Relational to RDF mapping using D2R for translational research in neuroscience

Rudi Verbeeck[*1], Tim Schultz[2], Laurent Alquier[3] and Susie Stephens[4]

Johnson & Johnson Pharmaceutical Research and Development
[1] Turnhoutseweg 30, Beerse, Belgium;
[2] Welch & McKean Roads, Spring House, PA, United States; [3] 1000 Route 202, Raritan, NJ, United States and
[4] 145 King of Prussia Road, Radnor, PA, United States

Abstract. Motivation: To support translational research and external innovation, we are evaluating the potential of the Semantic Web to integrate data from discovery research through to the clinical environment. This paper describes our experiences in mapping relational databases to RDF for data sets relating to neuroscience. Implementation: We describe how classes were identified in the original data sets and mapped to RDF, and how connections were made to public ontologies. Special attention was paid to the mapping of experimental measures to RDF and how it was impacted by the relational schemata.
Results: Mapping from relational databases to RDF can benefit from techniques borrowed from dimensional modeling. However, current tools like D2R are still evolving. Nevertheless, mapping data in RDF, if done properly and consistently, facilitates data integration efforts.

1. Introduction

Translational research has emerged over recent years as an important enabler of personalized medicine. It encompasses bridging the gap between discovery research insights in the molecular biology of a disease and predicted clinical response of an individual patient to a medicine. It also involves finding gene signatures or other biomarkers that separate responders from non-responders and understanding how these insights may contribute to disease mechanisms.

To counterbalance compound attrition and fill short or medium term pipeline gaps, pharmaceutical companies are seeking collaboration and licensing opportunities outside company boundaries. Internally and externally derived resources need to be viewed alongside each other in order to gain a comprehensive understanding of a company's development pipeline.

The translational medicine and external innovation trends are both leading to a more data intensive environment that requires well defined strategies for data integration and governance.

Relational database technology has been developed as an approach for managing and integrating data in a highly available, secure and scalable architecture. With this

approach, all metadata is embedded or implicit in the application or metadata schema itself, which results in performant queries. However, this architecture makes it difficult to share data across a large organization where different database schemata and applications are being used.

The Semantic Web offers a promising approach to interconnect databases across an organization, since the technology was designed to function within the distributed environment of the web. Resource Description Framework (RDF) and Web Ontology Language (OWL) are the two main Semantic Web standard recommendations. RDF represents data using subject-predicate-object triples, which connects data in a flexible piece-by-piece and link-by-link fashion that forms a directed labeled graph. The components of each RDF statement can be identified with Uniform Resource Identifiers (URIs). Alternatively, they can be referenced via links to RDF Schemas (RDFS), OWL ontologies, or to other (non-schema) RDF documents. Data in a Semantic Web representation can be queried using the SPARQL query language. Data can gradually be made available on the Semantic Web, without intensive coordination between data source providers [1,2]. Further, as semantics are added to the data, it becomes self-describing, so applications can be made agnostic of the data domain.

To verify if the Semantic Web can facilitate data integration, a Linked Data project [3] was established. The primary goal of the project was to enable scientists to answer novel translational questions related to Alzheimer's Disease (AD) by providing a flexible integrative data layer. The project hypotheses were that new, valuable scientific insights can be gained through the interrogation of Linked Data, and that Linked Data simplifies the incorporation of data sources from collaborators. This paper focuses on describing the mapping of data sources to RDF. More details regarding the Linked Data framework are described in reference [4].

In the next section we describe the data sources used in the Linked Data project. Section 0 reviews the modeling choices we took for mapping and translating the data sources to RDF. The final section discusses some considerations in the implementation of a successful data integration platform.

2. Methods

2.1. Data sources

An internal and a publicly available data source relating to AD were selected for the project.

- In 2005, the National Institutes of Health (NIH) and a number of partners started the Alzheimer's Disease Neuroimaging Initiative (ADNI). This multi-site, longitudinal study was designed to evaluate imaging and genetic biomarkers for the onset and progression of Mild Cognitive Impairment (MCI) and AD [5]. The study of around 800 subjects distributed over 3 cohorts (normal, MCI and AD) resulted in the collection of a wide variety of data, ranging from clinical, cognitive, functional and behavioral assessments, imaging derived anatomical volumes, and blood and Cerebro-Spinal Fluid (CSF) biomarker measurements.

- Internal clinical study data relating to AD that contains demographic and treatment information, vital signs, cognitive assessments and image derived measurements.

2.2. Data source formats

2.2.1. ADNI data.

The clinical data that are collected by participant sites in the ADNI study are deposited into an ADNI hosted, web accessible database according to published guidelines. Data sets reflecting the entry forms are made available to researchers in a flat file format (http://www.loni.ucla.edu/ADNI/).

A Microsoft SQL Server Database was used to host the ADNI data within Johnson & Johnson. The flat files were mapped to a star schema (Fig. 1), using SQL server integration services and Perl scripts. This has facilitated access to the data through SQL based query and analysis tools.

2.2.2. Internal clinical study data.

Clinical data was extracted from SAS files and loaded in an Oracle Database reflecting the original pivoted table structure of the files.

2.3. Ontologies

BioPortal (http://bioportal.bioontology.org) was used to identify public ontologies that best map to the entities in the clinical data sets. Selected ontologies included the Neuroscience Information Framework (NIF) [6], the National Cancer Institute's thesaurus (NCIt) and SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms).

Relevant terms from the ontologies were linked into a Common Resource Ontology (CRO) that was loaded into an instance of an openRDF triple store from Sesame.

2.4. The D2RQ platform

The SQL Server Database and Oracle Database were mapped to RDF using D2R server 0.7 (http://www4.wiwiss. fu-berlin.de/bizer/d2r-server/).

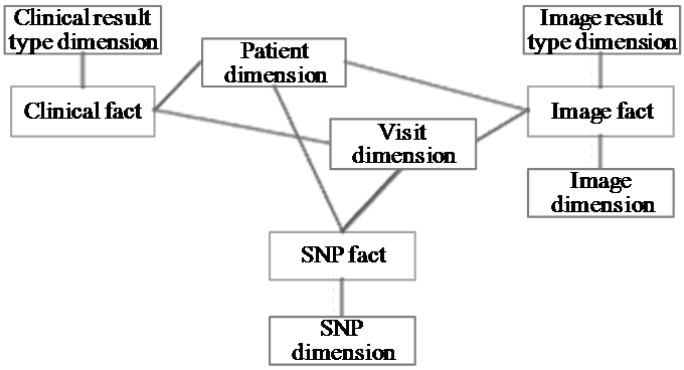

Fig. 1. High level representation of the star schema used to represent ADNI within Johnson & Johnson

3. D2R mapping patterns

There are many options as to how to publish relational data to RDF. For general guidelines, see reference [7]. In this section, we describe the patterns we used and design options we selected to develop the D2R mapping files to ease the integration of complex longitudinal data sources.

3.1. Identifying RDF classes

D2R provides an automated process to generate the mapping file, which converts every table into a class. This approach did not yield satisfactory results for a database with a normalized schema, largely because Third Normal Form modeling seeks to eliminate data redundancies, not reflect real world objects – such as patients, medical images, etc.

In dimensional modeling, a logical design technique for data warehouses [8], data are grouped into coherent categories[1] that more closely mimic reality. This makes the mapping of dimensional representations to RDF classes more straightforward, and enables the default D2R mapping process to yield better results. Further, hierarchies in the dimension tables may help to indicate RDF classes and their relationships.

The ADNI data were loaded into a star schema (Fig. 1). The Single Nucleotide Polymorphism (SNP) dimension contained hierarchical information relating to genes and chromosomes. By converting table column headers to classes, instead of the default literal values, they could be used to link to external ontologies. Fig. 2 shows the high level graph that was created when ADNI was mapped to OWL.

[1] In a star schema implementation, data are stored in fact tables, and categories in dimension tables. A fact table is joined to dimension tables creating a star-like representation.

3.2. Local namespaces and ontology mappings

The classes in Fig. 2 were defined in the CRO to avoid repeating class definitions for every data source. For classes available in public ontologies, the CRO builds a comprehensive representation of a domain by importing a standard set of complementary ontologies using the guidelines described in MIREOT [9]. Using an internal ontology presents some advantages:

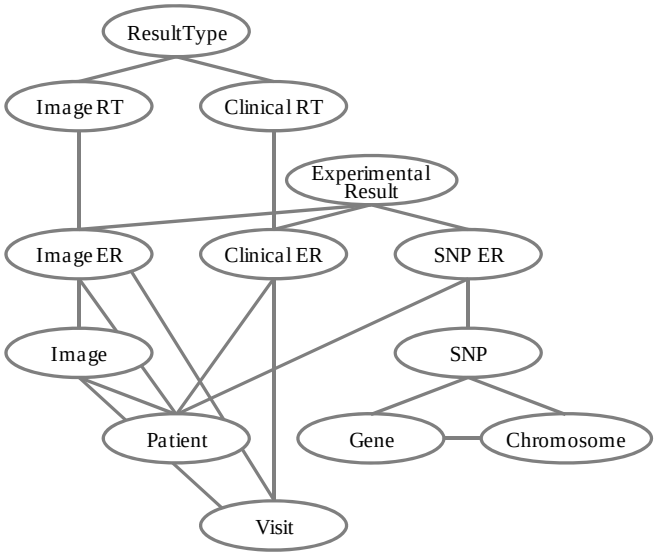

Fig. 2. Ontology used to represent content of the ADNI data (ER=ExperimentalResult, RT=ResultType).

- Scientists may have strong preferences for particular ontologies. When there is no general agreement about which ontology to use, we can include the definition of a proxy class in the CRO. The proxy can be linked to a number of public ontologies using URI aliases.
- Not all class definitions that were required for the mappings were available in public ontologies (e.g. subscores for the AD Assessment Scale Cognition). These definitions could be included within the CRO in anticipation of acceptance of the terms in public ontologies.
- Building a SPARQL query requires knowledge as to which ontology was selected during the mapping phase. This information can be retrieved from the CRO.
- Using Semantic MediaWiki technology, scientists can discuss CRO term definitions or suggest extensions.
- Data owners can use the wiki to enter metadata about their sources using terminology from the CRO. As Semantic MediaWiki stores its data in RDF it can be used as a metadata repository for data source discovery. This functionality is not well supported by SPARQL [10].

BioPortal is a valuable tool for searching for terms within public ontologies. Once the term has been identified, the mapping to public ontologies can be handled in a number of ways by D2R. For example, volume measurements of brain regions on MRI images were linked to the gross anatomy section of NIF using lookup tables. The lookup table can be stored in the D2R file (using `d2rq:TranslationTable`) or in the database (and used in a `d2rq:join`). We prefer the latter solution, but note that this approach restricts the lookup table to being in the same database as the data. When SNPs or genes were mapped to Bio2RDF (http://bio2rdf.org/), the database values were used directly to generate the URI of the object in the public ontology at runtime (using `d2rq:uriPattern` or `d2rq:uriSqlExpression`).

3.3. Experimental measures

To encode experimental results in RDF, the experimental conditions need to be uniquely specified. For example, to be able to correctly interpret a measured value, it needs to be clear which patient is being referred to, on which visit, and what exactly was measured.

One option is to define properties for the Patient class for every type of experiment. Reification[2] could be used to specify additional conditions (e.g. the visit and the imaging modality). However, this option was not selected because several levels of reification would be needed to specify the experimental conditions completely. This would lead to ballooning of the data and such queries are not well supported by SPARQL.

We decided to encode every experimental result (the measured value and the experimental conditions) in an ExperimentalResult class and link out to the corresponding Patient, Visit and Image classes (Fig. 2). However, this still leaves several options as to how to encode all of the details surrounding the experiment.

Defining a subclass of the ExperimentalResult class for every measurement type (e.g. ClinicalDementiaRating, HippocampalVolume, SystolicBloodPressure) was impractical due to the large number of observation types in the data sets. Alternatively, the measurement type can be encoded in an ExperimentalResult property name (e.g. hasClinicalDementiaRating). Contrary to subclass definitions, we can avoid writing D2R code for property definitions for a large number of measurement types using a `d2rq:dynamicProperty` statement, which specifies a pattern to generate the property URI at runtime. However, some experimental conditions are hard to describe in a property name[3] and are difficult to use in queries. We therefore took a different approach.

We decided to use two properties to specify the experimental conditions and measurement value of an ExperimentalResult, namely hasResultType and hasValue. The ResultType class can contain multiple properties to specify the experimental conditions fully and can be used as a bridge to public ontologies.

---

[2] Reification is a process that uses RDF to make statements about other RDF triples. The RDF vocabulary to describe a reified statement uses three triples to specify a single assertion, thus inflating the database.

[3] For example, take a property like hasStandardDeviationOfCorticalThicknessOfRightTransverseTemporalCortex.

3.4. Pivoted and depivoted tables

The depivoted format of the fact tables in Fig. 1 can be converted to RDF using the previously described techniques. Occasionally, a column in the fact table may contain values that can be used as predicates. In this case, using a `d2rq:dynamicProperty` may be sufficient to define all properties for the fact table at once. The mapping becomes independent of the properties listed in the fact table, and remains valid as rows introducing new properties are added to the table.

For statistical analysis, clinical data are mostly represented in a pivoted table format, where each patient is represented as a single row and columns represent clinical, laboratory and image results for each visit. Table columns can easily be mapped to properties connected to a Patient class. But as discussed above, we may have to introduce impractical property names to specify the experimental conditions.

Forming ExperimentalResult classes on a pivoted table requires that column names of the table are parsed and mapped to literal values or URIs. Where a D2R mapping would normally create an instance for every table row, this use case requires the mapping to create a new instance for every table cell (for selected columns). This is equivalent to depivoting the table before applying the mapping. The D2R release we used did not have this functionality. Consequently, we did the depivoting operation in the database instead.

4. Discussion

This paper highlights many design considerations that need to be taken into account when mapping relational databases to RDF. The approach taken for the mapping influences the ease with which data sources can be integrated, and the simplicity with which they can be queried. Further, although D2R is able to map most relational schemata to RDF, there are strong benefits to dimensional modeling over normalized approaches. This should be taken into consideration when designing schemata for data sources that are being brought into an organization.

Mapping data sources to public ontologies is a time consuming process. It also requires that subject matter experts are involved to ensure that the work is done accurately. This is especially the case when data sources are referencing brain regions, as the neuroscience domain does not have a common lexicon.

The Linked Data approach has the significant advantage that experts can incrementally and independently add data sources to the RDF graph. This enables the gradual creation of an integrated ecosystem of data. To allow domain experts to contribute requires an architecture for ontology curation and data source discovery, a strategy on data governance and stewardship and a culture of data caring and sharing.

In this paper, we focused on lessons learned using D2R to map clinical data to RDF for a Linked Data project. As more data sources are added, we will need to adapt the domain model in our CRO to accommodate new class definitions. Going forwards, it is likely that we will use the emerging Translational Medicine Ontology (http://esw.w3.org/HCLSIG/PharmaOntology) to meet our needs. This is because it includes a broader set of class definitions, and uses the Basic Formal Ontology (http://www.ifomis.org/bfo).

Acknowledgements

We would like to acknowledge the essential contributions of Michael Farnum, Victor Lobanov, John Stong and Xiang Yang (Mike) Xu. Michael and Victor went through the effort of converting the ADNI data from a collection of spreadsheets to a structured set of relational tables. John and Xiang-Yang did the same for the internal clinical study data, which were available as SAS data sets.

References

1. Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Scientific American*. May 2001;284(5):34-43.
2. Ruttenberg A, Clark T, Bug W, et al. Advancing translational research with the Semantic Web. *BMC Bioinformatics*. 2007:8(Suppl 3):S2.
3. Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far. *Int. J. Sem. Web Inf. Syst*. 2009;5(3):p. 1-22.
4. Alquier L, Schultz T, Stephens S. Exploration of a Data Landscape using a Collaborative Linked Data Framework. Paper presented at: Proceedings of the Workshop on the Future of the Web for Collaborative Science (WWW2010), 2010.
5. Molchan S. The Alzheimer's Disease Neuroimaging Initiative. *Business Briefings: US Neurology Review*. 2005:p. 30-32.
6. Bug WJ, Ascoli GA, Grethe JS, et al. The NIFSTD and BIRNLex Vocabularies: Building Comprehensive Ontologies for Neuroscience. *Neuroinform*. 2008:6:175-194.
7. Bizer C, Cyganiak R, Heath T. How to Publish Linked Data on the Web. July 27, 2007. Available at: http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/20070727/.
8. Kimball R, Ross M. *The data warehouse toolkit: the complete guide to dimensional modeling*. 2nd ed: Wiley; 2002.
9. Courtot M, Gibson F, Lister AL, et al. MIREOT: the Minimum Information to Reference an External Ontology Term. *Nature Precedings*. 2009. Available at: http://hdl.handle.net/10101/ npre.2009.3574.1.
10. Williams GT. SPARQL 1.1 Service Description. January 26, 2010. Available at: http://www.w3.org/TR/2010/WD-sparql11-service-description-20100126/.