Building an RDFized Life Science Dictionary

Yasunori Yamamoto and Shoko Kawamoto

Database Center for Life Science, Bunkyo, Tokyo, Japan
{yayamamo,shoko}@dbcls.rois.ac.jp

Abstract. There is a growing need for efficient and integrated access to databases provided by diverse institutions. Using Resource Description Framework (RDF) to publish a dataset makes it more reusable. Furthermore, providing a dictionary to translate words into another language in RDF is useful when we want to access datasets across a language barrier. Here, we built an RDF version of the Life Science Dictionary (LSD). LSD consists of various lexical resources including English-Japanese / Japanese-English dictionaries and a thesaurus. Since we believe that LSD is a useful language resource in the life science domain to use multilingual data seamlessly, we assumed that its RDF version enables us to make LSD more reusable and therefore contributes to the life science research community.

Keywords. Multi-lingual language resource, Dictionary, RDF

1 Background

To link heterogeneous databases and provide users with access to them in an integrated manner, publishing datasets using Resource Description Framework (RDF) has increasing appeal to database developers and users. It enables us to access raw data using the World Wide Web approach such as Uniform Resource Identifier (URI) and Hypertext Transfer Protocol (HTTP). In addition, the number of non-English RDF datasets is increasing [1]. Therefore, there are growing needs of cross language RDF resources to link monolingual RDF data sets of different languages [1]. An example is DBpedia [2], which has made the contents of Wikipedia available in RDF. Wikipedia is the largest open, collaboratively developed encyclopedia project, but it is not necessarily reliable to use it as a translation dictionary in a specific domain. For example, Wikipedia has 149 pages in the category of "World Health Organization essential medicines" in English, but has only 56 in Japanese [3].

The Life Science Dictionary (LSD) [4] consists of some lexical resources including English-Japanese / Japanese-English dictionaries, and a thesaurus using the Medical Subject Headings (MeSH) vocabulary, the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. LSD has been edited and maintained by the LSD project since 1993. Project members are experts in the domain. To be used as a complement to DBpedia in the life science domain, we built an RDFized LSD.

2 Methods and Results

We used the latest version (Mar. 2011) of LSD that contains 110k English and 120k Japanese terms, which consists of several tab-delimited plain text files. We developed an ontology of LSD to express its schema in RDF using the Protege ontology editor [5]. There are ambiguous column names across the tables and the relationships among them are not clear. The second author knows LSD well, and therefore her knowledge was used to disambiguate them. As a result, the numbers of the classes and the properties we created are eight and 16, respectively. We also used some Simple Knowledge Organization System (SKOS) terms in addition to basic ones from the Web Ontology Language (OWL) vocabulary.

Using this ontology, we built the RDF version of LSD, which has about 5.6M triples. We loaded them into the OWLIM triple store. To improve the reliability, we iterated the development cycle: building an ontology, RDFizing LSD, loading them into the triple store, and evaluating it to see if there are any undesirable or unexpected semantic relationships among terms. For example, we verified that two semantically different terms are not connected by a SKOS predicate.

3 Conclusions

We built an RDFized LSD, which can be freely accessible from http://purl.jp/bio/10/lsd/sparql under the license of Creative Commons Attribution-NoDereivs 2.0 Generic (CC BY-ND 2.0). We are using this dictionary to provide our cross language search service. We hope it to be used widely to utilize multilingual resources in life science seamlessly.

Acknowledgements. We thank Dr. Shuji Kaneko for permitting us to release LSD under CC BY-ND 2.0. This work is funded by the Integrated Database Project, Ministry of Education, Culture, Sports, Science and Technology of Japan and National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

References

1. Gracia, J., et al.: Challenges for the multilingual Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. 11, 63–71 (2012)
2. Lehmann, J., et al.: DBpedia—a crystallization point for the web of data. Journal of Web Semantics, 7 (3), 154–165 (2009)
3. http://en.wikipedia.org/wiki/Category:World_Health_Organization_essential_medicines
4. Kawamoto, T.,et al.: Life Science Dictionary: statistical and collocational analyses of life science English. 20th IUBMB International Congress of Biochemistry and Molecular Biology and 11th FAOBMB Congress, Kyoto (2006)
5. Protégé project, http://protege.stanford.edu