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

Use of Semantic Web technologies to publish data resources in life science

藤原 豊史[1]　山口 敦子[2]　山本 泰智[2]

Toyofumi Fujiwara[1], Atsuko Yamaguchi[2], Yasunori Yamamoto[2]

[1]株式会社インテック
[1] 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
[2] Database Center for Life Science, Research Organization of Information and Systems

Abstract: 生命科学分野におけるセマンティック Web 技術の利用は浸透しつつあり、データリソースの RDF 化、オントロジーの整備が徐々に進められている。我々は現在、生命科学分野の略語とその展開形を検索するためのサービス「Allie」を開発・運用しているが、そこでの利用を想定し、生物医学分野における書誌情報データベースの MEDLINE (約 2,000 万件) から抽出した略語・展開形のデータベースを作成している。今回の発表では、作成した Allie のデータベースを基にしたオントロジーおよびリンクトデータの構築について、作業過程および残された課題を報告する。

1　はじめに

1.1　セマンティック Web の利用状況

生命科学分野におけるセマンティック Web 技術の利用は徐々に浸透しつつあり、データリソースの RDF 化、オントロジー（統制語彙と階層データ構造）の整備が進められている[1]。

W3C SWEO Linking Open Data community プロジェクト[2]では、ウェブ上で利用可能なオープンデータを RDF 形式で作成すること、異なるデータリソースの項目間にリンクを設定することを目標としており、欧米を中心に W3C や様々なコミュニティによりデータの公開と利用が進められている。これまでに公開され関連づけられたオープンデータの全体像をみると、すでに多くの生命科学関連データが公開されており、全体の 1/4 程度を占めるまでになっている[3]。

そうした取り組みの先駆的な例としては、Universal Protein Resource (Uniprot) [4]や Biological Pathways Exchange (BioPAX) [5]が挙げられる。

*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

UniProt は、タンパク質のアミノ酸配列や機能などの特徴を包括的に収納するデータベースである。データの管理や更新に、専用に開発したソフトウェアを利用していたが、時代の経過とともにデータ構造が複雑化し、それに対応するためのプログラムの更新が困難になるという問題が生じた。そこで、OWL を用いてオントロジーを整備し、全てのデータを RDF 化することで解決を図った[6]。

BioPAX は、代謝ネットワーク、シグナル伝達ネットワーク、遺伝子制御ネットワーク、タンパク質間相互作用ネットワークなどのパスウェイデータの標準的な交換フォーマットを開発している。パスウェイデータは本質的に複雑な構造をもっているので、OWL でデータ交換フォーマットを定義し、同フォーマットに対応したデータの属性について曖昧さを極力排除している。現在では、BioPAX の交換フォーマットに対応したデータがさまざまな機関から提供されている (BioCyc、INOH 等) [7]。

また、既存公共データベースを RDF 化し、統合的な検索基盤の提供に取り組む Bio2RDF[8]、NeuroCommons[9] 、 LinkedLifeData[10] 、 Shared Names[11]などのプロジェクトが存在する。Bio2RDF プロジェクトでは、NCBI、EBI、PDB、KEGG など 40 以上の既存公共データベースの RDF 化に取り

組んでおり、それらデータを関連づけした統合的なデータベースを構築し、SPARQL による問い合わせが可能なシステムを公開している。NeuroCommons は、様々な公共データベースのデータおよびテキストマイニングした論文等の文献データを集積し、アルツハイマー病など脳科学に関する知識の問い合わせを SPARQL クエリによって行えるシステムの構築などを行っている。

1.2 MEDLINE データを用いたサービス

MEDLINE は生物医学分野における世界最大の書誌情報データベースであり、米国国立医学図書館 (National Library of Medicine, NLM) により維持管理がなされている。現在では、米国およびその他 80 カ国以上の国で出版される、39 言語 5,516 の学術誌に掲載された約 2,000 万件の書誌情報を収めている[12, 13]。

文部科学省「統合データベースプロジェクト」では、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 (DBCLS) を中心とし、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構築を行っている[14]。

その一環として、MEDLINE を利用したサービスをいくつか行っている。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」[15]や、文献中の英語表現を逐次検索できるサービス「inMeXes」[16]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」[17]があり、本研究は当該サービスで利用しているデータを対象として取り組んでいる。

1.3 略語／展開形検索サービス「Allie」

生命科学系の文献中に多く出現する略語は多義語であることが多く、特に専門外の読者には理解するのに困難を伴うことがある。また、MEDLINE を検索対象に含む文献検索サービス「PubMed」[18]を利用する際、クエリが略語の場合には、求めていない文献が多く検索結果に混じってしまうことがしばしば起こる。

この問題に対する一つの解となるよう、我々は生命科学分野において利用されている略語とその展開形のペアを検索するサービス Allie を開発・運用している。Allie は MEDLINE に含まれる全ての書誌情報(のうち、題目と要旨、以下 MEDLINE データと呼ぶ)を対象として略語とその展開形を検索し、略語／展開形ペア、そのペアを含む書誌情報などを返す。例えば、「RDF」をクエリとして検索をすると、RDF を略語とした場合の「radial distribution function (動径分布関数)」や「refuse-derived fuel (ゴミ固形燃料)」、「Resource Description Framework」などの展開形を得ることができ、また各ペアについて、出現する文献の書誌情報や、各 MEDLINE データ中で共起する略語、研究分野などの情報も得ることができる。

1.4 Allie データリソースの公開

我々は MEDLINE データ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データベース (以下 Allie データベースと呼ぶ) を構築している。既に、Allie データベースをテキストファイル形式で公開しているが、今回、さらに他のサービスやアプリケーションでより効率的に活用できるデータを提供するために、Allie データベースを基にしたオントロジーおよびリンクトデータの構築を行った。さらには SPARQL を用いて外部から検索をできるように、SPARQL エンドポイントを公開した。

2 Allie データベース構築

我々は生命科学分野で利用される略語およびその展開形を取得するため、全ての MEDLINE データを対象とし、ALICE ツール[19]を用いて略語とその展開形を機械的に取得した。取得した略語／展開形ペアに、それらが出現している文献の PubMed ID (MEDLINE 中の各書誌情報に付けられている ID) および出版年度を関連づけたデータについては、テキストファイル形式で公開している (ftp://ftp.dbcls.jp/allie/)。なお、日々新規文献が MEDLINE に格納されているため、毎週データを更新している。

取得した略語／展開形ペアには、例えば「RDF/ Resource Description Framework」と「RDF/ resource-description-framework」の様に、同じ概念に対して表記のゆれ (同義語) が存在する。そこで、互いに同義語の関係にあるペアをまとめるクラスタリングを行い、各クラスターの中から MEDLINE データ中での出現頻度が最も高い略語／展開形ペアをその代表表現として選定した。

各略語／展開形ペアおよび略語／展開形ペアクラスターには、共起略語および研究分野を関連づけた。なお、研究分野については、MEDLINE に納められている文献を出版する各学術誌に対し、その研究対

象を表すために生命科学分野の統制語彙「MeSH」[20]に含まれる「MeSH タ ーム」が NLM により付けられているので、その情報を基に関連づけた。また、Allie の検索用索引として用いるため、略語／展開形ペアクラスターには、正規化処理 (カンマ、ピリオド、ハイフン、及び空白文字の削除およびアルファベット大文字を小文字に変換) を施した略語を関連付けた。

これらのデータをリレーショナルデータベースに収納し管理しており、当該データベースのテーブル定義および ER 図を APPENDIX 1、APPENDIX 2 にそれぞれ示す。

3 オントロジー構築

Allie データベースを RDF で表現するために定義した語彙 (クラスおよびプロパティ) について述べる。全てのクラスおよびプロパティの定義は Protégé[21]を用いて作成した。ここで、名前空間 allie: は今回定義した語彙を表す。

3.1 略語および展開形データ表現モデル

略語および展開形それぞれについて URI を定義し、その URI を主語として属性を指定する。略語は allie:ShortForm クラスに属し、実際の記述は rdfs:label で行い、言語指定 @en を付ける。展開形は allie:LongForm クラスに属し、実際の記述は rdfs:label で行う。Allie では展開形の一部に日本語訳を含めているので、言語指定 @en / @ja をそれぞれ付ける。また、頻度属性を指定するための allie:frequency プロパティを用いて、略語または展開形が MEDLINE データに出現する頻度を指定する。

例えば、展開形 ID が 694079 であるデータにつき、

・展開形の英語表記 は「bacteriorhodopsin」
・展開形の日本語表記 は「バクテリオロドプシン」
・文献に出現する頻度 は「688」

という情報は、次のような形で表現される。

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix allie: <http://purl.org/allie/ontology/201102#>
<http://purl.org/allie/id/longform/694079>
   rdf:type allie:LongForm ;
   rdfs:label "bacteriorhodopsin"@en ;
   rdfs:label "バクテリオロドプシン"@ja ;
   allie:frequency "688" .
```

3.2 略語／展開形ペアデータ表現モデル

略語／展開形ペアについて URI を定義し、その URI を主語として属性を指定する。略語／展開形ペアは allie:Pair クラスに属し、allie:hasShortFormOf プロパティで略語の URI を指定し、allie:hasLongFormOf プロパティで展開形の URI を指定する。また、同じく allie:frequency プロパティを用いて、略語／展開形ペアが MEDLINE データに出現する頻度を指定する。更に、allie:inResearchAreaOf プロパティを用いて、略語／展開形ペアの研究分野を MeSH タームの URI で指定する。

略語／展開形ペアが出現する MEDLINE データを指定するために、まず allie:PubMedIDList クラスに属するインスタンスを allie:appearsIn プロパティで指定する。同インスタンスの属性として、当該ペアの出現する全ての MEDLINE データの URI を allie:hasMemberOf プロパティを用いて指定する。そして、略語／展開形ペアの共起略語を指定するために、まず allie:CooccurringShortFormList クラスに属するインスタンスを allie:cooccursWith プロパティで指定し、同インスタンスの属性として全ての共起略語の URI を allie:hasMemberOf プロパティを用いて指定する。

例えば略語／展開形ペア ID が 1 であるデータにつき、

・略語の ID は「1」
・展開形の ID は「389886」
・文献に出現する頻度 は「1」
・研究分野の MeSH ID は「D004358」
・出現する MEDLINE データの PubMed ID は「1331411」
・共起略語の略語 ID は「1279, 144828」

という情報は、次のような形で表現される。

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix allie: <http://purl.org/allie/ontology/201102#>
<http://purl.org/allie/id/pair/1>
   rdf:type allie:Pair ;
   allie:hasShortFormOf
      <http://purl.org/allie/id/shortform/1> ;
   allie:hasLongFormOf
      <http://purl.org/allie/id/longform/389886> ;
   allie:frequency "15" ;
   allie:inResearchAreaOf
      <http://www.nlm.nih.gov/mesh2011#D004358> ;
   allie:appearsIn [
      rdf:type allie:PubMedIDList ;
      allie:hasMemberOf
         <http://togows.dbcls.jp/entry/ncbi-pubmed/1331411>
   ] ;
   allie:cooccursWith [
      rdf:type allie:CooccurringShortFormList ;
      allie:hasMemberOf
         <http://purl.org/allie/id/shortform/1279> ;
      allie:hasMemberOf
         <http://purl.org/allie/id/shortform/144828>
   ] .
```

3.3 略語／展開形ペアクラスターデータ表現モデル

Allie データベースを更新する度にクラスターの略語／展開形が変化する可能性があるため、略語／展開形ペアクラスターの URI を定義せずに、主語はブランクノードとして属性を指定する。略語／展開形ペアクラスターは allie:PairCluster クラスに属し、allie:hasShortFormRepresentationOf プロパティでクラスターを代表する略語の URI を指定し、allie:hasLongFormRepresentationOf プロパティでクラスターを代表する展開形の URI を指定する。

略語／展開形ペアクラスターを構成する略語／展開形ペアを指定するために、allie:PairList クラスに属すインスタンスを allie:contains プロパティで指定し、同インスタンスの属性として全ての略語・展開形ペアの URI を allie:hasMemberOf プロパティを用いて指定する。また、略語／展開形ペアクラスターの全 MEDLINE データ中での出現頻度、研究分野、出現する書誌情報および共起略語の指定を、略語／展開形ペアの場合と同様に行う。

3.4 OWL を用いたオントロジー構築

OWL を用いて定義した語彙 (クラスおよびプロパティ) についての説明と、その全クラスの階層関係を APPENDIX 3 に示す。

4 リンクトデータ構築

今回定義した語彙を用いて、Allie データベースから RDF データを作成した。その結果、略語、展開形、略語／展開形ペアそれぞれの URI 数は、503,261、1,672,695、1,972,298 件となった。これらの URI にアクセスした際には、それぞれの関連情報を返すように Allie を実装した。また、書誌情報および研究分野のそれぞれについて既存 URI へのリンクを加えることで、より多くの情報をたどれるようにした。RDF データのトリプル数は 86,015,878 となり、それらを収納したテキストファイルのサイズは約 10,438 MB となった。RDF データは N-Triples 形式とし、RDF 化には CPAN モジュール RDF::Trine(Ver 0.133)[22]を用いた。

5 Allie データリソース公開

Allie データリソースを SPARQL を用いて外部から検索することができるように、今回作成した RDF データを Virtuoso v6 データサーバ[23]に収納し、SPARQL エンドポイントを公開した (http://data.allie.dbcls.jp/sparql/)。その結果、Allie を補完する形で、より効率的なデータ検索環境を提供することが可能となったので、ここではいくつかの検索例を示す。

検索例 (1)

同一研究分野内では多義性のあるペアが余り多くないことが想定されるが、それを調べるため、ある研究分野に関連する略語／展開形ペアの URI 一覧を取得するためのクエリを以下に示す。

```
PREFIX allie: <http://purl.org/allie/ontology/201102#>
PREFIX mesh: <http://www.nlm.nih.gov/mesh/2011#>
select ?X where {
   ?X a allie:Pair ;
      allie:inResearchAreaOf mesh:D005193 .
}
```

この結果、分野 D005193 (Family Planning Services) に特有の略語／展開形ペアを 259 件得ることができる。

検索例 (2)

あるキーワードに関連する文献を PubMed を用いて検索し、文献の PubMed ID リストを得る。それに対応した MEDLINE データに含まれる略語／展開形のペアとその出現頻度は、その文献セットの特徴を知る手掛かりとなりうる。そこで、構築したリンクトデータに収められている PubMed ID を示す URI を利用することで、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 を調べるためのクエリを以下に示す。

```
PREFIX allie: http://purl.org/allie/ontology/201102#
PREFIX pmid: http://togows.dbcls.jp/entry/ncbi-pubmed/
select ?X where {
  ?X allie:appearsIn ?Y ;
        rdf:type allie:Pair .
  { ?Y allie:hasMemberOf pmid:20022739 }
  union
  { ?Y allie:hasMemberOf pmid:19948798 }
  union
  { ?Y allie:hasMemberOf pmid:19745340 }
  union
  { ?Y allie:hasMemberOf pmid:19709868 }
  union
  { ?Y allie:hasMemberOf pmid:19561017 }
}
```

この結果、5 件の文献に関する略語／展開形ペアを 11 件 (10 種類) を得ることができる。

検索例 (3)

略語／展開形ペアとそれが含まれるクラスターの研究分野が異なる場合、当該略語／展開形ペアは、それを主に使う研究分野に特有の表現である可能性がある。そこで、ある略語／展開形ペアとそのクラスターの研究分野を調べるためのクエリを以下に示す。

```
PREFIX allie: <http://purl.org/allie/ontology/201102#>
select ?X,?Y where {
 [] allie:inResearchAreaOf ?X ;
    allie:contains [
      allie:hasMemberOf [
          allie:hasShortFormOf [
            rdfs:label "SMA"@EN ] ;
          allie:hasLongFormOf [
            rdfs:label "smooth muscle alpha-actin"@EN ] ;
          allie:inResearchAreaOf ?Y
    ] ].
}
```

この結果、SMA/smooth muscle alpha-actin ペアの研究分野 D003585(Cell Biology) は、それが属するクラスター (代表略語／展開形ペアは SMA/alpha-smooth muscle actin) の研究分野 D010336(Pathology) と異なり、その分野特有のペア表現である可能性がある。

6 おわりに

我々は生物医学分野における世界最大の書誌情報データベース MEDLINE を用いて、生命科学分野において利用されている略語とその展開形を基にしたデータベースを構築している。今回、他のサービスやアプリケーションでより効率的に活用できるデータを提供するために、Allie データベースを基にしたオントロジーおよびリンクトデータの構築を行い、それらのデータを SPARQL を用いて外部から検索できるように、SPARQL エンドポイントを公開した。これらの作業を通して、外部のデータリソースである MEDLINE と MeSH の既存 URI を利用した SPARQL での検索を実行できるなどの結果を示すことができたが、 今後の課題となる点も判明した。

Allie データベースのオントロジーを構築する際、PubMed ID のリストを表す PubMedIDList クラスや略語／展開形ペアのリストを表す PairList クラス、略語のリストを表す ShortFormList クラスを定義したが、それに加え、より詳細な記述を目指して owl:oneOf プロパティを用いた列挙を行うことも考えられる。しかし、PubMed ID の URI は 5,176,726 件、略語／展開形ペアは 1,972,298 件、略語は 503,261 件と件数が多く、そのため RDF データのトリプルの数が非常に増えてしまう。今後、このトリプル数増加によるクエリ処理速度への影響を調査し、厳密なオントロジー定義と検索パフォーマンスとのバランスを検討する必要がある。

また、2.93 GHz Quad-Core Intel Xeon, 32 GB RAM のマシンで稼働する Virtuoso v6 データサーバに約 8,600 万のトリプルを収納し、SPARQL エンドポイントを公開したが、複雑なクエリを実行する場合には、結果の取得に非常な時間を要することがわかった。例えば「検索例 (1)」では、検索条件にマッチする略語／展開形ペア 259 件を取得するのに約 50 秒要するが、その略語と展開形の URI も同時に取得するようクエリを変更すると、現状設定の Virtuoso の制限時間 (1000 秒) 以内に結果を取得できない。そこで、今回定義したオントロジーの複雑性や粒度がクエリ処理速度に与える影響を調査し、パフォーマンスを改善できるようなオントロジーを定義できないか、検討する必要がある。また、Virtuoso デー

タサーバの設定ファイル virtuoso.ini のパラメータ変更によるパフォーマンスのチューニングや他の実装なども検討する必要がある。

謝辞

中尾光輝博士には Virtuoso データサーバへのデータベース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] 中尾光輝, 片山俊明: 分散データの統合とセマンティック Web, 情報処理, Vol. 50, No. 9, pp. 836-844, (2009)
[2] W3C SWEO Linking Open Data community, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[3] The Linking Open Data cloud diagram, http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.png
[4] UniProt, http://www.uniprot.org/
[5] BioPAX, http://www.biopax.org/
[6] UniProt RDF, http://dev.isb-sib.ch/projects/uniprot-rdf/
[7] 福田賢一郎: BioPAX : パスウェイデータフォーマットの標準化とオントロジー, 生物物理, Vol. 47, pp. 179-184, (2007)
[8] Bio2RDF, http://bio2rdf.org/
[9] NeuroCommons, http://neurocommons.org/page/Main_Page
[10] LinkedLifeData, http://linkedlifedata.com/
[11] Sharred Names, http://sharedname.org/page/Main_Page/
[12] MEDLINE Fact Sheet, http://www.nlm.nih.gov/pubs/factsheets/medline.html
[13] MEDLINE : Number of Citations to English Language Articles; Number of Citations Containing Abstracts, http://www.nlm.nih.gov/bsd/medline_lang_distr.html
[14] 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)
[15] OReFiL, http://orefil.dbcls.jp/
[16] inMeXes, http://docman.dbcls.jp/im/
[17] Allie, http://allie.dbcls.jp/
[18] PubMed, http://www.ncbi.nlm.nih.gov/sites/entrez/
[19] Hiroko A., and Toshihisa T.: ALICE: an algorithm to extract abbreviations from MEDLINE, J Am Med Inform Assoc, Vol. 12, No. 5, pp. 576-586, (2005)
[20] MeSH, http://www.ncbi.nlm.nih.gov/mesh/
[21] Protégé, http://protege.stanford.edu/
[22] CPAN モジュール RDF::Trine, http://search.cpan.org/~gwilliams/RDF-Trine-0.133/
[23] Virtuoso, http://virtuoso.openlinksw.com/

Appendix 1: テーブル定義

略語情報を管理するテーブル　shortform_freq

カラム名（論理名）	カラム名（物理名）
略語 ID	id
略語	shortform
略語が MEDLINE データに出現する頻度	frequency

展開形情報を管理するテーブル　longform_freq

カラム名（論理名）	カラム名（物理名）
展開形 ID	id
展開形（英語表記）	longform
展開形（日本語表記）	longform_jp
展開形が MEDLINE データに出現する頻度	frequency

略語／展開形ペア情報を管理するテーブル　pairs

カラム名（論理名）	カラム名（物理名）
略語／展開形ペア ID	id
略語	shortform
展開形（英語表記）	longform
展開形（日本語表記）	longform_jp
略語／展開形ペアが MEDLINE データに出現する頻度	frequency
内部管理用の ID（略語／展開形ペアクラスター ID）	cluster_id

略語／展開形ペアクラスター情報を管理するテーブル　clusters

カラム名（論理名）	カラム名（物理名）
内部管理用の ID（略語／展開形ペアクラスター ID）	id
略語	shortform
略語正規形	shortform_normal
展開形（英語表記）	longform
展開形（日本語表記）	longform_jp
略語／展開形ペアクラスターが MEDLINE データに出現する頻度	frequency
略語／展開形ペアクラスターを含む文献 ID	document_id
共起略語（英語表記）	related_shortform_frequency
共起略語（日本語表記）	related_shortform_frequency_jp

文献情報を管理するテーブル　documents

カラム名（論理名）	カラム名（物理名）
内部管理用の ID（文献 ID）	id
PubMed ID	pubmed_id
文献の出版年度	year
文献のタイトル	title

文献の共起略語情報を管理するテーブル　document_shortforms

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
内部管理用の ID（文献 ID）	document_id
文献の共起略語	shortform

略語／展開形ペアの共起略語情報を管理するテーブル　raw_pair_shortforms

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
略語／展開形ペア ID	pair_id
略語／展開形ペアの共起略語	related_shortform
略語／展開形ペアと共起略語が MEDLINE データに出現する頻度	frequency

略語／展開形ペアの出現文献情報を管理するテーブル　raw_pair_documents

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
略語／展開形ペア ID	pair_id
略語／展開形ペアを含む文献の内部管理用の ID（文献 ID）	document_id

略語／展開形ペアの研究分野情報を管理するテーブル　raw_pair_AREAs

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
略語／展開形ペア ID	pair_id
略語／展開形ペアがその研究分野の MEDLINE データに出現する頻度	frequency
研究分野（英語表記）	AREA
研究分野（日本語表記）	AREA_ja

略語／展開形ペアクラスターの共起略語情報を管理するテーブル　pair_shortforms

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
内部管理用の ID（略語／展開形ペアクラスター ID）	cluster_id
略語／展開形ペアクラスターの共起略語	related_shortform
略語／展開形ペアクラスターと共起略語が MEDLINE データに出現する頻度	frequency

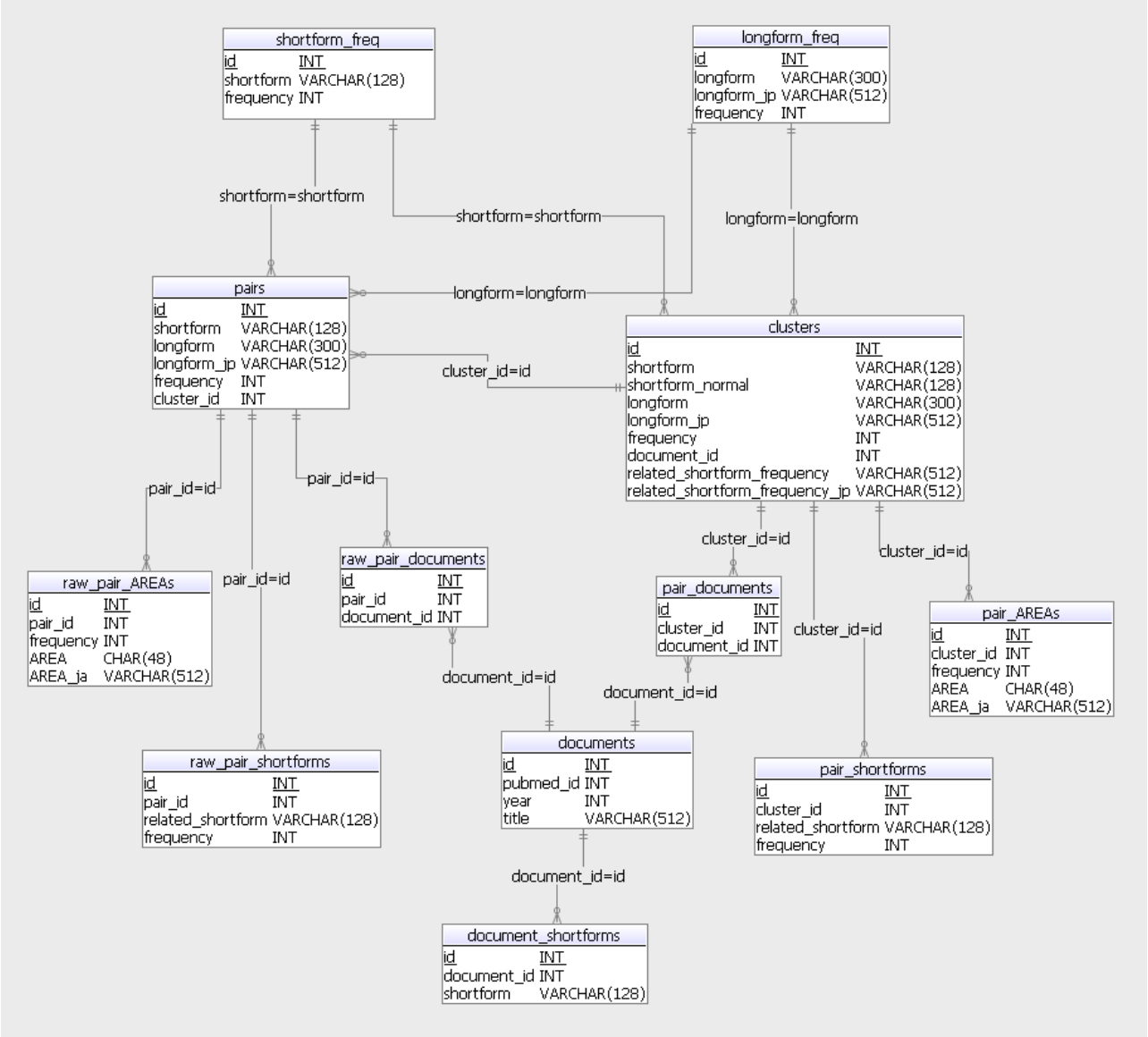
略語／展開形ペアクラスターの研究分野情報を管理するテーブル　pair_AREAs

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
内部管理用の ID（略語／展開形ペアクラスター ID）	cluster_id
略語／展開形ペアクラスターがその研究分野の MEDLINE データに出現する頻度	frequency
研究分野（英語表記）	AREA
研究分野（日本語表記）	AREA_ja

略語／展開形ペアクラスターの出現文献情報を管理するテーブル　pair_documents

カラム名（論理名）	カラム名（物理名）
内部管理用の ID	id
内部管理用の ID（略語／展開形ペアクラスター ID）	cluster_id
略語／展開形ペアクラスターを含む文献の内部管理用の ID（文献 ID）	document_id

Appendix 2: ER 図



Appendix 3: オントロジー

[Classes]

ShortForm : MEDLINE データベースから取得した略語

LongForm : MEDLINE データベースから取得した略語の展開形

Pair : MEDLINE データベースから取得した略語／展開形ペア

PairCluster : 略語／展開形ペアのクラスター

ExternalResource : Allie 外で構築、定義されているリソース

PubMedID : PubMed ID

MeSHTerm : MeSH タームの ID

ResearchArea : MeSH タームで表現した研究分野

List : リスト

PairList : 略語／展開形ペアのリスト

PubMedIDList : PubMed ID のリスト

ShortFormList : 略語のリスト

CooccurringShortFormList : 共起略語のリスト

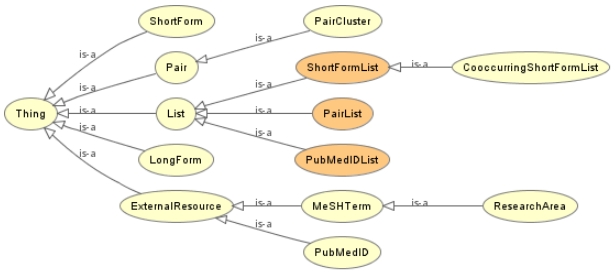


図 1　クラス階層図

[Object Properties]

hasShortFormOf: (属性として)略語を持つ

　ドメイン : Pair、レンジ : ShortForm

hasLongFormOf: 展開形を持つ

　ドメイン : Pair、レンジ : LongForm

hasShortFormRepresentationOf: 略語の代表表現を持つ

　ドメイン : PairCluster、レンジ : ShortForm

hasLongFormRepresentationOf: 展開形の代表表現を持つ

　ドメイン : PairCluster、レンジ : LongForm

inResearchAreaOf: 関連研究分野を持つ

　ドメイン : Pair、レンジ : ResearchArea

appearsIn: 出現する書誌情報のリストを持つ

　ドメイン : Pair、レンジ : PubMedIDList

contains: 略語／展開形ペアのリストを持つ

　ドメイン : PairCluster、レンジ : PairList

cooccursWith: 共起する略語のリストを持つ

　ドメイン : Pair、レンジ : CooccurringShortFormList

hasMemberOf: 構成メンバーを持つ

　ドメインが PairList の場合、レンジは Pair

　ドメインが PubMedIDList の場合、レンジは PubMedID

　ドメインが ShortFormList の場合は、レンジは ShortForm

[Data Property]

frequency: 出現する頻度

　レンジ : int 型