様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ５ 日現在

機関番号：８２４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００４２４

研究課題名（和文）生命科学分散メタデータの高効率な統合と検索のための新規ディレクトリ機能の開発

研究課題名（英文）Development of novel directory function for efficient integration and retrieval of life science distributed metadata

研究代表者
小林　紀郎（Kobayashi, Norio）

国立研究開発法人理化学研究所・情報システム本部・開発ユニットリーダー

研究者番号：２０４１５１６０

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：データを説明するデータであるメタデータはデータ利活用において重要な役割を果たすが、研究者や機械がデータを容易に発見し、データの種類や数を正確に把握しながら網羅的に解析するためには、メタデータのカタログをどのようなスキーマで記述するかを規定し、これをメタデータ検索機能に反映させることが不可欠である。本研究では、このようなスキーマの定義を行いLODSurfer Metadataとしてまとめ上げるとともに、生命科学メタデータを公開する主要なサイトについてメタデータカタログの作成に成功した。

研究成果の学術的意義や社会的意義
本研究は、生命科学系研究データの共有及び利活用促進に資するために、データにメタデータを付して公開するサイトを対象とし、機械による高度なデータ検索や解析の促進を図るものである。この成果は、オープンサイエンスや国際連携研究など、特にデータ駆動型研究に必要なサービスとして重要な役割を果たす。また本研究で開発された技術自体は生命科学に特化したものではないため、様々な分野のデータ利活用の促進のために応用可能であることも意義深い。

研究成果の概要（英文）：Metadata is the data that describes data, and plays an important role in data utilization. In order to easily discover data and to comprehensively analyze the data while accurately grasping the type and number of data, we defined data schema of metadata catalog which can be applied to develop metadata search function. More concretely, we defined such catalog data schema as "LOD Surfer Metadata" and successfully generated such metadata catalog for major sites that publish life science metadata.

研究分野：情報科学

キーワード：SPARQL　SPARQLエンドポイント　生命科学データ　メタデータ　セマンティックウエブ　RDF　連合検索　データカタログ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

生命科学において産出される計測データは、計測技術の発達に伴い、新たな種類のデータや大規模なデータなど多様化が進んでいる。これらのデータは研究成果としてファイルアーカイブやデータベースとして編纂、公開され、主に実験研究者にとっては自らが産出したデータの解析や解釈のために、また主に情報研究者にとっては種々のデータ統合、比較等の解析手法を用いて新たな生物学的知見を得るために活用される。データの多様化により、各研究者による論文やウエブ検索サービス等を駆使したデータの収集や解析では限界があり、情報機器や高度な情報学を駆使した、データの高度な利活用が研究を効率的かつ効果的に進めるために重要となる。そのための技術基盤として、データの保存形式に依存することなく、データそのものやデータが取得された実験やリソース等についての説明を、機械可読性を持たせて記述するメタデータの標準化がデータ利活用の推進役として求められる。メタデータの記述やそれをウエブ上で流通させる仕組みは、Resource Description Framework（RDF）や、RDF データへの問い合わせ言語やアクセス方法である SPARQL として世界標準化されている。SPARQL を通じて RDF データを公開するサーバである SPARQL エンドポイントは、主要な研究機関で構築され公開がすでに始まっていた。

しかし、このような大規模な組織に限らず、個々の研究者でも RDF に準拠したメタデータを、ウエブページを公開するがごとく容易に公開できる時代がやがて到来すると考えられる。この時予想される課題は、ウエブのどこで、どのようデータが、どのようなデータ構造で SPARQL エンドポイントとして公開されているかをデータ利用者が知る仕組み、すなわちデータ構造のディレクトリ機能（適切なメタデータとそれを利用した高度な SPARQL エンドポイント検索機能）の必要性である。既存のツールとして、データ集約型カタログサイトがあるが、生命科学分野の概念やデータ構造等のメタデータが標準化されていないことから、所望のデータを取得するための SPARQL エンドポイントの検索ができない問題があった。標準化されたメタデータの記法として W3C 仕様である VoID（https://www.w3.org/TR/void/）や生命科学系データのプロファイルを記載する HCLS Community Profile（https://www.w3.org/TR/hcls-dataset/）等があるが、これらはデータに含まれるオントロジーやデータクラス、それらの間の関係などの基本的な統計量の記述にとどまっていた。生命科学では大規模オミックス解析のような、ユーザや機械がデータを容易に発見し、データ数を正確に把握しながら行う網羅的解析を行う仕組みが必要なことから、既存のメタデータの記述、検索では不足であり、メタデータの改良と、これらから導出される複雑なデータ構造やデータ量等が提示できる検索機能が必要であった。

２．研究の目的

本研究では、セマンティックウエブに準拠した生命科学データベースを対象に、データベースの所在とデータクラス間の関係やデータ量等の統計量(メタデータ)を収集、蓄積して、人間のみならず機械(プログラム)が検索できる新たなシステムの開発研究を行う。具体的には、生命科学の様々なデータベースやデータ構造の高度な検索機能を実装・公開し、さらにその実現に必要なメタデータの仕様を取りまとめて公開する。

一つ目の研究目的は、生命科学におけるデータ解析に必要なデータセットのメタデータの仕様策定である。生命科学における SPARQL の活用事例等を詳細に調査、検討しながら、分散配置された SPARQL エンドポイントの連合検索が可能となるようなデータセットのメタデータ（以下メタデータ）の仕様を検討する。既存の SBM 等の仕様を基礎とし、これらの仕様では記述できない SPARQL エンドポイントを跨ぐクラス間関係やその関係に属するデータ数等の詳細な統計量記述について、メタデータ自体の情報量爆発やその取得の容易さ等を多面的に検討しながら仕様化する。

二つ目の目的は、生命科学関連 SPARQL エンドポイントのメタデータの構築とその試験公開である。上記メタデータ構築に必要なデータを確実に取得できる SPARQL クエリを、様々な SPARQL エンドポイントにアクセスしながら試行錯誤し作成することである。さらに、生命科学の主要な SPARQL エンドポイントに対して実行し、メタデータを収集する。さらに収集したメタデータや本研究で実装したプログラムを試験公開する。

３．研究の方法

上記の研究目的を達成するために、以下の手順を踏んで研究を推進した。

(1) 本研究に参画する研究者の所属機関である理化学研究所およびライフサイエンス統合データベースセンターを中心に RDF 活用の事例調査を実施し、本研究で開発するメタデータ仕様策定に必要な情報収集を行う。

(2) 既存のメタデータの仕様を基本とし、(1)の事例調査で不足している事項を洗い出して追加仕様を策定する。さらに策定した仕様を公開する。

(3) メタデータ構築用データを取得する SPARQL クエリを開発する。SPARQL エンドポイントに負担をなるべくかけず、かつ(2)で策定したメタデータ構築に必要なデータを確実に取得する、安定動作可能な SPARQL クエリを作成する。

(4) メタデータ構築用データを取得するプログラムを開発する。これは、(3)で開発した SPARQL

クエリを発行し、得られた結果からメタデータを構築するプログラムを開発するものである。
(5) メタデータの公開サイトの設計、プロトタイプ開発、およびその試験運用を行う。(4)で構築したメタデータを公開するためのウエブサイトを、理研のクラウド基盤等を用いて開発する。
(6) これまでの成果物の見直し及び改良を行う。

４．研究成果

(1) 生命科学新規ディレクトリ機能実現に必要なメタデータの定義

生命科学におけるデータ解析に必要なデータセットのメタデータの仕様策定に関して、分散配置された SPARQL エンドポイントの連合検索が可能となるようなデータセットのメタデータ（以下メタデータ）の仕様を検討し、さらにメタデータを SPARQL エンドポイントから取得するプログラムを作成、公開した。
その仕様策定に当たっては、RDF 活用の事例調査の結果を踏まえて以下の２点を考慮した。まず、メタデータを構成する基本的なデータ項は概念間（クラス間）の二項関係であるが、生命科学ではデータの網羅性が重要であるため、当該二項関係に含まれるインスタンスやトリプル数などの統計量を記述できるようにした。
また、① eagle-i などデータセット毎にエンドポイントが提供される場合と、② 理化学研究所など一つのエンドポイントが複数のデータセットを提供する場合があることから、メタデータの記載対象はエンドポイントとしながらも、②においてはデータセット毎のメタデータも記載できるようにした。
さらに、メタデータの仕様は単一のものではなく、SPARQL エンドポイントからメタデータ作成に必要な情報をクエリとして問い合わせ取得するコストや当該 SPARQL エンドポイントへの負荷を考慮し、単に概念間の二項関係を取得するレベルと、詳細な統計量まで取得するレベルを定義した。

(2) SPARQL エンドポイントからメタデータを作成するためのクローラプログラムの開発

公開中の生命科学データを提供する SPARQL エンドポイントからメタデータを取得するプロトタイププログラムの実装を行った。この実装を通して、データ量の多いエンドポイントにおいては、統計量の取得において誤った数値を返すことや、またタイムアウトする場合があることが分かった。この問題の解決のため、同一の目的を達成する数種類のクエリセットを準備し、それらを必要に応じて適用することにした。この改良により、完全ではないが、より確実なメタデータ作成に必要な情報の取得が可能になった。

(3) SPARQL エンドポイント評価サービス Umaka-Yummy との連携

Umaka-Yummy はライフサイエンス統合データベースセンターにより運営されている、生命科学データを公開している SPARQL エンドポイントを列挙するサイトである。このサイトでは、稼働率、データの新鮮度、運用度、活用性、適正度、処理速度の 6 種の評価基準に従って SPARQL エンドポイントを評価して Umaka Score と呼ぶスコアを算出するとともにこのスコアの順で SPARQL エンドポイントをランキングしている。本研究では、Umaka Score により評価された上位 50 件の SPARQL エンドポイントを対象に、上記 B)で開発されたクローラプログラムを実行し、メタデータの作成を行った。

(4) クラス間の 2 項関係の取得では解決しない連合検索時の問題点の解決試験

以上の過程により得られたメタデータを解析すると、RDF データに付されるクラスはデータセットにより異なることがあり、連合検索の障害となっていることが分かった。そこで、上位の概念をまとめたオントロジー（以下、上位オントロジーと呼ぶ）が与えられたときに、すべてのデータセットに含まれるすべてのクラスを上位オントロジーに集約するプログラムを開発することで問題解決の糸口を探ることにした。この問題を完全に解決するには、各インスタンス（データ項）について、そのインスタンスが属しているクラスをすべての SPARQL エンドポイントについて調べていかなくてはならず、インスタンスの数、クラスの数ともにこの作業を行うには膨大で現実的なことではない。この点から、ここで提案し試験した解決法は、問題を完全に解決するものではなく、簡易的な改善法であるが、その評価は実用性を測る上で意義深いと結論付けた。
まず、Umaka-Yummy に掲載された上位 50 の SPARQL エンドポイントを対象とし取得したメタデータを、ヒトの遺伝子と遺伝子によって規定される表現型のカタログ Online Mendelian Inheritance in Man(OMIM)を用いて、クラスを OMIM の概念に集約することに成功した。
次に、より多面的かつより多くのクラスに集約できるよう、Semanticscience Integrated Ontology (SIO)や Medical Subject Headings (MeSH)を含む６種のオントロジーに対応するよう拡張し、特に SIO と MeSH により多くのクラスが集約されることが分かった。しなしながら、上位オントロジーでまとめることができたクラス数は、今のところ上記 50 の SPARQL エンドポイントが持つクラスの約 10%にとどまっていた。この原因はクラス名のみを用いた単純なマッチングによるもので、表記ゆれ等の対応が必要であるという課題が残された。
上記のように得られる、上位オントロジーでまとめた結果もメタデータとして記述できるよう、A)のメタデータスキーマを拡張し、上記 6 種の上位オントロジーのクラスでまとめ上げた結果

も合わせてメタデータとして記述できるようにした。
以上の研究成果により、目標であった生命科学分散メタデータの高効率な統合と検索のための新規ディレクトリ機能に必要なメタデータの仕様を確立し、関連ツール類の開発に成功した。本研究でえられた成果物を、GitHub の LODSurfer メタデータリポジトリ―、および理研メタデータベースと呼ぶメタデータ公開基盤より公開した。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Satoshi Kume, Hiroshi Masuya, Mitsuyo Maeda, Mitsuo Suga, Yosky Kataoka, and Norio Kobayashi	４．巻 10675
２．論文標題 Development of Semantic Web-Based Imaging Database for Biological Morphome	５．発行年 2017年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 277-285
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/978-3-319-70682-5_19	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Atsuko Yamaguchi, Kouji Kozaki, Yasunori Yamamoto, Hiroshi Masuya, and Norio Kobayashi	４．巻 10675
２．論文標題 Semantic Graph Analysis for Federated LOD Surfing in Life Sciences	５．発行年 2017年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 268-276
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1007/978-3-319-70682-5_18	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yasunori Yamamoto, Atsuko Yamaguchi, and Andrea Splendiani	４．巻 Volume 2018
２．論文標題 YummyData: providing high-quality open life science data	５．発行年 2018年
３．雑誌名 Database	６．最初と最後の頁 bay022
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1093/database/bay022	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計18件（うち招待講演　1件／うち国際学会　13件）

１．発表者名 Norio Kobayashi, Yasunori Yamamoto, and Atsuko Yamaguchi
２．発表標題 UmakaData extension: Toward Realization of a Practical SPARQL Endpoint Discovery Service for Life Sciences.
３．学会等名 Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS) 2019（国際学会）
４．発表年 2019年

１．発表者名
Norio Kobayashi, Josh Moore, Shuichi Onami and Jason R. Swedlow

２．発表標題
OME Core Ontology: An OWL-based Life Science Imaging Data Model

３．学会等名
Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS) 2019（国際学会）

４．発表年
2019年

１．発表者名
山口敦子、小林紀郎、山本泰智、桝屋啓志、古崎晃司

２．発表標題
LOD Surfer: クラス間関係に基づく連合検索を利用したLOD探索

３．学会等名
2020年度　人工知能学会全国大会(第34回)

４．発表年
2020年

１．発表者名
Josh Moore, Norio Kobayashi, Susanne Kunis, Shuichi Onami and Jason R. Swedlow

２．発表標題
On Bringing Bioimaging Data into the Open(-World)

３．学会等名
Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS) 2019（国際学会）

４．発表年
2019年

１．発表者名
Yasunori Yamamoto and Atsuko Yamaguchi

２．発表標題
Finding the best RDF data by Umaka Suite.

３．学会等名
Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS) 2019（国際学会）

４．発表年
2019年

１．発表者名
Norio Kobayashi

２．発表標題
Novel trends in research data utilisation based on open science.

３．学会等名
6th INCF Japan Node International Workshop, Advances in Neuroinformatics AINI 2018（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Norio Kobayashi

２．発表標題
A Data Integration Platform for RIKEN's Cross-field Data-driven Researches.

３．学会等名
2018 OME annual meeting（国際学会）

４．発表年
2018年

１．発表者名
Norio Kobayashi, Satoshi Kume, Kenichi Ueno, Junichi Hata, Hiroshi Masuya, Yoko Yamaguchi, and Yosky Kataoka.

２．発表標題
An ontology-based OME data model for multimodal biomedical imaging.

３．学会等名
2018 OME annual meeting（国際学会）

４．発表年
2018年

１．発表者名
Satoshi Kume, Hiroshi Masuya, Yasuyoshi Murakawa, Yosky Kataoka, and Norio Kobayashi.

２．発表標題
Development of a Metadatabase for Electron Microscopy Microstructural Imaging Data.

３．学会等名
2018 OME annual meeting（国際学会）

４．発表年
2018年

１．発表者名
山口 敦子，小林 紀郎，桝屋 啓志，山本 泰智，古崎 晃司．

２．発表標題
LOD Surfer API: クラス間関係に基づく LOD探索のためのウェブAPI．

３．学会等名
2018年度人工知能学会全国大会

４．発表年
2018年

１．発表者名
Norio Kobayashi and Yasunori Yamamoto

２．発表標題
LOD Surfer Metadata: Essential LOD catalogue data for traversing life-science LOD amongst multiple SPARQL endpoints.

３．学会等名
11th International SWAT4HCLS Conference（国際学会）

４．発表年
2018年

１．発表者名
Norio Kobayashi, Satoshi Kume, Josh Moore, and Jason R. Swedlow.

２．発表標題
OME Ontology: A Novel Data and Tool Integration Methodology for Multi-Modal Imaging in the Life Sciences.

３．学会等名
11th International SWAT4HCLS Conference（国際学会）

４．発表年
2018年

１．発表者名
Atsushi Fukushima, Mikiko Takahashi, Nozomu Sakurai, Toshiaki Tokimatsu, Hideki Nagasaki, Hideki Hirakawa, Takeshi Ara, Masanori Arita and Norio Kobayashi.

２．発表標題
RIKEN Plant Metabolome MetaDatabase: an integrated plant metabolome data repository based on the semantic web.

３．学会等名
11th International SWAT4HCLS Conference（国際学会）

４．発表年
2018年

１．発表者名
山口 敦子，小林 紀郎，臼田 大輝，桝屋 啓志，山本 泰智，古崎 晃司

２．発表標題
LOD Surfer API: クラス間関係を用いたLODからの情報抽出Web API

３．学会等名
人工知能学会 セマンティックウェブとオントロジー研究会

４．発表年
2017年

１．発表者名
Atsuko Yamaguchi, Kouji Kozaki, Yasunori Yamamoto, Hiroshi Masuya, and Norio Kobayashi

２．発表標題
LOD Surfer API: Web API for LOD Surfing using Class-Class Relationships in Life Sciences

３．学会等名
10th International SWT4HCLS conference（国際学会）

４．発表年
2017年

１．発表者名
Norio Kobayashi, Satoshi Kume and Hiroshi Masuya

２．発表標題
Metadata-driven interdisciplinary research projects using RIKEN MetaDatabase

３．学会等名
10th International SWT4HCLS conference（国際学会）

４．発表年
2017年

１．発表者名
山本泰智，山口敦子

２．発表標題
より良いLOD利用環境の実現に向けて

３．学会等名
トーゴーの日シンポジウム2017

４．発表年
2017年

１．発表者名
山口敦子，小林紀郎，山本泰智，桝屋啓志，古崎晃司

２．発表標題
LOD上の情報をクラス間関係で切り取る基盤技術開発

３．学会等名
トーゴーの日シンポジウム2017

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
LODSurfer/lodsurfer-metadata
https://github.com/LODSurfer/lodsurfer-metadata/tree/develop

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山本　泰智 (Yamamoto Yasunori) (50470076)	大学共同利用機関法人情報・システム研究機構（機構本部施設等）・データサイエンス共同利用基盤施設・特任准教授 (82657)	