RDF Implementation of Clinical Trial Data Standards

Frederik Malfait[1] and Josephine Gough[2]

[1] IMOS Consulting, Switzerland
[2] Hofmann-La Roche, Basel, Switzerland

frederik.malfait@imos-consulting.com, josephine_anne.gough@roche.com

Abstract. Clinical trials pose increasing challenges to sponsors and regulators in terms of execution complexity, timing constraints, risk management, and quality of scientific output. We show how the systematic use of clinical trial data standards, combined with the application of model driven semantic technology within the context of a metadata registry can provide a radical but more effective way to start addressing these challenges. This approach has been implemented and deployed as a validated production system within a large pharmaceutical company.

Keywords: clinical trial data standards, metadata registry, model driven semantic technology

1 Problem and Solution Statements

The planning, execution, and analysis of clinical trials has many touch points throughout any large pharmaceutical company. The reference document of a clinical trial is the study protocol, but being (usually) a paper document, there are many hand-offs and labor intensive tasks in downstream processes to properly define, collect, tabulate, analyze, and submit clinical trial data to regulatory authorities. This touches on virtually every aspect of the clinical trial data life cycle and incurs significant cost.

Progress has been achieved by increased adoption of clinical trial data standards, most notably the leading industry standards developed and published by the Clinical Data Interchange Standards Consortium (CDISC). Many pharmaceutical companies implement at least some of these standards. There are also an increasing number of data standard departments to support standards implementation and governance within a pharmaceutical company.

By themselves data standards are useful, but still lack key features to achieve their full business value. Clinical trial data standards often lack a consistent background model, are published in PDF or other office document formats, are inconsistent across different areas, have insufficient versioning capabilities, and are difficult to reference, maintain, extend, and integrate with other information resources.

The issue of a consistent background model for data standards is already addressed by the ISO 11179 Metadata Registry (MDR) Standard. This standard introduces common terminology and a basic meta-model to define, register, and version data elements and related items. W3C semantic technology standards such as RDF, RDFS, OWL, and SKOS close the gap by introducing a standard language that can express an ISO 11179 type meta-model and to layer data standard instances on top of that.

2 RDF Implementation of a Metadata Registry

At Hoffmann-La Roche, an ISO 11179 metadata registry has been designed, implemented, and deployed using semantic technology. The system has gone through several iterations, making increased use of semantic capabilities.

The first milestone implemented a fully standards based information model for the definition, management and dissemination of clinical trial data standards throughout its life cycle from protocol to data collection (DC), data tabulation (DT), data analysis (DA), and submission as shown in Figure 1. Information is disseminated through a REST based API that serves data to a browser application, a search facility, and other standards driven applications.

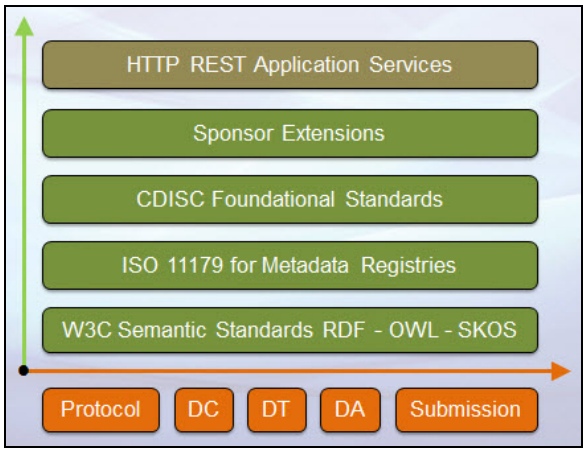

Figure 1 RDF Implementation of a Standards Metadata Registry

Increased use of the data standards repository led to demand for deploying additional schemas, new data sources, and REST interface extensions. Being a validated application, this turned out to be costly. The second milestone addressed this by introducing the capability for model driven services that are themselves expressed in RDF based on facet descriptions of RDF resources. These facets expose services that can act on representations of RDF resources or larger virtual RDF graphs. Currently deployed are services that include exposure of resources for GET requests, specification of searchable facets, and data validation rules expressed as SPARQL ASK queries. Currently being developed are services for model driven PUT/POST/DELETE requests, model driven UI, and model driven security.

Work has also started to reach the third milestone, which is the development of standards driven applications that are entirely specified through services expressed in RDF models. Use cases include a computable protocol representation, component based authoring, and downstream systems automation for setting up study databases and generating clinical trial submission data sets.

3 Industry Initiatives and Collaborations

Like with the web and RDF itself, the value of clinical trial data standards will increase significantly if the benefits of the network effect can be realized. Standards become more valuable with increased adoption. To this end, many of the RDF models have been shared as a starting point to represent most of the CDISC foundational standards in RDF. This work is pursued by the Semantic Technology (ST) working group of the PhUSE Computational Science Symposium (CSS). Last year, work was completed for the RDF representation of the CDASH, SDTM, SEND, and ADaM data standards, which may become subject for CDISC public review later this year. Currently, PhUSE CSS ST teams are working on a protocol representation model in RDF, analysis metadata representation in RDF (possibly using the RDF Data Cube Vocabulary), and linking Electronic Health Records (EHR) with CDASH data collection standards (KeyCRF project).