1N5-OS-10c-5

生命科学分野の学術文献情報からなるリンクトデータの構築

Publishing Linked Data toward integration of scientific literature in life science

藤原 豊史 [*1] Toyofumi Fujiwara
松宮 遼 [*2] Ryo Matsumiya
山本 泰智 [*3] Yasunori Yamamoto

[*1] 株式会社インテック INTEC Inc.
[*2] 電気通信大学 The University of Electro-Communications
[*3] 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

Recently, linked open data sets concerning biomedical literature have been constructed in some institutions and publishers, and it allows for easy exploration and retrieval of authors or titles or citations of literature by using SPARQL query. We have developed the Colil database which contains citations and their contexts of biomedical literature extracted from open access datasets. In this paper, we present the RDFized Colil database that uses newly compiled vocabularies defined as Colil ontology in addition to standard vocabularies such as BIBO, DC terms, and DoCO. Our aim is to provide users with easy exploration and retrieval of citations and their contexts of biomedical literature through the linked data set of Colil database.

1. はじめに

1.1 生命医学分野の学術文献情報に対する取り組み

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[Bono 2009]. その活動の一環として「MEDLINE」や「PMC」[*1] などの, 生命医学分野の学術文献に関するデータベースを利用したサービスの構築や, リンクトデータの提供を行っている.

MEDLINE は米国国立医学図書館(National Library of Medicine)により維持管理がなされている生物医学分野で世界最大の書誌情報データベースで, 米国およびその他 80 カ国以上の国で出版される, 39 言語 5,600 の学術誌に掲載された 2,200 万件以上の書誌情報を収めている[MEDLINE 2013]. DBCLS では MEDLINE を利用し, 例えば, 文献の題目や要旨から URL を抽出し, 当該アドレスで提供されるデータベースやツールなどを関連書誌情報とともに検索できるサービス「OReFiL」[Yamamoto 2007]や, 文献中の英語表現を逐次検索できるサービス「inMeXes」[*2] などがある. また, 同じく題目もしくは要旨に出現する略語と対応する展開形を検索するサービス「Allie」[Yamamoto 2011]が存在し, 同サービスについてはリンクトデータも提供している[藤原 2011].

NCBI(National Center for Biotechnology Information)により維持管理がなされている PMC は生物医学分野の文献アーカイブで, 約 260 万件の文献が保存されており, 本文を含めて無料での閲覧が可能である. DBCLS では PMC の中でも, Creative Commons もしくはそれと同様のライセンスのもとにある「PMC Open Access Subset」(以下 OA subset と呼ぶ)の文献を対象に(約 60 万件), 本文中に記述されている別文献への引用情報を抽出し, 書誌情報と合わせて収めたデータベース"Comments on literature in literature"(以下 Colil データベースと呼ぶ)と, それを利用した文献間の引用関係を検索・閲覧できるサービス「Colil」を開発している.

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.ncbi.nlm.nih.gov/pmc/
*2 http://docman.dbcls.jp/im/

1.2 生命医学分野の学術文献情報リンクトデータ

生命医学分野の学術文献情報に関して, 既に多くの利用可能なリンクトデータが提供されている. MEDLINE については DBCLS が提供するサービス「TogoWS」[Katayama 2010]にて, 文献ごとの書誌情報の RDF データを RESTful APIs を介して取得することができる. PMC については FSU Research Foundation が提供するサービス「biotea」[Alexander 2012]にて, 各文献の書誌情報とともに, 「NCBO Annotator」[*3] および「Whatizit」[*4] サービスを利用して取得した文献内容に対するアノテーション情報を含めた RDF データが, SPARQL エンドポイントとともに提供されている.

また, Nature Publishing Group(以下 NPG と呼ぶ) や Elsevier といった大手出版社からもリンクトデータが提供されている. NPG は 1845 年から出版した 45 万件以上の記事の基本的な書誌情報(title, author, publication date 等)と NPG オントロジーからなる RDF データと SPARQL エンドポイントを提供している[NPG 2013]. Elsevier も「Linked Data Repository」として, Elsevier の文献および関連する外部リソースの文献の書誌情報や, 文献間の関係を表すメタデータ, 統制語彙などを含む RDF データを作成しており, RESTful APIs を介してアクセスすることが出来る[Elsevier 2013].

各機関および出版社が学術文献情報をリンクトデータとして提供していることで, SPARQL クエリにより, 書誌情報の検索や取得, また文献間の関係および外部リソースとの関係を利用して複数の書誌情報データベースを横断的に検索することなどが, 簡単に実施できる状況にある.

今回, 我々は Colil データベースを RDF 化し(以下 Colil リンクトデータと呼ぶ), SPARQL エンドポイントを構築した. Colil リンクトデータを利用すれば, 文献間の引用関係や, ある文献がそれを引用する文献中でどのように引用されているかという文脈(以下 引用文脈と呼ぶ)を簡単に取得できる. これは, 自身が執筆する文献中で他文献を引用する際や, ある文献が他文献からどのように評価されているのかを概観する際などに役立つ.

*3 http://bioportal.bioontology.org/annotator
*4 http://www.ebi.ac.uk/webservices/whatizit/

尚, Colil リンクトデータと同様に引用文脈を提供する学術文献検索サービス「Microsoft Academic Search」[*5]が存在するが, 同サービスのデータベース自体はオープンアクセスではなく, Colil リンクトデータのようにデータを再利用することはできない.

2. Colil リンクトデータ構築

Colil リンクトデータを構築する際, 相互運用性を考慮し積極的に既存オントロジーの語彙を採用した. 既存オントロジーは Bibliographic Ontology(以下 BIBO と呼ぶ)[*6], DCMI Metadata Terms(以下 DC Terms と呼ぶ)[*7], Document Components Ontology(以下 DoCO と呼ぶ)[*8], TogoWS Ontology[*9] を利用した. 既存オントロジーに適切な語彙が存在しない場合は Colil Ontology[*10] として語彙を定義し, 公開した. Colil リンクトデータには外部リソース TogoWS, PMC, DOI System[*11] へのリンクを含めた.

基本的な書誌情報を表現するために TogoWS Ontology と Colil Ontology を利用した. 文献のセクション情報および各セクション中の引用文脈を表現するために DoCO, DC Terms, Colil Ontology を利用した. 文献間の引用関係および文献の識別情報を表現するために BIBO と Colil Ontology を利用した. 引用関係上の文献の属性および関連度の高い文献を表現するために Colil Ontology を利用した.

2.1 文献間の引用関係および関連を表す RDF データ

図1は文献間の引用関係および関連度の高い文献を表現する RDF データである. 引用されている文献(以下 参照文献 と呼ぶ)は colil:ReferencePaper クラスに属し, 参照文献を引用している文献(以下 引用文献 と呼ぶ)は colil:CitationPaper クラスに属す. 引用文献中に含まれる各セクションは doco:contains プロパティで指定し, セクションは doco:Section クラスに属し, セクション名は dcterms:title プロパティで指定する. 各セクション中の引用文脈を doco:contains プロパティで指定し, 引用文脈は colil:Context クラスに属し, 引用文脈を rdf:value プロパティで指定する. 引用文脈中で言及している参照文献について, 引用文脈から colil:mentions プロパティで参照文献を指定する. また, 引用文献から引用されている参照文献を bibo:cites プロパティで指定する.

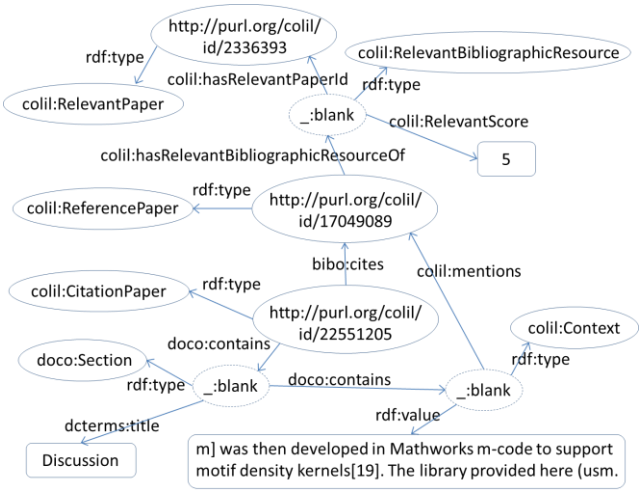

図１． 文献間の引用関係および関連を表す RDF データ

Colil データベースは, 共引用関係を基にした, 参照文献と関連度の高い文献(以下 関連文献 と呼ぶ)を関連スコアとともに収めている. すなわち, 参照文献 A を引用する文献群 A'と文献 B が引用する文献群 B'について, A'∩B'が 2 以上の場合, 文献 B を参照文献 A の関連文献と定義し, その要素数を関連スコアと定義する. 参照文献から関連書誌情報を colil:hasRelevantBibliographicResourceOf プロパティで指定し, 関連書誌情報は colil:RelevantBibliographicResource クラスに属し, 関連スコアを colil:RelevantScore プロパティで指定する. 更に, 関連書誌情報から関連文献を colil:hasRelevantPaperId プロパティで指定し, 関連文献は colil:RelevantPaper クラスに属する.

2.2 書誌情報を表す RDF データ

図2は書誌情報を表現する RDF データである. 参照文献, 引用文献および関連文献は TogoWS, PubMed, DOI System リソースへのリンクを rdfs:seeAlso プロパティで指定する. また, PMC ID を colil:pmcid プロパティで指定し, DOI を bibo:doi プロパティで指定する. rdfs:seeAlso プロパティで指定された TogoWS リソースは colil:PubMed クラスに属す. TogoWS リソースは文献のタイトルを togows:ti プロパティで指定し, 出版情報を togows:so プロパティで指定し, PubMed ID を togows:pmid プロパティで指定し, また著者情報を colil:writtenBy プロパティで指定する.

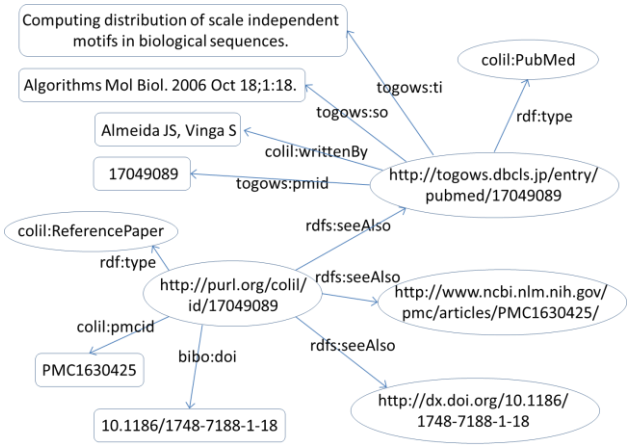

図２． 書誌情報を表す RDF データ

3. Colil リンクトデータについて

作成されたトリプル数は約 3 億 7000 万となった. Fuseki(ver 0.2.6)[*12]を利用し Colil リンクトデータを収めた SPARQL エンドポイント構築した. この SPARQL エンドポイントを介して, 例えば PubMed ID が"17049089"である文献が引用している文献の PMC ID 一覧を取得すること(Appendix 1)や, DOI が"10.1186/1748-7188-1-18"である文献と関連度が高い文献のタイトル一覧を取得すること (Appendix 2), また PMC ID が" PMC1630425"である文献の"Background" セクション内で引用している文献の DOI 一覧を取得すること(Appendix 3)や PMC ID が" PMC1630425"である文献を引用している文献のその引用文脈一覧を取得すること(Appendix 4)などが可能である.

*5 http://academic.research.microsoft.com/
*6 http://purl.org/ontology/bibo/ (abbr. bibo)
*7 http://purl.org/dc/elements/1.1/ (abbr. dcterms)
*8 http://purl.org/spar/doco/ (abbr. doco)
*9 http://hackathon3.dbcls.jp/wiki/RDF/TogoWS/PubMed (abbr. togows)
*10 http://purl.org/colil/ontology/201303# (abbr. colil)
*11 http://www.doi.org/
*12 https://jena.apache.org/documentation/serving_data/

4. おわりに

今後, 構築した Colil リンクトデータを利用した文献情報検索サービスの開発や, DBCLS が開発・運用するサービスでの Colil リンクトデータの活用などを予定している.

文献情報検索サービスでは, 文献間の引用関係情報を活用するだけではなく, 他データリソースとのリンクも活用させる予定である. 例えば, Colil リンクトデータは TogoWS リソースへのリンクを含んでいるが, TogoWS リソースは Allie リンクトデータからもリンクされているため, Colil リンクトデータの引用関係情報とともに Allie リンクトデータの略語とその展開形情報を利用することが可能である. 引用文脈中に略語が含まれている場合には, その文献の題目もしくは要旨に含まれる略語・展開形情報を提供することで, 引用文脈の理解に役立つかもしれない.

DBCLS が開発・運用するサービスでの Colil リンクトデータの活用として, ある学問分野・領域を広く総合的に取り上げ日本語でのレビューを公開する「ライフサイエンス 領域融合レビュー」サービス[*13] での活用を検討している. 各記事には参考文献が記されているが, これら文献を引用している文献や関連度の高い文献の情報を更に付加することで, レビュー記事をよりよく理解できると考えている.

謝辞

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

[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)

[MEDLINE 2013] MEDLINE Fact Sheet, http://www.nlm.nih.gov/pubs/factsheets/medline.html

[Yamamoto 2007] Y. Yamamoto and T. Takagi, OReFiL: an online resource finder for life sciences, BMC Bioinformatics, Vol. 8, No. 287, (2007)

[Yamamoto 2011] Y. Yamamoto, A. Yamaguchi, H. Bono and T. Takagi , Allie: a database and a search service of abbreviations and long forms, Database, bar03, (2011)

[藤原 2011] 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)

[Katayama 2010] T. Katayama, M. Nakao, T. Takagi, TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services, Vol. 38, W706-W711, (2010)

[Alexander 2012] Alexander Garcia, Leyla Jael Garcia, Casey McLaughlin and Stephen Flager, RDFising PubMed Central, Bio-ontologies, Long Beach, (2012)

[NPG 2013] NPG Linked Data Platform, http://developers.nature.com/docs/read/linked_data

[Elsevier 2013] Elsevier Linked Data Repository, http://data.elsevier.com/documentation/index.html

*13 http://leading.lifesciencedb.jp

Appendix 1

```
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX colil: <http://purl.org/colil/ontology/201303#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX togows: <http://togows.dbcls.jp/ontology/ncbi-pubmed#>
select ?CitationPaper ?PMCID
where {
  ? CitationPaper rdf:type colil: CitationPaper ;
     rdfs:seeAlso [
        rdf:type colil:PubMed;
        bibo:pmid "17049089"
     ] ;
     bibo:cites [ colil:pmcid ?PMCID ] .
}
```

Appendix 2

```
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX colil: <http://purl.org/colil/ontology/201303#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX togows: <http://togows.dbcls.jp/ontology/ncbi-pubmed#>
select ?RelevantPaperTitle where {
  [] bibo:doi "10.1186/1748-7188-1-18";
     colil:hasRelevantBibliographicResourceOf [
        colil:hasRelevantPaperId [
           rdfs:seeAlso [
              rdf:type colil:PubMed;
              togows:ti ?RelevantPaperTitle ]
        ]
     ] .
  FILTER ( lang(?RelevantPaperTitle) = "en" )
}
```

Appendix 3

```
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX colil: <http://purl.org/colil/ontology/201303#>
PREFIX dcterms: <http://purl.org/dc/elements/1.1/>
PREFIX doco: <http://purl.org/spar/doco/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
select ?DOI where {
  ?CitationPaper rdf:type colil:CitationPaper ;
     colil:pmcid  "PMC1630425" ;
     doco:contains [
        dcterms:title "background" ;
        doco:contains [
           colil:mentions [  bibo:doi ?DOI ]
        ]
     ] .
}
```

Appendix 4

```
PREFIX colil: <http://purl.org/colil/ontology/201303#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
select ?Context where {
  [] rdf:value ?Context ;
     colil:mentions [  colil:pmcid "PMC1630425" ] .
}
```