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 １９ 日現在

機関番号：８２６５７
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００４３４

研究課題名（和文）生命科学データの分散知識統合基盤に資する安定かつ高速な連合検索

研究課題名（英文）Stable and fast federated search for integration of distributed knowledge of life science data

研究代表者

山口　敦子（Yamaguchi, Atsuko）

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

研究者番号：１０３４６１０８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では，ウェブ上に分散したSPARQLエンドポイントで提供される生命科学分野のRDFデータ利用を安定かつ高速に行うために，概念間関係をたどるクエリに対する連合検索システムの開発を行った．各SPARQLエンドポイントをクロールして得られたクラスの情報およびクラス間関係の情報をグラフの形で保持する連合クラスグラフの構築を行い，さらに連合クラスグラフ上のパスの情報を用いることで，SPARQLエンドポイントをまたいで効率的な検索が可能な連合検索システムを構築した．

研究成果の学術的意義や社会的意義
本研究の成果は，ウェブ上に分散した生命科学分野のRDFデータを柔軟に切り出すシステムLOD Surferの連合検索エンジンとして利用されている．LOD Surferは概念間をたどる連合検索システムを利用するウェブAPIを提供しており，生命科学分野のRDFデータを利用するアプリケーションを，それぞれのSPARQLエンドポイントのデータの内容を意識することなく，容易に構築することができる．

研究成果の概要（英文）：In this study, in order to stably and efficiently use life science data provided by SPARQL endpoints distributed on the Web, we developed a federated search system for queries that trace the relationships between classes. To do so, by using information of classes and class-class relationships obtained by crawling each SPARQL endpoint, a federated class graph is constructed. We developed a federated search system that enables efficient searches across SPARQL endpoints by using the information of the path on the federated class graph.

研究分野：生命科学データ統合

キーワード：連合検索　SPARQL　生命科学データ統合　リンクト・オープン・データ

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

(1) 実験技術や測定機器の発達により，生命科学分野のデータはますます巨大かつ複雑になっている．これらのデータを統合的に扱うために，生命科学分野のデータベースにおいて，RDF をはじめとしたセマンティックウェブ技術を用いたデータ提供が広まっている．これらのデータはウェブ上では，それぞれ SPARQL エンドポイントとよばれるウェブ API を通じて提供されている．RDF データはそれぞれがグラフ構造をもつが，互いにリンクをはることでリンクト・オープン・データ(LOD)とよばれる巨大な一つのグラフとなり，統合的に利用することが原理的には可能となる．

(2) その一方で，LOD に含まれるデータはウェブ上に分散した状態で提供されているため，生命科学分野の RDF データを一つの巨大なグラフと見立てて利用するためには，膨大な探索空間を対象とした検索を行う必要がある．LOD 全体を対象にした検索は連合検索とよばれ，これまで様々なシステムが開発されてきたが，生命科学のような巨大なデータに対応可能なものはなかった．そのため，生命科学において実用的な連合検索クエリを実行可能とするには，ユーザが探索空間を絞り込む必要があるが，個々の SPARQL エンドポイントにどのようなデータがあり，また全体としてどのような構造をしているのか，探索空間の膨大さのために把握することが難しいという問題があった．また，実行可能な連合検索クエリをユーザが記述できても，個々のデータが巨大なため，応答に長い時間がかかる場合，あるいはタイムアウトによって応答がない場合が少なくなかった．

２．研究の目的

(1) LOD から解析に必要なデータを自在に切り出すために，各 SPARQL エンドポイントに含まれるデータベースから予め抽出された意味的繋がりを記載したメタデータを利用し，実用性を備えた新たな連合検索手法の研究を行う．

(2) 特に，メタデータを用いた探索空間の絞り込みに重点を置き，概念間の意味的繋がりを辿るクエリに対して，高速で安定した連合検索技術の確立を目的とする．開発した技術は，広く利用可能にするために，サービスとしてウェブ上で公開するとともに，ソースコードをオープンなライセンスをつけて GitHub で公開する．

３．研究の方法

(1) まず，各概念(クラス)およびクラス間のつながりを構成するクラス間関係について，どの SPARQL エンドポイントに問い合わせを行うべきか(ソース選択)を特定する手法を確立する．それを可能とするために，事前にクロールして取得すべき情報について検討し仕様としてまとめる．決定した仕様に沿って各 SPARQL エンドポイントをクロールし，メタデータをして蓄積する．

(2) クラス間の意味的繋がりを辿るクエリに対して，蓄積したメタデータを利用して，ソース発見やクエリ実行を高速に行う連合検索手法を検討し，プロトタイプシステムとして実装する．

(3) 実装したプロトタイプシステムに対し，実用上の性能を評価するため，具体的な生命科学の例題を適用する．そこで得られた評価やフィードバックを元に連合検索システムの改良を行う．プロトタイプシステムを，より実用性が高いシステムとして仕上げ，サービスとして公開する．また，広く利用可能にするためにソースコードを GitHub で公開する．

４．研究成果

(1) クラス間関係にもとづく連合検索において，どの SPARQL エンドポイントに問い合わせを行うべきか(ソース選択)を計算する基盤として，各データベースから意味的つながりを記載したメタデータをあらかじめ抽出し，共通クラスを重ね合わせて頂点とし，クラス間のつながりを辺とするグラフ(連合クラスグラフ)を構築した．さらに，連合クラスグラフを用いたソース選択を評価できるよう，連合クラスグラフ上のクラス間パスからソースが選択された SPARQL クエリを出力するウェブ API を実装し，適切にクエリが構築されることを確認した．

また，連合クラスグラフを用いた効率のよい手法を設計する前準備として，連合クラスグラフの構造について解析を試みた．その結果，a)グラフの連結成分の大きさはべき乗測に従い，最大の連結成分および二つ目の大きさの連結成分においてのみ SPARQL エンドポイントにまた

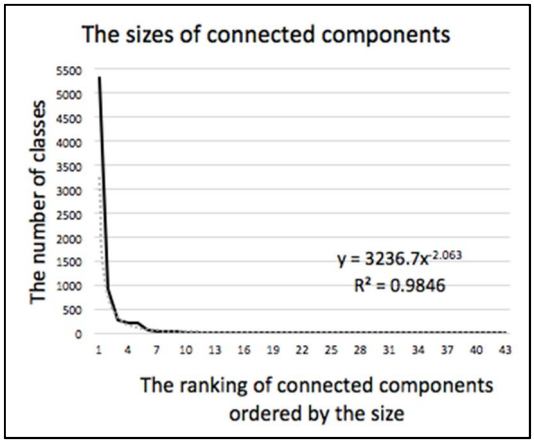

図１ 連合クラスグラフの連結成分のサイズ

がった検索が可能(図 1)，b)大きさ上位二つの連結成分は単連結であり，切断頂点に相当する概念は，生命科学分野における比較的上位概念である，の二点を示した．結果 a)より，連合クラスグラフに小さな連結成分が多いという問題が明らかになり，その問題に対処するため，上位クラスでまとめるようメタデータの設計を改良した．このことから，より多くの概念間がつながり，検索が可能となった．

(2) 連合クラスグラフを用いて，クラス間の意味的繋がりを辿るクエリに対して，ソース発見やクエリ実行を高速に行う連合検索システムを設計し，プロトタイプシステムの実装を行った．このシステムは LOD Surfer API とよばれる，ユーザがクラス間関係を利用して LOD から効率良くかつ柔軟にデータを切り出すためのウェブ API の機能の一部として実装された．この API を利用することで，対象となる SPARQL エンドポイント一覧，クラス一覧，あるクラスから到達可能なクラスの一覧，二つのクラスに対し，連合クラスグラフ上のクラス間パスのリスト，クラス間パスに対し，ソースを提示した連合検索クエリ，およびその結果を出力することができる．これらの機能は，連合クラスグラフおよび予め計算された連合クラスグラフの連結成分を用いて高速に実行できる．例えば，クラス間の到達可能性は連合クラスグラフの連結成分を用いて瞬時に計算でき，パスも連結成分内のみに探索空間が限定されるため，現実的な時間で計算できる．また，パスが提示されれば，連合クラスグラフの頂点(クラス)や辺(クラスをつなぐプロパティ)の情報を利用して高速に連合検索を行うことができる．

(3) プロトタイプシステムの実用上の評価のため，タンパク質配列のマルチプルアライメントビューワに組み込み，タンパク質アノテーションに適用することを試みた．その結果，現状の概念間パスからの連合検索を実用のアプリケーション上で利用する際の利点および課題を明らかにすることができた．
利点としては，ユーザは連合クラスグラフ上のクラス間パスを選ぶのみで，それらのデータがどこのデータベースにどのようなスキーマで格納されているかを考慮する必要がなく，柔軟なアノテーションが可能となった．課題は，メタデータを用いてクラスでまとめて高速化しているものの，タンパク質数が極端に増えると連合検索クエリによっては GUI 上では待てない検索時間になる場合があることが分かった．また，検索対象に頻繁に更新されるデータベースがある場合，メタデータを取得するクローラを走らせる頻度を上げる必要があるが，その対象が巨大なデータベースである場合，実行に時間がかかり，かつ SPARQL エンドポイントでデータを提供するサーバに多大な負荷をかけるという問題があった．そこで，RDF における標準的語彙である OWL や RDFS で定義されるクラスやクラス間関係，OWL や RDFS の語彙を通じて得られるメタデータを優先的に利用し，OWL や RDFS で定義されないクラスやプロパティについては従来の手法を用いるという方法で高速化を試みた．その結果，OWL の定義はメタデータの取得の高速化にほとんど寄与しないという結果が得られた一方で，RDFS で定義されるプロパティ群については，メタデータ取得を高速化することが可能であることが示された．

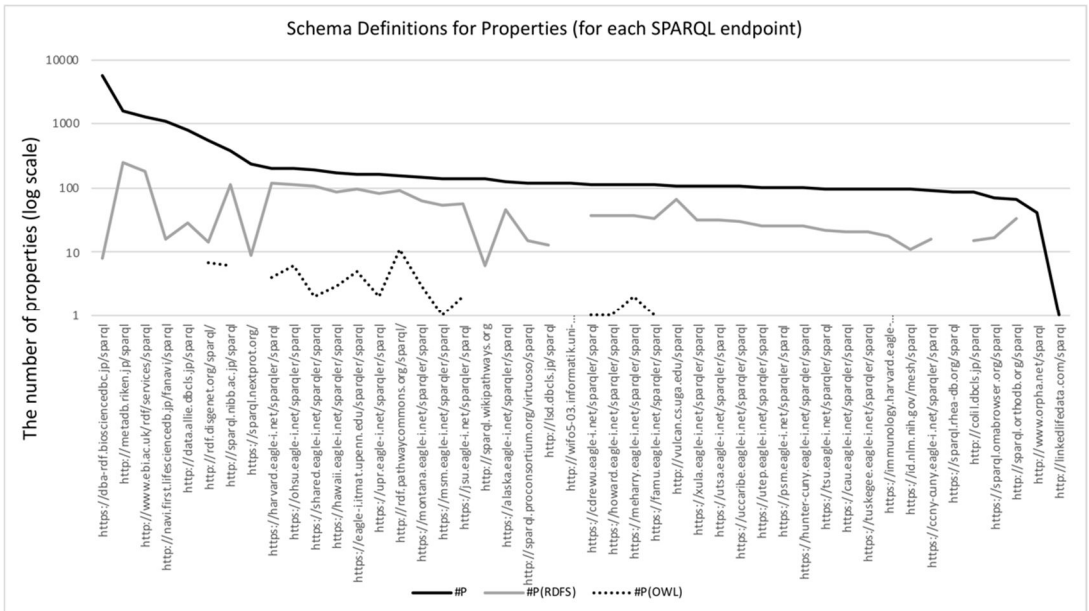

図 2 各 SPARQL エンドポイントに対する RDFS や OWL での定義状況

(4) 本研究でこれまで開発した，連合クラスグラフ上の概念間パスから連合検索クエリを生成する機能，取得したメタデータを元に概念間パスのクエリを高速に実行する機能などを LOD Surfer API の一部として組み込み，実行可能な状態で公開した．また，メタデータ取得のためのクローラ，メタデータから連合クラスグラフを構築するプログラム，連合検索機能を含んだ

LOD Surfer API，**全てのソースコードを** GitHub **より公開した.**

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Yamaguchi Atsuko、Kushida Tatsuya、Yamamoto Yasunori、Kozaki Kouji	４．巻 1157
２．論文標題 Investigating Schema Definitions Using RDFS and OWL 2 for RDF Databases in Life Sciences	５．発行年 2020年
３．雑誌名 Communications in Computer and Information Science	６．最初と最後の頁 137～144
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-981-15-3412-6_14	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 －

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　8件）

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（第32回）
４．発表年 2018年

１．発表者名 藤博幸，山口敦子
２．発表標題 linked open data を利用するアラインメントビューアの開発
３．学会等名 第18回 日本蛋白質科学会年会
４．発表年 2018年

１．発表者名 藤博幸，山口敦子
２．発表標題 Linked Open Dataを利用するアラインメントビューアの開発
３．学会等名 トーゴ―の日シンポジウム2018
４．発表年 2018年

１．発表者名
山口敦子，藤博幸

２．発表標題
LOD連合検索の課題と展望ータンパク質関連情報の取得を通じてー

３．学会等名
トーゴ―の日シンポジウム2018

４．発表年
2018年

１．発表者名
Atsuko Yamaguchi, Hiroyuki Toh

２．発表標題
Implementing LOD Surfer as a Search System for the Annotation of Multiple Protein Sequence Alignment

３．学会等名
The 8th Joint International Semantic Technology Conference (JIST2018)（国際学会）

４．発表年
2018年

１．発表者名
Norio Kobayashi, 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
The 11th International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4HCLS 2018)（国際学会）

４．発表年
2018年

１．発表者名
Atsuko Yamaguchi, Hiroyuki Toh

２．発表標題
Annotation of Proteins from LOD for a Viewer of Multiple Protein Sequence Alignment

３．学会等名
The 11th International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4HCLS 2018)（国際学会）

４．発表年
2018年

１．発表者名
山口敦子，櫛田達矢，山本泰智，古崎晃司

２．発表標題
ライフサイエンスのRDFデータベースにおけるスキーマ定義の現状分析

３．学会等名
第47回セマンティックウェブとオントロジー研究会

４．発表年
2018年

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

２．発表標題
Semantic Graph Analysis for Federated LOD Surfing in Life Sciences

３．学会等名
The 7th Joint International Semantic Technology Conference (JIST 2017)（国際学会）

４．発表年
2017年

１．発表者名
Atsuko Yamaguchi, Kouji Kozaki, Yasunori Yamamoto, Hiroshi Masuya, 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 Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS 2017)（国際学会）

４．発表年
2017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
第44回セマンティックウェブとオントロジー研究会

４．発表年
2018年

１．発表者名
Atsuko Yamaguchi, Hideki Hatanaka, Satoshi Fukuchi, Motonori Ota

２．発表標題
Semantic Integration of Intrinsically Disordered Proteins and Existing DBs

３．学会等名
12th International SWAT4HCLS Workshop（国際学会）

４．発表年
2019年

１．発表者名
Norio Kobayashi, Yasunori Yamamoto, Atsuko Yamaguchi

２．発表標題
Umaka data extension: Towards Realisation of Practical SPARQL Endpoint Discovery Service for Life Sciences

３．学会等名
12th International SWAT4HCLS Workshop（国際学会）

４．発表年
2019年

１．発表者名
Yasunori Yamamoto, Atsuko Yamaguchi

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

３．学会等名
12th International SWAT4HCLS Workshop（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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	小林　紀郎 (Kobayashi Norio) (20415160)	国立研究開発法人理化学研究所・情報システム本部・ユニットリーダー (82401)	

6．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	古崎　晃司 (Kozaki Kouji)		
研究協力者	桝屋　啓志 (Masuya Hiroshi)		
研究協力者	山本　泰智 (Yamamoto Yasunori)		