成果報告⑪ **10月6日16:35―16:55**

生命科学分野における質問応答システムの実現に向けて

金　進東
（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　特任准教授）

世界的に広く利用されている生命科学分野の学術論文検索サービス PubMed/MEDLINE は 1800 年頃以降に発表された 2600 万件を超える書誌情報を検索対象にしており、毎日平均 3000 件以上登録されている。このことから示唆される通り、生命科学は長い歴史を持ち、非常に多くの知見が蓄積され、そして今なお発展し続けている。これまでに得られた知見は主に文献として収められているほか、様々な切り口で多くのデータベースが構築されている。塩基配列情報やタンパク質情報、疾患情報などその数は 1600 をくだらない。生命科学という極めて広い研究領域において、生命の仕組みを解き明かすために必要な知識は増え続けており、様々なデータベースから効率よく必要な情報を取得しなくてはならない。

このような状況において複数のデータベースを横断的かつ合目的的に検索しやすくする基盤技術として、Resource Description Framework（RDF）というデータ表現方法が注目されている。各データは Uniform Resource Identifier（URI）を用いて識別され、互いに関係づけられることから、ウェブ上において曖昧なく情報を表現できるほか、機械によるアクセスと処理が容易になる。その結果、複数のデータベースから特定の値をキーとして必要な情報を効率的に収集できるなどの利点がある。このような基盤技術環境を実現するデータベースは Linked Data と呼ばれ、European Bioinformatics Institute（EBI）や National Center for Biotechnology Information

（NCBI）において提供されているデータベースもこれを採用し始めている。また、NBDC の RDF ポータルに集められている RDF データセットについても同様である。

これらの技術的優位性を生命科学研究の発展のために発揮させるには、より多くの生命科学研究者がその有用性を実感できる形で提供されなくてはならないが、現状では必ずしも実現されているとはいえない。たとえば、RDF データベースから必要な情報を取得するためには SPARQL と呼ばれる問い合わせ言語を用いることが一般的であるが、これを使いこなすまでには時間がかかるため、なかなか普及していない。そこで DBCLS では別の方法でアクセスできる技術の開発に取り組んでいる。その一つが質問応答システム LODQA である。すなわち自然言語による質問を機械的に SPARQL に翻訳して各種データベースに問合せし、情報を収集して結果を提示するものである。自然言語は言うに及ばず、SPARQL も汎用的で表現力の高い言語であるから、ひとたび RDF データとして表現しさえすれば、すべてのデータに対して自然言語で問い合わせできるようになる。本発表ではこの LODQA を紹介しながら、それを構成する様々な関連技術を紹介する予定である。