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(統合データ解析トライアル)
研究開発課題
「RDF ストア間データ連結フレームワークの開発およびオーソログ解析への適用」

研究開発終了報告書

研究開発期間： 平成26年9月 ～平成27年2月

研究代表者:千葉啓和
(自然科学研究機構　基礎生物学研究所
特任研究員)

§１ 研究開発の概要

様々なバイオデータベースの SPARQL エンドポイントが整備され、インターネット上の分散データベースが実現されつつある。しかしながら、SPARQL エンドポイントに対して問い合わせを行うためのクライアントソフトウェアは未だ未成熟である。そこで、まずは基礎的なクライアントを開発することが、データ解析を効率化するために必須であると考えた。

本課題では、UNIX コマンドラインで利用できる汎用的な SPARQL クライアント SPANG を開発した。SPANG は、簡単な操作で典型的な SPARQL を動的に生成する事ができる。また、プロセス間通信によってクエリ間でデータを受け渡す機構を持ち、複数のエンドポイントへの問い合わせを組み合わせる事ができる。

§２ 研究開発のねらい

様々なバイオデータベースの SPARQL エンドポイントが整備されることによって、インターネット上に分散型バイオデータベースが実現されつつある。これらを統合的に解析し知識発見するためには、RDF ストア間のデータ連結を可能にする技術が必須である。しかしながら現状ではそのための技術が未成熟であり、ツール開発が必要となっている。本課題では、RDF ストア間でのデータ連結を容易に行うことのできる汎用的なフレームワークを開発する。

本フレームワークを利用した解析の事例として、オーソログデータベースを中心とした遺伝子情報の統合を行うことによる知識発見を試みる。オーソログ関係は、種分化によって分岐した遺伝子間の関係を表しており、同様の機能を持つ遺伝子を対応付けるための情報として利用可能である。様々な生物種のデータを含むバイオデータベースを利用する際に、オーソログデータを情報のハブとして機能させることで、生物種間での情報の移転を促し、生物学的な知識の発見が期待される。

対象データベースは、統合化推進プログラムで開発された「微生物関連データベース」に含まれる MBGD SPARQL エンドポイント（http://sparql.nibb.ac.jp/sparql）、UniProt SPARQL エンドポイント（http://beta.sparql.uniprot.org）を想定したが、その他の SPARQL エンドポイントも対象となりうる。

§３ 研究開発計画

（1）当初の研究開発計画

本課題では、RDF ストア間でのデータ連結を容易に行うことのできる汎用的なフレームワークを開発する。一般に複雑な構造になりがちな federated query を記述する代わりに、対象とする RDF ストア毎に処理を分解して、単純化した SPARQL クエリを記述しておき、それらを連結させて実行できるようにする。分解した SPARQL クエリの連携は、UNIX コマンドラインのパイプとして実現する。

SPARQL の記述を煩雑にする要因として、リソース URI の prefix 記述がある。多岐に渡るデータを利用して統合解析を行う際には、多くの prefix を記述しなければならないが、その際に記述誤りを生じやすい。この問題を解決するため、よく用いられる prefix のリストを保存しておき、必要な prefix 記述を自動で行うようにする。

また、SPARQL の仕様には、定型的なコードパターンを記述する機能がないため、よく似た SPARQL コードを繰り返し記述する事態がしばしば生じ、生産性を下げている。この問題を解決するため、頻繁に記述する SPARQL コードの可変部をパラメータとして記述し、テンプレートの形で保存しておくようにする。この SPARQL テンプレートに対してパラメータを指定することで、同じような SPARQL を再度記述することなく実行できるようにする。

【使用言語】
Perl 言語を使用する。

【実行方法概要】
UNIX コマンドとして利用する。コマンドライン引数に、SPARQL エンドポイント、および SPARQL を記述したファイルを指定する。必要に応じて設定ファイルに記述を追加することで実行を簡便にする。標準出力として結果が得られる。コマンドラインオプションの指定によって、JSON やタブ区切り等の出力フォーマットを制御する。
複数の SPARQL エンドポイントへの問い合わせを組み合わせる場合には、UNIX のパイプでコマンドを連結させ、リストあるいはテーブル構造でデータを受け渡す。標準入力の内容を SPARQL の中に取り込む際には、SPARQL コード中に記述した$STDIN が特別な変数として認識され、標準入力の内容が内部的に実行時展開される。

【prefix リスト機能】
全てのユーザーで共通で用いられる prefix リストファイルを用意し、一般的によく用いられる prefix をあらかじめ記述しておく。さらに、ユーザー毎に用いることのできる prefix リストファイルも用意する。SPARQL の実行時に、SPARQL 中で用いられている prefix を認識し、必要な prefix 記述を SPARQL の先頭に自動で追加する。

【SPARQL テンプレート機能】
頻繁に記述する SPARQL コードにおける可変部分をパラメータとして記述し、SPARQL テンプレートとして保存しておくようにする。コマンド実行時には、SPARQL テンプレートとパラメータをコマンドラインにて指定する。

【研究開発の進め方の概要】
まず、Perl による UNIX コマンドの実装、およびその設定ファイルの作成を行う。続いて、対象データベースを使った解析に入るが、プログラムについても引き続きテスト・デバッグを行い完成度を高めていく。解析については、オーソログ情報をデータのハブとして効果的に利用した知識発見につながるような解析事例を作成する。さらに、これらの解析の結果に基づき、生物学的な解釈を行い、得られた知見をまとめる作業を行う。

(2)新たに追加・修正など変更した研究開発計画
中間報告会でのコメントをうけて、以下のような点を新たに研究開発計画に追加した。
より多くのユーザーに受け入れられるように、コマンドラインクライアントのインターフェースを整える。具体的には、よく用いる SPARQL エンドポイントに関して、SPANG システム内でニックネームを登録しておき、SPANG のコマンドラインでこのニックネームを指定すると、目的とする SPARQL エンドポイントに接続することができるようにする。
また、SPANG のユーザーコミュニティーを醸成することを目指した仕組みを導入する。具体的には、よく用いる SPARQL クエリがインターネット上に公開されれば、それを SPANG が読み込んで実行できるようにする。すなわち、SPARQL クエリをライブラリ化する方法を考案するとともに、それを SPANG のコマンドラインから指定して実行する方法を考案して、SPANG コマンドに実装する。

§４ 研究開発成果

上記研究開発計画に挙げた、SPANG に実装すべき機能は実装した。また、開発計画の中には詳細には記述していなかったが、SPANG コマンドには、ユーザーの利便性をはかるための数多くのコマンドラインオプションを設けた。これらのオプションは多岐にわたっており、かつ今後も追加される可能性があるため、SPANG document という Wiki サイトを立ち上げて、その中にまとめることにした。下記の URL からアクセスできる。

http://purl.org/net/spang

上記 URL は、purl サービスを用いて転送されるようになっており、現在は基生研のサーバー上に置いた Wiki サイトに転送するように設定してある。今後の SPANG のさらなる開発の展開によっては、転送先が変更されることも想定して、このような設定にした。

SPANG の簡単な使い方は、別紙に示した図のように、コマンドラインでオプションを指定することによって、典型的な SPARQL コードパターンを内部的に生成して、SPARQL エンドポイントへの問い合わせを実現するというものである。この使い方においては、SPARQL エンドポイントに対するニックネームを指定している。こうしたニックネームは、SPANG システムの設定ファイルにあらかじめ登録しておき指定することができるようにした。なお SPARQL エンドポイントの URL を直接指定してもよい。この SPARQL エンドポイントに対するニックネームに関しては、SPANG システムディレクトリ内の設定ファイルに記述できるほか、ユーザーのホームディレクトリの設定ファイルにもできるようにした。SPANG をある UNIX システムのユーザー間で共有して利用する際に、全ユーザーに共通の設定だけでなく、ユーザー特有の設定も行いたいという状況を考えて、このような仕組みを設けた。また、よく用いる prefix の設定についても、同様の理由から、ユーザー共通の設定に加えて、ユーザーごとの設定ができるようにした。

さらに、SPANG のユーザーコミュニティーを醸成することを目指して、インターネット上に公開された SPARQL コードを利用する仕組みを導入した。具体的には、SPARQL ライブラリを公開する仕組みを考案するとともに、SPANG コマンドから利用する仕組みを考案して実装した。以下の URL において、試験的に構築した SPARQL ライブラリを見ることができる。

http://purl.org/net/spang/library

SPANG コマンドを実行する際、ローカルの SPARQL ファイルを指定する代わりに、インターネット上の SPARQL クエリの URI を引数として指定して、実行することができる。この SPARQL クエリの URI もまた、定義済み prefix を使って短縮した表現で指定することができる。

SPANG を用いて SPARQL クエリを組み合わせて実行する方法についても、別紙の図に示した。複数の SPARQL クエリの間でのデータのやりとりは、プロセス間通信、特に、UNIX のパイプを使って実現している。データを受け渡す側の SPARQL クエリでは、SELECT 文によって受け渡したいデータを生成した上で、リストあるいはテーブル形式で標準出力に出力するようにする。データを受け取る側の SPARQL クエリの中で、特殊変数$STDIN を記述しておくと、標準入力から読み込んだデータが動的に埋め込まれる。さらに、SPARQL のコードの中に、コマンドライン引数を埋め込むことができる。このように、SPARQL のコードの中で、可変部分をパラメータ化したもの（SPARQL テンプレート）を作成すると、一度作成したクエリが特定の用途に限定されず、より汎用的なクエリとなる。すなわち、SPANG コマンドから SPARQL テンプレートを利用することによって、SPARQL を用いた研究開発の効率を高めることができ、RDF ストアを利用したデータ解析を促進すると期待される。

§5 研究開発計画に対する達成状況と将来展望

(1)達成状況

当初の研究開発計画は、フレームワークの実装に関しては達成することができた。また、中間報告会をうけて新たに追加・修正した研究開発計画に含まれる機能についても、実装することができた。知識発見につながるようなデータ解析への取り組みについてはまだ取り組む余地が残されている。

(2)ツール等の将来展望

SPANG コマンドに対して指定する SPARQL クエリに関して、ローカルのファイルだけでなく、インターネット上で公開された SPARQL ライブラリを実行できるようにした。現在は、試験的な SPARQL ライブラリを作成した段階であるが、様々な人が SPARQL ライブラリを公開すれば、一度作成したクエリを SPANG のユーザー間で容易に共有できるようになり、SPARQL を利用したデータ解析の効率をより高めることができると期待される。

§6 研究参加者

氏名	所属	役職	研究開発項目	参加時期
千葉啓和	基礎生物学研究所	特任研究員	ツール開発および解析	H26.9-H27.2

§7 成果発表等

(1)原著論文発表 (国内(和文)誌 0 件、国際(欧文)誌 0 件)

(2)その他の著作物(総説、書籍など)

なし

(3)国際学会発表及び主要な国内学会発表

① 招待講演 (国内会議 0 件、国際会議 0 件)
② 口頭発表 (国内会議 0 件、国際会議 0 件)
③ ポスター発表 (国内会議 0 件、国際会議 0 件)

(4)知財出願

①国内出願 (0 件)
②海外出願 (0 件)
③その他の知的財産権

なし

(5)受賞・報道等

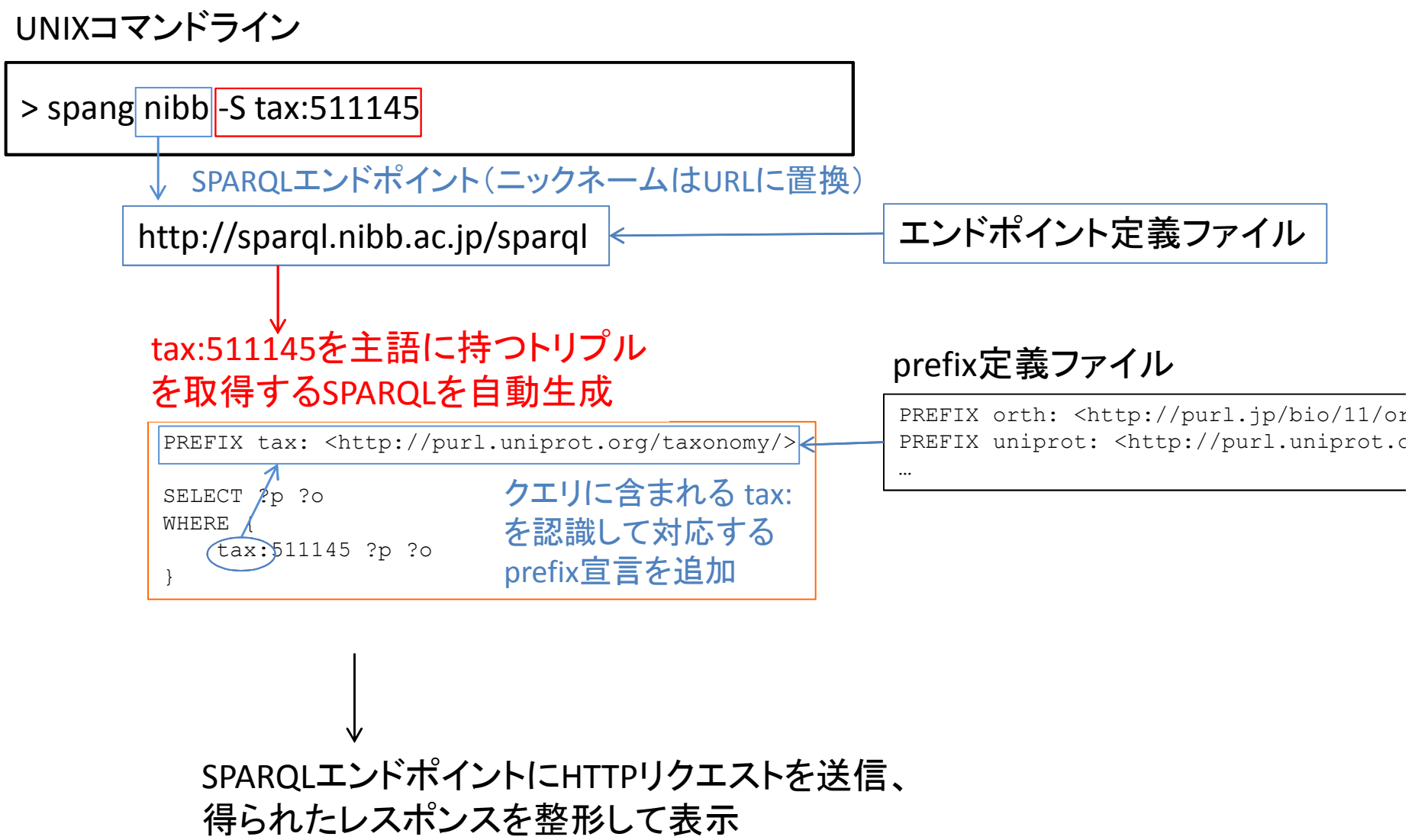
なし

§8 自己評価

RDF ストア間データ連結フレームワークの開発が、当初の研究開発計画における中心的なねらいであった。それは達成できたと考えているが、さらに、より広い意味で有用なツールができたと考えている。すなわち、必ずしも複数の RDF ストアを組み合わせるという用途に限定されるものでなく、単一の RDF ストアを対象とする SPARQL 問い合わせにおいても、特徴的な機能を持ったツールとなったと考えている。

また、中間報告会において、ツール利用者のコミュニティーを広げるにはどうしたらいいか、また、ツールの利用者を世界に広めるにはどうしたらいいか、という視点から貴重なコメントを頂き、中間報告会以降は、これらの点についても考慮しながら開発を進めるように努めた。その結果として、コマンドラインターフェースをより整ったものにすることができ、公開 SPARQL ライブラリを利用する機能などを備えた、より発展性のあるツールになったと考えている。

SPANGを利用した簡単な問い合わせ



SPANGを利用したSPARQLの組み合わせ