研究開発課題別事後評価報告書

１．研究開発課題名

データベース統合に関わる基盤技術開発

２．代表研究者名

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
センター長　小原　雄治

３．研究実施概要

本課題では、大規模集中型の統合ではなく、セマンティック・ウェブ技術を利用した「フェデレーション型」のデータベース統合を目指し、各種データベース、ツール等の基盤技術の開発を行った。

１）データベースのRDF による統合化

表形式の有用なライフサイエンスのデータを容易にデータベース化し公開するためのシステムである「TogoDB」について、データベース化されたデータをResource Description Framework (RDF)に変換し、統合的に利用しやすくするための機能追加を行い、さらに、SPARQLクエリ言語を容易に扱えるようにするため、自然文からSPARQLクエリを自動生成するシステムである「LODQA」を開発した。

２）解析プラットフォームによる統合利用環境の整備

RDF基盤を利用して複数のツールを組み合わせた解析をスムーズに行うため、解析プラットフォームである 「DBCLS Galaxy 」に RDF 入出力のフレームワーク Semantic Automated Discovery and Integration（SADI）対応機能を追加した。

３）インターネットを活用した高度検索技術の開発

「TogoGenome」において、生物種とゲノムに関連する多種多様な情報を集約し、ゲノム情報の統合的で新しい検索システムを構築するとともに、SPARQL 検索の結果を HTML に可視化するためのフレームワークで「TogoStanza」サーバーによる再利用可能なデータと可視化パーツを提供した。また、ゲノム情報のためのオントロジー（整理された共通語彙）とRDF（世界標準のデータ形式）データモデルの設計等を国際連携にもとづき実施した。さらに、統合化推進プログラムの各種データベースとの連携や、実験研究者用にLODネットワークを容易に利用可能とするための「TogoTable」の開発も行った。

４）RDF化に資するオントロジー、辞書、コーパスの整備、標準化

生命科学データのRDF化に資する重要な言語資源の管理のため、オントロジーについては、オントロジー登録に広く利用されるサイト「BioPortal」を活用の他、「BioPortal」オントロジーの可視化、再利用を助けるシステムとして「OntoFinder」、「OntoFactory」、「OntoCloud」といったシステムを開発した。また、辞書と文献アノテーションは、「PubDictionaries」と「PubAnnotation」という独自のプラットフォームを開発し、生命科学データのRDF化のため必要なオントロジー、辞書、文献アノテーションの整備を実施した。

５）大規模ゲノム配列データの利用技術開発
遺伝子発現のリファレンスデータセットの整備として、「RefEx （Reference Expression dataset）」を開発、公開した。さらに、公共データベース中の次世代シーケンサデータを、実験目的（ゲノム解読、発現解析、メタゲノム等）や測定機器、対象生物種によって分類し、簡便に検索、ダウンロードし、再利用を行えるようにした「DBCLS SRA」サイトについて、大幅な改良を行った。その他、遺伝子検索エンジン「GGRNA」や「GGGenome」等の開発・公開を行った。

６）情報統合化・知識発見のためのキュレーション支援
生命科学分野の略語情報を検索するシステム「Allie」について、日本語対訳を充実させるとともに、表記上のゆれを吸収して同義表現をまとめる技術の開発や、タンパク質名の定義(デフィニション)をするキュレーション作業について、テキスト処理技術を用いることで作業支援するツール「TogoAnnotator」を開発するなど、キュレーション支援システムの開発や、コンピュータ支援による協働作業手法 CSCW（computer-supported cooperative work）の研究分野で開発されているツールを組み合わせて、文献キュレーション、オントロジー開発、データマッピングの作業を支援した。

７）統合ＤＢにかかわるコンテンツの作成
チュートリアル動画として、「統合TV」を作成・公開し、さらに、「新着論文レビュー」や「領域融合レビュー」、生物並びに臓器形状3Dデータを公開するデジタル人体模型「BodyParts3D」の改良等を行った。

４．事後評価結果

４－１．当初計画の達成度
本課題は当初の研究計画で目標として掲げた１）データベースの RDF による統合化、２）解析プラットフォームによる統合利用環境の整備、３）インターネットを活用した高度検索技術の開発、４）RDF 化に資するオントロジー・辞書・コーパス整備・標準化技術開発、

５）大規模データの利用技術開発、６）情報統合化・知識発見のためのキュレーション支援、7）統合 DB に関わるコンテンツの作成・整備　の当初の研究計画をすべて達成した。統合化推進プログラムの複数の課題と連携し、特に微生物等のライフサイエンス分野の情報の整備と統合化に貢献した。当該課題が取り組んだ各項目は、いずれもライフサイエンス系データベースの統合化のための重要な基盤技術であり、本課題の統合化への貢献は非常に高く評価できる。

４－２．研究開発成果の公開および利用の状況等

本課題の成果となるデータベースやツール等の公開は順調に行われている。コンテンツは多岐にわたるため、月間のユニーク IP アクセス数で簡単に評価できないが、いずれも順調に利用されていると判断できる。

４－３．研究開発成果によるライフサイエンス分野への波及効果

データベース統合化により、データベースをまたいだデータ解析、知識発見が容易になり、ライフサイエンス研究が行われる素地が整備された。データベースの RDF による統合化の流れを作り出し、今後のライフサイエンス分野のデータ活用の標準仕様となることが期待できる。

４－４．広報・アウトリーチ活動等

論文発表や国内外の学会発表、講演等を通じて研究成果を周知するとともに、各種展示会にて成果データベースのデモ等を実施した。生命科学分野におけるセマンティック・ウェブ技術の研究開発を目的とし、チーム内外の関係者を集めた国内外の実務者会議の開催等、活発な活動が実施されたことは評価できる。

５．総合評価

本課題は、成果となるツール等の公開・提供の実施とともに、RDF 化によるデータベースの統合の手法を定着させた。統合化推進プログラムの課題との連携で成果となるツール等が活用され、開発された基盤技術がデータベース統合に有効に機能している。さらに、BioHackathon の主催を通じ、RDF 化によるデータベースの統合の手法について、国際的な広がりを与えることに大きく貢献したことなど、本課題の成果は非常に高く評価できる。今後、データベースの RDF 化並びにその活用を、更に強力かつ継続的に進め、データベースの統合化を促進していくことが必要である。それを実現するためには、本プロジェクトが、次年度以降はより安定した体制の下で運営されることが望ましい。