1N4-OS-10b-1in

マッピング生成・更新機構を持ったSPARQL処理フロントエンドによる異種LODの横断的検索システムの試作

A Preliminary Approach on Federated Querying to Heterogeneous LODs by SPARQL Front-end with Mapping Generation and Updating Mechanism

野口 宙毅[*1] 藤野 敬久[*1] 福田 直樹[*1]
Noguchi Hiroki Takahisa Fujino Naoki Fukuta

[*1]静岡大学大学院情報学研究科
Graduate School of Informatics, Shizuoka University

When we make a query for the semantic data by using SPARQL, we need to know how the data are modeled by their underlying ontology. We may consider multiple ontologies and their differences when we make a federated query to multiple LODs. In this paper, we present a preliminary approach on federated querying to heterogeneous LODs by SPARQL front-end with mapping generation and updating mechanism. Our system enables the application developers to construct effective queries without deeply considering the differences among underlying ontologies.

1. はじめに

SPARQL を利用して意味に基づくデータを検索するためには，その SPARQL クエリは検索対象のデータストアのオントロジーに基づいたものである必要がある．複数の LOD データストアから横断的な検索を行おうとした場合，それぞれのオントロジーが異なる概念化を行なっている可能性がある．そうした検索を可能とするためにはオントロジーマッピングが必要となる．本研究ではオントロジー間のマッピングの生成と更新機構を持ち，SPARQL による横断的な LOD 検索を実現する SPARQL 処理フロントエンドの試作について述べる．複数のデータセットを組み合わせて利用することは，データセットによって用いられているデータ形式が違うという問題と，オントロジーの差異という問題が存在するために困難である．

本研究では，これらの課題を解決するために，自動マッピング生成・更新機構を持った SPARQL 処理フロントエンドの実現を検討する．SPARQL クエリ内の処理に，マッピング生成時に得られた信頼度情報を効果的に利用するための手法として，藤野らによる SPARQLoid[Fujino 12a] がある．本研究では，マッピングの自動生成・更新機構と SPARQLoid を組み合わせることで，植物関連 LOD を横断的に検索できるようにし，その検索結果にマッピングの信頼度情報を効果的に利用できるようにすることで，マッピングの精度が必ずしも高くない状況でも，期待した問い合わせ結果を得られやすくなるようにすることを目指している．

2. 研究の背景

2.1 Linked Open Data とその提供事例

Linked Open Data（LOD）のデータモデルは，セマンティックウェブと同じく基本的に RDF で記述される．LOD の代表的なものに，Wikipedia を Linked Data 化した DBpedia がある．また，アメリカの data.gov を筆頭に海外では，「オープンガバメント」という理念のもと，様々なデータの LOD 化が進んでいる．その他にも多種多様なデータセットが存在しており，2011 年 9 月の段階で，295 のデータセットと 310 億以上の RDF トリプル，5 億以上の RDF リンクが存在[*1]し，現在も増加を続けている．

植物に関するオントロジーには，Plant Ontology Consortium による Plant Ontology[*2]や，理研による植物統合データベース[*3]，ウィスコンシン大学の Peter J. DeVries による GeoSpecies Knowledge Base[*4]などがある．Plant Ontology は，2012 年 12 月時点で 1609 個のクラスを保持し，OBO，OWL，TBL の 3 つのフォーマットで提供されている．植物統合データベースでは，44 個のクラスを保持し，OBO，NT，OWL，TTL で提供されている．GeoSpecies Ontology は 86 個のクラスを保持し，OWL で提供されている．

Wikipedia 上のデータを LOD 化した DBpedia には植物に関する情報も含まれており，他の LOD にも，植物に関わる内容を含んだものは，複数存在すると考えられる．DBpedia 上にある情報を利用したアプリケーションの例としては，川村の「花咲かめら」[Kawamura 12] などがあり，多くの応用が考えられるが，一方で，植物に関する学術的な知見などを含んだオントロジーなどと DBPedia などの大規模 LOD との相互利用は，必ずしも容易な状況ではない．

DBpedia では，DBpedia2.0 が公開された 2007 年 7 月から最新版である DBpedia3.8 が公開された 2012 年 5 月までの約 5 年間で 11 回更新されている．

2.2 Ontology Mapping

Semantic Web のような，広く開かれたシステムでは，データやオントロジーの異種性 (Heterogeneity) への対処が課題となる．その解決方法として，オントロジーマッピングに関する様々な検討が進められている [Euzenat 07]．オントロジーマッピングを行うための手法のみでなく，それを具体的に行うソフトウェアとしても様々なものが提供されており，たとえば LogMap[Jiménez-Ruiz 11] や，Alignment API[*5]などがある．

LogMap は，推論器（reasoner）として HermiT[*6]を採用し

連絡先: 野口宙毅, 静岡大学大学院情報学研究科, 〒 432-8011
静岡県浜松市中区城北 3-5-1, gs13025@s.inf.shizuoka.ac.jp

∗1 http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
∗2 http://www.plantontology.org/
∗3 http://ja.biolod.org/database/ria301i/The_integrated_database_of_plant_omics_data
∗4 http://lod.geospecies.org/
∗5 http://alignapi.gforge.inria.fr/
∗6 http://hermit-reasoner.com/

ており，推論に基づく高性能なマッピング生成を実現している反面，その適用には対象となるオントロジー自身が論理的に矛盾なく記述されている必要があり，OWL2 Datatype Maps 以外のデータタイプに対応させるために OWL2 Datatype とのマッピングを事前に定義しておく必要があるなど，対象となるオントロジーに事前の処理が必要になる場合がある．たとえば，上述の DBpedia オントロジーと Plant Ontology をマッチングさせようとした場合，DBpedia のオントロジーに存在する $http://www.w3.org/2001/XMLSchema\#gYear$ などいくつかのデータタイプが OWL 2 Datatype Maps に定義されていないことが原因で，マッピングの生成に失敗してしまうことも実際に確認した．また HermiT のような推論器をマッピング生成のために動作させる必要があるため，精度は高いものの処理に時間がかかることがあり，オントロジーの概念数が多いなど，スケールの大きいオントロジーのマッピングのために推論器を使用せずにマッピングを行う LogMapLT も提案されている．LogMapLT を用いれば，推論による処理時間の問題を回避できるものの，生成されたマッチングの精度は低下してしまう．

AlignmentAPI は，文字列の類似度などに基づいて API の利用者が柔軟にマッチング計算方法を設定してマッピングを行うことができるため，その計算は比較的高速であるものの，マッピング精度は必ずしも高くない．たとえば，特別な前処理やチューニングなどを行わずに AlignmentAPI を単純に適用した場合，大文字と小文字の違い以外は，ほぼクラス名が完全一致したペアしか返さない，といったような現象が生じる．

植物オントロジーのマッピングには，学名の曖昧性という問題がある．学名は統一されることが理想的で，一つの種に対し一つの固有の学名がつけられることが望ましいが，植物では同物異名 (シノニム)，異物同名（ホモニム）という問題が多く発生している．これには，植物の学名に関する議論や決定が，植物の「種」に関する研究の進展に伴って変化するという性質に起因するという理由もある．たとえば，クスノキ科スナヅル属の学名は *Cassytha* である．そして，サボテン科リプサリス属の学名は *Rhipsalis* であるが，同時に *Cassytha* という学名も使われる．また，地域によっても学名の不一致という問題があり，たとえばタデ科のハルタデは 6 つものシノニムを持っている．また，日本全土の植物を網羅した最新の植物誌は存在しない [米倉 11]．本研究のようにシノニムとホモニムを多く含むなど，上述のような課題を持つ植物のオントロジーを扱う場合には，単純に AlignmentAPI を適用するなどの方法では，必ずしも十分な精度を持ったオントロジーマッピングを生成できない．

図 1 に横断的検索の例を示す．ユーザが「日本に生えていて赤色の花が咲く遺伝子を持つ植物を知りたい」とする．従来の方法ではユーザは，それぞれのエンドポイント (今回の例では DBpedia と BioLOD) それぞれのオントロジーを理解して，そのオントロジーに基づいたクエリを書く必要がある．本研究で提案するシステムでは，ユーザが用意したオントロジーに基づいたクエリを変換することにより，横断的クエリを可能とする．

マッピング生成を支援するシステムとして，Euzenat は Alignment Server[David 11] を提案した．Alignment Server は，利用可能なアラインメントやアラインメントを探すメソッドを共有するのが目的である．

より精度・再現性の高いマッピングを発見することを目的に，Aguirre による Ontology Alignment Evaluation Initiative(OAEI)[*7]では，オントロジーマッピング手法の競技を行なっており，様々なデータを想定して競技を行なっている．OAEI2012[Augirre 12] では，23 種類のメソッドがエントリーされた．

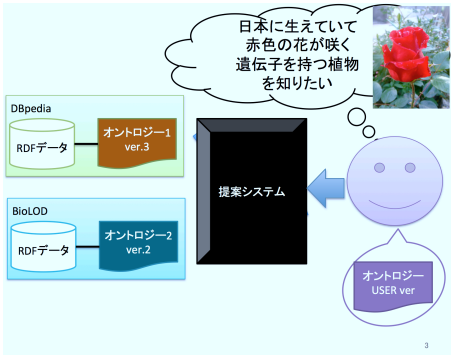

図 1: 横断的検索の例

3. SPARQLoid

SPARQLoid は，信頼度を用いたオントロジーマッピングを利用する SPARQL クエリを生成するシステムである．そのクエリをユーザのアプリケーションの中で容易に利用できるようにする目的で，Java のソースコードの断片も生成可能としている．また，変換されたクエリを実際に動作させて，その検索結果を確認できる実行・テスト環境も用意している．SPARQLoid を用いてクエリを実行させる際のシステム構成図を図 2 に示す．図 2 における Ontology Alignment Provider のためのマッピングを用意する際の課題として、本論文ではマッピングの生成・更新機構を試作している．

SPARQLoid の説明として，以下のクエリを取り上げる．このクエリは，二つのエンドポイントにまたがるデータを取り出すクエリである．このクエリの中で用いられている概念 *my:*オリジナル曲は，検索対象に用意されているものではない．こうしたクエリを行う際には，これらの概念に対するマッピングが必要となる．これらのマッピングを用意する目的で，本論文では，マッピングの生成・更新機構を提案する．SPARQLoid におけるマッピングのコントロールのために，RANKING 句と THRESHOLD 句を利用する．このクエリの場合，THRESHOLD 句における *my:*オリジナル曲=0.3 によって，マッピング生成時の信頼度が 0.3 以上のマッピングがクエリの際に用いられるようになる．また，RANKING 句では，それぞれのマッピングの信頼度を重み付けしており，*my:*オリジナル曲が最も重視されるマッピングとなっている．これらのクエリの詳細な記述方法については，[Fujino 12b] を参照されたい．

```
select ?x ?singer where {

?x rdf:type my: オリジナル曲.
?x my:hassinger ?singer

 SERVICE<http://otherEndpoint/sparql>{
        ?singer my:singerName ?singerName.
        FILTER regex(?singerName, "^ 初音ミク")
 }

RANKING{my: hassinger*0.4+
my: オリジナル曲*0.6+my: singerName*0.3}
THRESHOLD{my: hassinger=0.6,
my: オリジナル曲=0.3,my: singerName=0.2}
}
```

*7 http://oaei.ontologymatching.org/

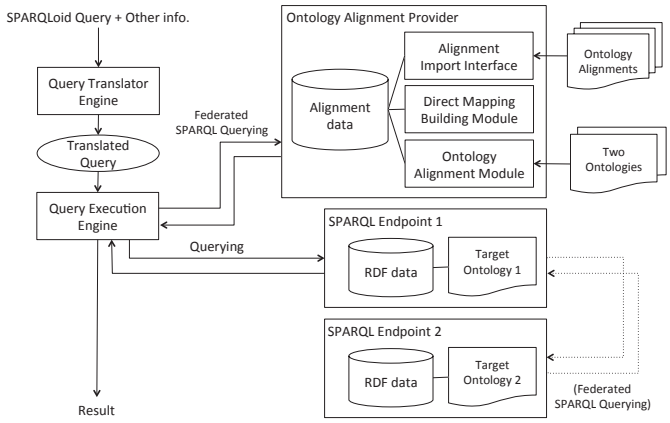

図 2: SPARQLoid Architecture

```
LIMIT 150
```

Listing 1: SPARQLoid query

4. 提案手法

我々は，マッピング生成・更新機構を持った SPARQL 処理フロントエンドによる異種 LOD の横断的検索システムを実現する．異種 LOD の横断的検索の実現にはマッピングの生成・更新が必要であるため本研究では，オントロジーマッピングの自動生成更新機構 [野口 12] を制作した．

SPARQL クエリの変換に用いるオントロジー間のマッピングを用意する手法としてオントロジーの生成更新機構の試作を行う．SPARQLoid は，異種オントロジーマッピングが存在する場合に，それらをユーザ自身がコントロールし利用するマッピングやランキングの指標を指定することを可能としている．

本システムの全体を図で表すと図 3 のようになる．

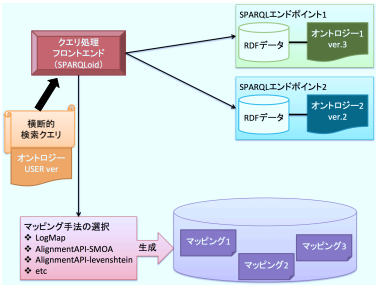

図 3: システムの全体図

本システムの全体図内のオントロジー生成構成を図で表すと図 4 のようになる．ユーザは自分が用意したオントロジーに基づいた横断的クエリを書き，フロントエンドへ送る．フロントエンドは，クエリ変換部により検索対象のオントロジーに基づいたクエリに変換される．変換に用いられるオントロジーマッピングは，マッピング更新生成部により用意される．クエリの変換部には，SPARQLoid を用いた．SPARQLoid はクエリを変換するエンジンである．マッピングの更新・生成部は，エンドポイントクロール機構がエンドポイントのオントロジーを監視することにより実現する．

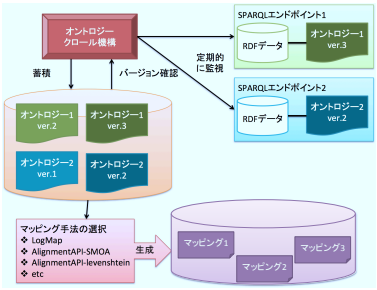

図 4: オントロジー生成更新部

従来の，エンドポイントが更新される度に新たなマッピングを 0 から生成するという手法では，時間的なロスが非常に大きいと考えられる．そのため本研究では，過去のオントロジーと最新のオントロジーの差分によってマッピングを生成することにより，短時間でのマッピング生成を可能とする．

この方法では，マッピングの精度の低下への対処と，生成されたマッピングに人手で必要な修正などを加えることが難しいという点が課題となる．また，複数のオントロジー間のマッピングが存在する場合，直接的なマッピングがなくても，オントロジー間でのマッピングから間接的なマッピングを得ることは可能であるが，その場合，オントロジーマッピングが複数通り存在することになり，どのマッピングをどの問題 (たとえば，SPARQL クエリの処理など) に用いればよいのかを，適切に選択する必要も出てくる．

オントロジーの更新を確認し，オントロジーが新しいものになっていた場合は，過去のオントロジーに対応した既存の信頼度の高いマッピングと比較して，新しいマッピングを自動的に生成する．

過去のオントロジーやマッピングと最新のオントロジーとを比較するために，本システムでは過去のオントロジーとマッピングは独自に本システム内に保持しておく．データのフォーマットは，元のデータと同じフォーマットで，タイムスタンプを添える．これは，大半のエンドポイントは過去のバージョンは公開されているが，それがなされていないエンドポイントの存在を考慮したからである．オントロジーの確認間隔は，オントロジーの更新間隔に比べて十分に短い n 間隔とする．本研

究では，n はここで用いるオントロジーの過去の更新の平均間隔の $n = 1/10$ とした．

本研究では，オントロジーマッピングの生成手法は，ユーザが選択可能としている．これは，オントロジーの内容や規模などによって適切な手法が異なるからである．例えば，上述の LogMap は推論に基づく高性能なマッピング生成を実現している反面，その適用には対象となるオントロジー自身が論理的に矛盾なく記述されている必要がある．本システムでは，どのような場合にどのマッピング生成手法が選ばれるのかについて，過去のユーザの選択履歴に基いてデフォルトの選択肢を学習しておき，ユーザが特別にマッピングの生成手法を指定しない場合は，システムが自動的に適当な生成手法を適用してマッピングを生成することを検討している．さらに，本システムでは，大規模なオントロジー間でのマッピング生成にかかる処理の負荷の大きさを考え，複数のコンピュータに接続し，負荷が最もかかっていないコンピュータをシステムが自動的に選択して利用する機構についても検討中である．

5. 実装

マッピング生成の手法には，AlignmentAPI と LogMap を用いた．AlignmentAPI は文字列一致度の計算方法を複数用意しており，場合によって違った手法を用いることを可能としている．例えば，N-Gram 法や，Levenshtein Distance 法，SMOA 法などである．本研究で試作したシステムでは，AlignmentAPI から Equal と Levenshtein と SMOA を選択することを可能とした．

図 5,6 にシステムの実行例を示す．図 5 に入力を行なってからオントロジーの更新を確認するまでのプロセスを示しており，図 6 にオントロジーの更新を確認してから新しいマッピングを生成するまでのプロセスを示している．

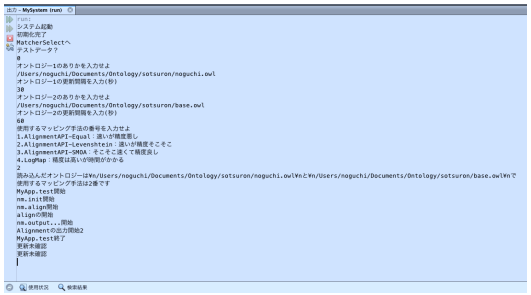

図 5: 動作例 1：入力〜通常時

6. まとめと今後の課題

本研究では，マッピング生成・更新機構を持った SPARQL 処理フロントエンドによる異種 LOD の横断的検索システムの試作を行った．クエリの変換に用いるオントロジー間のマッピングを用意するためにオントロジーの生成更新機構の制作を行った．横断的検索には SPARQL クエリの変換を行うことにより可能とした．

参考文献

[David 11] David, J., Euzenat, J., Scharffe, F. and Trojahn dos Santos, C.: The Alignment API 4.0, *Semantic web journal 2(1):3-10*, 2011.

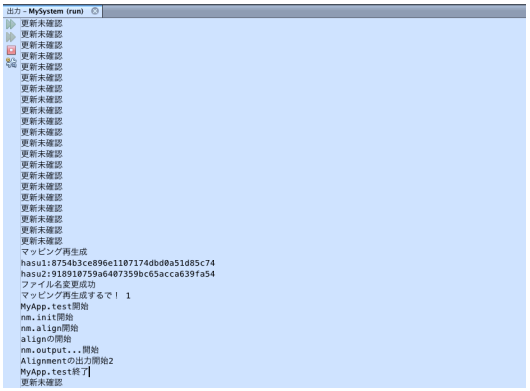

図 6: 動作例 2：オントロジーの更新を確認

[Euzenat 07] Euzenat, J. and Shvaiko, P.: Ontology Matching, *Springer-Verlag*, 2007.

[Fujino 12a] Fujino, T. and Fukuta, N.: SPARQLoid - a Querying System using Own Ontology and Ontology Mappings with Reliability, *Poster & Demo Notes of The11th International Semantic Web Conference 2012 (ISWC2012)*, 2012.

[Fujino 12b] Fujino, T. and Fukuta, N.: A SPARQL Query Rewriting Approach on Heterogeneous Ontologies with Mapping Reliability, *Proc. of the IIAI International Conference on Advanced Applied Informatics(IIAI-AAI 2012)*, pp.230–235, 2012.

[Jiménez-Ruiz 11] Jiménez-Ruiz, E. and Grau, B, C.: LogMap: Logic-based and Scalable Ontology Matching, *Proceedings of 11th International Semantic Web Conference (ISWC 2011)*, Part I, LNCS 7031, pp.273-288, 2011.

[Kawamura 12] Kawamura, T.: Toward an ecosystem of LOD in the field: LOD content generation and its consuming service, *Proceedings of 11th International Semantic Web Conference (ISWC 2012)*, Part II, LNCS 7650, pp.98-113, 2012.

[Augirre 12] Aguirre, J. L., Eckert, K., Euzenat, J., Ferrara, A., Van Hage, W. R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Zamazal, O. Šváb, Trojahn, C., Jiménez-Ruiz, E., Grau, B. C. and Zapilko, B. Preliminary results of the Ontology Alignment Evaluation Initiative 2012. In *Proc. of 7th Ontology Matching Workshop (OM2012), at International Semantic Web Conference (ISWC2012)*, 2012.

[野口 12] 野口宙毅，福田直樹：動的なオントロジーマッピング手法の適用による植物オントロジーの LOD 化の検討，情報処理学会第 *75* 回全国大会，2012.

[米倉 11] 米倉浩司：BG Plants 和名 – 学名インデックス（YList）の利便性と限界，*21* 世紀の生物多様性研究ワークショップ *2011*，2011.