Semantic Data Fusion: from Open Data to Linked Data

Haklae Kim, Jungsung Son, and Kisoog Jang

Samsung Electronics Co., Ltd.
Maetan dong 129, Samsung-ro, Yeongtong-gu, Suwon-si,
Gyeonggi-do 443-742, Korea
{scot.kim, junsung.son, kisoog.jang}@samsung.com

Abstract. Open government data is any data and information produced or commissioned by government bodies. Providing open government data is a matter of transparency and accessibility of government. However, this information is usually published in raw format, which is lack of specific guidelines. Linked data technologies, on the other hand, aim at transforming data published on the Web into a machine-readable format allowing them to be linked to other data sources. In this paper, we present a case study on the application of linked data technologies on government data in Korea. In particular, we focus on examining how this information can be transformed into linked data, and then how this linked data can be useful for providing new information across heterogenous domains or datasets.

1 Introduction

Open government data is about getting access to information held by government bodies [4]. The terms has come into prominence recently, becoming popular in 2008 after the publication of a set of principles and the launch of open data portals in the United States and the United Kingdom [3, 4]. Open government data are being established with increasingly solid evidence from many national and local governments. However, they are concerned with various issues including technical, social, and legal perspectives. From a technical point of view, open data is often published in raw format, which is need for being transformed into machine-readable formats for reusing these datasets. In addition, these data sets might have identical resources, though the details on these data sets can hardly be different. When identical resources may be existed, these can be linked each other correctly thanks to Linked Data technologies [1].

In this paper, we describe how resources among heterogeneous data sources can be connected using Linked Data, and then present a specific use case, which aims at delivering better recommendations via a combination within the linking open data datasets.

2 Methodology

2.1 Data collection

The data was obtained by more than one method and at different places via crawling websites, downloading a set of CSV/XLS files, or using through APIs. The following data sources are collected:

- Korea administrative divisions (24,000 records) - legal entities established for the purpose of government[1]
- National treasure (18,000 records with images and videos) - architecture, a landscape, document, or other artifact that is considered to be of national significance and an embodiment of the national heritage[2]
- Cultural Facilities in Seoul (6,000 records) - buildings, structures, parks or places for programs or activities [3]
- Subway Lines and Stations (17 lines and 600 stations): a terminal where subways load and unload passengers [4]

Some challenges of using collected data can be understood as technical problems addressing information representation, storage, and access. In particular, these data sets are stored in different formats without any connections among them. Linked data technologies act as a bridge to allow these data sets to map other data sets or relevant resources on each data set [2].

2.2 Data Transformation and Interlinking

Information about schemas/data structures used by other projects is also useful. Reuse of terms from well-known RDF vocabularies is highly desirable to ensure interoperability across heterogeneous datasets on Linked Open Data. It maximizes the probability of the data being consumed by applications tuned to well-known vocabularies, without further processing or modifying the application. In principle, we use the Schema.org[5] as a fundamental data model, and define additional vocabularies when all the requirements are not available.

Like other linked data, one of core functions is to make these data sets interlinked and freely available and accessible on the Web. The data is mapped to more than one schema or ontology: most of resources are mapped to instances of administrative divisions of Korea, and some of these data sets links to other external sources such as Dbpedia[6]. The data is stored in a triple store called 4Store and is published in the form of URIs with data in human readable and machine readable form. A SPARQL endpoint to an RDF description of data sets is developed.

[1] http://kostat.go.kr/kssc/board_notice/BoardAction.do?method=list&board_id=3&catgrp=kssc&catid1=kssc06&catid2=kssc06a

[2] http://www.cha.go.kr/korea/heritage/search/Detail_Result_new2012.jsp?mc=NS_04_03_01

[3] http://visitseoul.net

[4] http://data.seoul.go.kr/openinf/sheetview.jsp?infId=OA-1190

[5] http://schema.org

[6] http://dbpedia.org

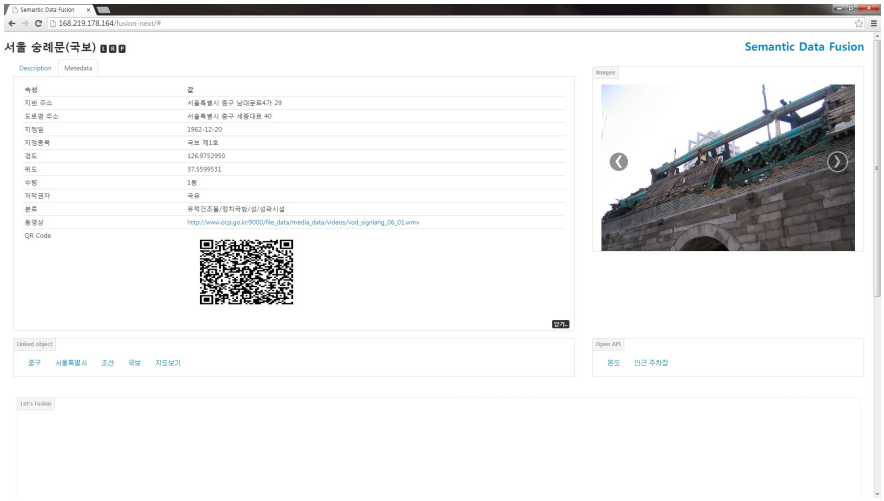

Fig. 1. Main interface of the Semantic Data Fusion

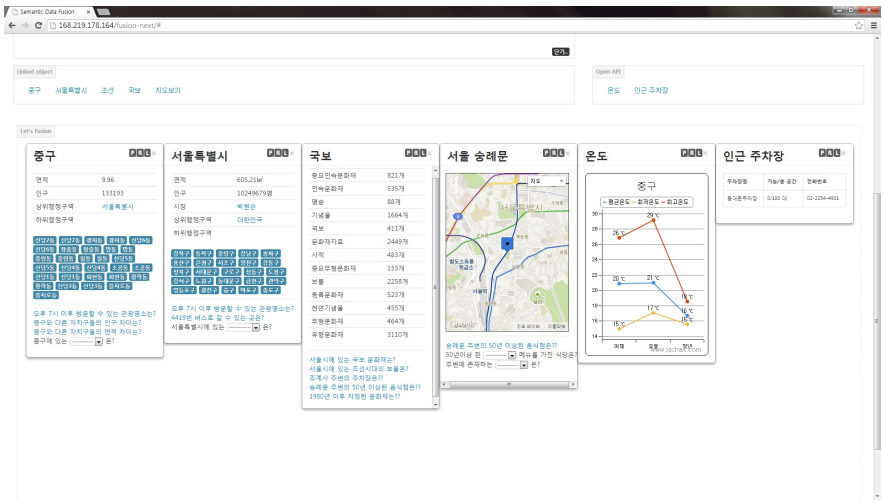

Fig. 2. Semantic Data Cards

3 The Semantic Data Fusion

This application is to allow end-users to deliver answers against complex questions, such as "Where are available restaurants which serve Italian foods near by a selected resource after 9 PM?" The Semantic Data Fusion excels at answering such complex questions via a combination of SPARQL queries via SPARQL endpoints.

– **Semantic Search**. It incorporates context, intent and concept of a given request via keyword-to-concept mapping. For example, if a user gives a simple keyword '*Gangnam*' in order to looking for restaurants closed to *Gangnam* in *Seoul*, all instances correspond to this keyword are returned and classified according to their concepts. Then users can choose only one classified resource from the search results.

– **Data Fusion**. It contains two components: 1) *the metadata viewer* provides easy read access to overall metadata of a selected resource with images. If metadata element has a URI as value, this one is located in the *Linked Objects* to perform a specific action, as shown in Figure 1. 2) *the semantic data card* is to make recommendations by answering user-initiated queries. As shown in Figure 2, several cards are implemented, including restaurant, population and area of individual administrative division, parking lot, temperature, national treasure, subway stations, cultural facility, etc. In particular, the Temperature and Parking lot cards deliver real-time information closed to a selected resource by delegating requests to a set of APIs from the Open Data Portal of Seoul city [7]. This approach would help users to get relevant information with their interests. It allows users to give their feedback for clarifying their intents. For example, it delivers better recommendations as results such as restaurants which can be served until 10 PM, restaurants located in Gate 2 of a subway station, the list of museums that have a specific national treasures, instead of simply providing overall information of a restaurant.

4 Conclusion

In this paper, we described a generalized approach to develop linked data by collecting and transforming open (raw) data, and addressed some features of the Semantic Data Fusion. This application focuses on delivering better recommendations of user-initiated requests using different kinds of data sources. Future work includes more publishing and interlinking of data sources, in order to bootstrap universal access across the Web. In addition, tools, which aim to publish and interlink data sources, will be developed.

References

1. Christian Bizer, Tom Heath, Tim Berners-Lee, Michael Hausenblas, and Sören Auer, editors. *Proceedings of the WWW2013 Workshop on Linked Data on the Web, Rio de Janeiro, Brazil, 14 May, 2013*, volume 996 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
2. Christos Koumenides, Harith Alani, Nigel Shadbolt, and Manuel Salvadores. Global integration of public sector information. In *Web Science Conference 2010*, March 2010. Event Dates: 26-27 April, 2010.

[7] `http://data.seoul.go.kr`

3. Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, and m.c. schraefel. Linked open government data: Lessons from data.gov.uk. *IEEE Intelligent Systems*, 27(3):16–24, 2012.
4. Barbara Ubaldi. Open government data: Towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance 22, OECD Publishing, 2013.

Annex: Challenge Criteria - Open Track Submission

Minimal requirements

- The application has to be an end-user application. The semantic Data Fusion provide intuitive interface for end-users to find out relation-based information.
- The information sources used should be under diverse ownership or control. The application integrates data from multiple sources of different ownerships.
- The information sources used should be heterogeneous. The data sources originally are in different formats, including CSV, HTML, JSON, or XLS, etc. The heterogeneity is resolved by transforming all data sources into RDF.
- The information sources should contain substantial quantities of real world data. All data used in the application are collected from government websites and the open data portal in Korea.
- Meaning must be represented using Semantic Web technologies. The whole data sets are represented in RDF. Each data set is described via own data model with additional vocabularies from Schema.org.
- Data must be processed in interesting ways to derive useful information.
- This semantic information processing has to play a central role. The application uses both RDF and SPARQL to query and answer user-initiated requests.

Additional Desirable Features

- The application provides an attractive and functional Web interface. The Semantic Data Fusion has an intuitive interface for end-users. The end-users do not need to know semantic web technologies to get any recommendations.
- The application should be scalable. Any data sources can be added and linked into existing data sources. We continue to integrate more data sources from different data sources into our application.
- Functionality is different from or goes beyond pure information retrieval. The search is based on keyword-to-concept mapping, and each card is implemented by deriving from diverse sources using subgraph queries.
- The application has clear commercial potential and/or large existing user base. The broad coverage of the proposed approach allows it to be used by mobile services.
- Contextual information is used for ratings or rankings. An instance of each card arranges via some contextual information such as time, location, etc.
- The results should be as accurate as possible (e.g. ranking of results according to context). Not applicable.