生物学的概念相互作用オントロジーを用いたライフサイエンスデータ統合
—化合物、遺伝子産物の機能推論の現状と課題—

Integration of Life Science Data Using Interlinking Ontology for Biological Concepts: Current Status and Issues of Inference of Chemical Compounds' and Gene Products' Functions

櫛田達矢[1]　古崎晃司[2]　山口敦子[3]　山本泰智[3]　川村隆浩[4]

Tatsuya Kushida[1], Kouji Kozaki[2], Atsuko Yamaguchi[3], Yasunori Yamamoto[3], and Takahiro Kawamura[4]

[1]科学技術振興機構 バイオサイエンスデータベースセンター
[1]National Bioscience Database Center, Japan Science and Technology Agency
[2]大阪大学産業科学研究所
[2]The Institute of Scientific and Industrial Research, Osaka Univ.
[3]ライフサイエンス統合データベースセンター
[3] Database Center for Life Science, Research Organization of Information and Systems
[4]科学技術振興機構
[4]Japan Science and Technology Agency

Abstract: We attempt to infer chemical substances' and gene products' functions, roles, applications, and involvements in diseases using life science RDF datasets and ontologies such as IOBC as a hub. In this study we share and discuss the current status of the preparation of the RDF datasets and issues of the integration by InChI/InChIKey and label matching using Lexical OWL Ontology Matcher (LOOM) algorithm.

1. はじめに

有用物質や高機能素材、創薬ターゲットの探索などを目的に、研究者や技術者が SPARQL 検索を使って、信頼性の高い化学物質および遺伝子産物の機能、ロール、アプリケーションの情報を推論するために必要となる RDF データやオントロジーの開発、整備を進めている。これまでに筆者らは疾患や薬物に関する情報を含む 8 万の概念を持つ生物学的概念作用オントロジー（Interlinking Ontology for Biological Concepts: IOBC）を開発し、is-a 関係および全体-部分関係などを使い、約 1,500 件の遺伝子産物など生体内分子に対してのべ 100 種類以上の機能およびロールの推論を可能にした[1]。その例を以下に示す。

e.g. 1　P タンパク質の機能として，その上位語にあたる ABC トランスポーターが持つ機能“生物学的輸送”を下位継承し推論。

e.g. 2　RNA-タンパク質複合体であるスプライセオソームの機能として、その部分構造であるスプライシング因子が持つ機能“RNA スプライシング”を推論。

さらに IOBC と、約 370 万の化学物質の情報を収録するリンクドオープンデータ NikkajiRDF を統合し、IOBC のオントロジー階層を用いることで 5000 件以上の NikkajiRDF の化合物に対して、のべ 400 種類の機能、ロール、アプリケーションの推論が可能になった[2]。しかし、このオントロジー階層を使った化学物質および遺伝子産物の機能等の推論に生物学的なパスウェイなどの情報が活用できていないなど、十分な推論、網羅的な情報提供ができていると

[1] kushida@biosciencedbc.jp

は言えない。

本研究では、上記のようなオントロジー階層やナレッジグラフ（KG）を使った化学物質や遺伝子産物の機能、ロール、アプリケーションおよび疾患との関係性の情報の収集、推論をより網羅的に実行可能にすることを目指し、必要になるライフサイエンス分野の RDF データやオントロジーの選定を目的に整備状況の調査を行いその結果を整理した。加えて、RDF データ間の統合方法などの課題を取り上げて議論する。

2. ライフサイエンスドメインの RDF データの整備状況

2.1. データベース RDF 化ガイドライン

ライフサイエンス統合データベースセンター（DBCLS）が主催する SPARQLthon は、国内のライフサイエンス分野のデータベースの研究、開発に携わる研究者、開発者が参加する研究開発会議である。毎月 1 回 2 日間の日程で開催され、その目的は SPARQL の勉強会にとどまらずデータベース統合、RDF 関係する幅広いトピックスが議論される。SPARQLthon は、年に 1 回、国内外のデータベース研究者、開発者が参加して開催される国際研究開発会議 BioHachathon と連動し、ライスサイエンス分野のデータベース RDF 化の最新の取り組み、技術の情報共有に一役買っている。

この SPARQLthon や BioHachathon で検討された RDF 化の進め方や知識の記述方法を整理して公開したものが DBCLS データベース RDF 化ガイドライン (https://github.com/dbcls/rdfizing-db-guidelines)である。このガイドラインは、データベース統合と、データの RDF 形式の適切なモデル化のために、ライフサイエンスデータでよく使われる測定項目および測定値の記述方法や遺伝子やタンパク質配列座標情報の記述方法の事例を紹介するとともに、以下のような RDF 化の指針を示している。(1) リソースには永続性の高い URI を利用する、(2) URI リソースはオントロジーのクラスのインスタンスとして定義する、(3) URI リソースにラベルや ID をつける、(4) 他のデータセットへのリンクをつける、(5) 既存のオントロジーを再利用する、(6) プロパティに domain と range を定義する、(7) ライセンス情報をつける、(8) スキーマ図を提供する、(9) SPARQL サンプルを提供するなどである。このガイドラインはライフサイエンス分野データベースを RDF 化する際に推奨される指針として作成されているが、材料工学など他の分野の RDF 化にもその多くの部分が適用可能であると考える。

2.2. NBDC RDF ポータル

NBDC RDF ポータル[3]は国内で産出され DBCLS データベース RDF 化ガイドラインに準拠した高品質なライフサイエンス分野の RDF データのレポジトリーである。2019 年 2 月現在、21 件のデータセット、5 億件のデータセット間のリンク、500 億件のトリプルが収録されている。Federated 検索が可能な SPARQL エンドポイントが提供されており、データセットのダウンロードも可能になっている。収録されるデータセットの多くは、JST バイオサイエンスデータベースセンター（NBDC）によるデータベース統合のファンディングプログラム「統合化推進プログラム」の成果のデータセットであるが、それ以外の RDF データセットでも、メタデータやスキーマ図が提供され、RDF 化ガイドラインに準拠していることが確認されれば登録が可能である。

一方、欧州には EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute) が提供する RDF データセットのポータルサイト RDF Platform (https://www.ebi.ac.uk/rdf/) がある。現在、EMBL-EBI が所有する 7 件 RDF データが、スキーマ図や SPARQL エンドポイントとも公開されている。上記のように日本および欧州ではライフサイエンス分野の RDF データに対するポータルサイトが用意され、メタデータ整備やスキーマ図および SPARQL エンドポイントを提供するなど RDF データの二次利用を促進する取り組みに積極的である。これに対して米国は、RDF データに対して日欧とは多少異なる対応を取っている。たとえば、National Center for Biotechnology Information (NCBI)が開発する化学物質の生理活性データベース PubChem (https://pubchem.ncbi.nlm.nih.gov/) では、スキーマ図の提供や RDF データセットのダウンロードは実現しているが、公式の SPARQL エンドポイントは公開していない。代わりに REST 形式の API によるデータ提供を行っている。

Linked Open Data Cloud (https://lod-cloud.net/) によると現在、ライフサイエンスドメインの Linked Open Data (LOD)は 300 件以上あるとされる。一方、ライフサイエンス分野のオントロジーのポータルサイト BioPortal (http://bioportal.bioontology.org/)には現在、700 件以上のオントロジーが登録されている。その他、様々な機関から多数の RDF データが公開されているが、詳細なメタデータやスキーマ図を付与された RDF データは、先に述べた日本の NBDC RDF ポ

ータルや欧州の RDF Platform を除きまだ多くない。今後、各機関から、自前のデータの二次利用が促進されること目的に、メタデータやスキーマ図の提供が進むことを期待する。

表 1. RDF データセットとその中で使われている外部データの種類．主にスキーマ図から確認した。一部情報は省略。

	RDF data/ Ontologies	Gene	Protein	Chemical compound	Disease	Biological Process
NBDC RDF Portal	NikkajiRDF			NikkajiRDF ChEMBL, PubChem, ChEBI, etc.		
	PDB		UniProtKB			GO
	jPOST		UniProtKB			
	OpenTG gates	NCBI	UniProtKB	PubChem		
	ICGC		UniProtKB			
	Refex	NCBI				
	MBGD RDF		UniProtKB			
	PDGBj		NCBI Protein			
	GGDonto	NCBI			MeSH, OMIM	
	DBKERO	NCBI				
RDF Platform	ChEMBL			NikkajiRDF ChEMBL, PubChem, ChEBI, etc.		
	Reactome	NCBI	UniProtKB	ChEMBL	NCIt	GO
	Ensambl	NCBI, Ensembl	UniProtKB		OMIM	
	DisGeNET	NCBI	UniProtKB		MeSH, OMIM, SNOMEDCT, NCIt, etc.	
BioPortal	IOBC			NikkajiRDF, ChEBI	MeSH, OMIM, SNOMEDCT, NCIt, etc.	MeSH
	ChEBI			ChEBI	MeSH, SNOMEDCT	
UniProt	UniProtKB		UniProtKB		MeSH	GO
NCBI	PubChem			NikkajiRDF ChEMBL, PubChem, ChEBI, etc.	SNOMEDCT NCIt, etc.	

3. RDF データ間の統合—リソース間のマッピング

3.1. 他のデータセットへのリンク

データベース RDF 化ガイドラインでは、他のデータセットへのリンクをつける場合、オリジナルのデータベースで使われている参照可能な URI (Polite URI)と、(もしあれば) ライフサイエンスデータベースを対象に永続的な URI を提供、リダイレクトするサービス Identifiers.org (http://identifiers.org/) の URI の両方に対してリンクを張ることを推奨している。これにより RDF データ間でのリソース同士の統合が進むと考えられる。

表 1 は、NBDC RDF ポータルや RDF Platform などに登録されている代表的な RDF データセットとそれの中で使われている外部（RDF）データやオントロジーを示している。データセット間で統合する場合、共通のデータやオントロジーがあれば、対象のデータセットだけで統合可能であるが、そうでない場合はそれらだけでは統合はできない。 たとえば、PGDBj (https://integbio.jp/rdf/?view=detail&id=pgdbj) と UniProt (https://sparql.uniprot.org/sparql)を統合する場合、タンパク質の URI として前者は NCBI Protein の URI を使っているため、そのままでは両者は統合できない。両者を統合するためには、別途、NCBI Protein と UniProt をマッピングする RDF データセットが必要になる。

3.2. InChI/InChIKey によるマッピング

IUPAC International Chemical Identifier (InChI)は、化学物質の構造情報に基づくユニークな識別子であり、InChIKey は InChI をハッシュ関数を用いて短縮された 27 キャラクターの化学物質の識別子である。InChI/InChIKey はいずれも文字列であるため RDF トリプルの中でリテラルとして記述することが可能である。InChIKey はリソース URI の一部として含められる場合もある。RDF データ間の統合において InChI/InChIKey は有効な情報になる。NikkajiRDF ではこの InChI/InChIKey の情報を使って他の化学物質 RDF データのマッピングを行っている。この時、同一の InChI/InChIKey を持つ化学物質同士をプロパティ skos:closeMatch を使ってその関係を記述している。

3.3. ラベルマッピング

遺伝子名、タンパク質名、生命現象名、疾患名などには、化学物質の InChI/InChIKey のようなに構造情報に基づき一意に決定される識別子がないので、これらを使って RDF データ間を統合する際には、参照可能なオリジナルの URI (Polite URI) や identifiers.org が提供する URI を使用するか、もしくは rdfs:label や skos:prefLabel を使って記述されるリテラルのラベル名の一致が必要になる。後者の場合、Lexical OWL Ontology Matcher (LOOM) algorithm [4] によるラベル一致がしばしば用いられる。これは区切り文字（例、スペース、アンダースコア、括弧、ハイフン）除いたラベル名を使って RDF データセットやオントロジー間で一致する用語を発見するものである。BioPortal に登録されているオントロジーであれば、このアルゴリズムを使用した BioPortal の Mapping サービスを利用してオントロジー間で一致する用語を見つけることができる。ただし、このアルゴリズムでも、同表記の多義語を同じ用語と判断する False positive error を完全に取り除くことはできない。著者らは、LOOM アルゴリズムを実行して得られた同一用語の候補に対して、専門家によるマニュアルキュレーションを実行し、False positive data の除外を試みた。

表 2. IOBC の用語と一致する各 RDF データ/オントロジーの用語数. *: BioPortal の Mapping サービス（ LOOM アルゴリズム）を利用。: LOOM アルゴリズムとマニュアルキュレーションを実行**

RDF data/ Ontologies	Number
ChEBI	30,811*
MeSH	15,946*
NikkajiRDF	10,576**

表 2 は BioPortal の Mapping サービスを使って IOBC と ChEBI および MeSH で一致することが示唆された用語の数および、IOBC と NikkajiRDF に対して著者らが LOOM アルゴリズムとマニュアルキュレーションを実施し一致することが確認された用語の数を示している。このとき後者の IOBC と NikkajiRDF の結果では、10,576 件の用語が一致することが示唆されたが、その後、専門家がマニュアルキュレーションを行い、68 件の False positive data が見つかったため、その修正を行っている [3]。LOOM

のみで一致することが示唆された用語間の関係は、skos:closeMatch を使って、LOOM とマニュアルキュレーションを組み合わせて一致することが確認された用語間の関係は、skos:exactMatch を使って記述した。

4. SPARQL 検索による KG 構築

著者らは IOBC を用いて、その中の任意の用語を始点に 3 ステップ以内で繋がる用語の集合を SPARQL 検索で収集し、その収集した集合を KG と定義した。得られた KG は、化学物質や遺伝子産物の機能、ロール、アプリケーションおよび疾患との関係の推論に使用される[3]。

本研究では、IOBC と収録したトリプルストアを基盤に、始点となる用語（例、線維素溶解）を指定すると、3 ステップ以内で繋がる KG のデータを生成、そのデータのダウンロードを可能にするツールのプロトタイプ（http://hozo.jp/IOBC/IOBCquery.html ）を作成した。ダウンロードしたデータは、ライフサイエンス分野でよく使用されるネットワーク解析ツールの Cytoscape で可視化できるようになる（図 1）。なお、KG を始点から 3 ステップ以内で繋がる用語の集合体と定義した理由の 1 つは、IOBC から KG を生成する際、生成された KG の構成要素が 3 ステップの場合 100 件前後になり、人間が Cytoscape など可視化ツールを使って全体を俯瞰および認識しやすいサイズであることが挙げられる。現在このプロトタイプでは、使用するプロパティが related (skos:related) や has function (http://semanticscience.org/resource/SIO_000225) などに限定されているが、今後ユーザーがプロパティの指定をできるように改良する予定である。

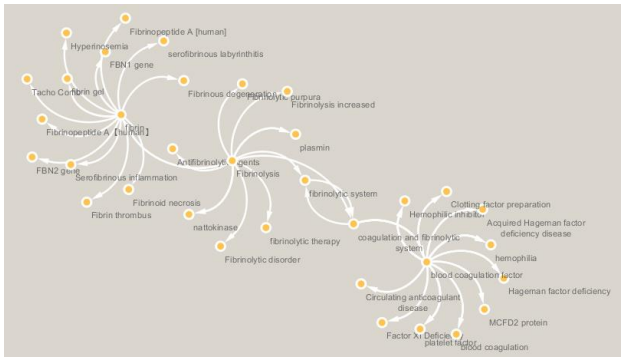

図 1. Cytoscape で可視化した Fibrinolysis ネットワーク

5. 結論

Linked Open Data Cloud によると現在、1,300 件以上ある LOD のうち、ライフサイエンス分野の LOD は 300 件以上あるとされ、ライフサイエンス分野では他の分野と比較して LOD 化や RDF 化が進んでいるように思われる。ただし、これら RDF データ間の連結、統合はまだ不十分と考える。5 Star Open Data (https://5stardata.info/ja/) では、最もオープン化が進んだデータの条件として、データが他者のデータへリンクを持つことが必要とされる。言い換えれば、他機関が生産したデータとのリンクの情報を持たないデータはデータ共有の取り組みとしては完全ではないことを意味している。

たとえば、遺伝子変異、遺伝子発現、タンパク質立体構造、タンパク質-化合物の相互作用、遺伝子-疾患の関係の各 RDF データを統合して、新しい知見の発見を試みようとした場合、それぞれの RDF データで遺伝子の表す URI の体系が異なっていたり、遺伝子とタンパク質の関係を表す RDF データがない場合には、これらのデータの統合は不可能である。データ生産者は、自分のデータにリソースの URI に polite URI や広く普及している URI（例、identifiers.org）を使用したり、他のデータとのリンク情報、たとえば UniProt などの主要な RDF データへのリンク情報を付与することが必要になる。

また今回の RDF データの調査では、NBDC RDF ポータルや RDF プラットホームなどの取り組みを除き、多くのライフサイエンス分野の機関で、自前のデータの二次利用促進のための詳細なメタデータやスキーマ図の整備、提供が進んでいないことが明らかになった。今後、ライフサイエンス研究者およびデータベース開発者は、上述のような自前のデータに対するリソース URI の適切な選定、リンク情報の拡充、メタデータやスキーマ図の提供などに積極的に取り組み、サイエンスの発展および産業界への貢献を意識して、ライフサイエンス分野だけでなく、他分野へのデータの二次利用の促進にまで踏み込んで取り組むことが望まれる。

筆者らは今後、これまで統合が進んでいなかった代謝パスウェイ、細胞内シグナル伝達パスウェイ、タンパク質間およびタンパク質-薬物間相互作用の RDF データと IOBC との統合を進め、新たに副作用の予測など、オントロジー階層や KG を使った化学物質および遺伝子産物に対する機能、ロール、アプリケーション、疾患との関係などのより網羅的な推論が可能になることを目指したい。

謝辞

本研究の一部は、科研費（17H01789：基盤研究(B)）の支援を受けて実施した。

参考文献

[1] Kushida, T., Masuda, T., Tateisi, Y., Watanabe, K., Matsumura, K., Kawamura, T., Kozaki, K., Takagi, T.: Refining JST thesaurus and discussing the effectiveness in life science research. In: Proc. of 5th Intelligent Exploration of Semantic Data Workshop (IESD 2016, co-located with ISWC 2016), pp. 1–14, Kobe, (2016).

[2] Kawashima, S., Katayama, T., Hatanaka, H., Kushida, T., Takagi T.: NBDC RDF portal: a comprehensive repository for semantic data in life sciences, Database: J. Biological Databases and Curation, Vol. 2018, bay123, (2018).

[3] Kushida, T., Kozaki, K., Kawamura, T., Tateisi, Y., Yamamoto Y., Takagi, T.: Inference of Functions, Roles, and Applications of Chemicals Using Linked Open Data and Ontologies. In Semantic Technology: 8th Joint International Semantic Technology Conference (JIST 2018), LNCS 11341, pp.385-397. Springer, Awaji (2018).

[4] Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedicine: simple methods work. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association, pp.198–202 (2009).