Linked Data
—Linking data and expanding service—

Fumihito Nishino Social Innovation Laboratories, Fujitsu Laboratories Ltd.

1 Introduction —What is possible with Linked Data?

Linked Data has recently been attracting attention. Although the precise definition of Linked Data will be explained later, briefly, Linked Data is a method of creating a "web of data" by linking data that are structured on the basis of simple rules. Namely, various pieces of data are placed on websites as machine-understandable structured data, rather than as text that people read, and such data are linked by semantic links. Consequently, users can retrieve information efficiently by specifying the meaning or exploring related information. The retrieved information can be used as knowledge and enables the easy development of new services.

For example, suppose you are trying to find information on a certain researcher, such as affiliation, contact address, and past research achievements. When you search for the information by entering the name of the researcher in the input field of a Web search engine, a list of pages (documents) that include the character string you entered is displayed as a search result. However, you cannot directly obtain the affiliation or contact address of the researcher (there is a possibility that pages including this information are obtained by chance).[*1] General search engines search for documents including the character string; therefore, pages that include the character string but are unrelated to the target researcher and pages that are related to the researcher but do not include his/her affiliation or contact address may be retrieved. If the information on researchers is structured according to attributes, such as affiliation, contact address, and past research achievements, on the basis of rules, and related data are linked to each other, you can find the necessary information of the target researcher by specifying a value for each facet. Namely, if data on particular entities (*e.g.*, people, organizations, documents, events) are structured and the related data are linked, users can efficiently retrieve related information.

Next, let us assume that you need to list up articles written by authors affiliated with financial companies. Whether an author's affiliated company deals with the finance business or not can be determined using Wikipedia and other websites. However, a manual search of all articles is extremely laborious. Searching Wikipedia for the category of business using the company name is also laborious even if this process is carried out using software, because the company names are not always written in a unified form. On the Internet, various types of information, such as on geography, music, and bibliography, as well as encyclopedias such as Wikipedia, are distributed. If such useful information is structured by a unified framework and we can refer to the information as if it were on a database, listing up articles written by authors affiliated with financial companies, as explained above, will be easy.

Furthermore, multiple information sources can be referred to by each other when they are mutually linked. Compounding data from multiple data sources will enable the development of new services (data mashup). For example, it will be easy to develop a service to display the location of a conference on a map using the data of the conference venue and map data.

Linked Data is expected to enable the new services described above to be developed. In the following sections, I will first explain the background and the mechanism of Linked Data, then introduce the Linking Open Data Project, which is aimed at making various data available in the form of Linked Data. Finally, I will introduce concrete examples of applications of Linked Data.

*1 In May 2012, Google (English version) implemented Knowledge Graph, which provides information similar to an encyclopedia regarding the keyword used in a query, causing this new Google to evolve into something distinct from conventional search engines.

2 Problems with current websites

Currently, there are vast amounts of information on websites. Such information is prepared to be read by people. For example, it is not a simple task to develop a program to extract the data on the time, location, and name of conferences from the page shown in Fig. 1 and register them on a schedule management tool. This is because current websites are typically designed under the premise that they are to be read by people (text), and hence, do not consist of data that allow for secondary uses such as editing, processing, and analyzing.

The measures taken after the Great East Japan Earthquake are a good example demonstrating the importance of releasing data that is easy to re-use. Various types of information were disseminated immediately after the Earthquake, but they are in machine-unreadable formats (data that is difficult to re-use) such as the portable document format (PDF) and graph format. Information centers operated by local municipalities and the Ministry of Economy, Trade and Industry requested information providers to release data in a machine-readable format that does not require specific software or system to access it.[1] Upon this request, Tokyo Electric Power Company released electric power supply data that is easy to re-use. As a result, a variety of original and ingenious data visualization software has been developed. This marks a milestone demonstrating that merely making data available on the web is not as meaningful as providing the data in a machine-readable format.

3 What is Linked Data?

Tim Berners-Lee, who is considered the father of the Web, defined the rules of releasing data on websites to increase the convenience of data. These rules make up Linked Data.[2] He outlined the following four principles.

1. Use URI(uniform resource identifier)s as names for things
2. Use HTTP(hypertext transfer protocol) URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL)
4. Include links to other URIs. so that users can discover more things.

URIs are used as a globally unambiguous identifier for properties and data. For example, Barack Obama, President of the United States, is assigned VIAF:52010985 on the Virtual International Authority File. Information on the target entities can be accessed via a general web architecture using the http (*e.g.*, http://viaf.org/viaf/52010985). The data obtained by dereferencing to the URI are machine understandable when expressed using the RDF(resource description framework) (a standard framework used for data description and exchange). RDF represents metadata as a collection of triples, each consisting of a subject, a predicate, and an object. A standard query language called SPARQL(SPARQL protocol and RDF query language) is also available. As an example, the network diagram of the RDF of this article is shown in Fig. 2.

In Fig. 2, ellipses indicate resources corresponding to objects such as events, people, and documents. The labels beside the arrows are predicates. The left resource represents a document, and includes the title and the date of publication. The predicate is, in fact, a URI. Predicates have been selected from widely spread vocabularies to unify the semantic description of data beyond the boundary of organizations. Such vocabularies include Dublin Core (for document metadata), vCard (for address books), Friend of a Friend (FOAF) (for information on people), iCalendar (for

August 7, 2012	IEICE will host EMC'14/Tokyo (to be held in May 2014 in Tokyo).
August 7, 2012	IEICE will cohost ICOIN 2013 (held in January 2013 in Thailand).
August 7, 2012	IEICE will cohost APNOMS 2012 (held in September 2012 in Korea).
August 7, 2012	IEICE will cohost the AsiaFI NV Workshop (held in August 2012 in Kyoto).
August 7, 2012	Information on international symposia was updated.

Fig. 1 Website written for people to read

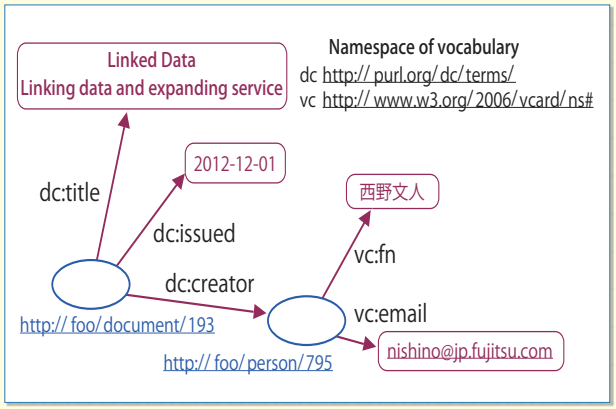

Fig. 2 Network diagram of RDF

★	make your stuff available on the Web (whatever format) under an open license[1]
★★	make it available as structured data (e.g., Excel instead of image scan of a table)[2]
★★★	use non-proprietary formats (e.g., CSV instead of Excel)[3]
★★★★	use URIs to denote things, so that people can point at your stuff[4]
★★★★★	link your data to other data to provide context[5]

Fig. 3 Rating of the level of Linked Data[(2)]

schedules), Geo (for location), and simple knowledge organization system (SKOS) (for knowledge system).

Note that Tim Berners-Lee proposed Linked Open Data 5 Star to rate the level of Linked Data (Fig. 3).

4 Linking Open Data Project

Linked Data can be accessed using a globally unambiguous URI via http, which is widely supported by standard web browsers, and can be shared by various users. In addition, data processing is standardized using the RDF, which is a standard semantic description. If data can be accessed beyond the boundary of fields, it will be easy to develop various types of service. In other words, Linked Data has the possibility to serve as a platform of knowledge.

Linked Data shows the technical interoperability of data and does not necessarily require the data to be open. However, the features of Linked Data will be fully utilized by opening data (in principle, users can freely use, reuse, and redistribute data). In 2007, advocating that data should be shared, not "hoarded", Tim Berners-Lee proposed to release various types of data through Linked Data. This was the beginning of the Linking Open Data Project. Figure 4 shows a Linking Open Data cloud.[(3)] The nodes are the linked open data (LOD) sets that have been released and registered. The arrows indicate reference relationships between Linked Data. In 2007, 12 sets of LOD were released. Since then, the number of LOD sets released has been increasing every year. As of September 2011, 295 LOD sets were released and registered (Fig. 5).[(4)] This figure indicates that the Linking Open Data Project has been accelerated worldwide.

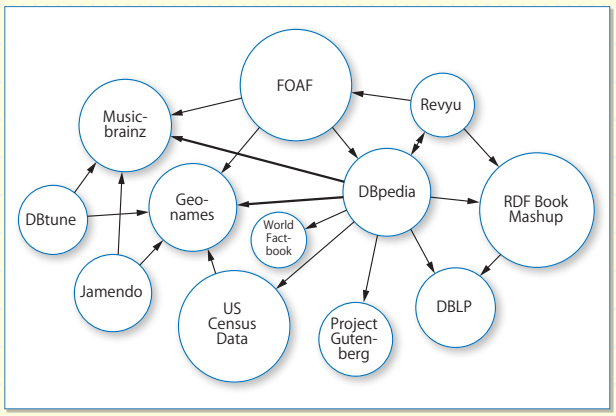

Fig. 4 Linking Open Data cloud diagram in May 2007[(3)]

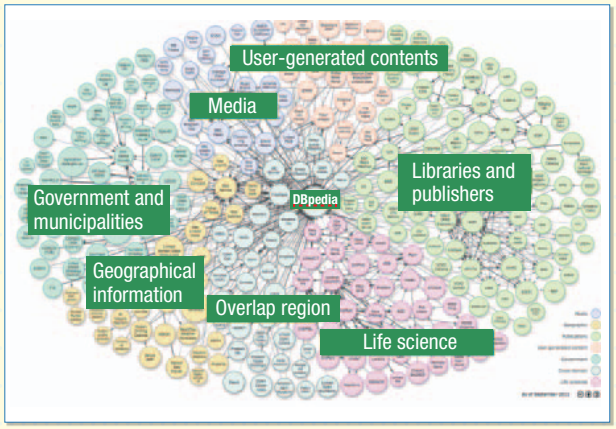

Fig. 5 Linking Open Data cloud diagram in 2011 (png file in ref. [(4)], revised by the author)

5 Concrete examples of LOD

In section 5, I will introduce several well-known examples of LOD.

(1) DBpedia (encyclopedia)[(5)]

Many of you have used Wikipedia. Wikipedia is an online encyclopedia designed to be read by people. In DBpedia, Wikipedia has been rewritten using the RDF so that it can be machine processed. However, not all information in Wikipedia has been converted; only semistructured information, such as infoboxes (tables in a prescribed format placed at the upper right of a Wikipedia article), external links, and affiliation categories, is converted to the RDF. For a company, data on the category of business, office location, and the number of employees are stored. For a person, data on occupation, date of birth, and his/her achievements are stored. DBpedia 3.9, released in September 2013 after approximately 13 months from the previous release, was based on Wikipedia from March / April 2013. The English version of the DBpedia knowledge base currently describes 4.0 million things, out of which 3.22 million are classified in a consistent Ontology, including 832,000 per-

```
SELECT ?company                          Select "?company" that satisfies the following conditions.
WHERE {                                  where
 ?company a                              "?company" is a company.
<http://dbpedia.org/ontology/
Company>.
 ?company                                The head office of "?company" is in Tokyo.
<http://dbpedia.org/property/
locationCity>
<http://dbpedia.org/resource/
Tokyo>.
 ?company
<http://dbpedia.org/property/
numEmployees> ?ne                        "?ne" is the number of employees.
FILTER (?ne >= 10000)                    "?ne" is 10,000 or more.
}
```

Fig. 6 Search expressions of SPARQL to list up companies with the head office in Tokyo and 10,000 or more employees

sons, 639,000 places (including 427,000 populated places), 372,000 creative works (including 116,000 music albums, 78,000 films and 18,500 video games), 209,000 organizations (including 49,000 companies and 45,000 educational institutions), 226,000 species and 5,600 diseases. The request to list up articles written by authors affiliated with financial companies as explained in section 1 is easily realized by a search using DBpedia. For example, if you need to list up companies with the head office in Tokyo and 10,000 or more employees, you should input the search expressions shown in Fig. 6 (this is written in a standard query language, SPARQL, which is not explained in detail in this article).

(2) GeoNames (geographical data)[6]

GeoNames includes over 10 million geographical names and related information including latitude, longitude, administrative district, altitude, time zone, and population.

(3) MusicBrainz (music data)[7]

MusicBrainz is an encyclopedia of music covering more than 820,000 artists and information on more than 15,000,000 pieces of music.

6 Examples of LOD in various fields and organizations

In this section, I will introduce examples of LOD in various fields and organizations.

(1) British Broadcasting Corporation (BBC)[8]

As many as 58 websites for television and radio broadcasting stations have been individually managed under BBC.[9] At that time, these websites were difficult to use because users could not find even artists using the name of a program as a search word. However, it was impossible to manually develop and manage websites for more than 1,000 programs a day. With the above background, BBC decided to make the contents available as LOD; ontological matching is carried out for all programs and these programs are linked to DBpedia. Now that the LOD on BBC is available, a user has access to abundant links to, for example, biographies, air dates on BBC, reviews, and related information of artists during a search. BBC says that various contents managed by BBC have been systematized and the management cost of its websites has been reduced. In addition, the BBC website has served as a kind of encyclopedia because of its increased usability, increasing the number of links from external websites, the visibility of the website, and the number of visits to the website (number of page views). BBC utilized the LOD technique in the London 2012 Olympics to provide information on athletes and game results.[10]

(2) The New York Times[11]

The New York Times endeavors to be a leader of innovation and needed a tool to manage its archives. In 2009, the New York Times decided to make the archive contents available as LOD.[12] In concrete terms, 1,000,000 words and terminology that have appeared in The New York Times over the past 150 years were classified according to the titles of articles, names of people, names of organizations, geographical names, and titles of pieces of work and productions, and released. In addition, their data were linked to DBpedia and Freebase[13] (a kind of encyclopedia) to form a base of new revenue such as revenue from online advertising. Thus far, descriptions on approximately 10,000 headings, including 5,000 people and 1,500 organizations, are available on LOD. The New York Times has also launched beta620,[14] a collection of projects in which anyone can propose new ideas and products or work collaboratively. Among these projects, Longitude[15] provides a mapping service whereby an article is mapped to a location on a map.

(3) Mass media in general

In January 2010, The Guardian, BBC, and the Media Standards Trust sponsored the News Linked Data Summit in London and invited those in the news industry.[16] At the summit, the possibility of Linked Data was discussed. It was confirmed that both editors and consumers can create more beneficial contents by linking data. Martin Moor compiled 10 reasons why news organizations should use "Linked Data".[17] In his article, he stated, 1) the management and development of the contents of your company be-

come more efficient, and services that cannot be provided using only the company's contents can be realized using Linked Data; 2) applications that link your company's contents to external Linked Data can be provided as the services of your company; 3) bidirectional services incorporating real-world information (*e.g.*, contents, users, developers) in the services of your company can be realized.

(4) Health care industry

Linked Data has also been actively used in the health care industry. Conventionally, information on pharmaceuticals, examples of side effects, diseases, clinical tests, proteins, genes, and medical insurance systems is not linked to each other. In addition, the terms are formatted and spelled in several different ways. In the LODD(Linking Open Drug Data) Project,[18] these data were made available as LOD to facilitate reference to the drugs used in clinical tests, details of diseases, and the relationships between drugs and diseases. In addition, end users can search for and access visual representations of these data using application software called TripleMap,[19] which helps to facilitate the research and development of pharmaceuticals and the reduction in the associated risks.[20]

(5) Government

Many countries all over the world are promoting the disclosure of data owned by governments and public institutions. The United States[21] and the United Kingdom[22] have been releasing these data as LOD, aiming not only to improve the transparency of data through data disclosure, but also to improve the efficiency of data use and stimulate the economy through the use of these data by public sectors. In Europe and the United States, contests for the development of application software using Open Data and for seeking new ideas for the use of Linked Data have been held. Consequently, various application software as well as data have been released. One such application software is Research Funding Explorer,[23] which visualizes the relationship between research funding and patents.

7 Situation in Japan

As seen in the Linking Open Data cloud diagram in 2011 (Fig. 5), only National Diet Library (NDL) Subjects[24] developed by the NDL were registered. However, interest in Linked Data has been increasing in Japan recently. Various projects for Linked Data are now under way, including CiNii[25] developed by the National Institute of Informatics, Kaken,[26] LOD in the field of museums by Linked Open Data for Academia (LODAC,[27] a project for the publication and sharing of academic information in Japan using LOD), Yokohama LOD Project[28] for local art and culture information under the direction of local communities and residents, and Integrated Database Project[29] managed by the Database Center for Life Science. In addition, Linked Open Data Challenge Japan has been held to provide opportunities to make presentations for those who are tackling the development of LOD schemes and preparation of data. Furthermore, special issues on Linked Data were published in March 2011 in the Journal of the Information Processing Society of Japan and in March 2012 in the Journal of the Japanese Society for Artificial Intelligence. As explained above, interest in Linked Data has been rapidly increasing in Japan.

8 Final remarks —toward knowledge infrastructure

Makoto Nagao, then Chief Librarian of the NDL, pointed out in 2010 that research and development in Japan lacks the perspective of incorporating individual systems into a large system. He also proposed that an overall integrated knowledge system (knowledge infrastructure) should be developed by systematically organizing academic information and knowledge contents, interrelating knowledge beyond the boundaries of fields, and centrally controlling the contents distributed all over Japan.[30] Nowadays, throughout the world, public sectors release data as a matter of course, and even private companies actively release data to differentiate their services on the basis of the analyzability of data rather than merely the availability of data. The inherent nature of Linked Data is that data are independent of individual tools and are distributed. In the future, a more advanced utilization of information will be realized by releasing various existing data using the standard access mechanism, Linked Data, and making the data owned by individuals available as Linked Data to link with other data throughout the world. I hope you will help to make data available as Linked Data.

References

(1) https://www.lasdec.or.jp/cms/12,22060,84.html
(2) http://www.w3.org/DesignIssues/LinkedData.html
(3) http://richard.cyganiak.de/2007/10/lod/lod-datasets_2007-05-01.png
(4) http://upload.wikimedia.org/wikipedia/commons/3/34/LOD_Cloud_Diagram_as_of_September_2011.png
(5) http://dbpedia.org/

(6) http://www.geonames.org/
(7) http://musicbrainz.org/
(8) http://www.bbc.co.uk/
(9) http://www.slideshare.net/metade/linked-data-on-the-bbc
(10) http://www.bbc.co.uk/sport/0/olympics/2012/
(11) http://data.nytimes.com/
(12) http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud/
(13) http://www.freebase.com/
(14) http://beta620.nytimes.com/ "beta620" was named after the address of the New York Times building.
(15) http://beta620.nytimes.com/viewer/longitude/
(16) http://www.iptc.org/download/mirror/IPTCMirror201002.pdf
(17) http://www.pbs.org/idealab/2010/03/10-reasons-why-news-organizations-should-use-linked-data073.html
(18) http://www.w3.org/wiki/HCLSIG/LODD
(19) http://triplemap.com/
(20) http://www.jcheminf.com/content/3/1/19
(21) http://www.data.gov/
(22) http://data.gov.uk/
(23) http://data.gov.uk/apps/research-funding-explorer
(24) http://id.ndl.go.jp/auth/ndla
(25) http://ci.nii.ac.jp/
(26) http://kaken.nii.ac.jp/
(27) http://lod.ac/
(28) http://ocdi.jp/?q=yokohama_lod
(29) http://lifesciencedb.jp/
(30) http://www8.cao.go.jp/cstp/tyousakai/seisaku/haihu05/nagao.pdf

Fumihito Nishino
(regular member)

received his master's degree from the Graduate School of the Department of Computer Science, Tokyo Institute of Technology, in 1981. He joined Fujitsu Laboratories Ltd. in the same year. He has been engaged in research and development of natural language processing including machine translation and information retrieval. He is currently anexpert senior researcher of Social Innovation Laboratories, Fujitsu Laboratories Ltd.