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

千葉啓和
基礎生物学研究所

背景と目的

- バイオデータベースのRDF化が進んでいる
  - インターネット上の分散データベースが実現しつつある
- SPARQLクライアントが未成熟なため、問い合わせる際の作業効率が低い
  - 特に、**RDFストア間**でのデータ連結はやりにくい

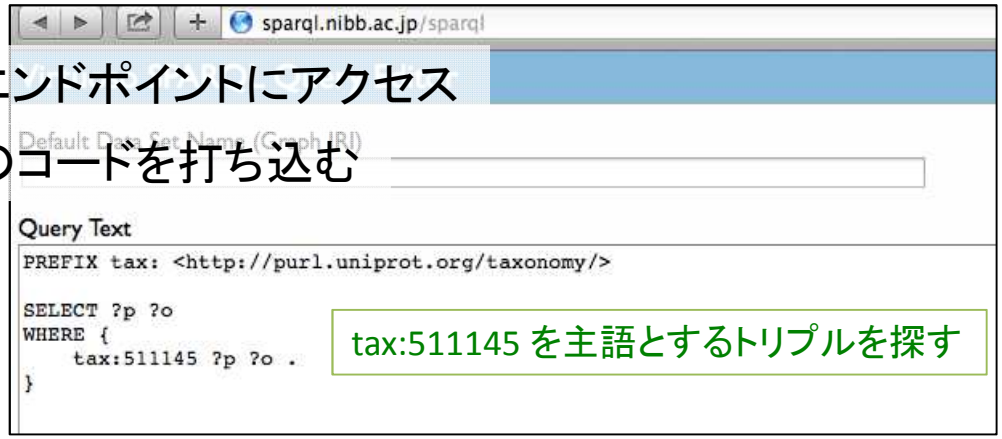
- RDFストアに効率よく問い合わせを行うための枠組みを開発
  - 複数のエンドポイントへのアクセスにも対応

- **オーソログ情報**を利用した遺伝子情報の統合的解析の試み

通常のSPARQL実行方法

SPARQLエンドポイントにアクセス

SPARQLのコードを打ち込む



p	o
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.uniprot.org/core/Taxon
http://www.w3.org/2000/01/rdf-schema#subClassOf	http://purl.uniprot.org/taxonomy/83333
http://purl.uniprot.org/core/otherName	"Escherichia coli MG1655"
http://purl.uniprot.org/core/otherName	"Escherichia coli str. K12 substr. MG1655"
http://purl.uniprot.org/core/otherName	"Escherichia coli str. MG1655"
http://purl.uniprot.org/core/otherName	"Escherichia coli strain MG1655"
http://purl.uniprot.org/core/partOfLineage	0
http://purl.uniprot.org/core/reviewed	0
http://purl.uniprot.org/core/scientificName	"Escherichia coli str. K-12 substr. MG1655"
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/1208340
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/1385755
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/694514
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/694515
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/694516
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/694517
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/694518
http://www.w3.org/2004/02/skos/core#narrowerTransitive	http://purl.uniprot.org/taxonomy/694519

- ✓ 単純な処理にも煩雑な操作が必要
- ✓ 検索結果を利用した解析がしにくい

SPANG: a command-line client supporting query generation for distributed SPARQL endpoints

- コマンドによってSPARQLを動的に生成
  - 典型的コードパターンの生成
  - prefix宣言の生成

  → 基本的な問い合わせ作業を効率化

- 複数のSPARQLクエリを組み合わせることが可能
  - クエリ毎にプロセスを生成
  - プロセス間通信により他の処理と組み合わせる

  → SPARQLコードのモジュール化による運用性の向上

SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -S tax:511145
```

SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -S tax:511145
```

SPARQLエンドポイント（ニックネームはURLに置換）

http://sparql.nibb.ac.jp/sparql

エンドポイント定義ファイル

SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -S tax:511145
```

SPARQLエンドポイント（ニックネームはURLに置換）

http://sparql.nibb.ac.jp/sparql

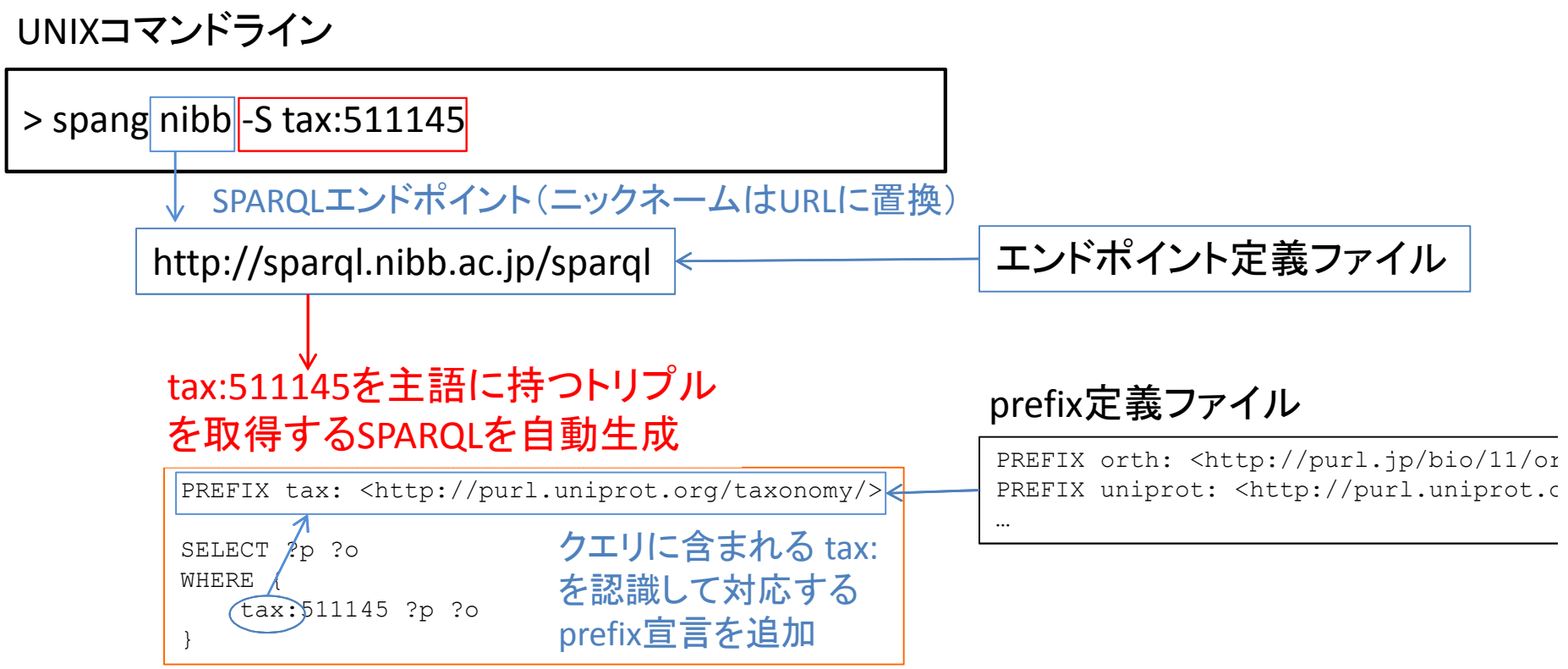
エンドポイント定義ファイル

tax:511145を主語に持つトリプル
を取得するSPARQLを自動生成

```
PREFIX tax: <http://purl.uniprot.org/taxonomy/>

SELECT ?p ?o
WHERE {
    tax:511145 ?p ?o
}
```

SPANGを利用した問い合わせ



SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -S tax:511145
```

SPARQLエンドポイント（ニックネームはURLに置換）

http://sparql.nibb.ac.jp/sparql

エンドポイント定義ファイル

tax:511145を主語に持つトリプル
を取得するSPARQLを自動生成

```
PREFIX tax: <http://purl.uniprot.org/taxonomy/>

SELECT ?p ?o
WHERE {
    tax:511145 ?p ?o
}
```

クエリに含まれる tax:
を認識して対応する
prefix宣言を追加

prefix定義ファイル

```
PREFIX orth: <http://purl.jp/bio/11/or
PREFIX uniprot: <http://purl.uniprot.c
…
```

SPARQLエンドポイントにHTTPリクエストを送信、
得られたレスポンスを整形して表示

SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -S tax:511145
```

↓ 標準出力

```
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>       <http://purl.uniprot.org/core/Taxon>
<http://www.w3.org/2000/01/rdf-schema#subClassOf>       <http://purl.uniprot.org/taxonomy/83333>
<http://purl.uniprot.org/core/otherName>        "Escherichia coli MG1655"
<http://purl.uniprot.org/core/otherName>        "Escherichia coli str. K12 substr. MG1655"
<http://purl.uniprot.org/core/otherName>        "Escherichia coli str. MG1655"
<http://purl.uniprot.org/core/otherName>        "Escherichia coli strain MG1655"
<http://purl.uniprot.org/core/partOfLineage>    "0"^^<http://www.w3.org/2001/XMLSchema#integer>
<http://purl.uniprot.org/core/reviewed> "0"^^<http://www.w3.org/2001/XMLSchema#integer>
<http://purl.uniprot.org/core/scientificName>   "Escherichia coli str. K-12 substr. MG1655"
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/12083
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/13857
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69451
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69451
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69451
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69451
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69451
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69451
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69452
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69452
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69452
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69452
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69452
<http://www.w3.org/2004/02/skos/core#narrowerTransitive>        <http://purl.uniprot.org/taxonomy/69452
```

SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -S tax:511145 -a
```

↓ prefix定義を利用し、省略形に変換

prefix定義ファイル

```
rdf:type          up:Taxon
rdfs:subClassOf   tax:83333
up:otherName      "Escherichia coli MG1655"
up:otherName      "Escherichia coli str. K12 substr. MG1655"
up:otherName      "Escherichia coli str. MG1655"
up:otherName      "Escherichia coli strain MG1655"
up:partOfLineage          "0"^^xsd:integer
up:reviewed       "0"^^xsd:integer
up:scientificName         "Escherichia coli str. K-12 substr. MG1655"
skos:narrowerTransitive tax:1208340
skos:narrowerTransitive tax:1385755
skos:narrowerTransitive tax:694514
skos:narrowerTransitive tax:694515
skos:narrowerTransitive tax:694516
skos:narrowerTransitive tax:694517
skos:narrowerTransitive tax:694518
skos:narrowerTransitive tax:694519
skos:narrowerTransitive tax:694520
skos:narrowerTransitive tax:694521
skos:narrowerTransitive tax:694522
skos:narrowerTransitive tax:694523
skos:narrowerTransitive tax:694524
skos:narrowerTransitive tax:694525
```

SPANGを利用した問い合わせ

主語と述語を指定した場合

```
> spang nibb -S tax:511145 -P up:otherName
```

```
"Escherichia coli MG1655"
"Escherichia coli str. K12 substr. MG1655"
"Escherichia coli str. MG1655"
"Escherichia coli strain MG1655"
```

```
> spang nibb -S tax:511145 -P rdfs:subClassOf+/up:scientificName
```

```
"Proteobacteria"
"Gammaproteobacteria"
"cellular organisms"
"Bacteria"
"Enterobacteriaceae"
"Escherichia"
"Escherichia coli"
"Escherichia coli (strain K12)"
"Enterobacteriales"
```

SPANGを利用した問い合わせ

UNIXコマンドライン

```
> spang nibb -O tax:511145 -a
```

↓ 目的語を指定した場合

```
tax:694525      rdfs:subClassOf tax:511145 .
tax:694517      rdfs:subClassOf tax:511145 .
tax:694532      rdfs:subClassOf tax:511145 .
tax:694524      rdfs:subClassOf tax:511145 .
…
```

SPANG：その他の主なオプション

```
> spang nibb -G
```

GRAPH句を生成する
→RDFストア内のグラフ一覧を取得できる

```
> spang nibb -F mbgdr:default -L 10
```

FROM句を生成して
グラフを限定

出力行数を制限
（LIMIT 10）

```
> spang nibb -F mbgdr:default -N
```

FROM句を生成して
グラフを限定

出力行数を数える
（COUNT関数）

```
> spang nibb -F mbgdr:default -f json
```

出力フォーマットの指定
tsv, json, n-triples (nt), turtle (ttl), rdf/xml (rdfxml), n3, xml, html

SPANG：その他の主なオプション

```
> spang nibb -F mbgdr:default -N -q
```

内部的に自動生成されたSPARQL
を出力して終了

```
PREFIX mbgdr: <http://mbgd.genome.ad.jp/rdf/resource/>

SELECT COUNT(*)
FROM mbgdr:default
WHERE {
    ?s ?p ?o
}
```

SPANG：その他の主なオプション

```
> spang nibb -F mbgdr:default -N -q
```

内部的に自動生成されたSPARQL
を出力して終了

```
PREFIX mbgdr: <http://mbgd.genome.ad.jp/rdf/resource/>

SELECT COUNT(*)
FROM mbgdr:default
WHERE {
    ?s ?p ?o
}
```

出力されたSPARQLを保存、編集して
再実行することも可能

```
> spang nibb ./query.rq
```

SPARQLファイルのパスを指定

SPANGを利用した問い合わせ（推論）

UNIXコマンドライン

```
> spang nibb -S tax:511145 -P rdfs:label -i ontologies -a
```

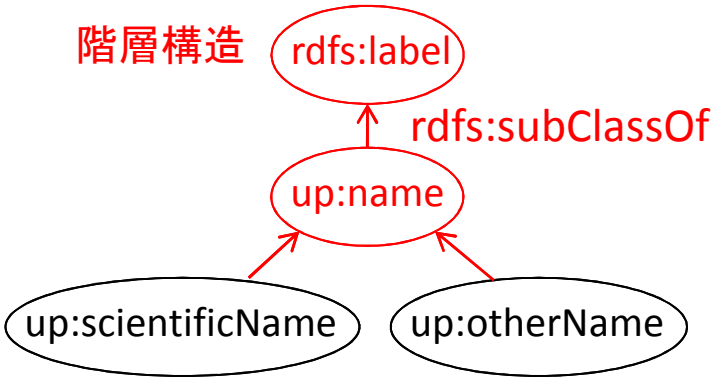
オントロジーを利用した推論を、RDFストア(Virtuoso)内で有効にする

SPARQL

```
define input:inference "ontologies"

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX tax: <http://purl.uniprot.org/taxonomy/>

SELECT ?o
WHERE {
    tax:511145 rdfs:label ?o
}
```



結果

```
"Escherichia coli MG1655"
"Escherichia coli str. K-12 substr. MG1655"
"Escherichia coli str. K12 substr. MG1655"
"Escherichia coli str. MG1655"
"Escherichia coli strain MG1655"
```

SPARQLライブラリを共有する機能

- ローカルなSPARQLライブラリを構築できる

  (System-wide) *SPANG_DIR/query/**

  (User-specific) *~/.spang/query/**

```
> spang nibb count_graphs.rq
```

ローカルのライブラリから見つける

```
> spang nibb count_graphs
```

.rqは省略可能

- 公開ライブラリをネットワーク越しに呼び出せる

```
> spang nibb http://purl.org/net/spang/library/count_graphs
```

クエリを取得するURI

```
> spang nibb lib:count_graphs
```

prefixを使って
省略できる

SPARQLクエリの汎用性を高める仕組み

- 引数を受け取ることができる

```
> spang nibb count_graph mbgdr:default
```

与えられたグラフ
のトリプル数をカウントする

- 標準入力を受け取ることができる

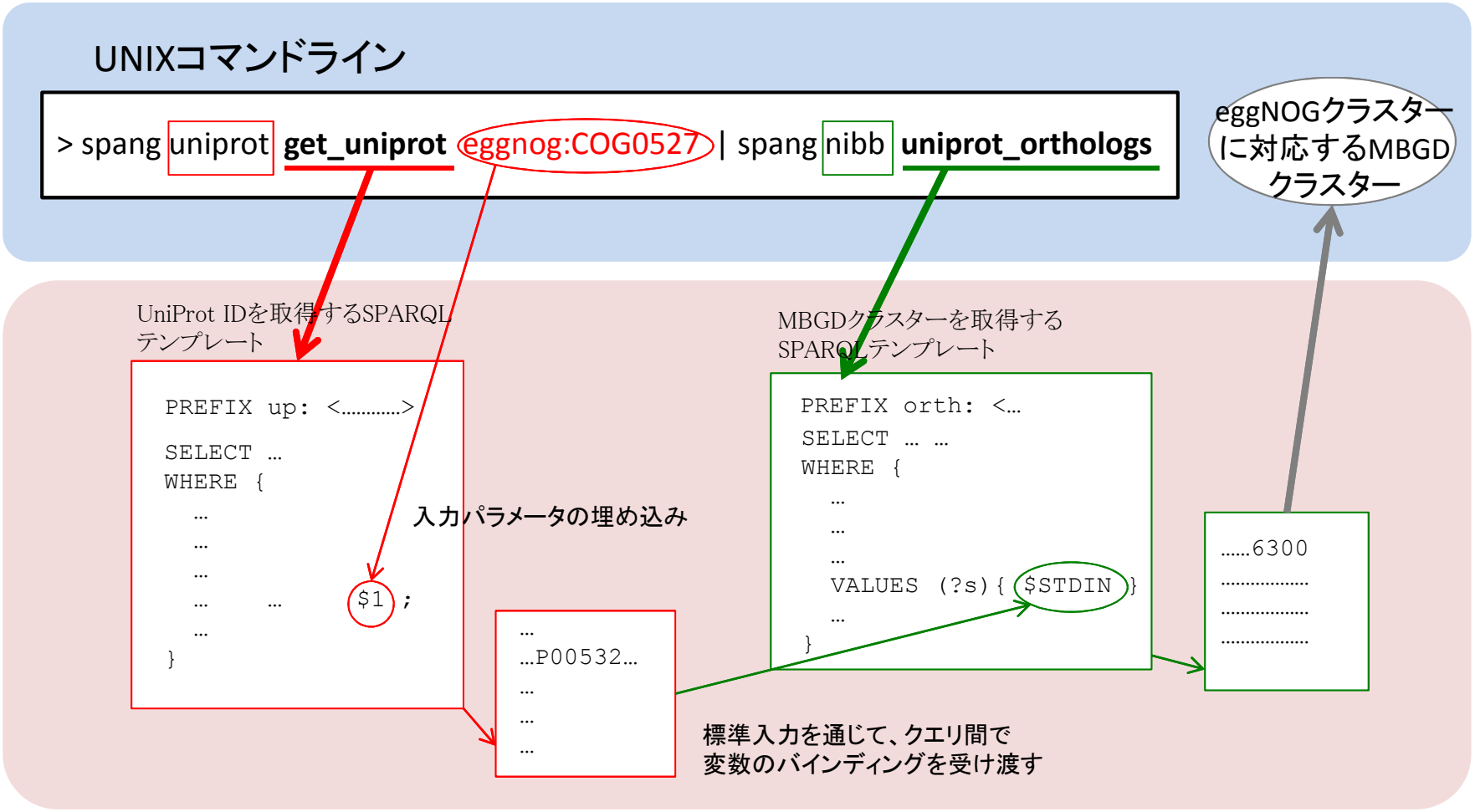
```
> spang nibb get_orthologs  rpoB | spang uniprot  get_annotation
```

```
> spang nibb get_orthologs  rpoB | spang uniprot  -S 1  -P rdfs:label
```

基生研にアクセスして、
オーソログ遺伝子リスト
を取得する(UniProt ID)

UniProtにアクセスして、
UniProt IDに対して付けられた
アノテーションを取得する

SPANGを利用したSPARQLの組み合わせ



SPANG 公開URL: http://purl.org/net/spang

SPANG

A command-line client supporting query generation for distributed SPARQL endpoint

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

Recently, an increasing number of biological databases have been made availabl
Network (**RDF**) and accessible through **SPARQL** endpoints, forming together a
data integration across the web. However, writing SPARQL codes for querying t
for biologists; thus, an easy-to-use querying interface is necessary. Here, we dev
client, **SPANG**. SPANG can **dynamically generate** typical SPARQL queries de
and arguments. SPANG supports interprocess communication to pass the variabl
execution of combination of queries and integration of data across the **multiple**
also search local RDF files besides remote data stores. These features provide th
distributed data described in RDF, thus enhances the **integrative analysis** of bio
platform.

Documentation

SPANG wiki

Program downloads

spang-0.2.3.tgz

Library

SPARQL Library

MBGD

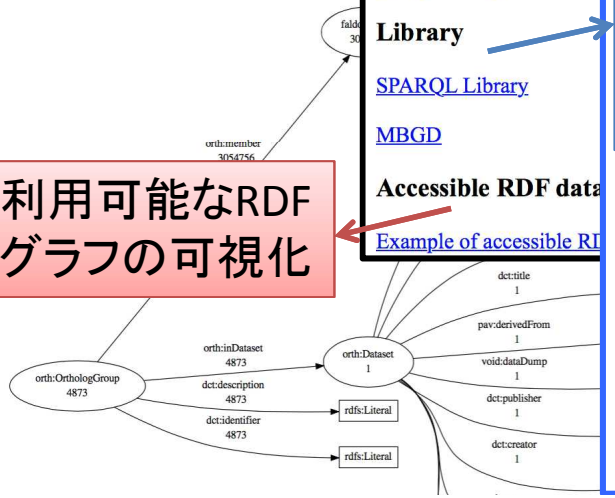
Accessible RDF data

Example of accessible RDF

Example of accessible RDF

(1) mbgd cog
(2) mbgd default
(3) mbgd eggnog_organism
(4) mbgd nog
(5) mbgd nucseq
(6) mbgd organism
(7) mbgd orthoxml_example
(8) mbgd protein
(9) mbgd uniprot-goa
(10) mbgd xref_uniprot
(11) uniprot citationmapping
(12) uniprot citations
(13) uniprot databases
(14) uniprot diseases
(15) uniprot enzymes
(16) uniprot go
(17) uniprot journals
(18) uniprot keywords
(19) uniprot locations
(20) uniprot pathways
(21) uniprot taxonomy
(22) uniprot tissues
(23) uniprot uniparc

(1) mbgd cog



利用可能なRDF
グラフの可視化

SPARQL Library

owing SPARQL queri
ql/

g eggNOG cluster

g eggNOG cluster wit

(3) cluster_to_go.rq
Description: Get GO annotation

(4) cluster_to_go_count.rq
Description: Get GO annotation

(5) cluster_to_organism.rq
Description: Get organism distribution

(6) cluster_to_protein.rq

(7) gene_ortholog.rq

利用可能な
SPARQLの
ライブラリ

ヘルプページ
（Wiki）

Usage [edit]

Usage of basic options [edit]

The subject, predicate, and the object are specified by options, **-S**, **-P**, **-O**,

pecifying the subject

```
S tax:511145
```

rm using prefix definitions (**-a option**)

```
spang nibb -S tax:511145 -a
```

Get sorted results

```
spang nibb -S tax:511145 -a | sort
```

Search by specifying the subject and predicate

```
spang nibb -S tax:511145 -P up:otherName -a
```

Search by specifying the object

```
spang nibb -O tax:511145 -a
```

Usage of other options [edit]

Get the list of graphs

```
spang nibb -G
```

Search from a specific graph (**-F option**) and limit the output (**-L option**)