# 研究開発課題別事後評価結果

## １．研究開発課題名

生命と環境のフェノーム統合データベース

## ２．代表研究者名

理化学研究所　情報基盤センター　統合データベース特別ユニット
ユニットリーダー　豊田哲郎

## ３．研究実施概要

フェノームとは個々のジェノタイプの表現形質である様々なフェノタイプの集合であり、環境と相互作用して変化する計測データの集合である。フェノームは生物種や分野によって研究者のコミュニティが分かれており、分野横断的な統合化は困難であるとこれまで考えられてきたが、フェノタイプ計測では、種や個体の違いを越えて同一の計測技術を適用できる場合も多く、種横断的な表現型の体系的な分類により分野の壁をこえて技術情報を共有する仕組みが必要とされている。そこで、本課題では、バイオリソースと計測技術の網羅的組み合わせとして、フェノーム統合データベースおよびフェノーム利用ワークフローを構築した。

１）フェノタイプデータの記述子の体系化

データ間の関係を属性として表す「プロパティ」を共通化させ、データ統合化のプロパティの標準化、および体系的なデータ記述を行うために必要な概念の関係性を表す「オントロジー」等、データ統合化の基礎となる識別子の体系化を行った。プロパティの共通化には、広く利用されている purl.jp システムを用い、また、オミックスデータ間の関連性を表現するのに有用なプロパティセットである biorel への対応付けも進めた。さらに、生物の表現型を網羅的に記述できる汎用フォーマット Resource Description Framework(RDF)モデルを開発し、同じく RDF によるデータ統合を推進する国際的な生物オントロジーの標準化コンソーシアム、Open Biomedical Ontology (OBO)コンソーシアムの提案する形式との相互関係（概念間の親子関係）を定義した。

２）フェノーム情報の統合化

バイオリソースに関連づけられるフェノームについて、情報の収集と整理を行うため、バイオリソースデータ収集プラットフォームを構築し、マウス系統、細胞株、微生物株、植物株などの表現型情報や有用性情報を統合化した。細胞株に関しては、細胞株の国際的な相互利用を目指し、ミシガン大学、Sanger 研究所、The European Bioinformatics Institute (EBI) 等との連携により、Cell line Ontology (CLO)コンソーシアムを設立し、RDF 準拠

の標準スキーマを開発した。シロイヌナズナフェノームについて、理研内で開発されたシロイヌナズナ変異株、および文献キュレーションにより収集したシロイヌナズナの表現型変異情報の２種類のデータセットを統合化した。

３）フェノーム利用ワークフロー開発
データ共有の規格として近年注目されているLinked Open Data (LOD)に準拠した標準形式で生命科学関連の公開データを提供する目的で、「BioLOD.org 」(Biological Linked Open Databases)を開設し、さらに、セマンティックウエブデータにアクセスするためのプログラミングインターフェイスとして semantic-JavaScript Object Notation (semantic-JSON)を新規に開発した。さらに、ユーザがデータ相互の関係性を把握し必要な情報を検索して取り出すことを容易化するための支援システムとして、「BioSPARQL」(Broadly Integrated Ontological SPARQL Protocol and RDF Query Language)を開発した。また、セマンティックウェブベースのデータを利用して、高速性、セキュリティへの配慮などの点に優れた大規模相関解析を実現する方法を模索し、結果としてSemantic-Web Association Study (SWAS )システムの開発を実施、RDFで記述されたマウスフェノタイプデータを分かりやすく配信するためのインターフェース開発を行った。

## ４．事後評価結果

### ４－１．当初計画の達成度
本課題は、１）フェノタイプデータの記述子の体系化、２）フェノーム情報の統合化、３）フェノーム利用ワークフロー開発、という当初の研究計画を達成した。フェノーム情報の統合化を進め、データの RDF 化を実施した。purl.jp システムにダウンロード可能なすべての RDF データファイルの URL を預け入れし、データの利用を促進しており、統合化は進んだといえる。

### ４－２．研究開発成果の公開および利用の状況等
新規に作成されたLOD形式のマウスのフェノタイプのデータベースである「BioLOD.org」は公開されており、月間ユニーク IP アクセス数は 1.6 万件程度と多く、広くユーザに利用されている。

### ４－３．研究開発成果によるライフサイエンス分野のデータ活用への波及効果
本課題は、表現型の記載データを体系化して記述することに成功し、データベース化を実施した。これは、種々の生物システムにおいて同様のアプローチをとる可能性に道を開いたことになり、さらなる生物表現系データの統合により、将来的にライフサイエンスの多くの分野への貢献が期待できる。

## ４－４．広報・アウトリーチ活動等

論文発表、招待講演はやや少ないが、口頭発表やポスター発表等において、積極的に紹介している。また、セマンティックウェブ推進委員会が推進しているLODの活動を通じ、研究データのオープン化に積極的に取り組んでおり、評価できる。

## ５．総合評価

本課題は、文献情報とバイオリソース情報を対象に、大量のフェノーム情報の統合に成功し、従来は網羅的な検索が困難であったフェノタイプデータの有用性を大きく高めた。今後、生物種を超えたデータベースの統合化を一層実現するため、オントロジーによるデータ標準化を国際連携のもとで進め、Cell line Ontology (CLO)を通した国際的な細胞リポジトリ連携や、標準化等が期待される。更なるフェノーム情報の統合化は他のオミクスデータベースの利用向上にもつながると思われ、今後、さらにバイオリソースの情報量を増やとともに、データベースの付加価値及び利用率の向上を目指していただきたい。