Linked Data and Graph Database

Project Scope

Investigate how Linked Data, Semantic Standards, Property Graphs and Graph Analytics can support the clinical and non-clinical trial data life cycle from protocol to submission.

Projects	**Overview & Resources**
Representing Clinical Program Design in RDF	2018: DIA Poster Presentation: The Clinical Development Design (CDD) Framework-Assisting and Improving Decision-Making for Product development 2017: Applied Clinical Trials Paper: Barriers and Solutions to Smart Clinical Program Designs 2016: White Paper: Introduction to the Clinical Development Design (CDD) Framework 2016: CSS Hirschfeld Clinical Trial Design Process. An Introduction 2015: Three Ws of Ontology 2015: Drafting the Information Model
CDISC Protocol Representation Model in RDF	2015: Draft version of model SPIRIT Statement Website

Representing CDISC Conformance Checks	**Project Rationale**: • ADaM standard includes validation rules • Identifying all validation rules associated with a SDTM (ADaM) domain or variable is a non-trivial, manual task • Vendor-agnostic representation of SDTM validation rules **Project Deliverables**: • Define ontology for validation • Identify version(s) of SDTM/ADaM validation rules for representation • Represent SDTM/ADaM validation rules in RDF • Link the RDF representations to CDISC Standards represented in RDF **2013: Project Related Documents:** • Validation Rules - Compiled list of validation rules for SDTM IG v3.1.2, SDTM IG v3.1.2, SEND IG v3.0 and ADaM IG v1.0 **Proposed Ontology:** • Classes ○ ValidationRule ○ ValidationRuleCategory (as a sub-class of mms:Classifier) ○ Research documentation • Predicates (of ValidationRule Class) ○ checkID: String literal ○ documentReference: Resource ○ documentReferenceText: String literal (this is the specific text in the document reference to which the rule refers. May need to think about a better way to model this ○ mms:dataset: Represent the Structural Group in the ADaM validation rule documentation ○ validationRuleCatergory ○ mms:variableGrouping: Note, the wording in the ADaM validation rules differs slightly from ADaM IG. This predicate may need to be re-evaluated during the modeling ○ failureCriteria: String literal ○ failureMessage: String literal (remove) **ADaM Checks in TopBraid Upload Format (Draft)**

Reusing Medical Summaries for Enabling Clinical Research	**Overview:** The goal of the keyCRF project is the creation of a semantically annotated electronic Case Report Form (eCRF) that can enable the pre-population of the eCRF from linked data elements in an EHR summary document, HL7's Continuity of Care Document (CCD). The project will draw on prior work of the Semantic Technology Work Group, specifically the RDF representation of the CDISC CDASH standard. The project will use the IHE Data Element Exchange (DEX) specification to create the annotated eCRF, the keyCRF, by drawing on metadata in a metadata repository (MDR) such as CDISC's SHARE or the SALUS MDR. The keyCRF can be used to create an extraction specification that pulls instance data from the CCD to pre-populate the eCRF. **Rationale:** The following use case describes the use of keyCRF through the eyes of an end user. A research forms designer is building a case report form for a particular research study. The designer refers to an on-line metadata registry of research data elements, e.g. SHARE, and selects the desired data elements from a set of research friendly elements such as CDASH, and, using a unique identifier for that data element, retrieves the metadata defined by the metadata registry into an annotated case report form. The metadata includes the exact specification, using XPath, to find the corresponding data element in the HL7 specification Continuity of Care Document (CCD) as extended in the IHE Clinical Research Document (CRD) profile. Using the XPath statements, the research system creates an extraction specification for all elements to be extracted from the CCD. This extraction specification provides a map that enables re-use of the proper data within a CCD with precision and without inappropriate access to extraneous information. The extraction specification could then be used with RFD and Redaction to pre-populate the case report form. **Resources:** keyCRF webinar The keyCRF team will present a webinar in February of 2015 with the following agenda: **1.** An animated illustration of how an application of keyCRF will transform data capture processes at a healthcare site conducting a clinical study. **2.** A walkthrough of the steps of the keyCRF process showing the role of the 'smart form', the metadata repository, and how the extraction specification applies to the electronic record's export document. XML snippets will explain the technical behind the scenes work. **3.** A discussion of future directions for the keyCRF work. How might RDF change the concept of an extraction specification? Mapping of HITSP C154 Data Dictionary Data Elements to RDF and XML Representation of CCD HITSP C32 (https://ushik.ahrq.gov/mdr/portals/hitsp?system=hitsp) describes the HL7/ASTM Continuity of Care Document (CCD) content “in order to promote interoperability between participating systems", in this case between an EHR and research data capture systems. HITSP C32 marks the elements in CCD document with the corresponding HITSP C154 data elements from HITSP Data Dictionary (https://ushik.ahrq.gov/mdr/portals/hitsp?system=hitsp) to establish common understanding of the meaning of the CCD elements. The native representation format of CCD documents are XML, while there are efforts to provide an RDF representation of HITSP C32 for enabling semantic interoperability across systems. The RDF model of HL7 CDA schema provided by SALUS Project is available from: http://www.salusproject.eu/ontology/hl7-cda-ontology.n3. In addition to this, there is a parallel effort to provide an RDF representation of FHIR (Fast Healthcare Interoperability Resources -http://hl7.org/implement/standards/fhir/index.html) Resources (http://www.w3.org/wiki/HCLS/ClinicalObservationsInteroperability/FHIR). We will maintain the data elements in HITSP C154 Data Dictionary in a metadata repository in conformance to ISO/IEC 11179 meta-model. In this metadata repository the extraction specifications of each HITSP C154 data element from CCD documents will also be stored: XPATH expressions will be given for XML representation of CCD documents, while SPARQL queries will be defined for being able to retrieve the data element instances from a medical summary in CCD RDF model. Through DEX profile, these extraction specifications will be retrievable in a machine processable manner as a part of data element metadata. Linkage of HITSP C154 Data elements to CDASH RDF This deliverable, the guts of the project, draws on the team's experts in both research and healthcare. The CDASH RDF model will be imported to a metadata repository, then the semantic links between the CDASH data elements and HITSP C154 Data elements will be defined and maintained in the metadata repository. This mapping will enable creation of an extraction specification from CCD documents which can be used to pull instance data into a waiting eCRF. We will also investigate to define and maintain the extraction specifications of HITSP C154 Data elements from XML and RDF serializations of FHIR Resources in the metadata repository. Demonstration of pre-population of an eCRF from a CCD An end-to-end demonstration of keyCRF creation, extraction specification creation, and pre-population of an eCRF will show industry the value of the approach. The demonstration will employ the well-known mechanism of RFD to define the necessary transactions between the EHR and the research system. **2015:** Key CRF Demo **2015:** Key CRF

Analysis Results Model	**Overview:** - Development of standard models and technical standards for the storage and usage of analysis results data and metadata to support clinical and non-clinical applications. **Rationale:** - To determine the logical model for the representation of analysis results and their associated metadata for clinical and non-clinical applications. Historically, the process of creating results in clinical and non-clinical development has been very labor intensive and inefficient. This team will be determining a semantic representation of the Analysis Results & Metadata model primarily based on RDF and OWL. The representation of analysis results in this manner will facilitate traceability and support broader process efficiency. **Resources:** - Assessment of using RDF data cube vocabulary for representing Analysis Results & Metadata - Proof of concept including Creation of a functional R package that creates RDF Data Cubes and associated documentation UPDATE 23-Sep-16: R package available on PHUSE GitHub here: https://github.com/phuse-org/rrdfqbcrnd Adaptation of a PHUSE Code Repository SAS program to use as input into the R package to generate an RDF Data CubeCreation of a SAS program that queries the RDF Data Cube using SPARQL to reproduce a table with the same layout as the PHUSE Scripting team. - Technical specification of the cube model **UPDATE 23-Sep-16:** Released Technical Specification Version 1.0: ARM-CubeStructureTechSpec-V-1-0.pdf The technical specification provides details of the RDF Data cube structure produced using the R Package. Use it as a reference for querying the cube or extending the existing model for your own purposes. Version 1.0 is considered a proof of concept. Additional development is required, specifically in the areas of codelist implementation and multi-cube/hypercube management. ** White Paper for considerations and benefits of modeling Analysis Results & Metadata in RDF* **Related Documents:** - W3C RDT Data Cube - CSS 2015 TT07 Supplementary Material - interactive summary tables - Semantic Technology Curriculum - Statistics Ontologies for representing Analysis Results & Metadata (see below) - AR&M Publications (see below) - CSS 2015 Files and Notes (see below)

Useful Content	Resources
Study Design Questions	1. Are the BRIDG extensions for the PRM included in the newer versions of the BRIDG Model? a. Yes, and more concepts 2. EPOCH vs Period - A treatment EPOCH can include multiple periods - can this be handled with visit (StudyEventDef) Types (eg Washout, Baseline, etc) 3. Alignment between PHUSE and CDISC 4. Does/Should the RDF version include concepts of changes and roles? 5. Need a selection of schedule of events to model 6. What is the alignment of the odm:MetadataVersion to the sdm:Protocol - different versions of the schedule of assessments? 7. What about modelling the actual text of the protocol?

Missing Elements in the Study Design Model	Each activity as defined by the SDM may have some associated sub-activities; as an example the activity of measuring a blood chemistry value could have the associated sub-activities
	`* Subject at site`
	`* Blood draw taken from Subject`
	`* Date and time of Sample taken`
	`* Blood sample labelled with a unique reference id`
	`* Blood sample sent to lab`
	`* Lab technician records comments on state of sample`
	`* Blood sample analysed (multiple subsequent activities lie here)`
	`* Result logged to Lab Information System`
	`* Result shared or entered into CRF`
	`* Result value checked against defined validation rule`
	`* Comment entered on clinical significance of lab result`
	`* ....`
	Each of these sub-activities could enter in a study workflow system, and be useful for trial scheduling, etc.
	Representation:
	The representation of Roles in ODM is not expansive enough for a full workflow. BPMN heavily uses swim lanes for representation of workflows, but there is no way to catch the full gamut of requirements using the current SDM. Roles may apply to Organisms (such as Site Staff), but can also apply to non-Organisms (such as a machine). Indication of a Role of MRI Machine would provide valuable insight for study site selection or protocol planning; as an example say the executable SoA is entered into a workflow system, but the site knows that an important piece of equipment is out of service for scheduled maintenance at some point, then recruitment could be influenced by following the workflow back to the start.

Analysis, Results & Metadata Publications	1. Hungria M. Delivering Statistical Results as an RDF Data Cube: A Simple Use Case to Illustrate the Process of an RDF Data Cube Creation and the Link to the RDF Representation of the CDISC Standards. North Bethesda, MD; 2014. Available here. Article: http://content.yudu.com/web/2htg1/0A2hthm/December2014/flash/resources/index.htm?referrerUrl=http%3A%2F%2Fcontent.yudu.com%2Fweb%2F2htg1%2F0A2hthm%2FDecember2014%2Findex.html - see page 8. 2. Williams T. A Primer on Converting Analysis Results Data to RDF Data Cubes using Free and Open Source Tools. London; 2014. Available from: https://phuse.s3.eu-central-1.amazonaws.com/Advance/Emerging+Trends+and+Technologies/TT03.pdf 3. Fleming I. The Application of Directed Graphs to Clinical Development. London; 2014 [cited 2015 Mar 14]. Available from: https://phuse.s3.eu-central-1.amazonaws.com/Advance/Emerging+Trends+and+Technologies/TT08.pdf 4. Andersen M. Linked data to support Clinical and Non-Clinical Reporting. Trentino; 2014 [cited 2015 Mar 14]. Available from: https://phuse.s3.eu-central-1.amazonaws.com/Advance/Emerging+Trends+and+Technologies/semstats2014_submission_5.pdf 5. Williams T., Andersen M. 'Dude. Where's My Graph?' RDF Data Cubes for Clinical Trial Data. PHUSE 2015. Paper https://phuse.s3.eu-central-1.amazonaws.com/Advance/Emerging+Trends+and+Technologies/TT07.pdf, presentation
CSS 2015 Content	Learning SPARQL Bob DuCharme Blog DBpedia SPARQL Query Page PREFIX dbo: <http://dbpedia.org/ontology/> SELECT ?city (SAMPLE(?name) AS ?cityName) (SAMPLE(?pop) AS ?cityPop) WHERE { ?city a dbo:Settlement . ?city foaf:name ?name . ?city dbo:populationTotal ?pop . ?city dbo:country ?country . ?city dbo:country dbpedia:Denmark . FILTER (?pop > 100000) } GROUP BY ?city SPARQL by Example BioPortal, the worlds most comprehensive repository of biomedical ontologies A SPARQL Endpoing: Apache Fena Fuseki SAS Program accessing SPARQL Endpoint: • Repository with all the programs • Example Localhost R Package

Statistics Ontologies for Representing Analysis Results Model

Vocabularies for the RDF Data Cube

The list of vocabularies is incomplete and subject to modification as our cube model matures. This list represents the current set of standard prefixes used in the Results Model work.

Prefix	URL	Use in Current Model
cts	https://www.cdisc.org/search?search_api_fulltext=http%3A%2F%2Frdf.cdisc.org%2Fct%2Fschema	Used when values are obtained from CDISC terminology files
mms	http://rdf.cdisc.org/mms#	A reference to the CDISC namespace. Used in the code value.
qb	http://purl.org/linked-data/cube#	Cube specification
rdfs	http://www.w3.org/2000/01/rdf-schema#	Labels, comments
xsd	http://www.w3.org/2001/XMLSchema#	Data types
dcat	http://www.w3.org/ns/dcat#	Distribution information
dct	http://purl.org/dc/terms/	Creator, issued date, title, description
prov	http://www.w3.org/ns/prov#	Provenance
owl	http://www.w3.org/2002/07/owl#	OWL2 Ontology Language
pav	http://purl.org/pav	Provenance, Authoring, Versioning