

A Study on the Open Platform Modeling for Linked Open Data Ecosystem in Public Sector

Suck-hyun Song*, Tae-Dong Kim**, Myung-Won SONG

*Department of Information Resource Planning, NIA(National Information Society Agency),
and Electrical & Electronic Engineering, Yonsei University, Korea

**Department of Information Resource Planning, NIA(National Information Society Agency)
and Information Science, Chung-Ang University, Korea

***NIA(National Information Society Agency

ssh@nia.or.kr, fellow33@nia.or.kr, smw@ieee.org

Abstract— This Study analyzes the cases of the U.S. and U.K., Wikipedia, and as well as domestic environment from the perspective of formulating a nation-wide data linking system for Korea. Based on this analysis, a linkage platform for public sector data, named “National DBpedia” is designed and implemented utilizing open platform and LOD (Linked Open Data). In addition a foundation for governance including LOD-based metadata basic design and registration method and operations is also designed. In addition, actual public and private sector data are applied in the actual field for a proof-of-concept. On the basis of this study, By expanding public and private sector data linkages, the research provides the cornerstone for building a public sector “wikipedia” for Korea.

Keywords— Open Platform, LOD, DBpedia

I. BACKGROUND

Recently, the U.S., the U.K, and the EU strongly promoted the data opening policy. The governments of the U.S. and the U.K. actively opened their data to people, using data.gov in the U.S. and data.gov.uk in the U.K. Various types of Web 2.0 based technologies are used, including raw data download, open API, and LOD (Linked Open Data), in order to provide the data to people so that they can use the data easily in their daily life. In Korea, the national database development project has been implemented since early 2000. It archives the data that is valuable at the national level. As a result, numerous quantitative and qualitative data is accumulated in the government, which laid the foundation of being selected as the world’s top e-government country by the U.S. for two years in a row. However, the related data policy remains at the data accumulation level only to keep the preservability of the national information. Accordingly, Korea is not advanced in the open policy that is creating a worldwide boom. Therefore, the Korean government is trying to actively open the data to the public, based on the recognition that the accumulated data can be a new growth engine, instead of simply realizing the people's right to know. This study is being carried out as a part of the policy that will become the smart infrastructure of the future in Korea.

II. ANALYSIS OF THE PRESENT STATUS AT HOME AND ABROAD

The present status at home and abroad was reviewed from the policy aspect and system aspect. Based on this review, the national knowledge intelligent link system was implemented by considering the current situations of Korea.

A. Policy aspect

1) U.S.

As the public data opening movement started in earnest at the beginning of 2009, the U.S. government began to open the machine readable data set that had been created by the federal government since the Obama government came to power. This has been done in order to improve government transparency and to create added value. To convert some of the developed data.gov into the RDF format and join the linked data cloud, the Tetherless World Team led by Professor Jim Hendeler, a founder of Semantic Web, implemented the project, together with Tim Berners-Lee. Up until now, they have converted 565 data sets into the RDF format and created 5,435,413,491.

2) U.K.

The U.K. government opened the public information opening site (www.data.gov.uk) on January 21, 2010, based on the “Power of Information” report, in order to create values by sharing and utilizing information in the public sector. The U.K. government serves more than 3,000 linked data sets. Major application services include the house price trend service; the private hospital location search and recommendation service; and the integration service, which shows the distributed research and investment information at a glance.

3) Other countries.

Many countries in the world are also participating in the data opening movement, including Australia, Timor, Spain, Singapore, Norway, New Zealand, Morocco, Moldova, Kenya, Italy, Ireland, Hong Kong, Finland, France, Estonia, Demark, Canada, and Austria.

4) Domestic trend

Korea is quite slow in developing the linked data service and technology diffusion, compared with other countries. It is hard to find a case of linked data, which is a practical case of Semantic Web. The following section describes the only two cases in Korea, which are KISTI LOD Cloud and bibleontology.

5) KISTI LOD Cloud

KISITI LOD Cloud refers to the linked data service that is operated by RKBExplorer, while the KISTI (Korea Institute of Science and Technology Information) provides the data to RKBExplorer in the University of Southampton. RKBExplorer serves the triple data that is converted from some of the data included in NDSL, which provides the thesis, patent, report, and trend information to domestic users.

6) Bibleontology

Bibleontology is the open linked data service (beta version) where Cho Myeongdae, a professor at the Graduate School of Convergence Science and Technology at Seoul National University, designed ontology regarding figures and events in the Bible and published using the linked data solution provided by TopQuadrantKorea.

B. System(website) aspect

1) www.data.gov (U.S).

The U.S. federal government operates the website and provides the service of retrieving the government's open data. Also, the website registers the meta information of the agency database, and performs RDF conversion after collecting the data first.

TABLE 1. DATA.GOV : IMPORTANT COMPONENT

Data	System	Service
Data information The meta information provided by individual agencies. The raw data possessed by individual agencies. The data possessed by the federal government. Data registration method Webmetadata registration. Computer registration via the API. Data format Converting the possessed data into the RDF format. RDF or structured as tables/spreadsheets (JSON, XML, CSV, etc.).	Website Search public data. Dataset Management System Managing the data registered by the agency. The metadata catalog Managing the meta information after classification. Performance tracking Analyzing the data use detail. Data discovery Hosting service Central data collection and participation. Provision to other agencies.	Sem Wiki Google Viz Tag Cloud RSS 1.0

2) DBpedia (Wikipedia)

DBpedia is the service designed to support the utilization of Wikipedia, which is called a user participating type dictionary, as more accurate and structured data on the web. The linked data technology is applied, and it is said that, "DBpedia is the Semantic Web mirror of Wikipedia." As of August 2011, DBpedia contains the information about 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species, and 5,400 diseases, which allows it to play an important hub role for the "web of data." The central data is created by user's participation, and RDF converted data expresses a relationship.

TABLE 2. DBPEDIA : IMPORTANT COMPONENT

Data	System	Service
------	--------	---------

Data information Data collection by web user sharing. User's voluntary participation. Web encyclopedia. Data registration method Media-wiki software is used. Wiki DB information extraction information is utilized. RDF data conversion design. Real-time, semi real-time data update. RDF change by Wiki change. Data saving format RDF data.	Wikipedia Data registration/collection channel role. Media-wiki. Wiki data extraction framework Extracting the Wiki information in various ways. RDF conversion. Data synchronization (Wikipedia update). Data browsing/retrieval Html/RDF/SPARQL interface.	LOD cloud Link relation service among the LOD data. Reference role among various domains Role of a reference site when connecting domains.
--	---	---

C. Analysis results and implications

1) Data aspect

The data format is structured in a machine-readable format to link, converge, and combine the data effectively. Then, the data is formalized (meta creation and registration by RDF dataset, and resource standardization) to prevent duplicated data creation and save on search time.

2) System aspect

To secure cost and performance efficiency, the low-capacity data management technology that operates and manages the metadata in an integrated manner, and raw data distribution type should be developed and operated, instead of centralized operation/management. In this way, the ever-increasing data can be combined and managed efficiently.

3) Service aspect

The RDF description is provided and the service is developed that can utilize SPARQL endpoint, in order to increase data accessibility, to develop a new application using increased accessibility, and to improve data reusability.

TABLE 3. ANALYSIS RESULTS

	www.data.gov	DBpedia.org	Domestic situations
Data	RDF conversion after registering the organization metadata. - Continuous occurrence of the conversion cost, time increase, keeping the minimum data, centered on the public data.	RDF conversion after collecting the web dictionary data. -Advanced conversion technology is required, time increase and data increase quantity need to be considered, centered on the open data	Most of the data possessed by agencies is unstructured data. -Data duplication occurs, and conversion costs increase.
System	Duplicated data management	Low-capacity data management	A silo is created among systems (closing/discon

	(before/after RDF) cost occurs. Proceeding with the central type data concentration method.	technology. Proceeding the distributed processing data method.	nection). Proceeding with the individual operation method.
Service	Including many services utilizing the collected data (Life service like a map and hospital). Holding a data utilization competition. Providing the API for data extraction by agency. Providing SPARQL for metadata retrieval.	Service via LOD cloud. Utilizing the reference data. Global data link expansion is required. Providing RDF description and SPARQL.	Immature LOD-based service (RDF description and SPARQL demonstration level). The reference data is processed by individual agencies.

III. DEVELOPMENT OF THE NATIONAL DATABASE PLATFORM

A. Platform design and implementation

1) System functions

Functions are provided that link the LOD data collected by agency and increase linkage, as well as play a role of the data hub that can increase data utilization.

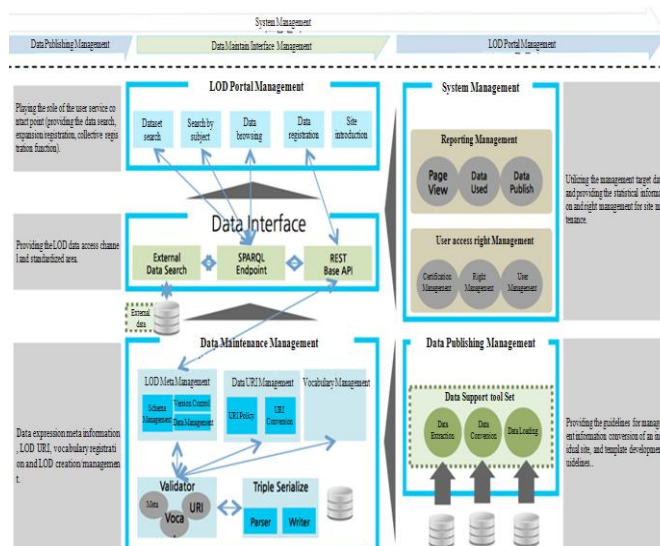
TABLE 4. NATIONAL DBPEDIA'S SYSTEM FUNCTION

Data	System	Service
Data information LOD dataset collected by agencies/organizations. Unit meta collection. Resource information by registered dataset. Partial collection. Data registration method Registering a Web-based interface RDF. Automatic dataset source collection. Data format RDF data. Centralized meta management.	LOD portal Web-based data retrieval/registration. Data link search. Dataset Interface External RDF access (SPARQL endpoint). Data registration/modification (REST Base API). Syndication module support. Data search and management. Data maintenance management Supporting RDF structure verification/conversion. Triple serialization.	Meta search and utilization Search by including the dataset's resource information. Domain independent data research. Data sharing among registration agencies/organizations.

2) System functions

Figure1. shows the overall composition of the national DBpedia platform

Figure 1. National DBpedia Platform's Conceptual Diagram

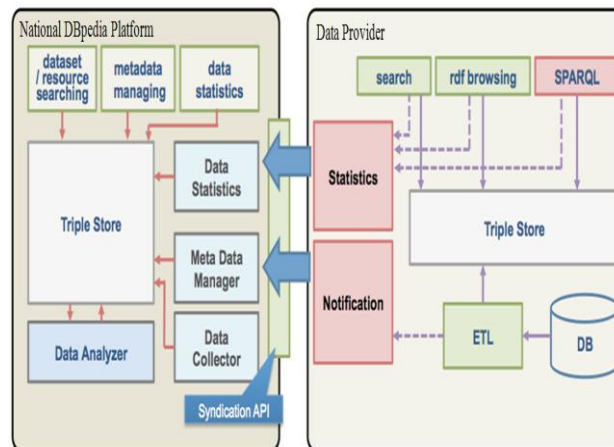


National DBpedia proposes standardized configuration to the data provider system (client), and makes a catalog of the meta-information from the linked data cloud that is participated in by many data providers. Therefore, National DBpedia enables the data user to search the linked data quickly on the web using the consistent interface, and provides the infrastructure platform where the data link among data providers can be accelerated.

3) System architecture

The architecture of National DBpedia and the data provider's system is as follows:

Figure 2. National DBpedia's System Architecture



1) Data provider system

The data provider's LOD system can be configured as shown in the example above. The modules in the red boxes in the data provider's system indicate the module that should be linked with National DBpedia. The data provider's system is composed of the SPARQL endpoint, the meta-information notification module, the statistics module, and the ETL (Extract-Transform-Load) module.

2) National DBpedia system

3) *Syndication API*

The syndication system is required between National DBpedia and the data provider's system to register the data set. As shown in the below sequence diagram, the dataset is registered using the Syndication API when the dataset is registered in National DBpedia for the first time.

The syndication system is required between National DBpedia and the data provider's system to register the data set. As shown in the below sequence diagram, the dataset is registered using the Syndication API when the dataset is registered in National DBpedia for the first time.

① The data provider describes the meta-information of the registering dataset by using the URI that can be referred reversely.

③ The platform reads the RDF of the dataset meta-information by reversely referring to the URI.

⑤ The platform accesses the SPARQL endpoint using the data meta-information and collects the meta-information of the resource.



Based on the data opened on the web (22 public data sets and 2 private data sets), 24 RDF dataset metadata were converted and then, 24 datasets and 12,781 resources were registered using the national DBpedia platform to verify the service function by checking the dataset relationship and search. The synchronization Syndication API that was proposed to prevent data inconsistency during dataset registration could express the useful data relationship in a distributed data environment by providing the two-way data synchronization function and

Figure 4. <Apartment information> dataset metadata

property	value
----------	-------

[Downloaded from ascelibrary.org by University of California, San Diego on 06/01/15. Copyright ASCE. For personal use only; all rights reserved.](#)



From the data production method aspect, the platform has high efficiency in project implementation, such as organization oriented data creation, deciding the LOD development scope and timing depending on the organization's environment, data ownership, knowledge data collection with specialized knowledge about the domain, and data lifecycle management (data change and management efficiency).

When the organization LOD is developed using the meta utilization LOD collection method, the dataset meta developing guidelines can be provided, and additional meta information can be collected after registering the meta information.

Also, distributing systems and storage provides stability and efficiency, the use and data increase can be responded to in a flexible manner, and system performance efficiency can be provided.

IV. CONCLUSIONS

In this study, the open platform (National DBpedia) was designed and implemented, including the LOD-based procedure and description, and architecture, based on the analysis of overseas data portals, the strengths of those portals, and the present conditions in Korea. The proof test was also carried out to verify validity from the system aspect. As a result, the pan-national platform was implemented, which can be reused by the private sector and government, by opening the data produced by the government systematically as the public asset. The developed platform will take the position of the data platform in the public sector that represents Korea through connection with DBpedia, which is linked with Wikipedia. In addition, the demonstration project will be carried out continuously to expand and advance the system, and additional basic studies will be conducted on the LOD-based data specification and detailed standardization, and on the development of the application service model.

REFERENCES

- [1] Kim Hakrae. 2012. Overview of Open Data, Linked Data and Web Science. 2012 Open Data Workshop presentation data, August 2012
- [2] Oh Wonseok. 2009. Linked data based e-Government promotion strategy to create soft power, TopQuadrant Korea, October 2009
- [3] Lee Jeonga. 2010. Smart government's public information opening and use promotion strategy, National Information Society Agency, CIO Report Vol.28
- [4] Cho Myeongdae. 2011. Linked Data for UCI. Korea Creative Content Agency, 2011 UCI utilization strategy presentation data, February 2011
- [5] Jeong Hanmin. 2010. KISTI Semantic web technology study, 2010 Korea Semantic Web Conference (KSWC) presentation data, March 2010
- [6] CKAN. 2012. Guidelines for Collecting Metadata on Linked Datasets in CKAN [cited 2012 Jul 21]
- [7] Georgi Kobilarov. DBpedia: An Interlinking-Hub in the Web of Data, 2010
- [8] Li Ding, et al. 2011. TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web 9(2011): 325 – 333
- [9] The White House Office. 2012. Digital Government: Building a 21st Century Platform To Better Serve The American People
- [10] UK Cabinet Office. 2012. Open data white paper: unleashing the potential [cited 2012 Jul 5]

- [11] W3C. Publishing Open Government Data.W3C Working Draft.Web page, 2009
- [12] Socrata, Unlock Government Data with Social Data Discovery, 2009
- [13] Socrata, 2010 Open government data benchmark study.



Suck-hyun, Song is director of information Resource Planning in National Information Society Agency and vice-chairman of Telecommunication Technology Association PG420 and Expert advisor of the Council of ISO SC38 in Korean Agency for Technology and Standards and Expert advisor of E-government Professional Committee in Presidential Council on National ICT strategies society. he received his master degree at the electrical electronics study in 2002 and in a doctoral Program in Electrical Electronic Engineering in Present. His significant area of interest is Government IT Planning. His significant areas of interest are NationalDB, Cloud computing, RFID/USN, u-City, E-government



Tae-Dong, Kim is senior researcher in the National Information Society Agency and technology sub-commissioner of UCI (Universal Contents Identifier), he received his master degree at the information science study in 2006. His significant area of interest is LOD(Linked open data) in semantic web sector



Muyn-Won SONG is Executive Director of Auditing Office in National Information Society Agency. He received PhD. in Computer Science in 2009, He is interest in Cloud Services in Public Sector and Open Platform