3C2-0S-13b-4

生命科学分野のデータベース統合化に向けた略語データベースの公開

Publishing an abbreviation database toward integration of databases in life science

藤原 豊史 [*1] Toyofumi Fujiwara
山口 敦子 [*2] Atsuko Yamaguchi
山本 泰智 [*2] Yasunori Yamamoto

[*1] 株式会社インテック INTEC Inc.
[*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 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 have released a linked open data set of the Allie database that stores abbreviations and their long forms extracted from titles and abstracts of the entire MEDLINE database. Here, to make this linked open data set more useful, we constructed another set of linked data that links Allie to a life science thesaurus. 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 (以下 DBCLS と呼ぶ) 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[1]. 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[2].

また，多様なデータを RDF 化してつないでいくだけでも実用的には十分なメリットがあるため[3]，既に生命科学分野の多くのデータベースが RDF 化され，リンクトデータとして公開されている[4]. 我々は生物医学分野における書誌情報データベースの MEDLINE [5] から抽出した生命科学分野の略語およびその展開形とその関連情報からなる Allie データベースを 2008 年に構築し，毎月更新しているが[6]，昨年から RDF 化も行い，リンクトデータ(以下 Allie リンクトデータと呼ぶ)としても公開している[7].

今般，LSD プロジェクト[8]で構築されている生命科学分野の電子辞書がリンクトデータ(以下 LSD リンクトデータと呼ぶ)として公開されたので[9]，Key Collision 法[10]を用いて Allie データベースと当該辞書を繋ぐリンクトデータを作成した. 本稿ではリンク作成方法およびリンクの有用性について報告する. また，Allie リンクトデータは他のデータベースと統合的にアクセス出来るようにするため，リンクトオープンデータのハブである DBpedia リンクトデータ[11]に対して，Allie の各展開形と DBPedia の各エントリーとを対応付けたリンクを既に含んでいる. そのリンクの有用性についても報告する.

2. LSD リンクトデータとの対応付け

2.1 Allie リンクトデータ概要

Allie データベースを RDF で表現するにあたり，用いる語彙を定義するオントロジーを OWL で記述し公開した[12]. Allie リンクトデータには，略語，展開形，略語／展開形ペア，ペアの表記のゆれを吸収するためのペアクラスター情報，およびそれらに関連する書誌情報や共起略語情報を含んでいる. 2011 年 6 月から Virtuoso データサーバ[13]を使用して SPARQL エンドポイントを公開している [14]. また，2011 年 12 月からは，Allie の展開形と DBPedia の各エントリーとのリンクを owl:sameAs で定義した RDF データ(トリプル数:95,280)を含めて公開しており，2012 年 4 月 時点での RDF データの総トリプル数は 102,820,269 である.

2.2 LSD リンクトデータ概要

LSD プロジェクトでは生命科学領域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 から 2.8 万語を統制語として採用し，それら統制語と日英対訳辞書とのすり合わせ作業を行い，20 万語の日英シノニム(同義語)辞書を作成している. また統制語のうち 2.5 万語について MeSH ツリーを利用して上位概念・下位概念の関係を整理しており，更に PubMed 抄録中での統制語の共起頻度を収集し，2.8 万語の統制語のうち，336 万組の共起関係を抽出している[15]. これらの中から，統制語情報，同義語情報，用語階層情報，共起情報が LSD リンクトデータとして提供されている(総トリプル数:6,569,066).

2.3 LSD リンクトデータとのリンク作成

LSD リンクトデータの統制語情報，用語階層情報，同義語情報の英語表記に対し(以下 LSD 対象表記と呼ぶ)，Allie リンクトデータの展開形英語表記(以下 Allie 対象表記と呼ぶ)を照合させた. 照合手法として完全一致および Key Collision 法の fingerprint，bi-gram fingerprint，tri-gram fingerprint を用いた.

Table 1 は各手法および fingerprint と bi-gram fingerprint を組み合わせた手法の結果である. それぞれ false positive および false negative が含まれている可能性がある. そこで LSD 対象表記から 100 個をランダムにサンプリングし，各手法の false positive および false negative を調査した. false positive については，照合されたペアの表記が同一概念であることを目視で確認した. false negative については，サンプルの文字列中でアルファベットが最も長く連続する部分の bi-gram を抽出し，Allie 対象表記からそれら bi-gram を含む表記のリストを取得して，リストにサンプルと同一概念の表記がないことを目視で確認した. その結果，各手法において false positive は存在しなかった. また，false negative については，exact match が 15，fingerprint が 0，

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

Table 1 :Allie 対象表記と LSD 対象表記との照合結果

手法	照合したAllie対象表記の数（総対象表記数:1,876,165）	照合したLSD対象表記の数（総対象表記数:142,683）	Allie対象表記 / LSD対象表記照合ペアの数
exact match	21,039(1.1%)	21,039(14.7%)	21,039
fingerprint	77,125(4.1%)	42,797(30.0%)	86,320
bi-gram fingerprint	100,274(5.3%)	41,908(29.4%)	114,409
tri-gram fingerprint	100,132(5.3%)	41,883(29.4%)	114,175
fingerprint + bi-gram fingerprint	103,080(5.5%)	44,397(31.1%)	122,136

※括弧内は照合した対象表記の割合

bi-gram fingerprint および tri-gram fingerprint が 3 という結果になり, ランダムサンプルにおいては fingerprint が最も取りこぼしが少なかった. 実際に Table 1 で, fingerprint と bi-gram fingerprint を組み合わせた結果を除いた場合, 照合された LSD 対象表記の数は fingerprint が最も多い. しかし照合された Allie 対象表記の数については fingerprint より n-gram fingerprint の方が多い. これは LSD 対象表記には表記ゆれがあまり含まれていないのに対して Allie 対象表記には多くの表記ゆれが含まれるため, n-gram fingerprint により照合されるものが多くなることに起因する. また tri-gram fingerprint での Allie 対象表記／LSD 対象表記照合ペアは bi-gram fingerprint の照合ペアに全て含まれていた.

これらの結果から, 取りこぼしの最も少ない fingerprint と Allie 対象表記の表記ゆれを吸収する bi-gram fingerprint の組み合わせで照合した結果を採用することにした. 照合された各表記の URI については owl:sameAs を用いて両者のリンク関係を定義した(トリプル数:122,136).

3. リンクの有用性について

LSD リンクトデータおよび DBPedia リンクトデータとのリンクの有用性を示すため, その利用例を述べる.

3.1 LSD リンクトデータとのリンク

Allie リンクトデータからある略語に対応する展開形リストを取得した場合, 各略語／展開形ペアの特徴を知る手掛かりとして, ペアを含む書誌情報や文献中でペアと共起する略語を取得することができる. 例えば ATP／adenosine triphosphate(略語／展開形)というペアに共起する略語として ADP, AMP, PCr 等を得ることができる. ここで, 今回作成した LSD リンクトデータとのリンクを利用することで, 各ペア(の展開形)についての同義語や共起語, 上位概念・下位概念も LSD リンクトデータから得ることでき, ペアに対するより多くの情報を得ることが可能となる. 例えば同義語として adenosine 5'-triphosphate, 共起語として Protein Kinase, 上位概念として Adenine Nucleotide 等を得る.

Allie リンクトデータと LSD リンクトデータが同じレポジトリーに収納されていることを前提とし, Allie 展開形についての同義語と共起語を LSD リンクトデータから得るための SPARQL クエリーの例を Appendix1.1 に示す. 展開形とリンクされている LSD リンクトデータの URI を利用して同義語を取得し, 更に LSD リンクトデータの概念 ID を利用して共起語を取得する. その結果, 36 件の同義語を取得し, 50 件の共起語を取得した.

3.2 DBPedia リンクトデータとのリンク

ある文献セットの特徴を知るための手掛かりとして, Allie リンクトデータからそれら文献に含まれている略語／展開形のペアとその出現頻度を得ることができる. さらに DBPedia リンクトデータとのリンクを利用すれば, 取得したペア(の展開形)について概要を DBPedia リンクトデータから取得でき, それら情報を解釈するうえでの手助けとなる.

この情報を得るための SPARQL クエリーの例を Appendix1.2 に示す. PubMed ID リストで与えられる文献群に含まれる略語／展開形ペアの URI を取得し, それら略語／展開形ペアおよび出現頻度を得る. 取得した展開形とリンクされている DBpedia リンクトデータ の URI を利用し, DBPedia SPARQL エンドポイントから展開形の概要を得る. 今回, DBpedia SPARQL エンドポイントから情報を取得するために, Federated Query をサポートした SPARQL1.1 を利用した. そのため, Allie リンクトデータを SPARQL1.1 に対応した OWLIM ver4.3[16]に収納しクエリーを実行した. その結果, 8 件のペアを取得し, うち 4 件の展開形について DBPedia リンクトデータから概要(英語表記)を取得した.

4. おわりに

今回, LSD リンクトデータとのリンクを含めた Allie リンクトデータを作成し, そのリンクの有用性を確認した. リンクの作成では, false positive および false negative の調査を 100 個のランダムサンプルに対して行ったが, 全体のデータ数を考えると, 今後より多くのサンプルで評価を行う必要があると考えている.

また, DBPedia リンクトデータとのリンクの有用性を確認し, さらに Federated Query を利用した DBPedia SPARQL エンドポイントからのデータ取得を実証できた. 今後は Allie SPARQL エンドポイントを SPARQL1.1 に対応するために, Virtuoso データサーバから OWLIM データサーバへの変更を検討する予定である.

LSD リンクトデータ, DBPedia リンクトデータとのリンクを利用することで略語／展開形情報に対してより包括的な情報を活用できるようになったので, 今後はこの特性を活かしたサービスを検討する. 例えば, DBCLS では生命科学分野の新着論文について日本語によるレビューを公開するサービス FIRSTAUTHOR'S[17]を提供しているが, 各記事には展開形が記されていない略語が多く存在する. これら記事に対し, 略語の展開形や, 展開形に関連する DBPedia リンクトデータの情報, LSD リンクトデータの情報を付加することで, レビュー記事をよりよく理解するための仕組みを検討している.

謝辞

呉紅艶博士には OWLIM データサーバへのデータのロードなどでご支援いただいた. ここに感謝の意を表したい. また, 本研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

[1] 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)

[2] 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

[3] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J, Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform, Oct;41(5):706-16, (2008)
[4] W3C SWEO Linking Open Data community, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[5] MEDLINE Fact Sheet, http://www.nlm.nih.gov/pubs/factsheets/medline.html
[6] Y. Yamamoto, A. Yamaguchi, H. Bono and T. Takagi, Allie: a database and a search service of abbreviations and long forms, Database, bar03, (2011)
[7] 藤原豊史, 山口敦子, 山本泰智: 生命科学分野におけるセマンティック Web 技術を利用したデータリソースの公開, 第24回セマンティックウェブとオントロジー研究会, (2011)
[8] LSD プロジェクト, http://lsd.pharm.kyoto-u.ac.jp/ja/about/about_lsd/index.html
[9] ライフサイエンス辞書（LSD）リンクドオープンデータ, http://www.semanticweb.jp/lod/LodOfLsd.html
[10] Key Collision Methods, http://code.google.com/p/google-refine/wiki/ClusteringInDepth
[11] DBPedia, http://dbpedia.org/
[12] Allie オントロジー, http://allie.dbcls.jp/ontology/201108
[13] Virtuoso, http://virtuoso.openlinksw.com/
[14] Allie SPARQL エンドポイント, http://data.allie.dbcls.jp/sparql/
[15] 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)
[16] OWLIM, http://www.ontotext.com/owlim
[17] FIRSTAUTHOR'S, http://first.lifesciencedb.jp/

Appendix 1.1

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX mesh: <http://www.nlm.nih.gov/mesh/2012#>
PREFIX allie: <http://purl.org/allie/ontology/201108#>
PREFIX pmid: <http://togows.dbcls.jp/entry/ncbi-pubmed/>
PREFIX lsd: <http://lifesciencedic.org/mo#>
select ?LSD ?SYN ?coEN
from <http://purl.org/allie>
from <http://purl.org/allie/lsdlod>
from <http://lifesciencedic.org/mo#>
where {
 [] allie:hasLongFormRepresentationOf [rdfs:label "adenosine triphosphate"@en] ;
    allie:hasLongFormRepresentationOf [owl:sameAs ?LSD] ;
    rdf:type allie:PairCluster .
 [] lsd:Term_Strings ?LSD ;
    lsd:Term_Strings [rdfs:label ?SYN] ;
    ?lsd_pid ?PID .
 [] rdf:type owl:Class ;
    owl:intersectionOf (?PID ?coID) .
 ?coJAID ?lsd_id ?coID .
 ?coJAID ?lsd_en [rdfs:label ?coEN] .
 ?lsd_pid rdfs:label "\u89AA\u6982\u5FF5ID" . # 親概念 ID
 ?lsd_id rdfs:label "\u6982\u5FF5ID" . # 概念 ID
 ?lsd_en rdfs:label "\u6982\u5FF5\u82F1\u8A9E\u8868\u8A18" . # 概念英語表記
}
```

Appendix 1.2

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX allie: <http://purl.org/allie/ontology/201108#>
PREFIX pmid: <http://togows.dbcls.jp/entry/ncbi-pubmed/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
select ?SF ?LF ?freq ?abs
from <http://purl.org/allie>
from <http://purl.org/allie/dbpedia>
where {
 ?X allie:appearsIn ?Y ;
    allie:hasShortFormRepresentationOf [rdfs:label ?SF] ;
    allie:hasLongFormRepresentationOf [rdfs:label ?LF] ;
    allie:frequency ?freq ;
    rdf:type allie:PairCluster .
 { ?Y allie:hasMemberOf pmid:22057056} union { ?Y allie:hasMemberOf pmid:22008769} union
 { ?Y allie:hasMemberOf pmid:21944663} union { ?Y allie:hasMemberOf pmid:21928700} union
 { ?Y allie:hasMemberOf pmid:21868617}
 OPTIONAL { ?X allie:hasLongFormRepresentationOf [owl:sameAs ?sa] .
  SERVICE <http://dbpedia.org/sparql> { ?sa dbpedia-owl:wikiPageRedirects [dbpedia-owl:abstract ?abs] . }
  FILTER ( lang(?abs) = "en" )
 }
 FILTER ( lang(?LF) = "en" )
} ORDER BY DESC (xsd:integer(?freq)) ?LF
```