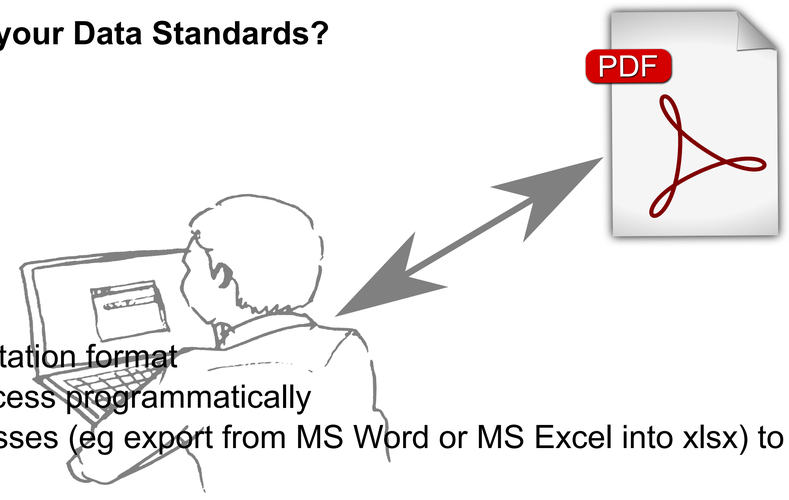
Representation of Foundation CDISC Standards Using RDF

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 (eg export from MS Word or MS Excel into xlsx) to import the metadata

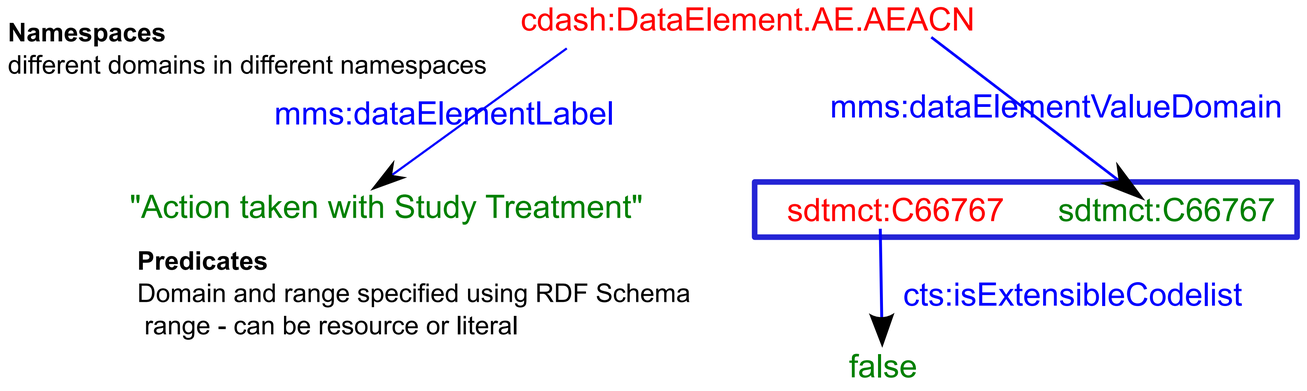
We seek to present the standards using a format that can support both presentation and processing formats.
- Use the technology behind the Semantic Web ==> RDF

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)

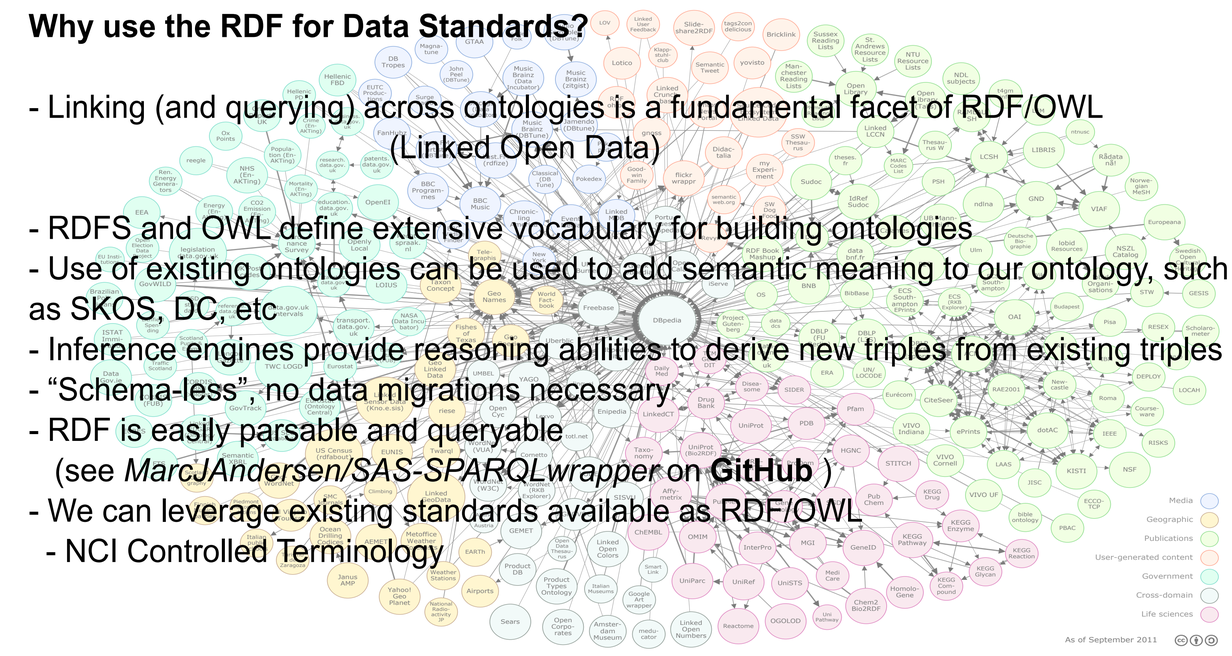
URIs
Resource is a Uniform Resource Identifier

Namespaces
different domains in different namespaces



Predicates
Domain and range specified using RDF Schema
range - can be resource or literal

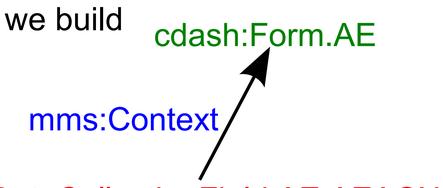
Why use the RDF for Data Standards?



- 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 *MarcJAndersen/SAS-SPARQLwrapper* on **GitHub**)
- We can leverage existing standards available as RDF/OWL
  - NCI Controlled Terminology

The Meta-model

-What is a meta-model?
  - A meta-model defines rules, constraints for the model we build
- We have used a subset of the ISO11179 specification
  - Everything is an Administered Item
    - Data Elements represent a generic unit of data
    - All Data elements exist within a Context
    - Codelists are represented as a ValueDomain
      - Codelist Items are modeled as PermissibleValues



- In the future we can extend to include Registration process (ISO11179/6)
  - An AdministeredItem is managed via an AdministrationRecord (which has associations to Roles, and Dates, etc)

Development Process - I

- To limit the scope we only worked with the standards in production at project initiation
  - 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

isFollowedBy

Development Process - II

- After initial modeling (ie the tables) was done we extended the model for each standard. We expanded the scope to include other elements like:
  - Document Sections - all enumerated so conceivably we could build the standards documents direct from the RDF model
  - Assumptions - assumptions asserted by CDISC in the preparation of the standards
- Once we had completed modeling the individual standards, we moved onto the consolidation phase. All the content modeling sheets were aggregated and handed to a master domain registrar.
- For each domain
  - The domain schema was evaluated to identify shared concepts which could be merged into the *cdisc* model and then the existing references were linked to the shared model.
  - General curation activities on content, following for consistency, etc
  - Built consolidated models for each standard, plus meta-model and cdisc model.
  - Put under source control!

isFollowedBy

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'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!