Information Integration and Utilization Technology using Linked Data

● Nobuyuki Igata ● Fumihito Nishino ● Terunobu Kume ● Takahide Matsutsuka

Fujitsu Laboratories is conducting R&D on technology to integrate and utilize data by using Linked Data, which is a standard methodology for publishing data on the Web that is being promoted by the World Wide Web Consortium (W3C). W3C is the main international standards organization for Web technologies. Linked Data uses a machine-readable structured data format that can be processed mechanically. Much information that has public value has recently been published in Linked Data format, and a global data space on the Web called Linked Open Data (LOD) has been created. This paper introduces an LOD utilization platform technology that gathers and stores data published around the world in Linked Data format and that provides a centralized search function. It was developed by the authors in collaboration with the Digital Enterprise Research Institute at the National University of Ireland, Galway.

1. Introduction

Activity related to Open Data has been on the increase recently. Many authors and other creators of data are publishing it in unrestricted form, making it available to anyone. In 2009, the U.S. government established the data.gov[1)] site as a centralized, public site for public data held by the government and other public institutions. As of May 2013, over 70 000 data sets from over 200 public institutions within the U.S. were published on the site. The wave of Open Data activity is spreading throughout the world, and governments in over 40 countries have established dedicated data-publishing sites.[2)] In 2012, the IT Strategic Headquarters of the Cabinet Office in Japan formulated the "e-Government Open Data Strategy," completing the legal groundwork for Open Data and reaching a stage where public data can gradually start to be published.

One reason governments around the world are undertaking Open Data initiatives is to ensure transparency in government activities, and other goals are to encourage secondary use of public data and to create new markets. The market for applications and services utilizing public data at research facilities in the EU has reached 28 billion euros, and the economic ripple effect of this, including increased efficiency of service users and strengthened industrial competitiveness, is estimated to be 140 billion euros. In Japan yen, this is a 1 trillion yen market with ripple effects estimated at 5.4 trillion yen.[3)]

The simple term "public data" includes various types of data, from national census data and other statistical data, to aerial photographs and other image data. This means that the data sharing sites described above need to use various formats, such as Excel, JPEG, and XML, depending on the type of data. Even if data is of the same type, the various organizations creating the data may format their data differently.

A mixture of data formats imposes the cost of converting formats on the users of the data, and this can be an obstacle to secondary use. For this reason, published data should not depend on a particular application and should be in a format that is machine processable.

For these reasons, the World Wide Web Consortium (W3C), an organization responsible for various Web standards, has recommended a methodology for publishing data on the Web called "Linked Data."[4)]

This paper gives an overview of Linked Data with examples of applications that use them. It then introduces a technology we developed in collaboration with

the Digital Enterprise Research Institute at the National University of Ireland, Galway, for storage and centralized search of data published in Linked Data format.

2. Linked Data and Linked Open Data

2.1 Linked Data

Heath and Bizer[5] introduced Linked Data as a structure for Evolving the Web into a global data space. The Web in its current form is mainly intended to be read by people, but Linked Data is a web of data intended for machine processing. The basic technical elements of Linked Data are the Hypertext Transfer Protocol (HTTP) and Uniform Resource Identifier (URI), with the same structure as for the current Web. However, there are also several differences, such as the Resource Description Framework (RDF), a structured format for data description. Also, URIs are assigned to "entities" rather than "documents," and the entity (URI) data structure is expressed in terms of attribute names and values. For example, the profile (name and gender) of person A would be expressed as the triple <URI>, <Attribute name>, and <Attribute value>:

<URI>	<Attribute name>	<Attribute value>
uri: person-A	name	"Alice"
uri: person-A	gender	"female"

Attribute values are not restricted to text and can also describe other URIs. For example, the fact that person A and person B have a friend relationship could be described using two triples:

<URI>	<Attribute name>	<Attribute value>
uri: person-A	friend	uri: person-B
uri: person-B	name	"Bob"

Creating relationships between URIs using attribute names in this way gives the data a linked structure. The links need not be restricted to objects created by the same author and can link with different types of data created by others. This enables data described with Linked Data to overcome differences in data type and create a huge network.

2.2 Linked Open Data

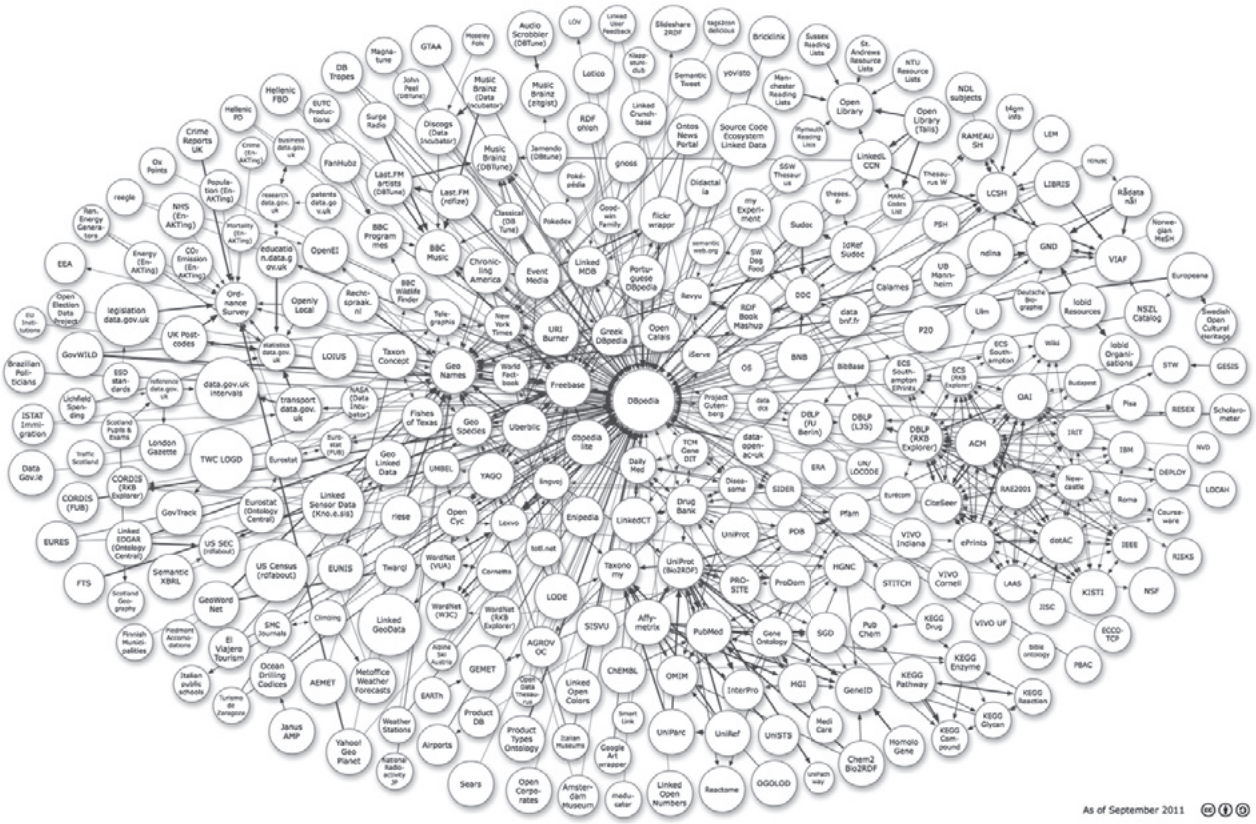
Data in the Linked Data format described above is now being published on the Web on a daily basis, and this data set is becoming a global data space on the Web that is being called "Linked Open Data (LOD)," merging the terms "Linked Data" and "Open Data."

Figure 1 is a diagram representing an overall view of the LOD published on the Web as of September 2011.[6] Each node in the diagram represents a single data set, and the arrows between nodes represent link relationships between data sets. Data sets from a wide range of fields have been published, including government, geography, cross-domain (common vocabulary), life sciences, academia, library, social media, and mass media. Examples include "DBpedia,"[7] a version of the Wikipedia Internet encyclopedia in Linked Data format, and "Data.gov.uk,"[8] published by the U.K. government. "Web NDL Authorities"[9] is a data set from Japan, provided by the National Diet Library. As of March, 2013, LOD comprised over 40 billion triples overall.

3. Examples of Linked Data utilization

Linked Data can be utilized in a variety of ways, and one typical way is management of content or knowledge. Documents, people, technical concepts (key words), events and other "entities" are managed as data. Various types of search and information provision are made available by attaching associations (links) among these entities. For example, we have created Linked Data for use within Fujitsu by attaching associations to data such as technical concepts, people, trade show information, press releases, patents, papers, technical keywords, and customer installation examples, and we are using it for knowledge management.[10] This makes it easier to find technologies, related people, examples of customer installations, and other information from technical keywords.

The Institute of Electronics, Information, and Communication Engineers (IEICE) provides a similar service with its I-Scover system.[11] The IEICE maintains information on journals, research reports, national convention papers, international conference proceedings, and other materials, and separate search systems were built for each of these data sets. It was not possible, for example, to perform searches spanning the data or to find papers related to a given paper. They thus created metadata in Linked Data form for information regarding books, periodicals, people, technical concepts, events, and facilities and built a system to search for related data and display it easily.[12] The Linked Data provided by IEICE is being used as a hub in the information and communications field, and, with increasing usage in the future, it will increase the presence of the Institute and is expected to contribute to increasing



Source: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/ [6]

Figure 1
LOD cloud diagram.

membership, improving scholarship, and strengthening industry.

4. LOD utilization platform

Currently, each LOD data set is provided on a separate Web site set up by the data provider. The desired data sets must be retrieved individually, which results in two problems in particular.

1) Aspects such as which public sites hold the desired data are unknown.
2) In some cases, data files are simply placed online, and the contents of a file cannot be accessed without actually downloading the file.

This is fine if the user has already identified the desired data set, but there is no way to search for a data set that is suitable for a given purpose. Possible ways to address these problems include enabling search for data sets according to purpose and enabling related data to be accessed without requiring the data to be downloaded.

As such, we developed, in cooperation with the Digital Enterprise Research Institute at the National University of Ireland, Galway, an LOD utilization platform (**Figure 2**) that collects the LOD published around the world, stores it, and enables centralized search of multiple data sets. The search interface enables application developers to find the data sets they need and supports the SPARQL[13] standard API, which was designed to handle data queries from applications.

With this LOD utilization platform, application developers do not need to search multiple data publishing sites individually and download data files separately. Instead, they can search and retrieve the desired data in a centralized manner. The standard API makes it much easier to develop applications that combine a wide variety of data.

The huge network created by the links must be handled when gathering this data centrally, and beyond simply increasing the amount of data, technology is needed to search the increasingly complex link

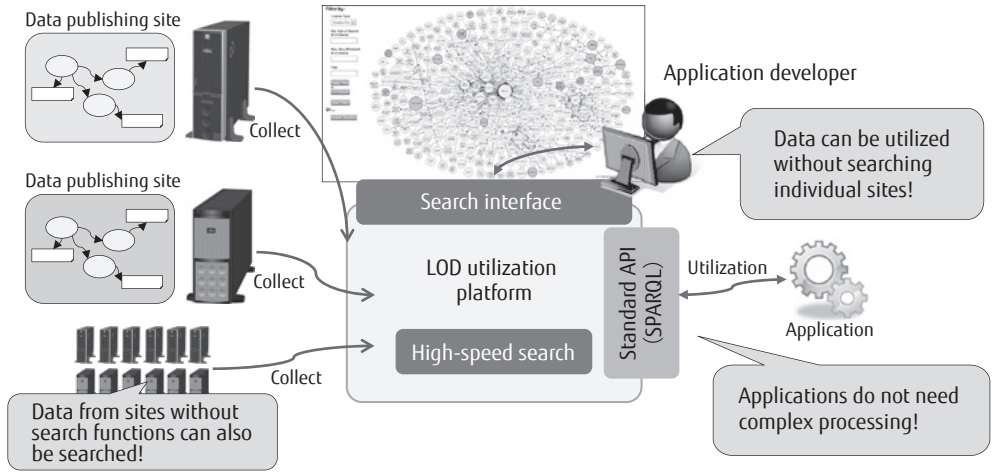


Figure 2
LOD utilization platform.

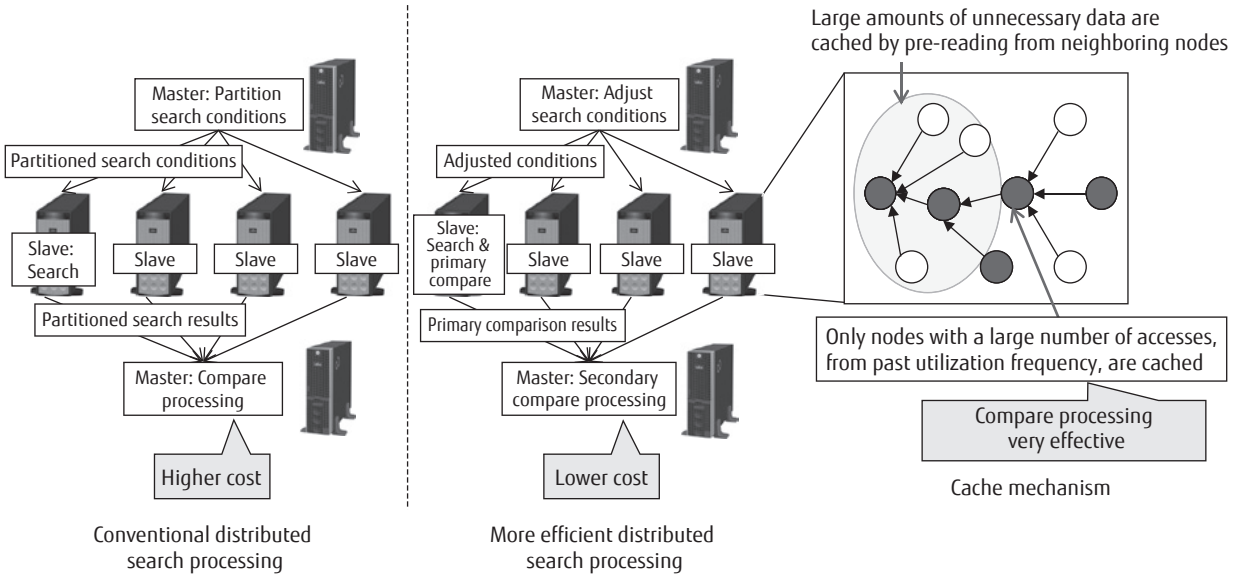


Figure 3
Search algorithm comparison.

structure at high speed. In particular, huge amounts of data must be processed through brute force (comparison processing) when searching in the data for common items with links between them, and this can degrade performance.

To meet this need for comparison processing, we combined more efficient distributed processing with a cache mechanism specialized for Linked Data, achieving performance five to ten times that of conventional processing.

An overview of the search algorithm developed is shown in **Figure 3**. The search conditions are adjusted, and a partial primary comparison process is performed on each slave server. This reduces the load of the secondary comparison process on the master server and thereby reduces the overall processing time. During the comparison processing, a cache mechanism caches only the frequently accessed data as determined

from the characteristics of the LOD link structure and past utilization frequency. The resulting reduction in the number of disk accesses results in higher search speeds.

5. Integration and utilization of LOD and public information

As a concrete example of an application using the LOD utilization platform, we prototyped an application that uses the Linked Data mechanism to combine LOD and public data for use in comparing and analyzing enterprises from various viewpoints. Specifically, basic profiles of enterprises obtained from data sets in LOD, such as DBpeida[7] and CrunchBase,[14] are integrated and analyzed with other public data, such as corporate financial reports and stock prices.

Normally these data sets use different enterprise codes. For example, in the U.S., a Central Index Key (CIK)[15] is used for financial reports, and a ticker symbol[16] is used for stock prices. We converted the public information into Linked Data and integrated (linked) it with LOD on the basis of the Legal Entity Identifier (LEI),[17] which is becoming standard in the financial field. This enables enterprises to be identified on the basis of their enterprise name or LEI, and a variety of public information available can be referenced by following successive links.

Figure 4 shows a screen shot of the prototyped enterprise comparison application. It shows actual companies in the U.S., compared on the basis of information from various data sets. The table at the lower right shows a comparison of various data values spanning different data sets, including financial indicators from financial reports, numbers of newspaper articles during a particular period of time, and stock information.

In this way, integration (linking) of different data sets can be expressed easily using Linked Data, and enterprises can be analyzed and compared from various aspects using a cross-section of this data from multiple data sets. This linking has made the task much easier.

6. Conclusion

The LOD utilization platform technology we introduced gathers and stores LOD published around the world and enables the data to be searched in a centralized manner. This technology has various potential applications.

We plan to open this LOD utilization platform, implemented in the cloud, free of charge to the public. Our goal is to accelerate the spread of Linked Data, with this LOD utilization platform as a core, and to invigorate the market for Open Data.

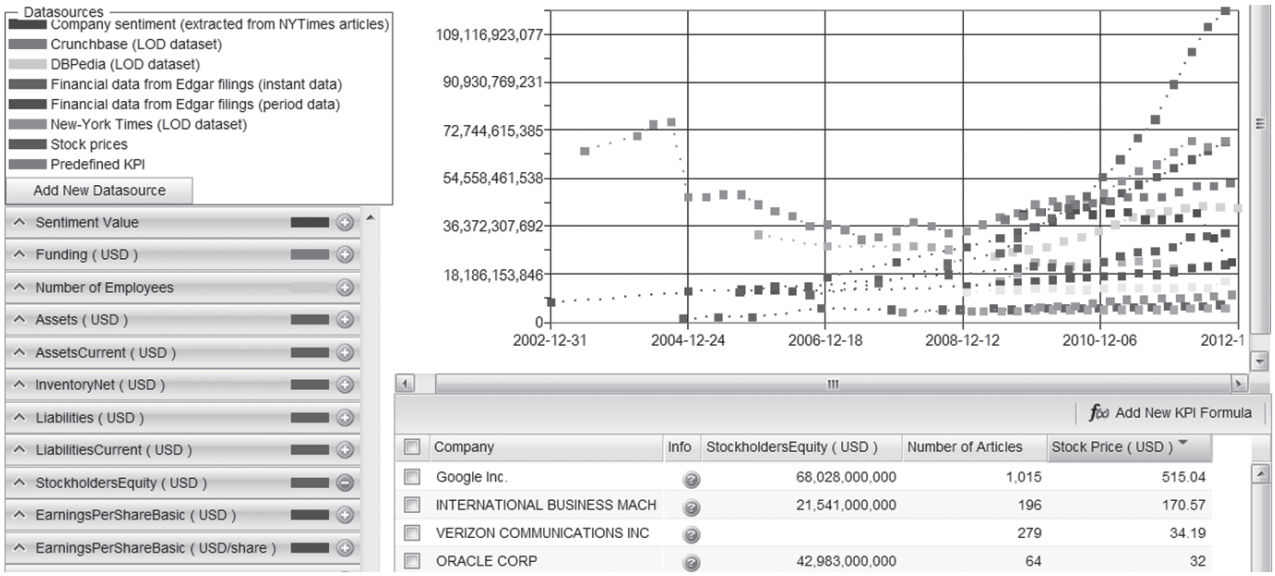


Figure 4
Screen shot of enterprise comparison application.

References

1) U.S. Government: Data.gov.
http://www.data.gov/
2) Open Data Site.
http://www.data.gov/opendatasites
3) NTT Data: Latest Trends in Open Data in Europe. 21st Electronics Policy Task Force materials (in Japanese).
http://www.kantei.go.jp/jp/singi/it2/denshigyousei/dai21/siryou1_2.pdf
4) Linked Data.
http://www.w3.org/DesignIssues/LinkedData.html
5) T. Heath et al.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers (2011).
6) R. Cyganiak and A. Jentzsch: Linking Open Data cloud diagram.
http://lod-cloud.net/
7) DBpedia.
http://dbpedia.org/
8) U.K. Government: DATA.GOV.UK.
http://data.gov.uk/linked-data
9) National Diet Library: Web NDL Authorities (in Japanese).
http://iss.ndl.go.jp/ndla/about/
10) F. Nishino et al.: Utilization of Linked Data in Enterprises. IPSJ SIG Notes, 2011-DD-82(2) (in Japanese).
11) I-Scover: IEICE Knowledge Discovery.
http://i-scover.ieice.org/?lang=en
12) F. Nishino: I-Scover: A System for Searching IEICE Documents Based on Linked Data. *IEICE B-Plus*, No. 25, pp. 49–53 (June 2013) (in Japanese).
13) SPARQL.
http://www.w3.org/TR/sparql11-query/
14) CrunchBase.
http://www.crunchbase.com/about
15) CIK: Central Index Key.
http://www.sec.gov/edgar/searchedgar/cik.htm
16) Ticker Symbol.
http://en.wikipedia.org/wiki/Ticker_symbol
17) FSB: A Global Legal Entity Identifier for Financial Markets.
http://www.financialstabilityboard.org/publications/r_120608.pdf

Nobuyuki Igata
Fujitsu Laboratories Ltd.
Mr. Igata is engaged in research on data integration and utilization platform technologies using Linked Data.

Terunobu Kume
Fujitsu Laboratories Ltd.
Mr. Kume is engaged mainly in research on knowledge processing and Linked Data utilization.

Fumihito Nishino
Fujitsu Laboratories Ltd.
Mr. Nishino is engaged in research on natural language processing, cognitive processing, and knowledge and content management systems utilizing Linked Data.

Takahide Matsutsuka
Fujitsu Laboratories of Europe Limited
Mr. Matsutsuka is engaged in development of large-scale distributed databases and processing platforms utilizing Linked Data.