平成 23 年度　研究開発実施報告書

ライフサイエンスデータベース統合推進事業「統合化推進プログラム」
平成 23 年度採択　研究代表者

豊田哲郎
独立行政法人理化学研究所生命情報基盤研究部門・部門長

生命と環境のフェノーム統合データベース

§1. 研究実施体制

（1）豊田グループ

① 研究代表者　豊田　哲郎　（理化学研究所生命情報基盤研究部門・部門長）

② 研究項目

フェノーム統合化・プロパティ標準化・先端計測データ統合化・フェノーム利用・ワークフロー開発

（2）桝屋グループ

①主たる共同研究者: 桝屋　啓志　（理化学研究所バイオリソースセンター・マウス表現型知識化研究開発ユニット・ユニットリーダー）

②研究項目

フェノタイプ記述子の体系化：識別子体系化と評価

フェノーム統合化：バイオリソースフェノーム

§2. 研究実施内容

課題1「フェノタイプ記述子の体系化」

プロパティ標準化と評価（豊田グループ）

今年度は、文献キュレーションによって収集したシロイヌナズナフェノーム情報をテストケースとし、上記方針に基づいて具体的な標準化方針について検討を重ねた。また、検討結果を反映したコンテンツ整備の基盤となる新たなデータベースを新設した。

まず、文献上の様々な表現型に関する記述を一つ一つ吟味して、RDF による適切な表現方法を検討した。2012 年 3 月末の段階で、23 種類のプロパティを定義し、6 種類の公開オントロジー（PO, GO, TO, ChEBI, PATO, EO）を利用することで、フェノーム情報の標準化と統合化が可能であるとの結論を得た。平成 24 年度は、これらの知見に基づきフェノーム情報の統合作業をすすめる（後述）。

一方で、現在世界中で公開されている多くの標準プロパティセット（たとえばBioGatewayサイトで使われているbiorel）について予備的な評価を行い、我々が定義した前述のプロパティとの関連付け作業を進めている。

識別子体系化と評価（桝屋グループ）

RDF での体系的なデータ記述を行うためのオントロジーや、インスタンスを国内外で使用されているものに共通化させていくことで、データ統合化の基礎となる識別子の体系化を行うことを目的として、平成 23 年度は、「上位オントロジーに基づくフェノタイプ識別子の、OWL による体系化」として、生物に普遍的な表現型データ形式を、国内で独自開発された上位オントロジー、Yet Another More Advanced Top-level Ontology (YAMATO) および、バイオ分野で広く使われている Basic Formal Ontology (BFO)それぞれの下位概念として定義した。このデータ形式策定にあたっては、ドメインオントロジーとして、OBO コンソーシアムのオントロジーを主に採用し、統合化プログラムの各プロジェクトとの議論を行い、相互運用性を確保する方向で調整を進めつつある。さらに、「識別子体系化作業用プラットフォームの開発」として、バイオリソースの表現型を、上記のドメインオントロジーを用いてアノテーションするシステムを作成し、最終的に BioLOD へエクスポートする OWL 化ワークフローを確立した。

課題2「フェノーム統合化」

バイオリソースフェノーム（桝屋グループ）

バイオリソースに関連付けられるフェノームについて、上記のフェノタイプ記述子に対応付けながら情報の収集と整理を行い、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する目的で、平成 23 年度は、「基盤データ収集のための、バイオリソースデータ収集プラットフォーム（マウス）構築」として、マウス用プラット

フォーム構築を完了し、月毎更新ワークフローを確立した。さらに「マウスリソースデータ収集作業」として、約 5000 リソース分、5 万付加情報を収集し、かつ、月ごとに最新情報を配信している。現在、細胞リソース収集プラットフォームを作成中である。

先端計測データフェノーム（豊田グループ）
文献キュレーションに基づきシロイヌナズナのフェノーム情報を収集し、前述の議論に基づいて策定した標準化方針に基づいてそれぞれ標準化し、フェノーム統合データベースを新たに作成して収納した。2012 年 4 月時点で収納されたフェノーム情報は 824 件であり、それぞれが適切なオントロジーターム、TAIR の遺伝子情報、PosMed の文献レコードなどへのリンクを伴っている。このデータベースは現在内容の最終チェックを進めており、近日中に、後述する公開データ共有サイトである BioLOD（http://biolod.org)を通じて公開する予定である。

課題3「フェノーム利用ワークフロー開発」
平成 23 年度は、新たに BioLOD.org (Biological Linked Open Data)を開設し、データ共有の規格として近年注目されている LOD(Linking Open Data)に準拠した標準形式で生命科学関連の公開データ提供を開始した。本サイトの検索機能は、結果を様々な観点からランク付けして示す仕様となっており、目的に適した良質なデータセットをユーザに容易に提供できる。また、データは様々な形式でダウンロードでき、生物情報学的研究への適用も容易である。2012 年 3 月現在で、BioLOD.org には 205 データベース、764 クラス、826 万インスタンスのバイオデータが統合されている。
こうして統合されたデータにアクセスする為の論理はSemantic-JSON.orgによって提供される[1]。一方、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(バイオスパークル；Broadly Integrated Ontological SPARQL Protocol and RDF Query Language)を開発した。本システムは、BioLOD で扱っている RDF/OWL データ構造を解析して適切な SPARQL クエリ（検索対象となるテーブルやデータの抽出条件、並べ方などを指定する文字列）を構築し提供するためのフレームワークである。ローカル環境においてユーザの非公開データと公開データとを組み合わせたクエリを実行することもできる。現在、アルファバージョンのシステムが公開されており、有用性の検証と改良作業を進めている。

§3. 成果発表等

(3—1)データベースおよびウェブツールの構築と公開

① 公開中のデータベース・ウェブツール等

データベース名： BRC　マウス　リソース

概要：本 DB は、生物遺伝材料としてのマウス系統を収録している。系統の持つ遺伝子の変異、生物学的な特性が、公共データやオントロジーにリンクされている。リンクの仕方は上位オントロジーに従っており、他のデータベースと連携している。ここに登録されているマウス系統は、理研バイオリソースセンターより提供されている。

公開日：H24 年 1 月 16 日

URL：http://biolod.org/class/cria315s1i/BRC_Mouse_Strain

アクセス数：図 1 のとおり

データベース名： BRC　細胞　リソース

概要：本 DB は、生物遺伝材料としての培養細胞株を収録している。ヒトやマウス等の哺乳類をはじめ様々な生物種にわたる、多種の細胞株を公開している。ここに登録されている細胞株は、理研バイオリソースセンターより提供されている。

公開日：H24 年 3 月 10 日

URL：http://biolod.org/class/cria322s1i/BRC_Cell_Resourse

アクセス数：図 1 のとおり

サイト名:BioLOD.org

概要:生命科学関連の公開データを、W3C の LOD プロジェクト(World Wide Web Consortium Linking Open Data project)に準拠した標準形式で提供している。2012 年 3 月現在で 205 件のデータベース、986 件のクラス、8,773,671 件のインスタンスを統合している。

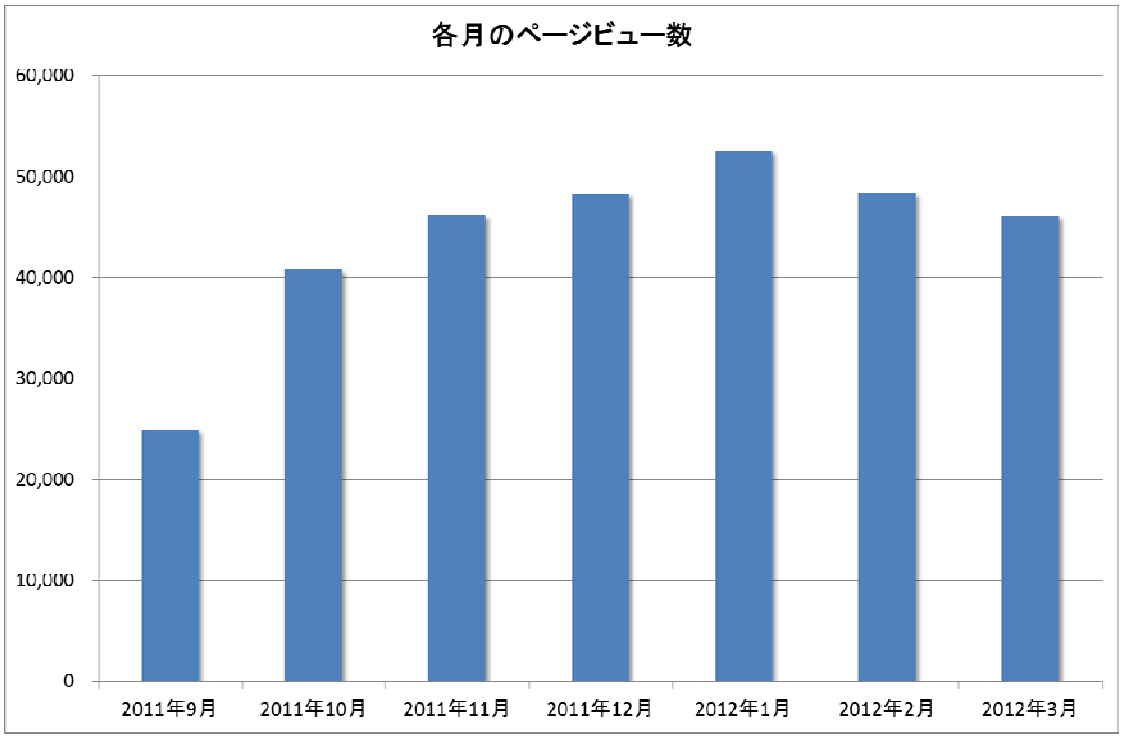
公開日: H23 年 7 月

URL: http://biolod.org

アクセス数：図 1 のとおり

サイト名:BioSPARQL
概要:　BioSPARQL は、RDF/OWL データ構造を解析して適切な SPARQL クエリを容易に構築するためのフレームワークである。既存のクエリ生成補助ツールとは異なり、BioSPARQL は RDF/OWL データを論理的に解析して、適切なクエリのひな形をグラフィカルにユーザに提示する。生成されたテンプレートに必要に応じてキーワードを入力することで、ユーザは実行可能な SPARQL クエリを容易に得られる。
公開日：H23 年 11 月
URL: http://biosparql.org
アクセス数：図 1 のとおり

②　未公開のデータベース・ウェブツール等
データベース名: フェノーム統合データベース（仮称）
概要:本研究で模索したフェノーム情報の統合化手法に基づき、文献キュレーションによって収集したシロイヌナズナのフェノーム情報を標準化し、集積している。2012 年 4 月時点で 23 種類のプロパティを定義し、これらのプロパティと 6 種類の公開オントロジー（PO, GO, TO, ChEBI, PATO, EO）を用いて標準化した 824 件のフェノーム情報を収録した。公開に向けて、コンテンツの最終的な校正作業を進めている。
公開予定:H24 年 5 月



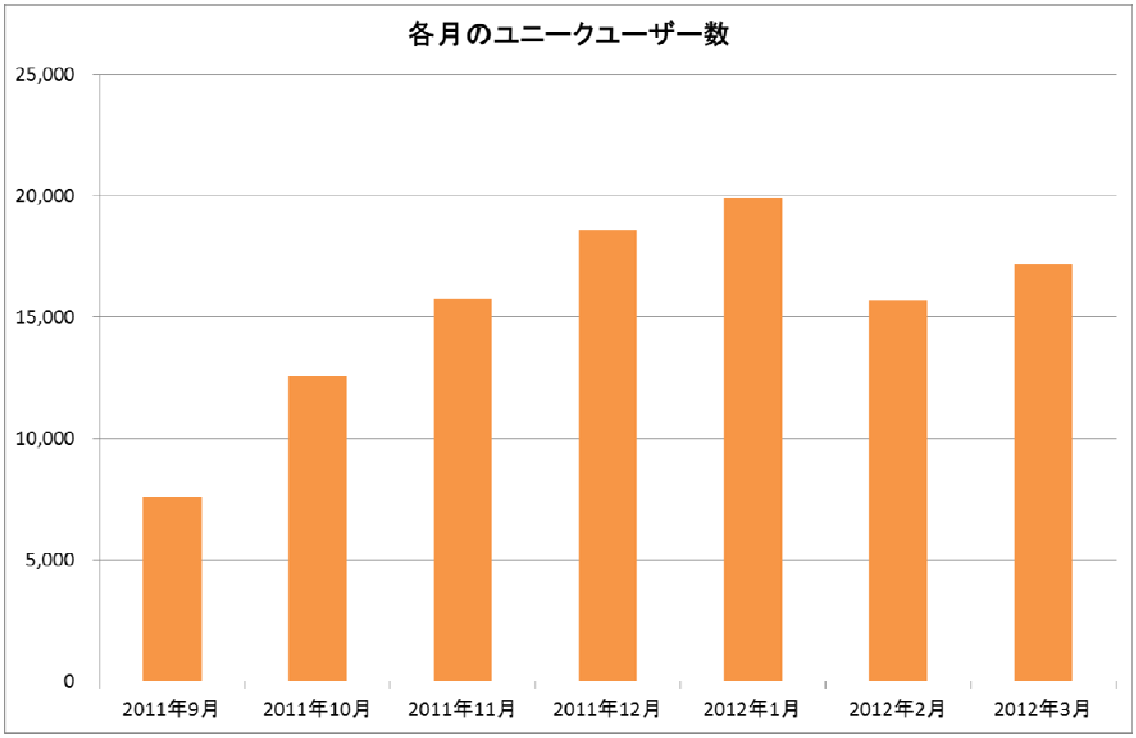


図 1:公開中のデータベースおよびウェブツールの月別ページビューとユニークユーザー数の推移。

(3－2) 原著論文発表

① 発行済論文数(国内(和文) 0 件、国際(欧文) 1 件):

② 未発行論文数("accepted"、"in press"等)(国内(和文) 0 件、国際 (欧文)0 件)

③ 論文詳細情報

1. NorioKobayashi, Manabu Ishii, Satoshi Takahashi, Yoshiki Mochizuki, Akihiro Matsushima, Tetsuro Toyoda"Semantic-JSON: a lightweight web service interface for Semantic Web contents integrating multiple life science databases." Nucleic Acids Research 39: W533-40. doi: 10.1093/nar/gkr353