Representation of CDISC Foundational Standards Using RDF

Authors:

Scott Bahlavooni, Biogen-Idec
Geoff Low, Medidata Solutions
Frederik Malfait, Roche

Abstract:

Over the past two years the PhUSE CSS Semantic Technology (ST) project has made considerable progress to represent existing CDISC standards in RDF, the established web standard for semantic technology, based on a ISO 11179 type meta-model. RDF representations are available for CDASH, SDTM, SEND, ADaM, and their controlled terminologies; a CDISC public review for these deliverables is in progress for the fourth quarter of 2014. The presentation will give an overview of the different models, how they relate to each other, and how they can be used to manage data standards in an ISO 11179 type Metadata Registry.

Introduction

The mission of CDISC is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare.[1] Interoperability is a fundamental part of this mission, definitions of standards that can interoperate allow organisations to share information freely internally and externally.

In the most part these standards are released in the Portable Document Format (PDF). PDF is a well-understood read-only document format and many amongst us are used to dealing with these files. Unfortunately it is not a good format for electronic processing, extract table data and metadata can be complex and error prone.

CDISC Standards are not based on a formal model and have no common standard format. There is no innate way to semantically link the different standards together and implementers must rely on User Guides or Training Courses to draw this insight out. There is not currently any automated way for users to be able to electronically import the standards so most people have to adopt a read-extract-use approach.

W3C Semantic Standards offer a valuable possibility for holding study data and metadata. They have several advantages over the current format of the CDISC standards.
W3C Semantic Standards are based on a formal model, provide a standard format and language, are machine readable, directly support semantic interoperability and directly support Linked Data.

[1] http://www.cdisc.org/CDISC-Vision-and-Mission (retrieved 3 Sep 2014)

Each of these aspects provide a good foundation upon which a truly interoperable standard set can be developed, curated and published.

As part of the PhUSE Computational Science Symposium, a working project was put together under the Emerging Technologies Group to commit to the representation of the Foundation CDISC Standards using the RDF. This paper will discuss the development of the standards under the banner of the PhUSE Semantic Technologies group.

The Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a general graph-based method for the conceptual description or modelling of resources. The RDF has been associated with Web resources, but the scope of its adoption has far expanded beyond this. It is a W3C standard, and is the core technology beneath advances such as the Linked Open Data Cloud[2], Google's Knowledge Graph[3] and DBPedia[4].

ISO11179 Metadata Management

ISO11179 is an international standard for the management for representing metadata for an organization in a metadata registry.[5] It is a model common to many of the commercial metadata repository tools available in the clinical industry. In the interest of assured alignment in the future, ISO11179 OWL classes are used to manage the resources in the RDF representations of the CDISC standards.

The Meta-model Schema

The meta-model schema (MMS) was defined to provide a common language and infrastructure for all the CDISC operational standards. The MMS defines generic concepts that are used in multiple standards, examples of which include *mms:DataElement* and *mms:ValueDomain*. Using the meta-model schema aided in development of the models for the individual standards by providing a top-level ontology.

Controlled Terminology

One of the first components to be modeled completely was the CDISC controlled Terminology. This was done in cooperation with the NCI Enterprise Vocabulary Service. The existing CDISC controlled vocabulary sets were defined with an OWL extract, and an associated RDF Schema. The scheduled generation of controlled vocabulary subsets for CDISC was augmented to generate content in RDF which this project could leverage.

---

[2] http://linkeddata.org/ (retrieved 16 Sep 2014)
[3] http://www.google.co.uk/insidesearch/features/search/knowledge.html (Retrieved 16 Sep 2014)
[4] http://dbpedia.org/About (Retrieved 16 Sep 2014)
[5] http://metadata-standards.org/11179/ (retrieved 21 Sep 2014)

Development Methodology

Each of the Standards was split off to a separate group with nominated SMEs leading the development for their team. Each team took the responsibly for consuming, and pulling out the fundamental object classes from the standards documentation as the core concepts.

Once the fundamental concepts were defined, a content template spreadsheet was put together. The content template spreadsheet was an important tool in the development of the standards. Each template spreadsheet has one tab for each of the core concepts, and these concepts map across to RDF classes in the model. Each class attribute is assigned a column within a tab and instances are created as rows in the template.

The content template spreadsheets were reviewed for consistency. Any changes that were required to ensure that the different standards would be as interoperable as possible by clarifying class or attribute names/definitions were made to the model and the template was updated accordingly.

Following this, the teams focused on transcribing the data from the standards documents into spreadsheets. The working groups followed a minimal viable product approach; each team focused on populating the model with the minimum amount of content to represent the core components of their nominated standards; examples of the minimum viable content included the transcribed domain tables and model definition. After the core content was assembled the each team expanded their model to include additional metadata such as enumeration of assumptions for the each domain in the SDTM.

At the completion of the modeling and curation by the individual teams, the concept template spreadsheets were collated together by the lead modeller. Common concepts that had hitherto been modelled in the individual template sheets were drawn up into the CDISC Schema, such as the *cdiscs:DataElementType* and *cdiscs:Assumption* common concept definition document with the cdisc namespace.

The final proposed model was presented to the team and reviewed for consistency. Following any feedback, the package was published on GitHub under the phuse-org organisation for public use.[6] Discussions with CDISC ensued to assist them in taking ownership for the publishing and maintenance. A reviewers guide has been published and the models should be going out for Public Review, last quarter of 2014. Further discussions are also ongoing at this point in relation to the development of RDF representations of the CDISC SHARE MDR.

Conclusion

The PhUSE Semantic Technologies group undertook the task of representation of the Foundation CDISC Standards. By ensuring that teams were kept to small, focused and proactive groups a pleasing amount of throughput was achieved. This project was a first step, on the path of wider-scale development of semantically-rich data standard representations.

[6] https://github.com/phuse-org/rdf.cdisc.org (retrieved 22-Sep-2014)