2C5-OS-21b-1

RDF 技術を用いた疾患モデル動物の表現型データの統合と利用拡大

Integration and promotion of reuse of phenotype data of disease model animals using RDF

桝屋 啓志[*1] 高月 照江[1] 斎藤 実香子[1] 高山 英紀[1]
Hiroshi Masuya Terue Takatsuki Mikako Saito Eiki Takayama

大島 和也[1] 田中 信彦[1] 戀津魁[2] 小林紀郎[2]
Kazuya Ohshima Nobuhiko Tanaka Kai Lenz Norio Kobayashi

[*1] 理研バイオリソースセンター RIKEN BioResource Center
[*2] 理研情報基盤センター Advanced Center for Computing and Communication, RIKEN

Experimental animals play crucial roles in the basic medical research aiming elucidations of pathogeneses and development of treatment protocols of diseases. “Phenotype” of experimental animals and “Symptom” of human diseases are key concepts, which should be mutually mapped and disseminated as the informational source for the basic medical research. Therefore, it is desired that development of the informational infrastructure to promote collection, reuses and intelligent processing of phenotype/symptom data. We have started “J-phenome” (http://jphenome.jp) project, a trial of data integration, reuse and dissemination of phenotype data produced in Japan using the Resource Description Framework. In this report, we describe overview of J-phenome and discuss advantages of Resource Description Framework related technologies.

1. はじめに

生命は分子の部品で構成された複雑なシステムである。また、各生物種は、個別のシステムでありながら、祖先を同一としており、生命全体としては、多様に分岐した分子システムを共有していると見ることができる。人類の健康を改善する医療が発展してきた背景には、人類(ヒト)だけでなく様々な生物種で得られた基礎研究の知識が生かされている。

近年の生命科学では、莫大な個別知識が蓄積され、これらを最大限利用して、イノベーションにつなげることが求められている。中でも人類の健康に直接的に貢献する医学や医療研究への情報活用は極めて重要であり、分子間相互作用とその結果起こる症状、疾患に至る広範囲の情報を大規模に収集し、活用できるような新たな情報基盤が必要である。特に、巨視的で複雑な現象の記述でもある「疾患」の定義や付随情報をいかに効率的、普遍的に共有するかは大きな課題である。

バイオインフォマティクス分野では、このような高次の情報の共有に、オントロジーや、Resource Description Framework (RDF)をはじめとしたセマンティック Web 技術が用いられ始めている。RDF は Web 上のデータの記述と扱いの標準を提供するだけでなく、Web Ontology Language (OWL)で書かれたオントロジーを用いた語彙の標準化、意味記述の基盤を提供することで、上記の「高次」な情報共有に有効である。

また最近では、疾患概念の共有のために、表現型(Phenotype)概念の有用性が認められるようになってきている。表現型とは、生物が遺伝因子や環境因子の結果として示す形質、あるいはその特性である。この概念は「疾患」という一括りの概念に含まれる個別の「症状」(Symptom)にも対応させやすい粒度を持つ。つまり表現型に着目することで、疾患を分解し、疾患同士の症状の重なりを示すことや、表現型を生物種横断的にマッピングして表現型と症状を繋ぎ、非ヒトで得られた知識を、医療のために活かす(疾患モデル生物)ことが、よりシステマティックに可能になると考えられる。Open Biomedical Ontology (OBO)のコミュニティでは、ヒト表現型オントロジーである Human Phenotype Ontology (HPO) [HPO]が症状の語彙リストの役割を演じるようになってきており、疾患の語彙として用いられる Disease Ontology (DO) [DO], Online Mendelian Inheritance of Man (OMIM) [OMIM]とのマッピングデータが利用可能となっている。このデータを用いることで、複数の症状で構成される疾患の情報を得ることができる[Köhler 2014]。また、HPO は UberPheno オントロジーを介すことで、Mammalian Phenotype Ontology, Zebrafish Phenotype Ontology 等の多種の生物の表現型とマッピングすることができるため、たとえばマウスの表現型に相当するヒトの表現型、さらにはその表現型を含む疾患、というように、非ヒト生物の表現型と関連する疾患を紐付けることができる[Köhler 2013]。

我々は、国内外の複数の表現型解析研究プロジェクトのデータを幅広い研究コミュニティから収集し、研究分野の垣根を超えて標準化・統合化・体系化してオープンに公開する事を目的として、”J-phenome”プロジェクト(http://jphenome.jp)を進めている。このプロジェクトでは、これまでのオントロジー研究に基づいて作成した共通の表現型データ記述スキーマに沿って、各表現型データを RDF 化し、シンプルな GUI でのデータ閲覧、ダウンロード、SPARQL エンドポイントを公開している。本論文では、表現型データ統合における RDF の利点について論じる。

2. RDF データの作成

2.1 表現型データの概要

理研バイオリソースセンターが公開するマウス 5,850 系統(http://www2.brc.riken.jp/lab/animal/search.php)、基礎生物学研究所が公開する NBRP メダカ 298 系統(https://shigen.nig.ac.jp/medaka/) 国際マウス表現型解析コンソーシアム(http://www.mousephenotype.org)の公開する 4,916 系統についてのそれぞれ 13,066、581 及び約 12 万個の表現型データを対象とした。

連絡先:桝屋啓志, 理化学研究所バイオリソースセンター, 茨城県つくば市高野台 3-1-1, hmasuya@brc.riken.jp

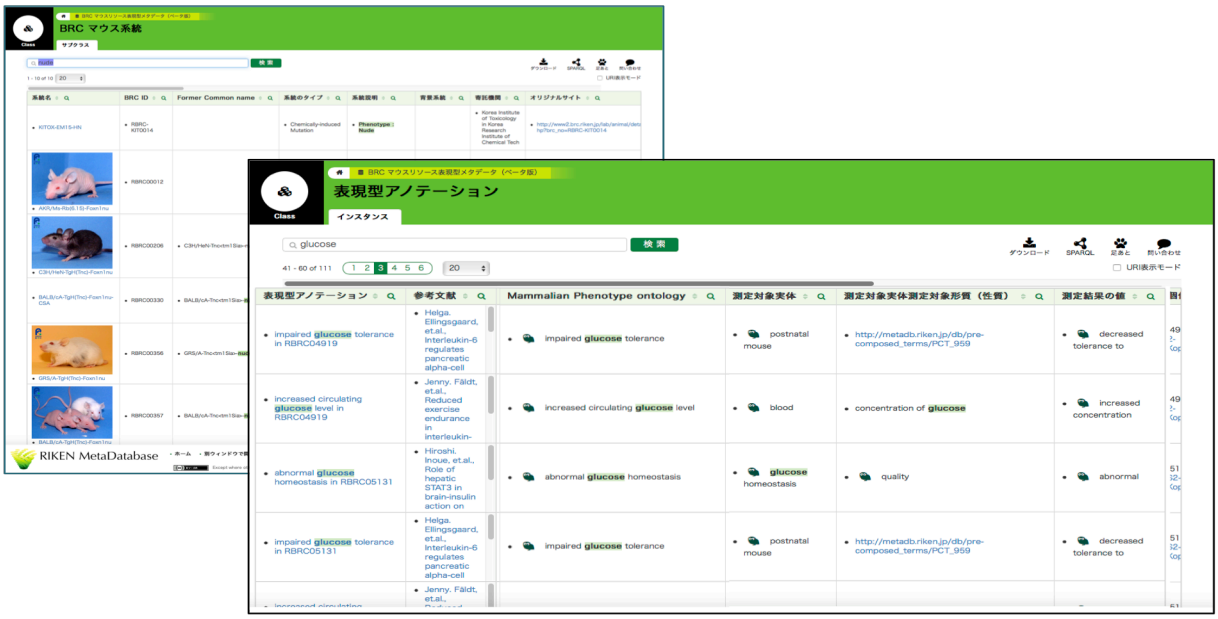

図 1 理研メタデータベースによる表現型データの表示（表画面）

簡易的な検索機能により、表現型の部位、形質、値等での絞り込みが可能。

2.2 スキーマの設計

我々は、表現型を普遍的に記述するためのデータ構造について検討を行ってきた[桝屋　2013]。概要としては測定対象である実体（Entity:マウスやラットの部位や器官組織）、性質タイプ（Attribute: 各計測のパラメータである形質）、値（Value: 測定値）を、オントロジー概念のインスタンスとして記述し、さらに、実体が特定の性質を持つ事を記述する Entity-Attribute-Value3つ組形式の表現のインスタンスとして個々のデータを記述するものである。

今回は、以前の研究で用いた法造[太田　2011]及び、Yet Another More Advanced Top-level Ontology (YAMATO) [YAMATO]ではなく、RDF 及び OBO のオントロジーを略式に利用する目的で、上記データの RDF 化のための共通部分のスキーマを、以前の研究に準じて設計した（[桝屋 2014]を参照）。

2.3 理研メタデータベースを用いた RDF データ作成と公開

RDF データ作成、公開、統合の基盤として理研メタデータベース（http://metadb.riken.jp/metadb/）を用いた。理研メタデータベースは、理研で生産されたデータの利活用促進を目的に開発された DB 運用の共通基盤システムである。異なるデータセットの横断検索や、統合を実現するために、基盤技術として RDF を採用している。トリプルストアとしては、Virtuoso（OpenLink Software ink.）を用いている。

(1) 理研メタデータベースの RDF メタデータ作成ワークフロー

理研メタデータベースでは、既存のデータベースやデータそのものに対して、メタデータを RDF 形式で作成し、データベースの枠を超えて連結、統合することで、データへのアクセスを高めようとしている。RDF に親しみの少ないユーザーに対応するため、表形式（MS Excel 及びタブ区切りテキスト）から変換して RDF を生成するツールを提供しており、今回のデータの変換作業もこれを用いることで、プログラミングスキルのない作成者でも一度表を作るというステップを踏むことで比較的容易に RDF を作成することができた。

(2) データの表示、ダウンロード及び任意検索

上記ワークフローによって作成された RDF データは、理研メタデータベースにアップロードされ、同システムで一律に提供される閲覧機能（表及びカード型表示）、ダウンロード、及び SPARQL エンドポイントによる任意のデータ検索機能が適用される。各々1 個のデータベースに由来するデータセットは、virtuoso の 1 グラフに格納され、上記の表／カード型インターフェースによって、リレーショナルデータベース的な使用感が提供される。表現型データは上記スキーマ構造により、表インターフェース上で表現型を示す部位、形質等での絞り込みが可能であり、表の各セルに表示されている RDF リソースだけでなく、そのリソースを主語とする目的語も絞り込み対象としていることも相まって、表現型データを検索する生命科学者に一定の利便性をもたらしている。

2.4 生物種横断的な同等表現型検索システム

複数の生物の表現型の同等性を示す UberPheno オントロジー[Köhler 2013]を用いて、RDF 化した上記の表現型データ間の同等性、及び疾患との関連を示すためのアプリケーションを試作した。開発工程の簡略化を図るため、UberPheno オントロジー及び、上記 RDF データを、SciGraph を用いて Neo4J データベース (Neo Technology, Inc.) にインポートし、Tomcat を用いてアプリケーション開発を行った。主な機能は以下の通りである。

(1) ヒト疾患名を記入して、疾患モデル動物の候補を検索する機能

ヒト疾患名（Disease Ontology:DOID, OMIM 等）を記入することで、＝＞Human Phenotype Ontology＝＞UberPheno オントロジーの生物横断的関連性データ＝＞各種動物の表現型オントロジー（Mammalian Phenotyp 等）＝＞表現型を示す系統、とい

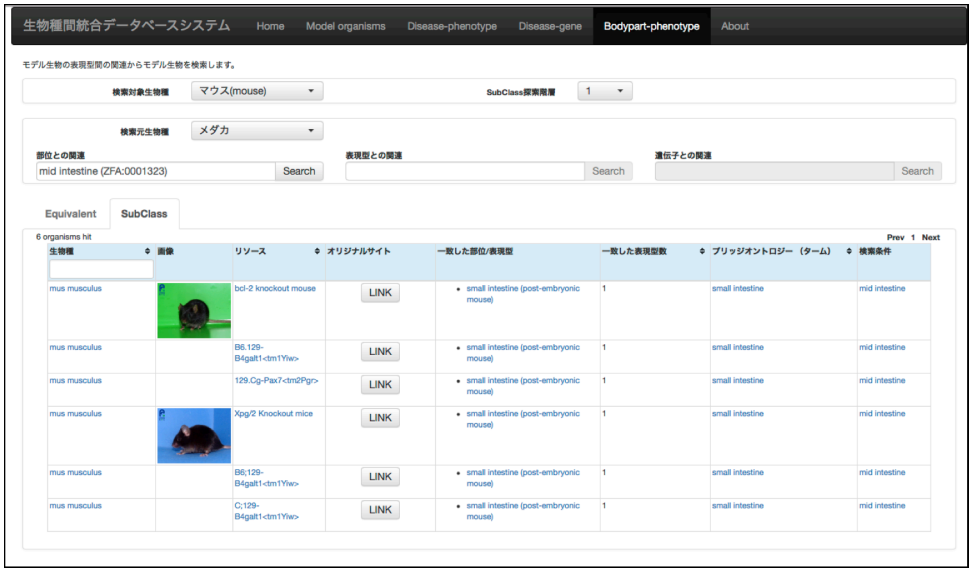

図 2　部位名から、別種のモデル動物表現型を検索する機能を用いて、メダカ中腸に相当する部位に表現型を示すマウスを検索した例

う関連性の経路を辿って、疾患と関連する表現型を示す系統をリストアップする。

(2)　部位名から、別種のモデル動物表現型を検索する機能

部位を示す語彙は、対象とする生物の分野によって用語が異なる。この違いを吸収して、ある分野の研究者が他の分野の表現型データを検索しやすくするための機能を提供する。例えば、魚類部位（Zebrafish Anatomy）＝＞Uberon 部位相同性データ＝＞Uberon 全体部分関連（part_of リンク）データ＝＞マウス部位（Mouse Anatomy）のという関連性の経路を辿って、魚類の「中腸：mid intestine」に相当する小腸の部位に異常を示すマウスをリストアップする。

3. RDF を用いた表現型データの利用拡大

3.1 生物種横断的検索

理研データベースの SPARQL エンドポイントでは、グラフを跨ぐことで、同基盤に格納されているデータベースの横断的な検索が可能である。表現型を扱うデータベースは、通常生物種や、研究単位ごとに分かれているが、これらのデータについて表現型データ部分のスキーマを共通化することで、生物種横断的な表現型データ検索を行うことができた。しかしながら、生物種毎に表現型性質や部位を示すオントロジーが異なる（例えば、マウスは MP オントロジー、メダカでは ZP オントロジー）ため、その語彙の違いを吸収して検索を行うことが求められる。

このようなニーズを可能にするのが前述の UberPheno オントロジーである。このオントロジーのリンクを介した検索はリンクが複雑になるため、本プロジェクトでは 2.4 に述べた専用のアプリケーションを用意した。本アプリケーションを利用することで、「パーキンソン病」に関連する表現型を示すマウス、つまり、当該疾患の研究モデル動物候補となるマウスの絞り込み、メダカ中腸に相当する部位（マウスでは小腸）に表現型を示すマウスの検索が可能となった。このような機能は、多様な生物材料を用いて研究を行う生命科学全体の知識統合に役立つと考えられる。

3.2 Federated search

遺伝的要因に基づいて生物の表現型が発現する分子メカニズムを推測するためには、分子レベルの事象を格納したデータベースへのアクセスが有効である。SPARQL エンドポイントが提供する Federated search 機能を利用して、理研メタデータベースの国際マウス表現型解析コンソーシアムの表現型データと、欧州バイオインフォマティクス研究所（EBI）にある、分子パスウェイのデータベース Reactome [Fabregat et al 2016, Reactome]、およびタンパク質データベースである Uniprot [UniProt Consortium, UniProt] の 3 つのエンドポイントにまたがるクエリを実行した。その結果、ある分子パスウェイに含まれるたんぱく質をコードする遺伝子をそれぞれ破壊した場合の表現型リストを取得することができた（結果は省略）。このような結果は生物中の分子ネットワークが表現型や疾患にどのように影響するかを示唆するため、IMPC 表現型データの利用として一般的な例の 1 つとなると考えられる。

4. 考察と今後の展望

以上、J-phenome プロジェクトでの RDF を用いた表現型データ統合の概要を述べた。本プロジェクトでは、RDF データの可視化、検索に関して極めてシンプルなアプリケーションを用いているにもかかわらず、比較的容易にデータ統合を実現できた。

表による RDF データの表示は、グラフ型のデータに馴染みのないユーザーにとって、比較的わかりやすく個々のデータを表示できる利点があった。各生物の表現型データを一度に表示するには、SPARQL エンドポイントを用いることができるが、上位のクラスを指定して、各サブクラスのインスタンスを表示することは、表現型データ以外のデータの統合にも有効だと考えられるので、理研メタデータベースにもそのような画面が望まれる。

生命科学では、多様なデータを結合して解析を行うことで、新たな知見や、生命システムについての示唆的なデータを得ることが求められる。SPARQL エンドポイントは、このようなデータの結合を行う際に極めて有効である。同一エンドポイント内に複数のデータベースを別グラフとして格納した場合に加え、複数のエンドポイントに跨る Federated search を用いることで、データベースを横断して異種データの結合を比較的容易に行うことができる。このように、今回のシンプルな試行においても、異種データの結合が容易に行えたことから、RDF が Web 上に存在する生命科学データの統合に極めて有効であることが示唆された。

今後は大規模データ間のつながりを網羅的に検索し、あらゆる角度から統計を取ることが求められていくと考えられる。そのためには、1) Web 上に散在する SPARQL エンドポイントに、利用者が関心のあるデータ、および関心のあるデータからリンクす

るデータがどのようにあるかを知ること、2) 複数の大規模データを総なめにするような検索のパフォーマンスが格段に向上することなどが重要と思われた。今回の試行では、1)に関しては、表現型と結合して解析したい分子パスウェイのデータについて、あらかじめ調査が必要だったこと、2)に関しては、Reactome の全パスウェイと、IMPC 表現データの関連検索がパフォーマンスの問題で行えなかったことなどが挙げられる。これらの問題が解決されれば、Web 上のデータをフルに用いた生命科学知識の知識抽出がより現実的になると期待される。

4.1 UberPheno、および Human Phenotype Ontology を用いた疾患と表現型情報との統合

我々のグループでは、法造形式のオントロジーを基盤に作成された、PATO2YAMATO オントロジー [桝屋 2011]および、臨床医学オントロジー[大江 2009, 溝口 2011]を用いた生物表現型情報と、疾患情報をつなげるデータベースの開発も行っている[桝屋 2015]。これらの方法に対して今回の RDF と UberPheno, HPO オントロジーを用いた方法は、1)標準技術を直接的に用いるため、開発コストが小さい点、2) 国際的に広く用いられているオントロジーに準拠しており、データの利用拡大につなげやすい 3)HPO の研究グループによりもたらされる疾患と表現型(症状)の関係性データ量が多く、今後も増加が見込まれる点などが挙げられる。例えば、同じく UberPheno オントロジーを用いる Monarch initiative プロジェクト [Monarch initiative] では、同じく RDF 基盤のデータベースであり、類似のスキーマを用いているという理由で、J-phenome の持つ日本国内の表現型データを提供することで合意している。このようにデータをプラットフォームを跨いで移行できることは、国内で作成されたバイオリソースの利用を海外に広げるという意味で有効な手段であると考えている。

一方、法造と臨床医学オントロジーを用いた統合では、データ記述に用いられる、定量値定性変換を、コンテキストに従って行えること、臨床医学オントロジーは、疾患を異常状態に分解している点で HPO と類似しているが、さらに異常状態間の因果連鎖の関係を持っていることで、より有用な情報を提供しうると考えられる。今後は、両データの相互利用を実現し、互いの利点を最大化するような、マッピングデータや推論、データ可視化システムの開発が期待される。

謝辞

本研究は独立行政法人科学技術振興機構(JST)、バイオサイエンスデータベースセンター (NBDC) の助成による.

参考文献

[DO] http://disease-ontology.org

[Fabregat et al 2016] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P., The Reactome pathway Knowledgebase., Nucleic Acids Res. 44:D481-487.,2016.

[HPO] http://human-phenotype-ontology.github.io

[Köhler 2013] Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, Robinson PN, Mungall CJ.: Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research., Version 2. F1000Res. 2013.

[Köhler 2014] Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park SM, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AO, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BB, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN.: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 42, D966-974, 2014

[桝屋 2011] Masuya H., Gkoutos G.V., Tanaka N, Waki K, Okuda Y, Kushida T., Kobayashi N, Doi K, Kozaki K, Hoehndorf R., Wakana S, Toyoda T., and Mizoguchi R.: An Advanced Strategy for Integration of Biological Measurement Data, Proc. of 2nd International Conference on Biomedical Ontology (ICBO2011) ,pp.79-86 (2011)

[Monarch initiative] https://monarchinitiative.org

[OMIM] http://www.omim.org

[Reactome] https://www.ebi.ac.uk/rdf/services/reactome/

[SciGraph] https://github.com/SciGraph

[UniProt] http://sparql.uniprot.org

[UniProt Consortium] The UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43, D204-212, 2015

[YAMATO] http://www.ei.sanken.osaka-u.ac.jp/hozo/onto_library/upperOnto.htm

[大江 2009] 大江和彦: 病名用語の標準化と臨床医学オントロジーの開発, 情報管理, Vol. 52 , No. 12 p.701-709. (2009)

[太田 2011] 太田 衛, 古崎 晃司, 溝口 理一郎:実践的なオントロジー開発に向けたオントロジー構築・利用環境「法造」□拡張 ― 理論編 ―人工知能学会論文誌, Vol.26 No.2,pp.387-402, (2011)

[桝屋 2013] 桝屋啓志, 古崎晃司, 大江 和彦, 溝口理一郎:コンテキストに依存した定性値を扱う生物表現型統合データベースの試作, 第27回人工知能学会全国大会予稿集, 3I1-2 (2013)

[桝屋 2014] 桝屋啓志, 高月照江 斎藤実香子, 高山英紀, 吉田有子, 蒔田由布子, 望月芳樹, 土井考爾,小林紀郎, 豊田哲郎, セマンティック Web 技術を用いた、生物表現型統合データベース, 第28 回人工知能学会全国大会予稿集, 1G3-3 (2014)

[桝屋 2015] 桝屋啓志, 高山英紀, 古崎晃司, 大江 和彦, 今井健, 溝口理一郎, 物表現型情報と、疾患情報をつなげるデータベース, 第29 回人工知能学会全国大会予稿集, 1G3-3 (2015)