【課題番号】課題 3

【課題名】トリプルストアのドキュメントやサンプルクエリ生成支援システム

１）課題とその背景

バイオサイエンスデータベースセンター(NBDC)の統合化推進プログラムでは、生命科学分野のセマンティック・ウェブによるデータベースの標準化が進められている。その結果、様々なデータベースがセマンティック・ウェブの標準データモデル Resource Description Framework (RDF)を用いて構築されてきており、それらのデータに対する柔軟なアクセス手段として RDF の標準検索言語 SPARQL Protocol and RDF Query Language (SPARQL)による検索サービス(エンドポイント)を提供することになっている。エンドポイントを利用することで利用者は自由に SPARQL クエリを記述して目的のデータだけを効率良く取得することができるが、従来の RDB におけるスキーマに相当する情報(メタデータ)が分からないことには適切なクエリを書くことが出来ない。このため、利用者がエンドポイントを介して容易にデータベースにアクセスするためには分かり易いメタデータの提供が必須となる。しかし、RDF データベースの構築とともに分かり易いメタデータを提供し、データの更新に併せて最新の状態に保つのは手間ひまのかかる作業である。対象データベースが複雑であったり多数であったりするとなおさらである。

実際には RDF のデータでは述語 (predicate) に rdf:type や owl:Class などの標準的な語彙が使われていることが多く、これを手がかりにすれば機械的な処理だけである程度対象 RDF データの構造やそれに付随する統計情報を取得できる。そこでこれらの情報を基にして自動的にメタデータを提供する効果的な方法を開発することにより、SPARQL クエリの構築を容易にできると考えられる。さらに、データ構造を可視化したり、対象 RDF データベースに併せたクエリ例を構築したりすることが自動的に行えるようになれば統合化推進プログラムのデータが有効に再利用しやすくなる。

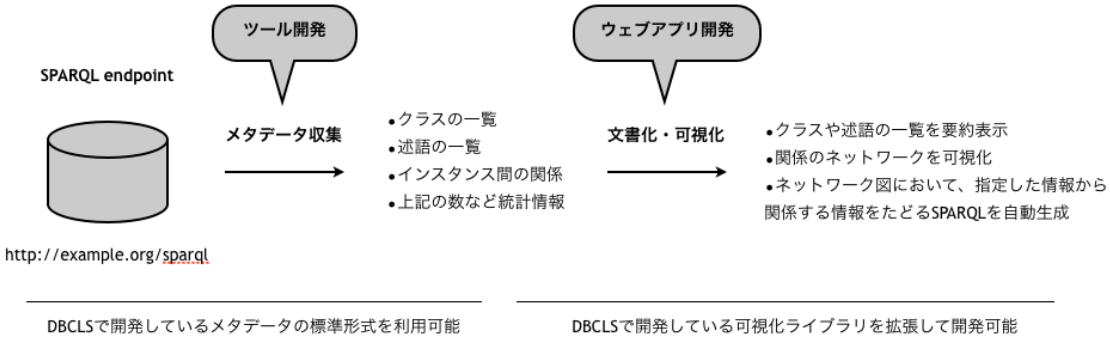
２）課題の解決方法の概要

RDF データベースに対するメタデータの自動構築に関するツールは Neologism や LODE、そして広く使われているオントロジーエディタの Protege に対するプラグインなど幾つか存在しているが､統合化推進プログラムにて開発されつつある RDF データセットのように、多数のクラスがある場合のスケーラビリティや、インスタンスをスキーマとして扱えないなど、表現能力の点で問題がある。また、特定のクラスに含まれるインスタンスの数などの統計情報を考慮したものも存在しないほか、SPARQL クエリ例の自動構築まで行うツール

は存在していない。

本トライアルでは、対象 RDF データベースの構造を解析し、利用者に有益なメタデータを構築するツールを開発する。具体的には、対象データベースに含まれるクラスの一覧や、それらの間を結ぶ述語に関する情報といった、典型的なデータ表現を可視化するとともに、テキストを用いた説明を生成するツールの開発を目指す。一般的にデータベースは、ある固有の ID がつけられた概念(例えば遺伝子)に対して、複数の属性(例えば生物種やゲノム上の位置やその機能アノテーションなど)がつけられていることから、これらのデータに関する上記のようなメタデータを定形的な表現で利用者に提示できれば、SPARQL クエリを生成する際に有益であると考えられる。

３）課題の解決方法の概略図



※DBCLS：ライフサイエンス統合データベースセンター