and institutions. The distributed nature of LD expansion poses serious challenges in enhancing LD quality since the former depends on the degree of connectivity based on common identifiers. Even though we have many well-established identifiers at both class and property and instance levels, the actual deployment of them has been lacking in a systematic manner.

2.2. LLD (Library Linked Data)

LLD (Library Linked Data) refers to a dataset published on the Web in accordance with the four LD principles recommended by Tim Berners-Lee. Library data is typically made available in the MARC21 or MODS formats. This data is usually converted into LD through database mapping or through conversion modules. Assigning unique identifiers in the form of Uniform Resource Identifiers (URIs) is the core of Linked Data. As library data already has standard identifiers and links among items based on subject authority, it has the advantage of being ready for publication as LD of higher quality than other datasets. The clear identification systems and cross-domain subjects of LLD make library data much more valuable when integrated with other datasets. In order to integrate LLD with the global LD, we need a smart search system that can help one to find identifiers associated with classes, properties, and instances of the worldwide LD. It is crucial that this system is regularly updated to provide seamless linking to the existing data.

2.3. Open Data

Much like open source or open access, open data is data free from copyright, patents, and similar control systems; it is available to anyone who wishes to access and republish it (Auer et al., 2007). Although the

concept of open data is not new, its popularity began with the advent of the Web. In recent years, the United States, United Kingdom, and major research institutions (including government agencies and businesses) have chosen to open their data to the public. The amount of data being published on the Web has been increasing exponentially each year. Most of the data made open via the Web are available in the formats of CSV, Excel, plain text, or XML/ JSON via OpenAPI. These kinds of open data can be easily converted into LD, enabling them to be integrated into other data since its data formats are compatible.

3. OPEN DATA REPOSITORIES (ODR)

The most important process in building a successful LLD is to identify appropriate data sources to which to link. One of the most effective ways to identify such data sources is to use open data repositories equipped with search capabilities. As mentioned in Section 2, in order to expand and enhance LD quality, we need a support system that allows us to easily find the information regarding an external dataset and its components. Along this line, there have been many attempts to implement data repositories with search capabilities. The major representatives of these kinds are Datahub.io, LODStats, Datacatalogs, and EU Open Data Portal. They allow one not only to search for published open data in diverse formats under a single interface, but also major LD published in the world. These experiences can be the bases for implementing an optimized LLD repository. We intend to use all of the above to find ways to expand LOV functionalities, thus enhancing LD quality.

3.1. Datahub.io (http://datahub.io/)

Datahub is a data registry service launched in 2007. It was developed using the CKAN (Comprehensive Knowledge Archiving Network) open platform maintained by the organization OK (Open Knowledge). CKAN is a Web-based open-source data management system for data repository. Datahub provides information on diverse datasets and details on licensing, dataset sizes, data owners, and SPARQL endpoints. Datahub is used as government data catalogs for UK's data.gov.uk and US's data.gov services. Open Knowledge, formerly known as the Open Knowledge Foundation, promotes open data. This site provides about 10,000 datasets and search services to them via diverse conditions. For example, searching can be done by a publishing entity, data format, usage condition, and keyword. However, it is very difficult to find detailed information regarding each dataset and its relationship to others.

3.2. LODStats (http://stats.lod2.eu/)

LODStats tracks the structure, coverage, and coherence of datasets on the Web (Demter, Auer, Martin, & Lehmann, 2012). LODStats gathers statistics about datasets related to the Resource Description Framework (LODStats, 2014). LODStats is comprised of a LODStats core module that can handle diverse LOD requests and a frontend module capable of providing the user with the Web interfaces, search classes, and data properties contained in LOD. Ermilov (2013) has developed a LODStats Web application that makes it possible to gather comprehensive statistics on LOD status and investigate the usage of vocabularies, classes, and properties. The LODStats system can be used to monitor real time changes of large CKAN data. Figure 1 shows the Python-based architecture of CKAN (Ermilov, Martin, Lehmann, & Auer, 2013).

As can be seen in Figure 1, LODStats holds major components needed for LLD repositories. It provides a SPARQL module to support queries for LD dump and a related module to confirm any updated status of data stored in the CKAN platform. LODStats also supports a search mechanism to find detailed information regarding classes and properties of each dataset.

3.3. Datacatalogs.org (http://datacatalogs.org/)

DataCatalogs.org is the most comprehensive list of open data catalogs in the world. The purpose of Datacatalogs.org is to provide users with a list of trustworthy ODR (open data repositories) selected by open data experts from all over the world. The first version of Datacatalogs.org was released at OKCon 2011 in Berlin, Germany. As of November 2014, the list includes detailed information on approximately 390 ODR. The strength of this service is its association with maps so one can easily find who is running each repository and its regional information, which is usually not available at most repositories.

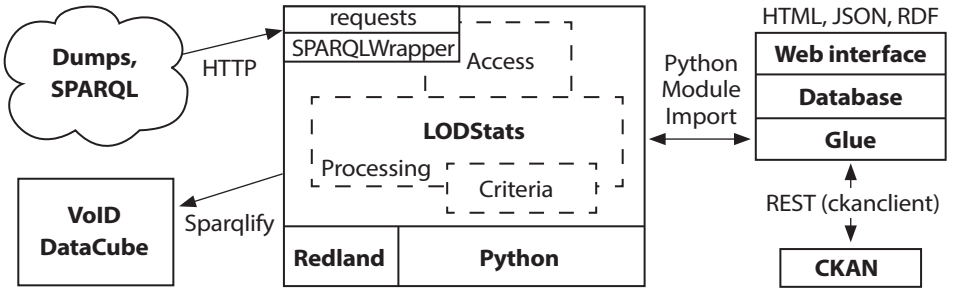

Fig. 1 LODStats system architecture

3.4. European Union Open Data Portal (https://open-data.europa.eu/en/data/)

The European Union Open Data Portal is the main access point for data from institutions and other organizations of the European Union (EU). The EU Open Data Portal consists of a catalog of major EU and EU-related organizations that provide data in LD format. Users can take advantage of this data for either commercial or non-commercial purposes. SPARQL endpoints may be used to search the data. As of November 2014, more than 7,600 linked datasets have been made available through the EU Open Data Portal.

4. ANALYSIS OF LOV SYSTEMS

4.1. Introducing LOV: http://lov.okfn.org/dataset/lov/

Linked Open Vocabularies (LOV) is a portal for RDFS and OWL ontologies developed by Bernard Vatant and Pierre-Yves Vandenbussche as a crucial part of the framework for the Datalift (http://datalift.org) project (Scharffe et al., 2012). The service has constantly been updated and refined and was selected as an official OKF project in 2012. LOV aims to achieve the same purposes as most ODR, furnishing users with a visual rendering of how LOD is used in Linked Data cloud settings and helping them design their own ontologies using available classes and properties. LOV uses the terms "vocabulary" and "ontology" interchangeably, with an emphasis on reusability, integration, and association.

The following are core criteria for inclusion in LOV:

- Appropriate data size
- A low-level normalized constraint (RDFS or partially described in OWL)
- Provision of instance information
- Provision of detailed documentation for users (labels, comments, definitions, descriptions, etc.)

LOV provides detailed descriptions of ontologies formulated in RDFS or OWL, both of which are widely used in Linked Data. The following are the characteristics required for an ontology to be included in LOV:

- Described using a Semantic Web language (RDFS or OWL)
- Published on the Web without usage limitations (cost, etc.)
- Content negotiation enabled by namespace URIs and searchable contents
- Appropriate data size for easy integration and reuse with other ontologies

According to 2011 criteria, an (ideal) appropriate data size is defined as 10 classes and 20 properties. More than 80% of LOV vocabularies have less than 100 elements. The largest LOV vocabulary is schema.org, with more than 500 elements.

The requirements presented above could render LOV more advantageous for the library community than other open data systems. LOV not only provides the user with diverse search capabilities concerning registered vocabularies but also focuses on datasets freely available under Creative Commons licenses for all purposes (including commercial usage). The following is the description of the four functions provided by the LOV system.

Table 1. The Functions Provided by LOV

Functions	Description
Aggregator	Provision of all vocabularies via either endpoint or dump file
Search	Search capability for finding classes or properties in ontology and vocabulary
Stat	Statistical display of current status of LOV vocabularies
Suggest	Capability of registering new vocabulary to LOV system

4.2. LOV System Analysis

The LOV system architecture is shown in Figure 2. The modules are integrated and work as a single Web application.

The functions of the main modules in the LOV architecture are shown in Table 2.

All LOV modules are implemented using Java language. The LOV system provides search API calls which can easily be interfaced with external systems. When a search API call is made to find elements or ontologies using the HTTP GET format, result values are returned in JSON format. The core of LOV is composed of Bot and Aggregator modules, which regularly update LOV for the latest versions of registered ontologies using SPARQL rules. The LOV modules also provide all statistics regarding updated ontologies relevant to LD disciplines.

4.3. Analyzing LOV Data

As of December 2014, LOV provides information regarding 10 categories and 421 ontologies. Table 3 is a summary of the sub-categories and number of ontologies that belong to each of these 10 categories. The major 10 categories are further divided to help users easily find representative ontologies in a given subject area.

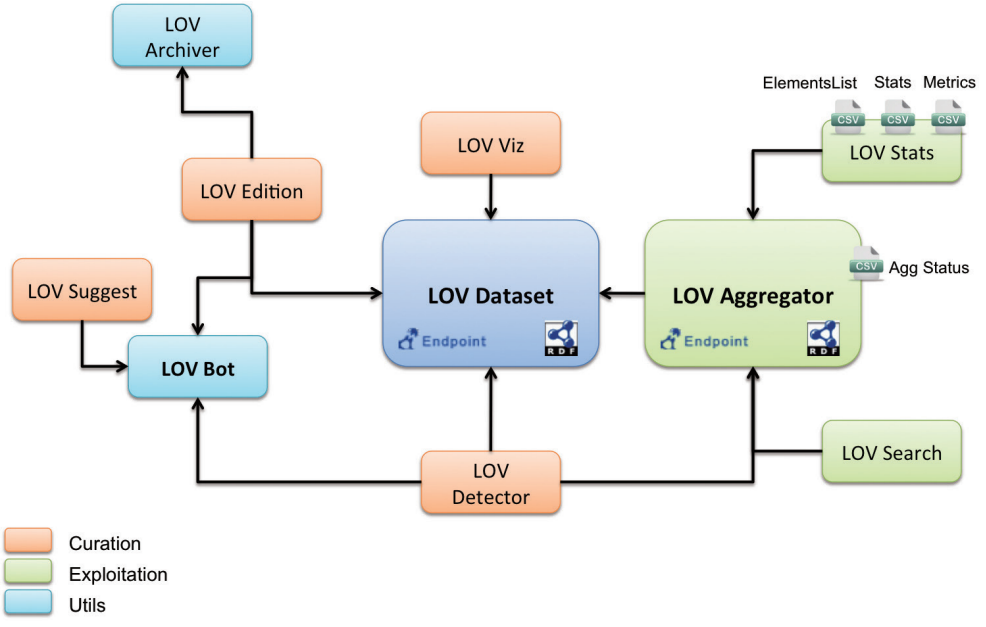

Fig. 2 LOV architecture

Table 2. Module Functions of LOV System

Module	Function
LOV Bot	Extracts the number of classes and ontologies based on SPARQL queries
LOV Aggregator	Downloads the latest version of ontology to maintain information regarding time line and to search all required information using a single SPARQL Endpoint
LOV Viz	Creates a new website based on a dataset collected in LOV
LOV Edition	Provides back-office supporting tool for data curator to edit and review LOV contents
LOV Search	Supports text-based searches for classes and attributes and targets all of the elements. Performs a search for all literal values.
LOV Suggest	Registers new ontology to LOV system

Table 3. LOV Ontology by Category

Category	Sub Category	# of Ontology	Description	Total
Data & System	SSDesk	5	Ontologies developed by the NEPOMUK Project	63
	Security	7	Security, Network, attacks and countermeasures	
	PLM	8	Product Lifecycle Management	
	RDF	17	RDF data bases, named graphs, RDB to RDF	
	API	26	Protocols, Services, API	
Science	Methods	30	Methods and Protocols, Units, Measures	46
	Life	9	Biology and Life Sciences	
	Health	2	Vocabularies for healthcare and medicine	
	Environment	5	Energy, Natural resources, Biodiversity	
Where & When	Geography	20	Geographical entities and features	50
	Events	10	Vocabularies for all kind of events: political, cultural, sports	
	Geometry	8	Geometry and spatial relationships	
	Time	12	Time intervals, timeline, time zones	
Media	Multimedia	10	Video, TV, Broadcasting…	18
	Image	3	Image and video description	
	Press	5	Press and news vocabularies	
Market	Industry	6	Industrial products and process	19
	Contracts	5	Calls, contracts, payments	
	eBusiness	8	Vocabularies supporting online business	

Library	Vocabularies	19	Vocabularies used to describe and organize vocabularies, languages, and terminologies	67
	Catalogs	27	Vocabularies used to describe and organize library and archives resources	
	SPAR	11	Semantic Publishing and Referencing Ontology	
	FRBR	10	Vocabularies built on the FRBR model	
City	People	16	Personal Information: vcards, career, genealogy, interests	53
	Society	19	Social relations, collectivities, organizations	
	Government	9	Government, Administrative, Organization	
	Academy	9	Vocabularies for academics and research	
General	General & Upper	11	Ontologies for general use in various domains	34
	Support	17	Miscellaneous general support vocabularies	
	PROTON	4	Ontologies developed in the PROTON project	
	Schema	2	Schema.org and extension	
Entertainment	Food	6	Food and drinks	17
	Travel	3	Transport and Tourism	
	Games	1	Games of any kind or any type	
	Rec	2	Review and Recommendations	
	Music	5	Music, Sound, Audio files	
Metadata	-	22	Vocabularies used for metadata and annotations such as Dublin Core	52
	Tag	7	Vocabularies describing tags, folksonomies, tagging events	
	Quality	23	Quality, Provenance, and Trust	

The categories with the largest number of ontologies are 'Data & System' and 'Library,' both of which are solid references to consult when implementing LLDs. The 'Science' ontology can be used when publishing LLD in the natural sciences, just as the 'Media' ontology is applicable to LLD on library multimedia resources. As of 2011, the vocabularies tracked by LOV have shown a high reusability rate, resulting in LOV providing quality data and sufficient mechanisms for linkage with other data (Poveda Villalón, Suárez-Figueroa, & Gómez-Pérez, 2012).

4.4. LOV Comparison

As shown in Table 4, Braunschweig (2012) has compared diverse open data repositories published by various governments across the world and proposed nine criteria for repository evaluation. The criteria are intended to measure data size, data quality, and degree of usability.

The results of the LOV data investigation using the criteria depicted in Table 4 are available at http://ww-wdb.inf.tu-dresden.de/opendatasurvey. Table 5 shows the results of applying the nine criteria developed by Braunschweig (2012) to LOV data and other systems mentioned in previous studies for the purpose of testing the overall usefulness of the LOV system.

As can be seen in Table 5, LOV and the EU Open Data Portal satisfy most of the criteria specified above.

Table 4. Comparison Element of Data Repository (Braunschweig, 2012)

Evaluation Criteria	Description
Number of published datasets	The most important questions about the data are whether the set has information about useful datasets. It is possible that the dataset can be assumed to have a high quality if the data holds many datasets.
Existence of standardized metadata attributes	Many platforms provide functions that add metadata to their datasets; the crucial question is whether a platform has standard properties that can be easily referenced by others.
Standardized file formats	In terms of reusability, providing the data in standardized file formats is the key.
Standardized domain categories	The grouping of datasets into a specific domain provides the user with a great advantage when searching. Automatic processing can also be greatly improved domain categories are standardized.
Standardized spatial/ temporal metadata	Data reuse will be maximized if spatial and temporal metadata are described using standard formats.
Existence of an API	Whether or not the open data platform supports API.
API granularity	Whether or not immediate access to platforms that support API is allowed or not.
Curation	Whether or not curation that supports the opposite of Wiki-style editing and uploading is available.
Latest date of activity	To judge whether data is actively being used and constantly updated.

Table 5. Comparison of LOV and Major Open Data Repositories

Criteria	LOV	Datahub	LODStats	Datacatalogs	EUODP
Number of published datasets	**421**	9,074	2,122	384	6596
Existence of standardized metadata attributes	**O**	O	O	X	O
Standardized file formats	**O**	X	O	X	O
Standardized domain categories	**X**	X	X	X	X
Standardized spatial/temporal metadata	**O**	X	X	O	O
Existence of API	**O**	O	O	O	O
API granularity	**O**	O	O	O	O
Curation	**Δ**	O	O	Δ	Δ
August 20, 2014 (As of 2014-05)	**2014-12**	2014-12	2014-12	2014-12	2014-12

While the coverage of LOV is relatively small compared to other repositories, this is due to the strictness of the criteria applied before any institution can register data to LOV. Due to the meticulous maintenance of LOV ontology quality, LOV is able to provide better quality data and exhibit a wider range of applicability compared to other open data repositories. The major strength of LOV is in its ability to specify how data are connected among different ontologies. The LOV provides visualization of relationships among internal and external ontologies when one searches for a particular vocabulary. In addition, the LOV provides all the versioning information regarding before or after ontology updates in the system, which is not offered by other systems.

5. WAYS TO UTILIZE LOV

The LOV system contains practical ontologies and provides straightforward methods to search ontology elements. The fact that all of the system's elements are provided further facilitates the examination of detailed items. Ontology search systems can accordingly be implemented in local environments. This study proposes several ways to utilize LOV with regards to LLD extension.

5.1. Implementing an Open and Specialized LLD Repository

Studies on and examples of LD and repository construction have tended to focus on government agencies and industrial sectors. Similar case studies in the library community are comparatively lacking. While a few major libraries have attempted to issue and publish LLD on the Web, these instances were made possible by the efforts of individual institutions rather than a collective trend. With the construction of various LD-based application services underway, an LLD repository system will be required for the library community to be able to conduct sophisticated searches and filter for data with greater ease. Trained professionals must implement LLD by actively utilizing standard identifiers, thereby increasing the quality of LD and making the linkage of library data to other open repositories more meaningful.

From this perspective, LOV can be used to build LLD-specialized repositories. Based on our analysis of the LOV system, LOV is equipped with all the functionalities needed to run an LLD repository. In addition, the LOV open source codes can be modified to maximize the repositories for the needs of the library community.

5.2. Applying LOV to Library Systems

Most libraries use commercial or open-source data management systems to organize their bibliographic and authority metadata. With the need to publish bibliographic data as LD on the rise, libraries should be prepared to publish LLD and integrate LLD with external LD. One approach is to implement an LLD-specialized repository system. The LLD-specialized repository system can enable the library community to query or import its data into library data, with LOV being utilized as a separate repository associated with a library system or as a module with the system. In other words, data mapping may be achieved by searching for LLD-specialized repositories to ingest metadata into a library system. Should this be accomplished, library systems would be able to build models with the capacity to publish data as LLD and automatically integrate said LLD into repositories. Such real-time ingestion from and publishing into repositories would in turn maintain the efficiency of LD publishing updates.

5.3. Applying LOV to Big Data Archiving

Recently, both academia and the industry sector have used Big Data analytics to study and derive meaningful results from massive amounts of data (i.e., social data or internal data). Traditionally, ontologies have been widely used for mapping, identifying, and integrating meaning in data analysis—a role expected to extend to big data analytics. As libraries have extended our knowledge of data collection and preservation, it is likely that they will play an important role in big data management. With such anticipations in mind, the LOV system presents the promising possibility of classifying, mapping, and analyzing data to be collected and preserved. LOV can be utilized in the analysis of mapping data types, major concepts, and data elements. In addition, LOV can be useful in enhancing the quality of LD and LLD by creating previously nonexistent associations. In order to produce continuous values from big data, we should be able to feed refined data based on the criteria from diverse contexts into big data analytics. In an effort to do so, discerning which data or ontology needs to be consulted for big data analysis at different contexts will be important and LOV could be employed meaningfully in this step. When archived big data linked with LOV are fully utilized, it will add value to big data and provide an incentive for archiving them.

6. CONCLUSIONS

The purpose of this study was to review the potential use value of the LOV system in the library community. The study conducted technical and data analyses of current LOV systems and compared them with major open data repository systems in order to identify ways to utilize them in the library community. The suggested use of the LOV system in the library community is briefly described and a more extended investigation of LOV

use is foreseen from an LD service perspective. LOD is rapidly expanding, and with it increases the need for more studies on how to increase LOD and LLD-based ontology design and implement a search system that provides users and application developers with easy access to LLD. As clearly demonstrated in the comparative analysis of LOV and open data repositories, LOV displays the functionalities for LLD search and use required by the library community. In order to capitalize on the potential of LLD, it is expected that future studies will be conducted in this area by comparing current repositories and optimized repositories and employing evaluations of repository use to improve the usefulness of LLD.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Berlin Heidelberg: Springer.

Berners-Lee, T. (2006). Linked Data. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS), 5*(3), 1-22.

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The state of open data. Limits of current open data platforms. *Proceedings of the International World Wide Web Conference*, WWW 2012, Lyon, France.

Demter, J., Auer, S., Martin, M., & Lehmann, J. (2012). LODStats—An extensible framework for high-performance dataset analytics. Paper presented at the EKAW 2012.

Ermilov, I., Martin, M., Lehmann, J., & Auer, S. (2013). *Linked open data statistics: Collection and exploitation*. Berlin Heidelberg: Springer.

Oh, S., Kim, S., & Jang, W. (2011). Analyzing current state of library linked data, designing an integrated LLD, and thoughts on ways to expand LLD. *Journal of Korean Library and Information Science Society, 42*(4), 331-351.

Poveda Villalón, M., Suárez-Figueroa, M. C., & Gómez-Pérez, A. (2012). The landscape of ontology reuse in linked data. 1st Ontology Engineering in a Data-driven World (OEDW 2012) Workshop at the 18th International Conference on Knowledge Engineering and Knowledge Management, Galway, Ireland.

Scharffe, F., Atemezing, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., & Vatant, B. (2012). Enabling linked-data publication with the Datalift platform. Paper presented at the AAAI Workshop on Semantic Cities.