Building DBpedia Japanese and Linked Data Cloud in Japanese

Fumihiro Kato[1,2], Hideaki Takeda[1,3], Seiji Koide[1,2], and Ikki Ohmukai[1,3]

[1] National Institute of Informatics, 2-1-2, Chiyoda-ku, Tokyo, Japan
[2] Research Organization of Information and Systems, Tokyo, Japan
[3] The Graduate University for Advanced Studies, Kanagawa, Japan
email: {fumi, takeda, koide, i2k}@nii.ac.jp

Abstract. Wikipedia is one of the most valuable language and ontological resources covering wide domains so that DBpedia, LOD based on Wikipedia, plays the important role in the LOD cloud by connecting various resources. DBpedia Japanese is the LOD created from Wikipedia Japanese just as DBpedia data sources in other languages like English and German. We here describe how the conversion could be carried out with the efforts to fit DBpedia software and show the results of the dataset. We also describe how the created DBpedia Japanese is used by other Linked Data and show the Linked Data Cloud in Japanese.

1 Introduction

Today, various datasets are interconnected to each other under the concept of Linked Data[1]. In particular cross-media data such as encyclopedia plays a hub to connect data in various fields. In order to promote Linked Data in Japan, we have started to offer DBpedia Japanese which is generated from Wikipedia Japanese since 2012. It is expected to be a hub for Japanese resources. It has links to DBpedia English [2] via cross-language links so that it is also exected to connect Japanese resources with other international resources.

We also created RDF version of Japanese WordNet[3][4] and connected it to DBpedia Japanese[5]. Although we have an electoric dictionary called EDR Electronic Dictionary[4] originally developed from the scratch [6]. It is still continued updating but not open-free. So the Japanese WordNet is the first free online dictionary and therefore its RDF version is also the first LOD resource for Japanese. Both DBpedia Japanese and the RDF version of WordNet Japanese are important as primary resources for Japanese.

In this paper, we focus on DBpedia Japanese. In Section 2, we describe how DBpedia Japanese has been created, in particular show language-related issues in using the software. Then we show the Linked Data Cloud in Japanese in Section 3. We describe what we have learnt during the process in section 4, and conclude the paper at Section 5.

[4] See `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html`

2 DBpedia Japanese

DBpedia Japanese is the DBpedia generated from the Japanese Wikipedia. It is an internationalization of DBpedia where all parts fo the software used to build it from Wikipedia are those developed for English DBpedia. Therefore most of the building process were done except some language-specific treatments like the configuration for information extraction, and ontology mapping.

DBpedia Japanese is built and maintained in the activity of LODAC Project[5] where various data resources such as museums and biology are published as Linked Data.

Currently it contains 79,423,068 triples including links to the Japanese WordNet and the Japanese Wikipedia Ontology[7][8], and the statistics for ontology mapping is shown in Table 2. Roughly speaking, the rate for ontology mapping is a half of English DBpedia. There is still room to improve. In the following sections, we describe what we have done to build DBpedia Japanese and the results we achieved.

2.1 Customization of DBpedia Information Extraction Framework

DBpedia Information Extraction Framework (DIEF) [6] is the package of the software to extract information from Wikipedia. It is basically applicable to Wikipedia sources in any languages but is needed to customize in order to adapt information written in the language[9]. In building DBpedia Japanese, we set up the following extraction modules for Japanese;

1. DisambiguationExtractor
2. HomepageExtractor
3. ImageExtractor
4. PersondataExtractor

Some modules are needed to customize for Japanese Wikipedia as follows.

DisambiguationExtractor We added the following line in the configuration file for DisambiguationExtractor for processing disambiguation pages.

```
  "ja" -> "(曖昧さ回避)"
```

HomepageExtractor This module is used to extract official web pages. We added three patterns to process for hyperlinks with the following strings.

```
propertyNamesMap
"ja" -> Set("homepage", "website", "ホームページ", "ウェブサイト", "Web
サイト", "Ｗｅｂサイト")
externalLinkSectionsMap
```

[5] http://lod.ac

[6] http://wiki.dbpedia.org/Documenation

```
"ja" -> "外部リンク"
officialMap
"ja" -> "公式"
```

They mean Web page, Outgoing page from Wikipedia, and Official respectively.

ImageExtractor This module is used to extract links to images. To avoid non re-usable images, we added the following line;

```
  "ja" -> """(?i)\{\{\s?(Non free|Non-free pubart)\s?\}\}""".r
```

PersondataExtractor This module is used to extract information on persons. We added "ja" in the supported languages and defined the following patterns.

1. Names of templates for personal information
2. "名前"(name)
3. "別名"(alias)
4. "概要"(abstract)
5. dates and places for birth and death

DBpedia Japanese uses IRIs to identify resources now so that there are no prolems to include Japanese characters. But there was the problem for coding since URI instead of IRI was used in Virtuoso before version 6.1.4.

2.2 Statistics of the conversion

The results of the conversion is shown in Table 1. It is the statistics by the first build of DBpedia Japanese on January, 2012. Wikipedia Japanese contains 1,558,754 fragments including article, templates, image descriptions, and primary meta-pages. We can conclude ca. 90% of the original data is converted into DBpedia. The bold figures indicate the improvement by the configuration for Japanese Wikipedia.

2.3 Ontology Mapping

In DBpedia, an ontology is created and shared among different languages. A class in the ontology typically corresponds to templates which are used to generate Infobox in Wikipedia. Properties in a class corresponds to parameters in templates. Ontology mapping in DBpedia is to create classes and their properties and to create mapping from templates in Wikipedia. Ontology mapping is not easy task indeed since it is required to understand both background knowledge in domains and structures in templates. In order to stimulate the ontology mapping activity, we hold two so-called "mapping party" where people gathered and created mapping together on August 2012 and March 2013 in which 10 and 25 people participated. The current results of mapping is shown in Table 2. The

Table 1. Statistics of the conversion

Type	No. w/o configuration	No. w configuration
label	1,409,191	1,409,191
geo	34,368	34,368
infobox_properties	7,664,573	**7,664,575**
infobox_properties_definitions	33,214	33,214
infobox_test	7,192,853	**7,192,855**
page_links	44,421,598	44,421,598
wikipedia_links	4,227,573	4,227,573
article_categories		**2,153**
disambiguation		**106,386**
homepages		**49,355**
personadata		**1,811**
images		**843,170**

numbers of mapped templates are still low in comparison with English DBpedia. One of the reasons is that we did not create new classes yet, rather added mapping from the existing classes to templates in Wikipedia Japanese. There are some original classes like “武士” or Samurai but we have not created such classes yet. We need more improvement on ontology mapping.

Table 2. Statistics for Ontology Mapping in DBpedia (October 10, 2013)

	Japanese		English	
rate of all templates in Wikipedia are mapped	4.67%	(81 of 1,733)	6.33%	(369 of 5826)
rate of all properties Wikipedia are mapped	2.47%	(1581 of 62,679)	3.47%	(6,169 of 177,599)
rate of all template occurrences Wikipedia are mapped	47.99%	(286,858 of 597,696)	82.24%	(2,2435,773) of 2,728,357
rate of all property occurrences Wikipedia are mapped	38.75%	(3,128,208 of 8,071,982)	54.95%	(27,283,343 of 49,654,072)

3 Linked Data Cloud in Japanese

In this section, we describe Linked Data resources in Japanese. Here Linked Data resources in Japanese means Linked Data resources where their important or significant part is data presented in Japanese or data is mainly about Japan. An example for the former is DBpedia Japanese and the latter is the statistics data about Japanese Industry. The distinction between them is sometimes important but they mostly overlapped in our case since Japanese is mostly used in Japan.

Linked Data resources in Japanese have been limited but have increased rapidly in a couple of years. As mentioned in Section 1, LODAC project initiated to create datasets Linked Data in Japanese. We create and maintain LODAC Location, LODAC Museum, LODAC SPECIES, and WorldNet-J, and connect them to each other.

Semantic Web Community in Japan started the contest to promote LOD, namely the LOD Challenge. LOD Challenge was hold twice in 2012 and 2013, and collected over 70 and 200 applications respectively. It should be noted that the category of LOD challenge includes Dataset section as well as Idea and Application sections. Through the LOD challenge, several Linked Data resources are published.

We collect major Linked Data resources in Japan and map them in a figure (see Figure 3 and 1). The criteria to include datasets in it is as follows;

1. providing more than 10,000 triples,
2. providing either derefernece, data dump or SPARQL Endpoint,
3. providing labels in Japanese, and
4. open license is recommended but not mandatory.

The last criteria may seem odd but we think currently that forcing clear open licenses is not suitable at the beginning of the emergent community of LOD in Japan since there are a lot of datasets which are widely used just like with open licenses but do not have licenses yet.

3.1 Overview of the Linked Data Cloud in Japanese

The Linked Data Cloud in Japanese is still small, i.e., only 21 datasets are included which is just one tenth of Global LOD cloud. The proportion of the cloud is similar. Major category is publication-related datasets. Then some datasets are from geographic, life science and government domains. The clear difference is that there are many datasets without open licenses. It is mainly because the original datasets from which Linked Data datasets are generated often lack licenses. Some look very open but there are no clear mentions for licenses yet.

We pick up some of the datasets that are important resources connected to DBpedia tightly.

3.2 Japanese WordNet and its Linking to DBpedia Japanese

The efforts for multilingual WordNet has been made worldwide based on the Princeton's English WorldNet so far. In 2008, the Japanese WordNet (WN-ja) was developed and released by the National Institute of Information and Communications Technology (NICT) in Japan [3][4]. Currently WN-ja is built using the structure of the English WorldNet so that the synsets in WN-ja is equivalent to those in English WordNet and only words in Japanese are added as translation of words in English. We used version (1.1) of the Japanese WordNet with 187,000 senses (word-synset pairs), 57,000 concepts (synsets) and 94,000 unique Japanese words. For up-to-date information on the Japanese WordNet see `nlpwww.nict.go.jp/wn-ja`.

There are some attempts to convert WorldNet into RDFs. In 2006, W3C published the Working Draft for the representation in RDF of WorldNet 2.0 [10], in which the OWL representation of WorldNet and an OWL schema for WorldNet were introduced. Then, we applied the proposal to WorldNet 2.1 with the two extended pointer properties for the new version, i.e., `instanceHypernymOf` and `instanceHyponymOf` in 2006 [11][12], and subsequently for WorldNet 3.0. Up to now, several attempts to represent Princeton's WorldNet in OWL were made along with updating the WorldNet. Whereas the team members of the W3C Working Draft had actually converted WorldNet 2.0 to OWL representation [13], and then from the viewpoint of Linked Open Data (LOD), de Melo and Weikum has made the word search web pages [14]. The latest WN-ja is built on Princeton's English WorldNet 3.0. Appropriate Japanese words are added and linked to synsets via wordsenses as usual in the WorldNet manner. Thus, according to the W3C proposal of RDF representation of WorldNet, we have made the conversion of WN-ja to OWL [15].

Since both WorldNet and Wikipedia are the most famous comprehensive languages resources, there are many studies how combining them can contribute to build better languages resources (e.g., Yago [16]). Currently we here link entities in both datasets literally, i.e., link entities which share the same strings since we want to provide the basic dataset for Japanese language resources. We pick up nouns from WorldNet-ja and map names of resources and properties.

Table 3 shows the statistics of linking data of WN-ja to DBpedia Japanese, and Table 4 shows the statistics of linking data of DBpedia Japanese to WN-ja. In this attempt of linking by words in WN-ja and resource and property names in DBpedia Japanese, we made the connection by literally exact matching. Therefore, the mapping is exactly one by one and inversely equivalent in this case.

Table 3. Number of Linked Data from WN-ja to DBpedia Japanese

DBpedia	# of linked	# of WN nouns	rate
resources	33,017	65,788	50.1%
properties	1,245	65,788	1.9%

Table 4. Number of Linked Data from DBpedia Japanese to WN-ja

DBpedia	# of linked	# of IRIs	rate
resources	33,017	1,456,158	2.3%
properties	1,245	16,020	7.8%

3.3 Japanese Wikipedia Ontology

Takahira Yamaguchi and his group.[7][8] created Ontology for Wikipedia, i.e., it categorizes articles in Wikipedia. It provides the class hierarchy with properties, and is built by combining different techniques to analyze texts of articles, infobox, and categories. Currently Japanese Wikipedia Ontology and DBpedia Japanese are connected via `owl:sameAs` but we are planning to integrate them just as DBpedia and Yago.

4 Lessons Learnt

We have worked to create DBpedia Japanese as LOD resources in Japanese which can be used by other LOD datasets. What we have learnt through the process are as follows;

1. The language-dependent software problems are been solved mostly, but some still exist. They are not serious but it takes time and techniques to solve them. In particular, some software and tools are not completely compatible with IRI or UTF-8, and some lack language tags when describing literals. Anyway it is important to share demands from local communities with the global community.
2. People are willing to link to our resources. Within a year and so, we have 13 datasets which link their resources to those in DBpedia Japanese. The problem is that it is difficult to find datasets linking to us. We collect the datasets by personal communication or by local academic meetings. We need some techniques to find them automatically.
3. Having the hub for linking is not just good to make LOD cloud but also good to connect people. We had several academic meetings and tutorials for LOD. DBpedia Japanese is always a good example to demonstrate LOD since it covers various topics including daily issues and they are of course written in Japanese. People were motivated to join the LOD community in order to connect their data to DBpedia Japanese.

5 Conclusion

Wikipedia isthe precious ontology resources covering wide domains so that DBpedia play an important role in the LOD cloud by connecting various LOD resources. Building DBpedia for an lanugage is not so difficult task since all parts of the software are ready to use except some minor issues. We strongly recommend people in other languages to build own DBpedia. Building DBpedia is not just adding an useful resource but stimulating people to connect their data. It is the first step to make the society to be connected by data.

References

1. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool (2011)
2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3) (2009) 154–165
3. Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.: Development of japanese wordnet. In: Sixth international conference on Language Resources and Evaluation (LREC 2008), Marrakech (2008)
4. Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., Kanzaki, K.: Enhancing the japanese wordnet. In: Proceedings of the 7th Workshop on Asian Language Resources. ALR7, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 1–8
5. Koide, S., Takeda, H.: Rdfization of japanese electronic dictionaries and lod. In: 2nd Workshop on Linked Data in Linguistics: Representing and linking lexicons, terminologies and other language data Pisa, Italy, 23rd September 2013. Collocated with GL2013. (2013)
6. Yokoi, T.: The edr electronic dictionary. **38**(11) (1995) 42–44
7. Morita, T., Sekimoto, Y., Tamagawa, S., Yamaguchi, T.: Building up a class hierarchy with properties from japanese wikipedia. In: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE Computer Society (2012) 514–521
8. Tamagawa, S., Sakurai, S., Tejima, T., Morita, T., Izumi, N., Yamaguchi, T.: Building up a class hierarchy with properties from japanese wikipedia. In: Proceedings of the The 2010 IEEE/WIC/ACM International Joint Conferences on Web Intelligence, IEEE Computer Society (2010) 279–286
9. Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., Metakides, G.: Internationalization of linked data: The case of the greek {DBpedia} edition. Web Semantics: Science, Services and Agents on the World Wide Web **15**(0) (2012) 51 – 61
10. van Assem, M., Gangemi, A., Schreiber, G.: Rdf/owl representation of wordnet. W3C Working Draft 19 June 2006 (2006) http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/.
11. Koide, S., Morita, T., Yamaguchi, T., Muljadi, H., Takeda, H.: Owl expressions on wordnet and edr. In: SIG for Semantic Web and Ontology, SIG-SWO-A601-03, The Japanese AI Society (2006) (In Japanese).
12. Koide, S., Morita, T., Yamaguchi, T., Muljadi, H., Takeda, H.: Rdf/owl representation of wordnet 2.1 and japanese edr electronic dictionary. In: ISWC2006, Poster. (2006)
13. Van Assem, M., Gangemi, A., Schreiber, G.: Conversion of wordnet to a standard rdf/owl representation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06), Genoa, Italy. (2006)
14. De Melo, G., Weikum, G.: Language as a foundation of the semantic web. In: Proceedings of the 7th International Semantic Web Conference (ISWC 2008). Volume 401. (2008)
15. Koide, S., Takeda, H., Ohmukai, I.: An lod approach toward wordnet japanization. In: SIG for Semantic Web and Ontology, SIG-SWO-A1103-05, The Japanese AI Society (2011) (In Japanese).

16. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web **6**(3) (2008) 203–217

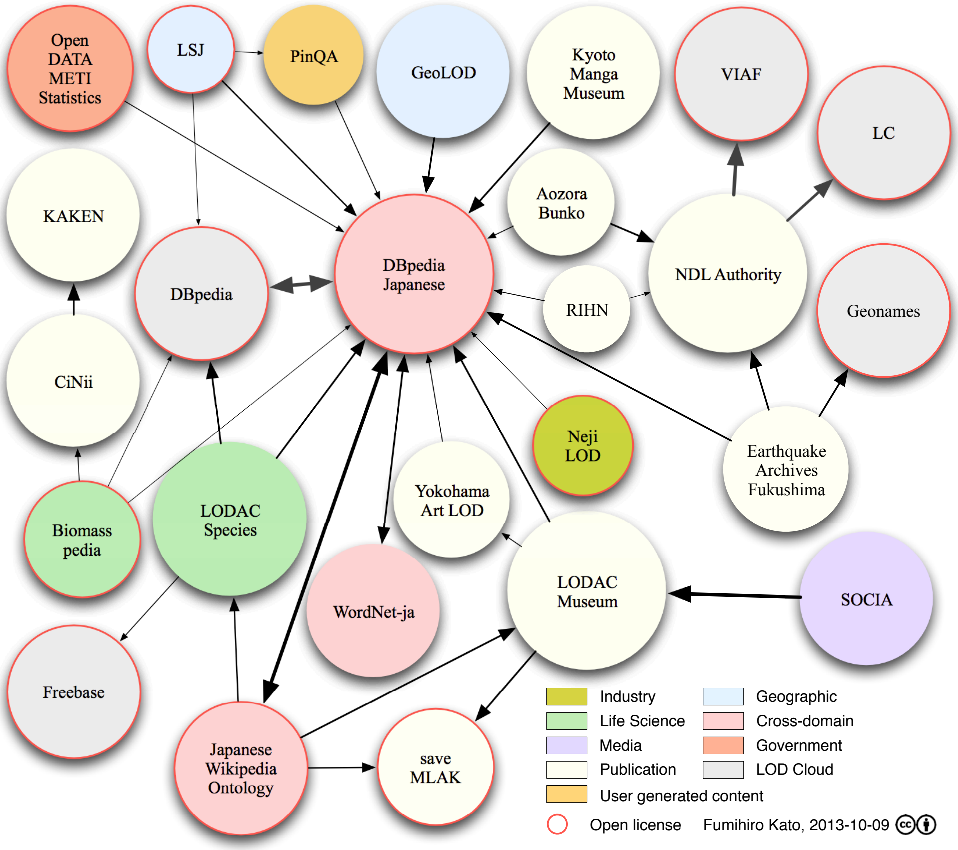
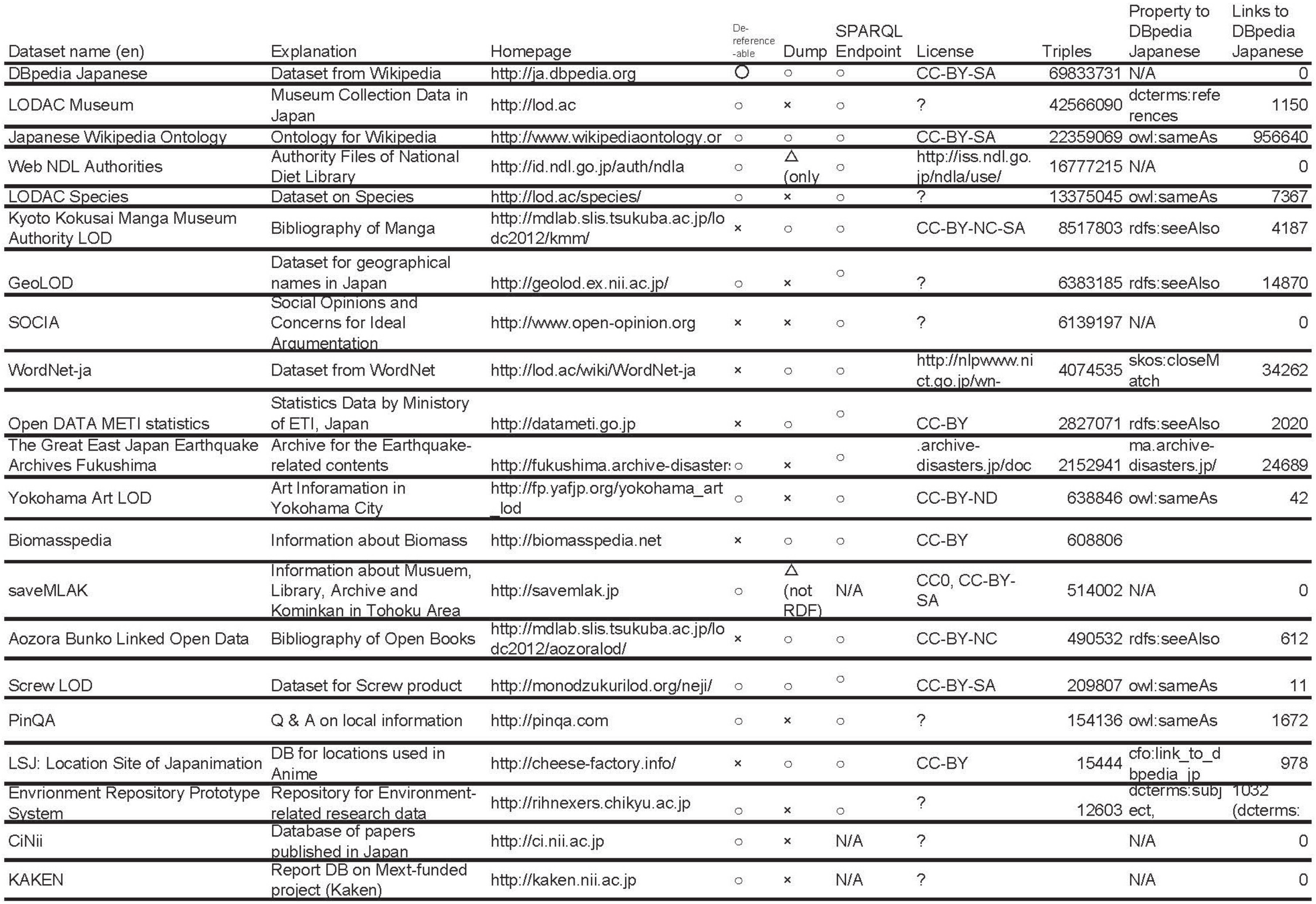


Fig. 1. Linked Data Cloud in Japan



Dataset name (en)	Explanation	Homepage	De-reference-able	Dump	SPARQL Endpoint	License	Triples	Property to DBpedia Japanese	Links to DBpedia Japanese
DBpedia Japanese	Dataset from Wikipedia	http://ja.dbpedia.org	○	○	○	CC-BY-SA	69833731	N/A	0
LODAC Museum	Museum Collection Data in Japan	http://lod.ac	○	×	○	?	42566090	dcterms:references	1150
Japanese Wikipedia Ontology	Ontology for Wikipedia	http://www.wikipediaontology.or	○	○	○	CC-BY-SA	22359069	owl:sameAs	956640
Web NDL Authorities	Authority Files of National Diet Library	http://id.ndl.go.jp/auth/ndla	○	△ (only	○	http://iss.ndl.go.jp/ndla/use/	16777215	N/A	0
LODAC Species	Dataset on Species	http://lod.ac/species/	○	×	○	?	13375045	owl:sameAs	7367
Kyoto Kokusai Manga Museum Authority LOD	Bibliography of Manga	http://mdlab.slis.tsukuba.ac.jp/lodc2012/kmm/	×	○	○	CC-BY-NC-SA	8517803	rdfs:seeAlso	4187
GeoLOD	Dataset for geographical names in Japan	http://geolod.ex.nii.ac.jp/	○	×	○	?	6383185	rdfs:seeAlso	14870
SOCIA	Social Opinions and Concerns for Ideal Argumentation	http://www.open-opinion.org	×	×	○	?	6139197	N/A	0
WordNet-ja	Dataset from WordNet	http://lod.ac/wiki/WordNet-ja	×	○	○	http://nlpwww.nict.go.jp/wn-	4074535	skos:closeMatch	34262
Open DATA METI statistics	Statistics Data by Ministry of ETI, Japan	http://datameti.go.jp	×	○	○	CC-BY	2827071	rdfs:seeAlso	2020
The Great East Japan Earthquake Archives Fukushima	Archive for the Earthquake-related contents	http://fukushima.archive-disaster:	○	×	○	.archive-disasters.jp/doc	2152941	ma.archive-disasters.jp/	24689
Yokohama Art LOD	Art Inforamation in Yokohama City	http://fp.yafjp.org/yokohama_art_lod	○	×	○	CC-BY-ND	638846	owl:sameAs	42
Biomasspedia	Information about Biomass	http://biomasspedia.net	×	○	○	CC-BY	608806		
saveMLAK	Information about Musuem, Library, Archive and Kominkan in Tohoku Area	http://savemlak.jp	○	△ (not RDF)	N/A	CC0, CC-BY-SA	514002	N/A	0
Aozora Bunko Linked Open Data	Bibliography of Open Books	http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/	×	○	○	CC-BY-NC	490532	rdfs:seeAlso	612
Screw LOD	Dataset for Screw product	http://monodzukurilod.org/neji/	○	○	○	CC-BY-SA	209807	owl:sameAs	11
PinQA	Q & A on local information	http://pinqa.com	○	×	○	?	154136	owl:sameAs	1672
LSJ: Location Site of Japanimation	DB for locations used in Anime	http://cheese-factory.info/	×	○	○	CC-BY	15444	cfo:link_to_dbpedia_jp	978
Envrionment Repository Prototype System	Repository for Environment-related research data	http://rihnexers.chikyu.ac.jp	○	×	○	?	12603	dcterms:subject,	1032 (dcterms:
CiNii	Database of papers published in Japan	http://ci.nii.ac.jp	○	×	N/A	?		N/A	0
KAKEN	Report DB on Mext-funded project (Kaken)	http://kaken.nii.ac.jp	○	×	N/A	?		N/A	0

Fig. 2. Linked Data resources in Japan