Representation of Foundation CDISC Standards Using RDF

PhUSE Semantic Technologies Group, Emerging Technologies

How would you like your Data Standards?

- Relevant
- Well-managed
- Timely
- Easily Accessible
  - By Humans
  - By Machines

PDF is a good presentation format
- very difficult to process programmatically
- rely on other processes (e.g. export from MS Word or MS Excel into xlsx) to import the metadata

We seek to present the standards using a format that can support both presentation and processing formats.
- Use the technology behind the Semantic Web ==> RDF

Why use the RDF for Data Standards?

- Linking (and querying) across ontologies is a fundamental facet of RDF/OWL (Linked Open Data)
- RDFS and OWL define extensive vocabulary for building ontologies
- Use of existing ontologies can be used to add semantic meaning to our ontology, such as SKOS, DC, etc
- Inference engines provide reasoning abilities to derive new triples from existing triples
- Schema-less - no data migrations necessary
- RDF is easily parsable and queryable
  (see MarclAnders/SAS-SPARQL wrapper on Github)
- We can leverage existing standards available as RDF/OWL
- NCI Controlled Terminology

Development Process - I

- To limit the scope we only worked with the standards in production at project inception
  - CDASH 1.1
  - SDTM 1.2, SDTM IG 3.1.2
  - SDTM 1.3, SDTM IG 3.1.3
  - SEND IG 3.0
  - ADaM 2.1, ADaM IG 1.0
- We had teams working in parallel to model the standards
- Read and understand CDISC Standards
- Identify elements to model
- Define predicates for the elements
- Identify required terminology elements (using the agreed NCI representation)
- Define a draft model schema for each standard
- Use the draft schema to define modeling documents (as xlsx)
- Aggregate content into modeling spreadsheets

Review and Publication

- The Standards have been published in a GitHub repository for public review
  https://github.org/phuse-org/rdf.cdisc.org
- We include te following components:
  - import-files - the content for the model as xlsx
  - resources - copies of component model
  - schemas - shared schemas for the models including the meta-model, cdisc schema and the controlled terminology schema
  - std - the standards themselves, one Turtle (ttl) file per standard
- We are currently working with CDISC to undertake a formal review process. We intend for CDISC to own these standards moving forward, but we will prepare for their review.
- Discussions are underway for CDISC to publish the SHARE metadata using models based on work prepared by the PhUSE ST Working Groups

What is the RDF?

- The Resource Description Framework is a specification that allows for conceptual description or modeling of information.
- We make statements about “resources” using triples
  (Subject-Predicate-Object)

Namespaces
different domains in different namespaces

URIs
Resource is a Uniform Resource Identifier

Predicates
Domain and range specified using RDF Schema

"Action taken with Study Treatment"

What’s next for the PhUSE Semantic Representation of Standards Group?

- We are currently working on a mapping Protocol Representation Model (PRM)
- Developed a model for the Study Design Model (SDM)
- Using the SDM to model real protocols, to be assured of the strength of the underlying model
- Adding extended Protocol Metadata in the next Phase (to include common elements from the PRM, Study Design: Structured Document)
- Once we have a review process confirmed with CDISC, we will work on adding new releases of existing standards (SDTM 3-3, CDASH E2B, etc)
- Work on using the RDF to provide formal model of traceability between the different standards
- Sample use cases for machine incorporation of RDF models

How can you use this work?

We would like to thank all the people who have volunteered to assist and their employers for granting them the time to work on this project. We are always looking for new volunteers!