1G4-OS-19a-2

ライフサイエンス辞書のリンクトデータ構築

Building Linked Open Data of the Life Science Dictionary

藤原 豊史 *1　　山本 泰智 *2
Toyofumi Fujiwara　　Yasunori Yamamoto

*1 株式会社インテック　*2 情報・システム研究機構　ライフサイエンス統合データベースセンター
INTEC Inc.　Database Center for Life Science, Research Organization of Information and Systems

Recent innovative progress of the experimental technologies in life science has produced large and diverse data, and many databases have been constructed. We are developing an environment where researchers can access these databases in an integrated manner by using Semantic Web technologies. In an effort to realize this environment, we built a Linked Open Dataset of the Life Science Dictionary (LSD) with links to DBpedia. A dictionary to translate words into another language in Resource Description Framework (RDF) is useful to cross a language barrier such as English and Japanese when we want to access datasets in multiple languages. LSD consists of various lexical resources including English-Japanese / Japanese-English dictionaries and a thesaurus using the MeSH vocabulary. The latest version of LSD contains 110 thousand English and 120 thousand Japanese terms. Our aim is to provide users with a useful language resource in the life science domain to process Japanese and English text data seamlessly, and linking LSD to DBpedia enables us to find related knowledge more easily. We report these works and usefulness of the linked data sets.

1. はじめに

大学共同利用機関法人 情報・システム研究機構 ライフサイエンス統合データベースセンターでは，生命科学分野でこれまでに蓄積された知見やプロジェクトの成果を研究者がより効率的に利活用できる環境の構築を行っている[Bono 2009]．その活動の一環として，生命科学分野の研究により生み出される多様かつ膨大なデータから必要な情報を効率的に得るために，ばらばらに構築されているデータベースを統合的に利用可能とする技術開発をセマンティックウェブ技術に着目して進めている[Yamaguchi 2011]．

リンクトオープンデータ*1（以下 LOD と呼ぶ）は，散在するデータセットに対して，それぞれに含まれる互いに同一の概念を表す要素同士を結ぶことでデータの共有を容易にするが，互いに異なる言語からなる RDF データセットに対して，同一概念同士を結ぶ場合には，対訳辞書が必要となる．現在，日本の国立国会図書館が国立国会図書館典拠データ*2 を RDF データで提供するなど英語以外の言語を含む RDF データセットも増加しており，対訳辞書情報を含む RDF データに対する需要が増加している[Gracia 2012]．

そこで我々は，生命科学分野の電子辞書であるライフサイエンス辞書（以下 LSD と呼ぶ）*3 の RDF データを構築し，LSD に含まれる語彙から対応する DBpedia の英語記事へのリンクセットを作成した．LSD は 1993 年から LSD プロジェクトにより構築・維持されており，プロジェクトのメンバーは生命科学分野の専門家である．LSD は，生命科学分野で用いられる英語および日本語の専門用語について，対訳関係を定義した日英対訳辞書を作成している．その他に，MeSH*4 から採用した統制語と日英対訳辞書とのすり合わせ作業を行い，日英シノニム（同義語）辞書を作成している．また MeSH ツリーを利用して統制語の上位概念・下位概念の関係を整理しており，更に PubMed*5 抄録中での統制語の共起頻度を収集し，統制語の共起関係を抽出している[Kaneko 2014]．

対訳辞書情報を含む RDF データとしては，DBpedia*6 を利用することも考えられる．DBpedia は多言語オンライン百科事典プロジェクトである Wikipedia*7 の内容を RDF データとして提供し，LOD クラウドで最も大きなハブとなっている．ある言語の記事について，それに対応する他言語の記事が紐付けられているので，日本語記事のタイトルと，それに対応付けられた英語記事のタイトルを日英対訳辞書として利用することができる．しかしながら，特定領域の対訳辞書として DBpedia を利用する場合，必ずしも有用とは言えない．例えば，執筆時点（2014 年 3 月 11 日）で Wikipedia の英語サイトにおける”World Health Organization essential medicines”カテゴリには 147 の記事が存在するが，一方，日本語サイトには 57 の記事しか存在しないことからも，利用できる日英対訳情報が限定されることがわかる．

我々は，作成した LSD リンクトデータが生命科学分野において日英対訳情報を含む RDF データとして利用され，また DBpedia を補完するものとして利用させることを目的としている．構築した LSD リンクトデータおよび Key Collision 法[Google 2013]等の文字列比較アルゴリズムを用いて DBpedia の各記事のタイトルと対応付けた結果と，そのデータセットの有用性について報告する．

2. LSD リンクトデータ構築

2.1 LSD および DBpedia データ概要

108,213 の英語の見出し語と 121,851 の日本語の見出し語からなる LSD の最新バージョン（Mar. 2013）と，8,826,375 トリプルからなる DBpedia version 3.7（labels_en.nt.bz2）を利用した．

2.2 LSD と DBpedia とのリンク作成

連絡先：株式会社インテック
〒136-8637 東京都江東区新砂 1-3-3
E-mail: fujiwara_toyofumi@intec.co.jp

*1 http://www.w3.org/standards/semanticweb/data
*2 http://id.ndl.go.jp/auth/ndla
*3 http://lsd.pharm.kyoto-u.ac.jp/ja/about/about_lsd/index.html
*4 http://www.ncbi.nlm.nih.gov/mesh
*5 http://www.ncbi.nlm.nih.gov/pubmed
*6 http://dbpedia.org/
*7 http://www.wikipedia.org/

英語の見出し語の表記と DBpeida の英語記事タイトルの文字列を，完全一致やコサイン類似度[Cohen 2003,Okazaki 2010]等を利用して比較し，リンクを作成した．各 DBpedia の英語記事タイトルに対して，全ての表記を完全一致，Fingerprint Key Collision（FKC），bi-gram FKC，tri-gram FKC およびコサイン類似度の順で対応するものを探索し，対応するものが見つかった時点で次のステップに移行する．これは，例えば bi-gram FKC で対応するものが見つかった場合は，それ以降の tri-gram FKC およびコサイン類似度は利用しないことを意味する．コサイン類似度での対応については，広く対応する見出し語を探索するために，類似度の閾値を 70%と設定した．一方，interleukin-1 と interleukin-2 などの望まない対応を除くためのルールをいくつか設定し，それらを除去した．対応付けられた 60,348 の英語の見出し語と 81,065 の DBpedia の英語記事は，skos:exactMatch [Miles 2009] を用いて両者のリンク関係を定義した（トリプル数：81,065）．

60,348 の英語の見出し語に DBpedia とのリンクが付加された一方で，DBpedia 英語記事と対応付けられている日本語 DBpedia の記事は 390,994 あり，そのうち英語の見出し語とリンク付けられた DBpedia 英語記事は 9,816 である．つまりこれは，生命科学分野において DBpedia を日英対訳辞書として利用するには，情報が限定されていることを意味している．

2.3 LSD リンクトデータの構築

(1) 英語および日本語の見出し語を表す RDF データ

図 1 は英語の見出し語を表す RDF データである．英語の見出し語は lsd:EnglishCode クラスに属し，表記を rdfs:label プロパティで表現する．英語の見出し語と対応付けられた MeSH の統制語を lsd:MeSHUniqueID プロパティで指定し，類義語となる LSD 内の英語の見出し語を skos:closeMatch プロパティで指定する．また，DBpedia リソースへのリンクを skos:exactMatch プロパティで指定する．

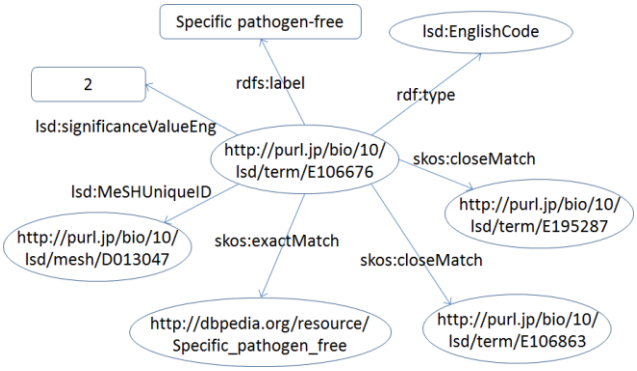

図 1. 英語の見出し語を表す RDF データ

(2) 見出し語の対訳関係を表す RDF データ

図 2 は英語の見出し語と日本語の見出し語との対訳関係を表す RDF データである．ある見出し語について，対訳が複数存在する場合や対訳となる見出し語についての付加情報（品詞など）も持つ場合があるため，対訳となる見出し語間に対訳と付加情報をもつ lsd:EnglishEntry クラスおよび lsd:JapaneseEntry クラスに属する概念を設定し，それらは見出し語から lsd:hasEntry プロパティで指定される．lsd:EnglishEntry クラスに属する概念から日本語訳となる見出し語を lsd:hasJapaneseTranslationOf プロパティおよび skos:related プロパティで指定する．また，lsd:JapaneseEntry クラスに属する概念から英語訳となる見出し語を lsd:hasEnglishTranslationOf プロパティおよび skos:related プロパティで指定する．

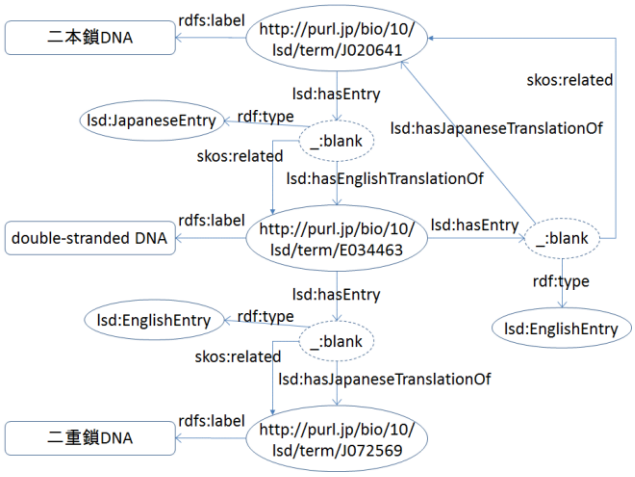

図 2. 見出し語の対訳関係を表す RDF データ

(3) MeSH 統制語の共起関係を表す RDF データ

図 3 は PubMed 抄録中での MeSH 統制語の共起頻度を表す RDF データである．共起関係を表す概念は lsd:CooccurrenceCode クラスに属し，PubMed 抄録中での tf-idf を lsd:TfIdfvalue プロパティで指定し，共起関係にある MeSH 統制語を skos:member プロパティで指定する．また，MeSH 統制語の日本語表記および英語表記を rdfs:label プロパティで表現する．

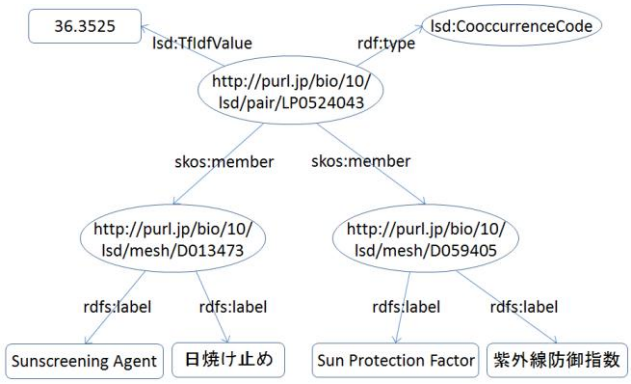

図 3. MeSH 統制語の共起関係を表す RDF データ

3. LSD リンクトデータの有用性について

LSD リンクトデータおよび DBpedia とのリンクセットの有用性を示すため，その利用例を述べる．

日本語の見出し語”片頭痛”に割り当てられている MeSH を取得し，その MeSH が割り当てられている全ての日本語の見出し語とその類義語について，対訳となる英語の見出し語と DBpedia へのリンクを取得する．この情報を得るための SPARQL クエリーを Appendix 1 に示す．その結果，15 件の英語の見出し語を取得し，うち 4 件について DBpedia へのリンクを取得した．取得した英語の見出し語は，”片頭痛”に関連した文献情報の検索などに利用でき，DBpedia へのリンクは”片頭痛”に関連した情報を DBpedia から取得する際に役立つ．

(abbr. rdf) http://www.w3.org/1999/02/22-rdf-syntax-ns#
(abbr. rdfs) http://www.w3.org/2000/01/rdf-schema#
(abbr. lsd) http://purl.jp/bio/10/lsd/ontology/201209#
(abbr. skos) http://www.w3.org/2004/02/skos/core#

MeSH 統制語”片頭痛”と共起する MeSH 統制語について，共起頻度上位 10 件の統制語を取得する．この情報を得るための SPARQL クエリーを Appendix 2 に示す．取得した MeSH 統制語は上記と同様に，”片頭痛”に関連した文献情報の検索などに利用できる．

4. おわりに

今回，我々は LSD リンクトデータとともにそこで利用するオントロジー（http://purl.jp/bio/10/lsd/ontology/201209#）を作成した．これが生命科学分野で，多言語リソースを円滑に利用するために広く使われることを望んでいる．データは CC BY-ND 3.0 のもとに，またオントロジーは CC0 1.0 のもとに SPARQL エンドポイント（http://purl.jp/bio/10/lsd/sparql）から利用および取得が可能である．

謝辞

金子周司博士には LSD を CC BY-ND 3.0 でリリースすることを許可していただいた．ここに感謝の意を表したい．また，本究は文部科学省委託研究開発事業「統合データベースプロジェクト」の助成による．

参考文献

[Bono 2009] 坊農秀雅: ライフサイエンス統合データベースセンターと統合データベースプロジェクト, 情報の科学と技術, Vol. 59, No. 4, pp. 165-169, (2009)

[Yamaguchi 2011] 山口敦子, 片山俊明: データベースを統合利用するための基盤としてのセマンティックウェブ技術, 我が国のデータベース構築・統合政略, National Bioscience Database Center , http://events.biosciencedbc.jp/article/02

[Gracia 2012] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. Challenges for the multilingual Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, 11, 63-71, (2012)

[Kaneko 2010] 金子周司, 藤田信之, 鵜川義弘: 生命科学知識の連想検索における提示語の最適化, 言語処理学会 第16回年次大会 発表論文集, (2010)

[Google 2013] Key Collision Methods, https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

[Cohen 2003] Cohen, W., Ravikumar, P., & Fienberg, S.: A comparison of string metrics for matching names and records. In KDD Workshop on Data Cleaning and Object Consolidation, Vol. 3, 73-78 (2003)

[Okazaki 2010] Okazaki, N., Tsujii, J.: Simple and efficient algorithm for approximate dictionary matching. In Proceedings of the 23rd International Conference on Computational Linguistics, 851-859 (2010)

[Miles 2009] Miles ,A., Bechhofer, S.: SKOS-Simple Knowledge Organization System Reference, W3C Recommendation (=2009)

Appendix 1

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX icd10: <http://purl.jp/bio/10/lsd/icd10/>
PREFIX lsd: <http://purl.jp/bio/10/lsd/ontology/201209#>
PREFIX mesh: <http://purl.jp/bio/10/lsd/mesh/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX pair: <http://purl.jp/bio/10/lsd/pair/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX term: <http://purl.jp/bio/10/lsd/term/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpedia_ja: <http://ja.dbpedia.org/resource/>
SELECT DISTINCT ?synonymLabelEn ?dbpedia WHERE {
  ?code1 rdfs:label "片頭痛"@ja ;
        lsd:MeSHUniqueID ?meshID .
  ?code2 lsd:MeSHUniqueID ?meshID ;
        skos:closeMatch ?synonymJa .
  ?synonymJa rdfs:label ?synonymLabelJa ;
           lsd:hasEntry [lsd:hasEnglishTranslationOf ?synonymEn].
  ?synonymEn rdfs:label ?synonymLabelEn .
  OPTIONAL{
    ?synonymEn skos:exactMatch ?dbpedia .
  }
  FILTER(lang(?synonymLabelJa) = "ja")
  FILTER(lang(?synonymLabelEn) = "en")
} ORDER BY ?synonymLabelJa ?synonymLabelEn
```

Appendix 2

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX icd10: <http://purl.jp/bio/10/lsd/icd10/>
PREFIX lsd: <http://purl.jp/bio/10/lsd/ontology/201209#>
PREFIX mesh: <http://purl.jp/bio/10/lsd/mesh/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX pair: <http://purl.jp/bio/10/lsd/pair/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX term: <http://purl.jp/bio/10/lsd/term/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpedia_ja: <http://ja.dbpedia.org/resource/>
SELECT DISTINCT ?meshLabel ?tfidf WHERE {
  ?cooccurrence skos:member <http://purl.jp/bio/10/lsd/mesh/D008881> ;
    skos:member ?comesh ;
    lsd:TfIdfValue ?tfidf .
  ?comesh rdfs:label ?meshLabel .
  FILTER(<http://purl.jp/bio/10/lsd/mesh/D008881> != ?comesh)
  FILTER(lang(?meshLabel) = "ja")
} ORDER BY DESC(?tfidf) LIMIT 10
```