既存 RDB を効率的に RDF 化する支援ツール D2RQ Mapper

D2RQ Mapper to efficiently RDFize existing RDBs

山本泰智*　片山俊明

Yasunori Yamamoto and Toshiaki Katayama

*〒277-0871 千葉県柏市若柴 178-4-4
東京大学柏の葉キャンパス駅前サテライト 6 階
E-mail: yy@dbcls.rois.ac.jp
ライフサイエンス統合データベースセンター
Database Center for Life Science

Abstract: We developed D2RQ Mapper (http://d2rq.dbcls.jp/), a web application to edit a mapping file of D2RQ, which is a middleware to bridge Relational Database (RDB) and Resource Description Framework (RDF). A D2RQ mapping file defines how to map data stored in an RDB to RDF in the turtle format, and to write it with a text editor is cumbersome. D2RQ Mapper assists you to edit it by contextualizing input forms in the target RDB schema and the mapping language. In addition, D2RQ Mapper supports the R2RML format to output a mapping definition. For users who need to access a target RDB within their intranets, we provide a Docker image of D2RQ Mapper.

背景

生命科学における知見は主に文献に記述されているが、そこから、遺伝子やたんぱく質、疾患との関係など、特定の観点から情報を抽出し、各観点での最新の研究成果を効率良く取得できるデータベースを構築する作業や、特定の研究目的のためにデータベースを構築し、得られた成果を文献で発表することも広く行われている．このため、生命科学の発展は関連するデータベースの発展と表裏一体であり、特に情報通信技術の発展にともない、誰でも自由にインターネットを介してアクセスできる形でのデータベースは大幅に増加している[１].

このように生命科学の研究に必要不可欠である各種データベースであるが、その多くが各々独自のアクセス方法とデータ表現形式を採用しているため目的のデータを取得するためにはデータベース毎にそれらを習得しなければならない．従い、複数の関連するデータベースを横断的に検索したい場合には特に手間隙のかかる作業が必要になる．研究活動を効率良く進めるためには過去の知見を効率良く再利用できる環境が必要になるが、近年、セマンティックウェブ技術がこの問題に対処するための一つの手段として採用されつつある．Resource Description Framework (RDF)[２]と SPARQL[３]という標準技術を採用することで、アクセス方法の問題に対処するとともに OWL 形式でのオントロジーを提供することでデータ表現形式の効率的な習得がしやすくなる．執筆時点で SPARQL エンドポイント経由でのアクセス手段を提供するデータベースとしては UniProt[４]や ChEMBL[５]、MBGD[６]などがある．その一方で、関係データベースをバックエンドで利用した従来通りのアクセス手段を提供している事例は依然として多い．生命科学分野のデータベースを毎年特集号として紹介する Nucleic Acids Research (NAR) の Database Issue では 2015 年の最新号で 172 のデータベースを紹介する文献を掲載している[１]が、これらの文献中では「relational database」や「MySQL」という表現の出現頻度がそれぞれ 28、45 であるのに対し、「SPARQL」は 5 に過ぎない．

この問題の原因としては、上述のようなセマンティックウェブ技術を用いた場合の利点が現時点では実証されていないことや、既にデータベースを公開している場合に改めてセマンティックウェブ技術を採用したデータベースを構築することの手間隙があるのではないかと考えられる．このため、現行の関係データベースはそのままに、比較的簡単にセマンティックウェブ技術を試すことのできる環境が提供されれば、両者の比較検討が容易に行えるとともに、より多くのデータベースが従来のアクセス手段に加えて SPARQL あるいは Linked Data によるアクセスを提供しやすくなるのではないかと考えられる．

このような環境を実現するために既に複数のツールが開発されており、例えば、D2RQ[1]や Ontop[2]などがオープンソースとして公開され、自由に利用できる．特に前者についてはこのような目的のツールとしては 2006 年から開発されている最古参になり、また、関係データベースに対して動的に SPARQL での問い合わせを可能としたり、RDF データセットとしてファイルの出力を可能としたりするなど関連する機能を備えた一つのパッケージとして提供されていることもあり、広く使われている．

D2RQ

D2RQ は関係データベースを RDF データセットとして参照可能にするためのツールであり、次に掲げる事項を可能にする．

- 関係データベースに対して SPARQL で問い合わせ
- 関係データベースのデータにウェブを介して Linked Data としてアクセス
- 利用者の指定したマッピングルールに基づいて関係データベースのデータを RDF データベース(トリプルストア)に格納可能な RDF データとして出力
- 関係データベースのデータに Apache Jena API としてアクセス

D2RQ は Java パッケージとして配布されており、利用するのに先立ち整える必要のあるソフトウェア環境が git や Java、Apache Ant であるため、容易に試すことができる．さらに、RDF データとしてアクセスするためのマッピングルールを全く記述しなくても、対象関係データベースのスキーマに基づいて自動的にマッピングルールを構築する機能を持つため、URI の構造や RDF としてデータベースを公開するためのオントロジーなどについて予め詳細を決めておかなくても、RDF データに変換することができる．このため、初めて RDF データを構築する場合でも、実際に自身に興味のあるデータが RDF として生成されることから、自身の管理する関係データベースを RDF や Linked Data として公開する動機がより働くと考えられる．これは、ワールドワイドウェブ(WWW)が急速に広く普及した理由の一つとして、Instant Gratification という特徴があることにつながる[7]．すなわち、WWW においては誰でも HTML ファイルを編集でき、その結果をブラウザでつぶさに確認できるという特徴である．

ただ、自動構築された D2RQ のマッピングルールは必ずしも望ましいものとはなりにくい．というのも、上述の通り、URI の構造についての情報を必要としないことから、Linked Data の原則 2 から 4[8]に記述されている事項を満たすことができないからである．さらに、RDF により表現されるデータは、その再利用性を高めるためにはオントロジーが併せて提供されることが重要であるが、自動生成されたマッピングファイルには含まれていない．

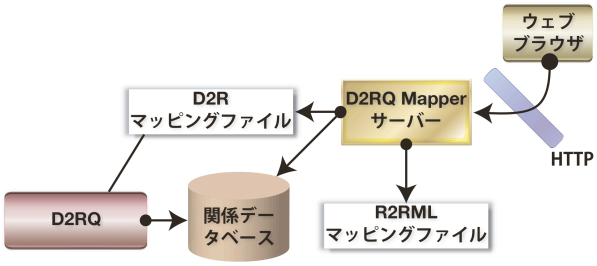

図 1:D2RQ と D2RQ Mapper との関係

その一方で、マッピングファイルは D2RQ マッピング言語の仕様に従い、Turtle 形式で記述しなくてはならず、その習得には手間暇がかかる．このため、より簡単にマッピングファイルを編集できる環境があれば望ましい．これは、同じく WWW の普及に貢献した事項として簡単に HTML を編集できる環境が提供された点が挙げられている[7]ことにつながる．

なお、D2RQ の開発は現在精力的に行われておらず、github において最新版が配布されてから執筆時点で既に 3 年が経過している[3]．また、D2RQ は MySQL や PostgreSQL など様々な関係データベース管理システムに対応しているが、SQLite には対応していないなどの課題が残るほか、動的に生成される SQL は冗長な場合がある．

D2RQ Mapper

D2RQ Mapper は上述の問題点、すなわちマッピングファイルの記述には手間暇がかかることに対応するために開発している．D2RQ Mapper を利用することでテキストエディタを使わず、ウェブブラウザ上での各種設定だけで所望の RDF データを生成するマッピングルールを構築できる．また、マッピングルールに対応するマッピングファイルは D2RQ マッピング言語だけでなく、W3C 勧告である R2RML 形式[9]でも取得できる．図 1 に D2RQ と D2RQ Mapper との関係を示す．

D2RQ Mapper でマッピングルールを構築するためには対象の関係データベースにアクセスするため

[1] http://d2rq.org/
[2] http://ontop.inf.unibz.it/
[3] https://github.com/d2rq/d2rq

の情報を必要とするが、これにより、スキーマに基づいた入力フォームを提示できる．結果として、対象関係データベースのスキーマと矛盾のないマッピングルールを構築することが可能になる．

さらに、D2RQ および D2RQ Mapper は RDF を生成するため対象関係データベースに SQL で問い合わせするが、イントラネット内からのみ許可としている事例が多いと想定される．その一方で D2RQ Mapper を利用するために必要な作業は最小限であることが望ましい．この問題に対応するために、D2RQ Mapper はライフサイエンス統合データベースセンターのサーバー上で稼働するウェブアプリケーション[4](以降、DBCLS ウェブアプリ)としてだけでなく、Docker イメージとしても提供されている．その結果、Docker 環境が予めインストールされていさえすれば、利用者は D2RQ Mapper のイメージをダウンロードするコマンドとダウンロードしたイメージを起動するコマンドを実行するだけで使える状態になる．

利用

D2RQ Mapper は構築するマッピングルールに関する情報の永続性を確保するために利用者管理機能を備えており、DBCLS ウェブアプリでは、当該アプリ内でのみ有効なアカウントを用意できるほか、Twitter および Facebook で利用しているアカウントも使うことができる．Docker イメージの場合は、イントラネット内でのみ利用されることが想定されるため、当該アプリでのみ有効なアカウントを使うことになる．

Docker イメージを利用する際には、Docker のターミナルにおいて以下のコマンドを実行する．

```
$ docker pull d2rqmapper/d2rq-mapper
$ docker run -d -p 80:80 d2rqmapper/d2rq-mapper
```

以上の操作のあと、Docker 起動時に提示された IP アドレスにブラウザでアクセスすることで D2RQ Mapper が DBCLS ウェブアプリと同様に利用可能になる．

ウェブブラウザで D2RQ Mapper にアクセスするとトップページが表示され、そこからサインインするページに移動できる．サインインが成功すると構築する一連のマッピングルールセットに対応した識別子(名前)の設定とともに対象関係データベースにアクセスするための各種設定を行う．問題なく対象関係データベースにアクセスできることが確認されると、続いて対象データベースに含まれるテーブルの一覧が表示されるので、RDF データに含めるデータを含むテーブルを選択する．複数テーブルにまたがる関係を RDF データとして適切に表現できるよう、JOIN の設定も可能であり、JOIN の設定をすると、それを一つのテーブルと同等に扱え、マッピングの設定ができる．

一つのテーブルを選択すると、各列に含まれるデータを RDF の主語、述語、目的語のいずれかに対応させる設定を、その表現方法の指定とともに行えるページが表示される．各列について、URI あるいはリテラルとして表現する方法を指定する．URI については属するクラスを、リテラルについてはそのデータ型を併せて指定できる．設定を行う際の参考として、ページ冒頭に対象テーブルに実際に収められているデータのサンプルが表示される．

以上の設定が終わると、その時点でのマッピングルールセットが D2RQ マッピング言語あるいは R2RML に従うマッピングファイルとして取得できる．さらに、SPARQL クエリによる問い合わせと RDF データセットとしてのファイルの取得も可能である．

考察

現在の D2RQ Mapper では D2RQ マッピング言語および R2RML の全ての言語仕様を満たしておらず、生成できるマッピングファイルは限られたものとなる．このため、現在、より柔軟なマッピングルールを構築できるように拡張する開発計画を進めている．

さらに、現状のインターフェースでは、RDF データとして表現した際のオントロジーに関する情報について、その構造を俯瞰できないため、生成された RDF データの構造を把握しにくい問題がある．このため、他のオントロジーエディタなどで構築された OWL ファイルを読み込んでそれを視覚的に表示するとともに、関係データベースの各データとのマッピングを視覚的に行えるようなインターフェースを検討している．

RDF はデータ構造を規定しないことから RDF でデータを表現しただけで直ちに再利用性が高まるわけではない．ライフサイエンス統合データベースセンターでは生命科学分野における RDF データベースの再利用性を高め、複数のデータベースを効率良く横断的に利用できる環境を実現するため、RDF データを生成する際に参照可能な RDF ガイドラインを構築している[5]．そこで、D2RQ Mapper を用いて生

[4] http://d2rq.dbcls.jp/

[5] http://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline

成した RDF データが可能な限りこれに従うようにする予定である.

結論

関係データベースに格納されているデータを効率的に RDF として扱える環境を実現するために、D2RQ マッピング言語と R2RML に従うファイルを編集するウェブアプリ D2RQ Mapper を開発した. D2RQ Mapper を利用することで、次の事項が可能あるいは容易になる.

- ウェブインターフェースによる D2RQ マッピング言語及び R2RML に従うマッピングルールの構築
- 必要最小限の文字入力や対象関係データベースのスキーマを意識したフォーム生成によるマッピングルールの誤りの低減
- 利用環境を選ばないマッピングルール構築環境の構築

今後は機能向上を目指して開発を継続する予定である.

謝辞

本研究は独立行政法人科学技術振興機構(JST)、バイオサイエンスデータベースセンター (NBDC) の助成による.

参考文献

[1] Galperin, M. Y., Rigden, D. J., Fernández-Suárez, X. M.: The 2015 Nucleic Acids Research Database Issue and molecular biology database collection, Nucleic Acids Res., Vol. 43, Database issue, D1-5 (2015)

[2] RDF 1.1 Concepts and Abstract Syntax. http://www.w3.org/TR/rdf11-concepts/

[3] SPARQL 1.1 Overview. http://www.w3.org/TR/sparql11-overview/

[4] UniProt Consortium: Activities at the Universal Protein Resource (UniProt), Nucleic Acids Res., Vol. 42, Database issue, D191-8 (2014)

[5] Jupp, S. et al.: The EBI RDF platform: linked open data for the life sciences, Bioinformatics, Vol. 30, No. 9, pp. 1338-9 (2014)

[6] Chiba, H., Nishide, H., and Uchiyama, I.: Construction of an Ortholog Database Using the Semantic Web Technology for Integrative Analysis of Genomic Data, PLoS ONE, Vol. 10, No. 4, e0122802, (2015)

[7] McDowell, L. et al.: Mangrove: Enticing Ordinary People onto the Semantic Web via Instant Gratification, In Proc. 2nd International Semantic Web Conference (ISWC2003), (2003)

[8] Berners-Lee, T.: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html

[9] R2RML: RDB to RDF Mapping Language. http://www.w3.org/TR/r2rml/