RDFによるデータ統合と相互運用性のための技術開発

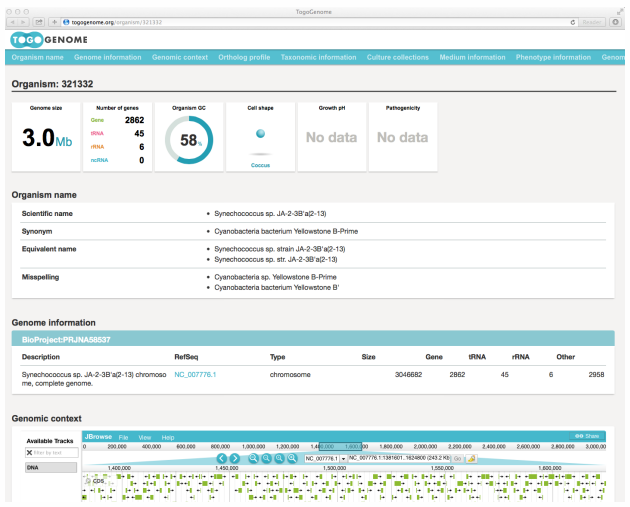
片山 俊明 (ライフサイエンス統合データベースセンター)
三島 博之 (長崎大学原爆後障害医療研究所)
岡本 忍 (ライフサイエンス統合データベースセンター)
西澤 達也 (情報数理バイオ)
川島 秀一 (ライフサイエンス統合データベースセンター)
藤澤 貴智 (国立遺伝学研究所)

基盤的データ・公共データ・小規模データを、セマンティック・ウェブ技術で統合するための最近の技術開発について報告する。生命科学のDBに共通する基盤的なデータについては、統合化推進プログラムの参画機関と月１回開催のSPARQLthonなどの会議を通じて、必要なオントロジーの開発、URIやRDFモデルの整備、データのRDF化について技術的な支援を行っている。特に、ゲノム情報においては、これまで微生物を対象にTogoGenomeを開発してきたが、今年度DDBJ, SIB, EBIなどのグループとともに、INSDCの塩基配列DBとEnsemblゲノムDBで共通のデータモデルを開発し真核生物も対象として相互運用性の向上を図った。一方で、まだRDF化が進んでいない公共DBについては、TogoWSサービスを通じて、データ取得とRDF化をオンデマンドで提供している。昨年度からUCSCのゲノムDBにも対応し、REST APIを用いて容易にデータアクセスが可能になった。さらに、研究者のもつ小規模データについてはTogoDBサービスを提供しており、表形式のデータを容易に高機能なDBとして公開できるだけでなく、RDF生成やSPARQLエンドポイントの公開、オントロジー編集機能の追加、D2RQの採用によるタイムラグのないデータ管理への対応などを進めている。

http://togogenome.org/

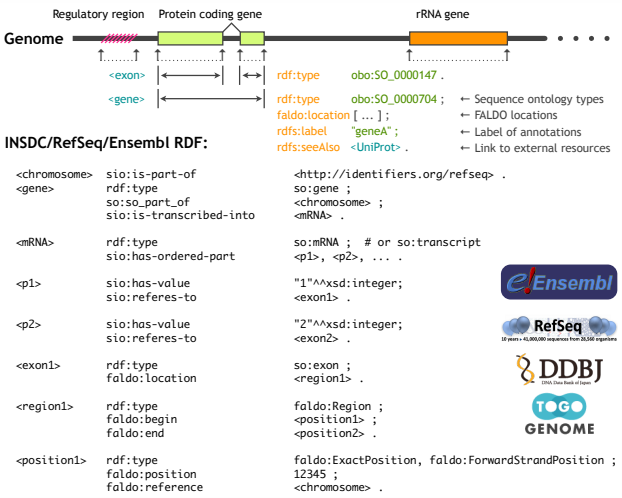
基盤的データ

現在の生命科学はゲノム情報が基盤となっているが、ヒトやモデル生物のゲノムデータベースで集積されている情報は、必ずしも再利用性の高い形で統合化されていない。ライフサイエンス統合データベースセンター (DBCLS) のTogoGenomeでは、生物種、 環境情報、 培養条件、ゲノム配列、遺伝子アノテーション、タンパク質の機能と立体構造、糖鎖やメチル化などの修飾、進化系統など、多様なデータのセマンティック・ウェブ技術での統合を進めている。



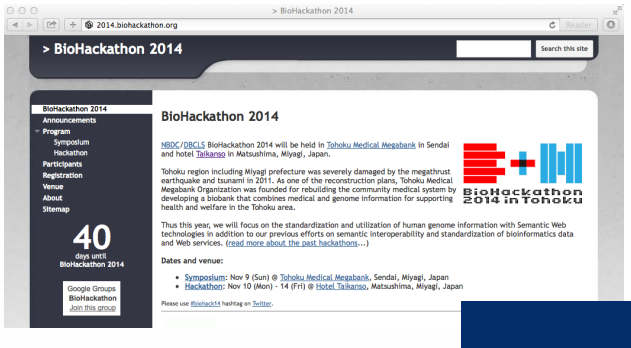
RDF summit

2014年5月にDBCLSにおいてRDFサミットを開催し、DBCLS/DDBJ共同でTogoGenomeのために開発してきたゲノム情報のRDF化モデルと、欧州バイオインフォマティクス研究所 (EBI) で開発されているゲノムデータベースEnsemblのRDFモデルの共通化について合意した。



BioHackathon / SPARQLthon

ライフサイエンスのデータベースをセマンティック・ウェブ技術を用いて統合化するため、国際標準化の促進と相互運用性向上に必要な技術開発を行う国際会議BioHackathonを年に1回 (今年で7回目)、国内の統合化推進プログラムのデータをRDF化し再利用性を向上するため技術開発合宿SPARQLthonを毎月1回 (先月で24回目)、それぞれ継続的に開催し、オントロジー・RDFデータ・アプリケーションの開発などを行ってきている。



http://togows.org/

公共データ

多くの公共データベースは独自のウェブページやREST APIによってデータ提供されていることが多い。それぞれ利用方法が異なるこれらのデータベースを統一的な方法で検索し、エントリを取得するため、DBCLSではTogoWSを開発してきた。TogoWSではデータ形式の変換機能を提供しているが、ここでRDF化への対応も進めている。

検索API http://togows.org/search/database/query+string[/offset,limit][.format]

TogoWSでは http://togows.org/search/ のあとに、データベース名とキーワードを指定することでエントリのリストを取得できる。

取得API http://togows.org/entry/database/entry_id[,entry_id2,...][/field][.format]

エントリのリストを検索したら、/entry のあとに、データベース名とエントリIDを指定するだけで、該当するエントリを取得できる。この際、エントリをJSONやRDFなど様々な形式に変換したり、指定した部分だけを取得することが可能となっており、解析ワークフローに組み込むといった相互運用性の高いデータの利用が実現している。

変換API http://togows.org/convert/data_source.format

オープンバイオ (BioPerlやBioRubyなど) のライブラリやツールを活用し、生命科学で使われる様々なデータ形式を相互に変換する機能を提供しており、手持ちのデータをRDFに変換することもできる。

外部API http://togows.org/api/ucsc/db/table/column[!]=value[/offset,limit][.format]

MySQLで公開されているUCSCゲノムデータベースなど、ウェブサービスのない外部データベースにも統一的なAPIを提供している。

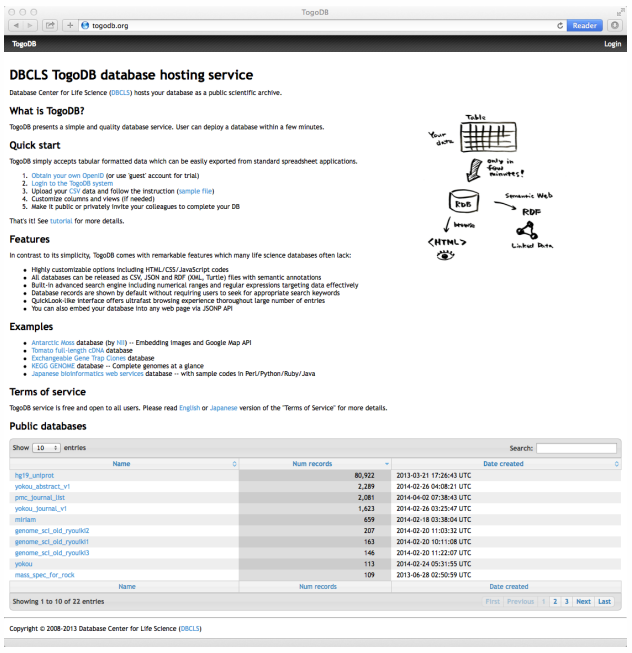
例：

```
% curl http://togows.org/api/ucsc/hg38/refGene/name2=UVSSA
% curl http://togows.org/api/ucsc/hg38/chr1:12,345-12,500.fasta
```

http://togodb.org/

小規模データ

研究者や研究室の持つデータは、これまで論文のサプリメントとして公開されることが多く、データベースとして永続化されることは少なかった。DBCLSで提供しているTogoDBを利用すると、小規模なデータベースであっても容易に構築しインターネットで公開することができる。TogoDBでは画面デザインを含め様々なカスタマイズが可能だが、既存または独自のオントロジーを利用したRDF化にも対応しており、再利用性の高いデータベースを容易に提供することができる。



また、バイオサイエンスデータベースセンター (NBDC) の生命科学系データベースアーカイブでもTogoDBが活用されている。

D2RQ

すでにデータベースが関係データベース (RDB) で構築されている場合には、D2RQを使うことによってSPARQL検索可能なRDFのデータベースにすることができる。DBCLSでは現在、D2RQの設定を容易にするためのアプリケーションを開発中である。

CC-BY © 2014 片山俊明 licensed under Creative Commons Attribution 2.1 Japan (CC表示2.1日本)