5S-9

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

野口　宙毅†　　福田　直樹†

†静岡大学情報学部

1　はじめに

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．

本研究では前者の問題に対しては，データをコンピュータが処理しやすい形で共有するためのLinked Open Dataという形式にし，後者の問題に対しては，異なるオントロジー間の対応関係を自動的に導出するOntology Mapping技術の手法を，グラフに基づく手法や，知識資源に基づく手法など複数を動的に適用して植物の統合データのLOD化の検討を行う．

2　研究の背景

2.1　植物における Linked Open Data

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．

Plnt 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には当然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，多くあると考えられる．DBPedia上にある情報を利用したアプリケーションとしては，たとえば川村の「花咲かめら」[4] 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．その理由を次に述べる．

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 [1]．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[3] や，Alignment API[5]などがある．

LogMapは，推論器（reasoner）として HermiT[6]を採用しており，推論に基づく高性能なマッピング生成を実現している [3] 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 map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．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，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

†Hiroki NOGUCHI　†Naoki FUKUTA
†Faculty of Informatics, Shizuoka University

[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/

にマッピングを行うLogMapLTがある．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，先の例では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．学名は統一されるべきもので，一つの種に対し一つの固有の学名がつけられるべきであるが，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[5]．

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．

3　提案手法と今後の課題

Web上のオントロジーは，それぞれ独立に管理され，それぞれが個別に随時更新されている．それらをLOD化して使用しようとすると，複数のオントロジーを繋ぐマッピングを事前に用意しておくとしても，オントロジーの更新の内容によっては，事前に用意していたマッピングが使用できなくなる場合がある．特に植物の分野では，新たな発見により種の分類体系が大きく見直される可能性があるため，この課題についての対処は重要である．この課題に対処する1つの方法としては，オントロジーの更新に追従して自動的にマッピングを再生成するという方法が考えられるが，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．

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考える．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[2]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．現在、本提案手法の実装を進めており，提案手法の評価及び他のアプリケーションへの具体的な利用は，今後の課題である．

参考文献

[1] Jérôme Euzenat and Pavel Shvaiko.: Ontology Matching. Springer-Verlag. 2007.

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

[3] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau: LogMap: Logic-basedandScalableOntologyMatching, Proceedings of 11th International Semantic Web Conference (ISWC 2011), Part I, LNCS 7031, pp.273-288, 2011.

[4] Takahiro Kawamura. 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.

[5] 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．