論文の内容の要旨

論文題目

A study on semantic integration of biological databases with the Semantic Web technologies.

（Semantic Web 技術を用いたバイオデータベースの意味的統合に関する研究）

氏名：川島秀一

生物学は、アリストテレスの時代より、対象を観察し記載する学問であった。20 世紀半ば以降、分子生物学が大きく発展し、また、様々な観察・測定技術が開発されることで、現代に至るまで記載される生物学分野の情報の量は加速度的に増加している。生物を要素還元主義的に理解しようとするアプローチから、近年ではシステム生物学のように細胞や個体を様々な要素から構成される複雑系として理解しようとする方法や、大量のデータから機械学習技術等を駆使して新しい知識を発見する分野も台頭しているが、それゆえに、情報を適切に記載する必要性は以前にまして重要になっている。1970 年代以降、生命科学の情報は、データベースとして編集されるようになり、また 1990 年代以降インターネットが著しく普及してからは、World Wide Web (WWW)上でデータベースを公開することが一般的になってきた。もともと、生命科学データベースは、Protein Data Bank (PDB) や GenBank に代表されるように、個々のタンパク質や遺伝子などに関わる情報を一つのエントリーとして編纂し、複数のエントリーを集めたフラットファイル形式で配布することが一般的であった。データベースをウェブ上で提供することが一般的になってからは、リレーショナルデータベース管理システム（RDBMS）上に複数のテーブルとして格納した情報から、エントリーに対応するようなビューを、ウェブブラウザ上に動的に生成することも普通になってきた。すでに数万という数でデータベースが開発されており、生命科学研究を遂行する上で必要不可欠な存在になっている。その一方で、生命科学データベースの数や種類が多すぎる上、使われている情報技術や記述フォーマットも多岐に渡り、使われている語彙も共通しておらず、データに関するデータ（メタデータ）も不十分なことが多い等、様々な問題があることも指摘されるようになった。このことが、複数のデータベースを統合して使うことの妨げになっている。そこで、セマンティック・ウェブ技術を応用することで、これらの問題点を解決しようとする研究が、近年精力的に行われている。セマンティック・ウェブとは、一連

の標準技術仕様を基に、意味のレベルで機械的な処理が行えるような情報を WWW 上に記述・流通させるための概念である。

本論文では、セマンティック・ウェブ技術を用いて、複数のデータベースを統合して利用するための手法に関する研究を行った。まず、はじめに、古典的なフラットファイル形式のデータベースを紹介するために、具体例として、20 種の生体アミノ酸に関する物理化学的な指標を集めた AAindex データベースについて解説する。AAindex はそれ自体有用なデータベースとして一定の評価を得ているが、独自の統合検索システム上に実装されている以外には、他のデータベースと統合して利用する仕組みはない。

次に、複数のデータベースを統合して利用するという観点から、セマンティック・ウェブ技術の利点を紹介する。セマンティック・ウェブでは、情報の記述を Resource Description Framework (RDF)を用いて行う。RDF では、ものを識別するのに Uniform Resource Identifier (URI)を用いる。URI は、世界的にユニークな識別子なので、記述する情報が曖昧になることがない。また、Web Ontology Language (OWL) で記述されたオントロジーを利用することで、データベース間で語彙のレベルにおいても共通化することが図れる。しかし、セマンティック・ウェブでは、データベースの相互運用性を高めるための仕組みは準備されているが、RDF やオントロジーを用いて、どのように情報を記述するのかについての規則はない。そのため、RDF データセット間で、(1) 現実のデータベースでは、しばしば複数の異なる URI が同じリソースを指すことがあり、その場合、同じリソースが別の URI で記述されうる、(2)同じ概念に対して、異なるオントロジーのクラスやプロパティが存在する、(3) 同じ情報が、異なる RDF モデルで記述される、等の問題がある。このことが、RDF データセット間で本質的には可能なはずの検索が、実際にはできないという問題を引き起こしている。そこで、本研究では、(1)〜(3)の問題点を解決する上で実用上重要な点に関して、ガイドラインを提案した。RDF を構築する際に本ガイドラインに準拠することで、RDF データセット間が意味のレベルにおいても標準化されるようになり、検索の精度を高くすることができる。

さらに National Bioscience Database Center（NBDC）の協力を得て、NBDC RDF portal という名前の RDF データセットのリポジトリを開発し、インターネット上でサービスを公開した（https://integbio.jp/rdf/）。NBDC RDF portal は、投稿された RDF データの一覧と、RDF ファイルのダウンロードサービスおよび SPARQL エンドポイントの提供を行う。本サービスでは、データセットの登録に際し、事前に上記ガイドラインへの準拠という観点からレビューを受けることを条件としている。これまでに、NBDC RDF portal に登録した、21 の RDF データセットは、すべてレビューを行い、必要に応じて修正を行ったものになる。また、レビュー作業を網羅的に行うために、一部のガイドライン項目に対する検証ツールを開発し

自動化も行った。その結果、異なる研究グループが個別に構築した RDF データセット間、さらには既存の RDF データセットとも、より高い相互運用性を実現することができた。NBDC RDF portal は、2018 年 11 月現在、遺伝子配列、タンパク質立体構造、エピゲノム、ガンゲノム、糖鎖生物学、化学化合物、トキシコゲノミクス等、様々な研究分野の RDF データセットを収録しており、その RDF トリプル数は 455 億件を超える巨大なサービスに成長している。今後、データサイエンスの有用な基盤リソースとして貢献できると期待している。