PhUSE US Connect 2019

Paper TT08

Next Generation Data Management: Case Studies for Ontology-driven, Graph Databases

James T. LaPointe, Cambridge Semantics, Inc., Boston, MA USA

ABSTRACT

Clinical Data Interchange and Standards Consortium (CDISC®) data standards greatly streamline data definitions for individual studies but harmonizing clinical patient data across studies still presents many data management challenges. Patient data definitions and formats vary from study-to-study, across and within Therapeutic Areas and often utilize different data standard versions.

Ontology-driven, graph database technologies offer a highly flexible mechanism to extend prior study data models to new study data definitions which preserve the contextual definition of the original data while extending and harmonizing new study data concepts. This approach results in a dynamic, harmonized data model across multiple studies which can be used to rapidly access data and ready it for analysis.

An ontology-driven, graph database platform was used by several leading BioPharmaceutical companies to integrate, merge, access and analyze clinical patient data to simplify and streamline their clinical data management processes. Lessons learned for these implementations will be discussed including technology benefits and risks.

INTRODUCTION

Effective and efficient clinical trial patient data analyses are essential to prove the efficacy and safety of new investigative products and therapies. To achieve this, clinical data management practices and tools have been developed and continue to evolve to keep pace with an ever-changing landscape of more complex study protocols, increased regulatory constraints, and globally diverse and collaborative data management teams. However, these practices and tools have largely focused on managing a single study with little to no consideration of prior studies and poorly serve the rapid adoption of Real World Data (RWD) healthcare data and insights.

The industry understands the need for new data management and analytic techniques. We see the Clinical Data Interchange and Standards Consortium (CDISC) and Health Level 7 (HL7©) industry groups investigate and adopt new ontology-driven and graph-based data standards. Why? What are the benefits of these new data standards and how have early-adopters utilized these capabilities for both collaborative and competitive advantage?

CLINICAL DATA MANAGEMENT CHALLENGES

The definition of clinical data management according to the DAMA International Data Management Body of Knowledge (DAMA DMBOK™) might be defined as - clinical data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of clinical trial data and information. More simply stated, clinical data management is responsible for getting the right clinical data (trusted), to the right person (authorized), at the right time (timely according to a schedule). This has been the primary goal of CDISC and the development of their foundational standards to define data collection, analysis, and reporting standards. But data standards evolve to keep pace with unique Therapeutic Area (TA) requirements and more complex protocol definitions. As shown in Figure 1, standards change.

So, clinical data management must be able to implement many different CDISC standards. It is not unusual for sequential clinical trials of the same compound to utilize different data standards. When this happens, biostatisticians are needed to transform and harmonize the data into a consistent definition and format which takes time and adds cost to clinical trials.

With the adoption of Electronic Data Capture (EDC) systems over the last 20 years, automation of the entire clinical trial patient data lifecycle has become possible. However, clinical system integration amongst EDCs, Clinical Data Repositories (CDRs), statistical analyses and programs, and electronic reporting and submission systems remains poor. These systems all rely on well-defined data definitions and data formats. Unfortunately, the definitions and formats for these systems are different

which requires the introduction of clinical MetaData Repositories (MDRs) solutions and manual data management processes to harmonized and transform clinical data which again takes time and adds more study costs.

The ability to analyze data across multiple clinical trials, or meta-analyses, relies on well-defined, harmonized data. However, data is rarely defined for cross-study purposes and must be manually harmonized and transformed by the biostatistician team. This makes meta-analysis slow and costly unable to keep pace with changing business requirements.

Like meta-analyses, combination therapies require the ability to harmonize study data across several sponsors using different data definitions and formats. Again, the biostatistics team must manually harmonize and transform the data to investigate the benefits and risks of a new, combination therapy introducing complexity and delays into possible business collaborations.

Lastly, the insights from real-world healthcare data promises to help improve clinical trial outcomes through better understanding of messy longitudinal patient populations leading to improved protocol designs. However, relating clinical trial data to healthcare records is difficult with existing, relational technologies and cannot adequately scale to support efficient repeatable processes.

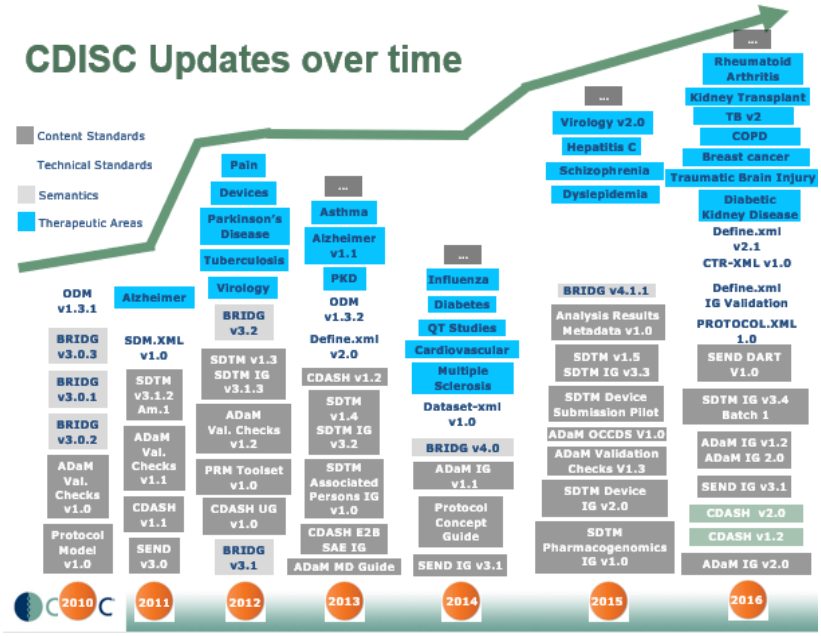

Figure 1. CDISC Foundational Standards Over Time

CLINICAL DATA MANAGEMENT OBJECTIVES

Responding to the myriad of clinical data management challenges previously stated, it is clear we need a new clinical data management framework which supports the following key objectives;

- Consistency with current CDISC foundational standards to promote data reuse
- Relate past CDISC foundational standards to current standards for maximum reuse and to support traceability
- A flexible, extensible clinical and healthcare data model based on harmonized data concepts and definitions
- Data governance process support for data model and data item changes
- Automated data transformation for interoperability amongst disparate clinical systems (vendor agnostic)
- Legacy clinical systems support while enabling future technology adoption
- User role security to protect confidential and sensitive data
- Scale to support global business volume
- Enable new use cases, such as impact analyses and enterprise-wide controlled terminology
- Support business transformation initiatives driven by data

RESOURCE DESCRIPTION FRAMEWORK (RDF) BASICS

CDISC recognized the need for a more flexible and extensible clinical data standards framework and introduced their foundational standards [(1)] in RDF format in 2015 [(2,3)]. RDF is the fundamental graph-based data structure from the World-Wide Web Consortium (W3C) to define and relate real-world concepts (resources) in a rigorous, unambiguous way based on naming conventions from an authoritative source, such as CDISC, via a unique Internationalized Uniform Resource Identifier (IRI) [(4)]. The following Adverse Event Outcomes (AEOUT) IRI is an example of CDISC's SDTM version 3.1.2 data resource in RDF format;

```
http://rdf.cdisc.org/std/sdtmig-3-1-2#Column.AE.AEOUT
```

where; `http://` is a standard internet communications protocol

`rdf.cdisc.org/std/sdtmig-3-1-2` is the authoritative source path name

`Column.AE.AEOUT` is the resource fragment identifier

By applying a namespace identifier, we can simplify the IRI as follows;

where: `xmlns:sdtmig-3-1-2#="http://rdf.cdisc.org/std/sdtmig-3-1-2#` is a namespace identifier

`sdtmig-3-1-2:Column.AE.AEOUT`

Now that we are able to uniquely define a resource, we want to capture some attributes about that resource. In RDF, we relate resources to their attributes or other resources using the following simple syntax called a *RDF triple*;

`subject → predicate → object (value)`

Example CDISC RDF triples are;

`xmlns:sdtmig-3-1-2#="http://rdf.cdisc.org/std/sdtmig-3-1-2#` namespace for SDTM 3.1.2

`xmlns:mms="http://rdf.cdisc.org/mms#` namespace for the Meta Model Schema (MMS)

`xmlns:sdtmct="http://rdf.cdisc.org/sdtm-terminology#` namespace for SDTM Controlled Terms

subject	→ *predicate*	→ *object (value)*
sdtmig-3-1-2:Column.AE.AEOUT	mms:dataElementName	"AEOUT"
sdtmig-3-1-2:Column.AE.AEOUT	mms:dataElementLabel	"Outcome of Adverse Event"
sdtmig-3-1-2:Column.AE.AEOUT	mms:dataElementValueDomain	sdtmct:C66768

Now that we have a flexible and extensible data structure, we need to define and document the vocabulary of the business domain of clinical trials. The W3C created a mechanism to define these vocabularies in an RDF format called a RDF Schema (RDFS) or ontology. These models are created by Subject Matter Experts (SMEs) using a W3C standard modeling language called the Ontology Web Language (OWL) [5]. OWL models or ontologies define and document the clinical data standards in familiar concepts, such as CDASH, SDTM and ADaM their corresponding domains, variables, code list, code list items, controlled terminology, etc.

Now we have a powerful mechanism to define, store, change and extend clinical data standards as needed. Resources and their attributes are defined and related to other resources as required to capture the entire domain of business information and often grows to millions of RDF triples and can be visualized as an inherent graph structure. Figure 2 illustrates a small portion of the CDISC SDTMIG 3.1.2 data standard in RDF format.

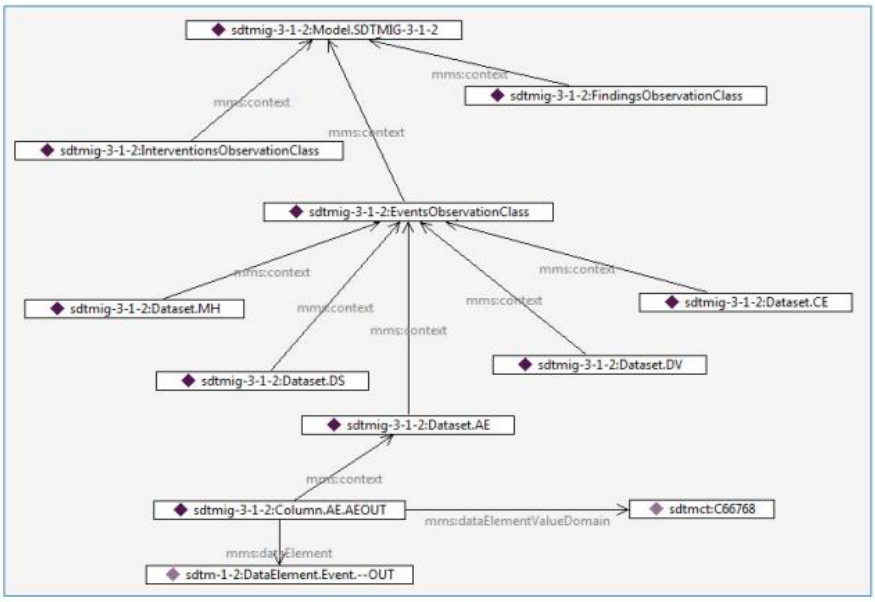

Figure 2 CDISC SDTM Foundational Standard Fragment

RDF is a simple, yet sophisticated directed labeled graph data structure capable of representing and persisting diverse types of data. OWL is the RDF representation of domain specific vocabularies and taxonomies. To extend RDF-OWL mechanisms with full database capabilities, the W3C adopted and recommended the SPARQL query language to manipulate RDF data [6]. SPARQL provides many common data manipulation operations typical of modern databases including complex queries for union, conditional queries, filtering, value aggregation, path expressions, and nested queries. The SPARQL query language completes a well-defined, industry standard technology stack capable of implemented powerful, data-driven applications based on the integration and harmonization of diverse sources of data.

But why is the W3C (RDF-OWL-SPARQL) technology stack better at handling the challenges of clinical data management than existing relational or document-based NoSQL technologies? For three fundamental reasons – first, RDF is machine readable which enhances interoperability and rapid system integration based on industry data standards; second - RDF and OWL ontologies are inherently n-dimensional, where relational technologies are only 2-dimensional, and flexible enough to capture complex data structures and their relationships to each other. These features address many of the clinical data

management challenges previously described to rapidly capture the complexity of the ever-changing clinical data landscape; and third, new resources (data items) can be related or mapped to prior resources to provide continuity of evolving data definitions and full traceability to prior data items.

CLINICAL DATA MANAGEMENT USE CASES

Cambridge Semantics, Inc. ® (CSI) has developed an end-to-end, W3C technology-based data integration and real-time data analytics platform called Anzo ®. The Anzo platform has been successfully utilized in numerous clinical data management settings. The following clinical data management use case highlights are described as evidence of the technology's ability to address many clinical data management challenges.

Use case 1 – Integrate an existing clinical MetaData Repository (MDR) with Oracle clinical systems. In this clinical operational setting, clinical metadata needed to be shared amongst a series of Oracle applications including Central Designer, the InForm EDC and the Data Management Workbench (DMW) CDR based on metadata managed and governed by Accenture's Quantum MDR. Using Anzo, RDF-based ontologies were created for each application to be integrated and harmonized into an umbrella canonical data model based on CDISC's Operational Data Model (ODM-XML) data standards. This canonical model was then used to identify and define data transformations needed to integrate data flow between the Accenture MDR and Oracle clinical applications. ODM was crucial to establishing the canonical data model for this solution. This Anzo-based solution was developed over several months and helped the clinical IT team setup and monitor data flows for select trials.

Use case 2 – A MDR solution for Oracle clinical systems. The precise and consistent definition of clinical metadata is essential for the efficient flow of clinical trial patient data throughout the entire data lifecycle. In this operational clinical setting, an Anzo-based MDR was developed to integrate clinical metadata needed by the suite of Oracle clinical applications including Central Designer, InForm EDC and DMW. CDISC-based data standards were promulgated into progressively refined sponsor data standards for global, therapeutic area, program and eventually study-level metadata managed by a rigorous data governance workflow and shared across a global clinical operations organization. A custom user interface was developed to manage all data items, their versions and support data change governance.

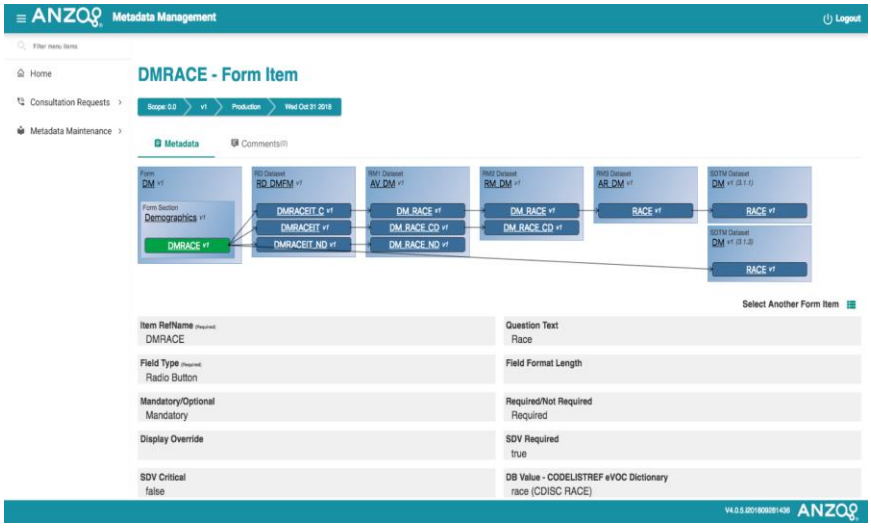

Figure 3 SDTM Item DMRACE Traceability across Oracle systems

A key feature of this solution was the introduction of an umbrella meta-meta model to describe and harmonize different CDISC data standard version and to establish the automated mapping between different versions. Instead of over-writing prior data standards, unchanged prior data items could be versioned to preserve the original lineage of the data item. Data item transformation versions were also managed to permit the forward and backward data flow mapping shown in Figure 3 and shows the cascading series of data item versions throughout the data lifecycle and across Oracle application data structures. This feature promoted the maximum reuse of data items and helped support traceability reports.

Change requests to recommended data standards were supported and managed as different versioned entities for full traceability and possible future reuse. Over time, high use data standards were identified and recommended as the preferred data standards and little used or one-off data items could be avoid or de-authorized.

Use case 3 – A Clinical Data Repository (CDR) for Oncology Data Management. Clinical trial patient data collected over several years for several investigative oncology therapies with expected effects for multiple indications and multiple anatomical sites needed to be managed and supported across many trials and to support sophisticated cross-trial analytics or meta-analyses. An Anzo-based CDR was developed to manage and transform patient data into a harmonized canonical data model for rapid retrieval and consistent analysis. As depicted in Figure 4, a CDISC ADaM-based ontology was used as the framework for the canonical data model. SAS™ data from completed clinical trials was exported as Comma Separated Values (CSV) files and mapped into the RDF-based canonical data model. Numerous reusable analytic dashboards were created to quickly screen large amounts of data and to produce interesting answer sets of data for further analysis by Biostatisticians. A

high-speed data conduit was developed to pipe data to SpotFire® for further analysis and visualization. In addition, several advanced end-users utilized self-generated, ad-hoc queries to explore and visualize data across 100+ clinical trials. The volume of data and speed of analysis required the use of Anzo's industry leading, in-memory RDF-graph database – AnzoGraph®. AnzoGraph enabled the real-time analysis of over four (4) billion RDF triples loaded as Turtle files (a human readable representation of RDF) and persisted in S3 buckets within the Amazon Web Services (AWS) infrastructure. Several new or enhanced AWS capabilities were needed to support the deployment of the solution. With the use of AWS, integration to internal sponsor systems to authenticate users and grant their role privileges was added.

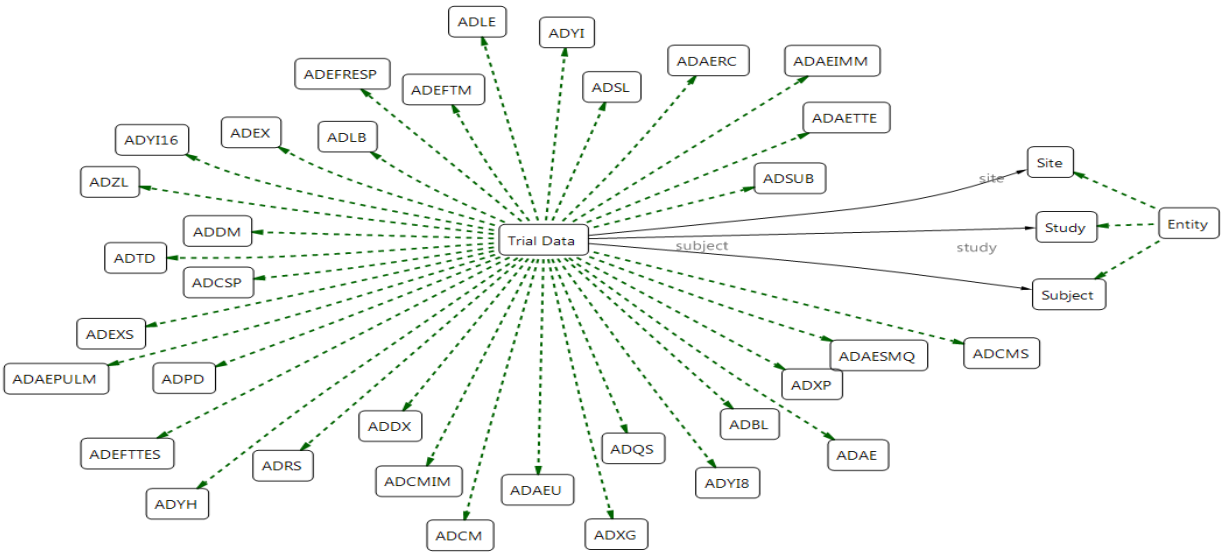

Figure 4 CDISC ADaM-based Canonical Data Model

Use case 4 – An automated Protocol Schedule of Activities (SoA). The Protocol Schedule of Activities (SoA) provides a simple mechanism to depict the arms, visits and activities of a Protocol including any unusual visit requirements or constraints. An automated Protocol SoA also provides an opportunity define the series of electronic Case Report Forms (eCRFs) needed to collect visit specific information about a patient. Using an Anzo-based Protocol SoA solution, standard visit profiles were defined along with their associated eCRF. The CDISC Protocol Representation Model (PRM) was used as the basis to develop the ontology supporting this solution. Figure 5 shows the high-level view of the SoA , a typical visit – by – activity matrix described in most Protocols. An RDF-based implementation of the PRM was developed to support the definition of Protocol SoA information and the user interface to quickly setup SoAs based on prior Protocols. This approach offers the potential to reuse portions of prior Protocols and will establish the foundation for standard eCRFs for standard activities.

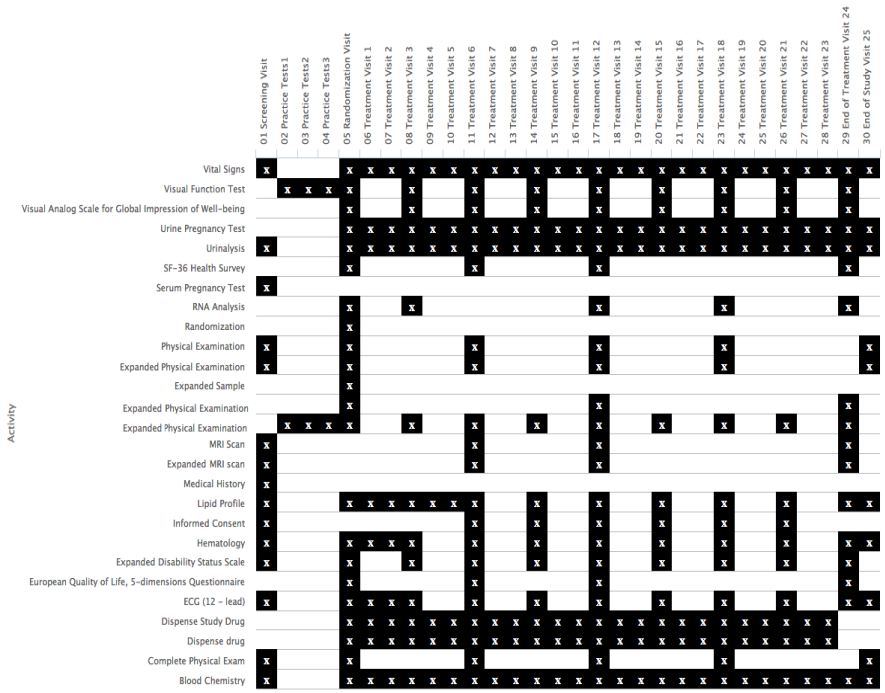

Figure 5 Anzo-based Protocol Schedule of Activities (SoA)

Use case 5 – Integrate clinical data with healthcare Real World Data (RWD). Leading sponsors recognize the tremendous data difference between their well-structured and often well-behaved clinical trial data and the complex and messy real-world heathcare data and have launched healthcare Real World Data (RWD) initiatives. However, visit-based clinical trial data and longitudinal healthcare data has proven difficult to integrate using traditional relational technologies.

Using previous experience gained from managing clinical trial data in CDISC-based RDF ontologies, an ADaM-based ontology was used to harmonize four (4) clinical trials for SAS-generated CSVs data files. A series of special-purpose analytic dashboards were created to support meta-analysis objectives and compared to prior Biostatistician developed meta-analyses.

A new, Electronic Medical Record (EMR) ontology was automatically created using Natural Language Processing (NLP) techniques to parse the EMR data dictionary and create the EMR ontology. The EMR ontology was used to map approximately 3.6 million patient records into RDF. The creation of the EMR ontology was more complicated and took longer than desired due to a lack of a pre-existing EMR data model. The creation of the EMR ontology would have been greatly accelerated had the selected EMR used the existing HL7 Fast Healthcare Interoperability Resource (FHIR®) data standards.

The two ontologies were joined through a series of canonical classes shared by both ontologies for Medication, Observation, Patient, Study, and Complaint. The combination of the clinical trial, EMR and canonical classes were used to define the purposefully built canonical RWD data model shown in Figure 6. The RWD data model was used to drive cross-data source analytics and capitalized on the processing power of the in-memory AnzoGraph RDF database. The RWD RDF database totaled about sixteen (16) billion triples and was hosting in an eight (8) node AWS cluster.

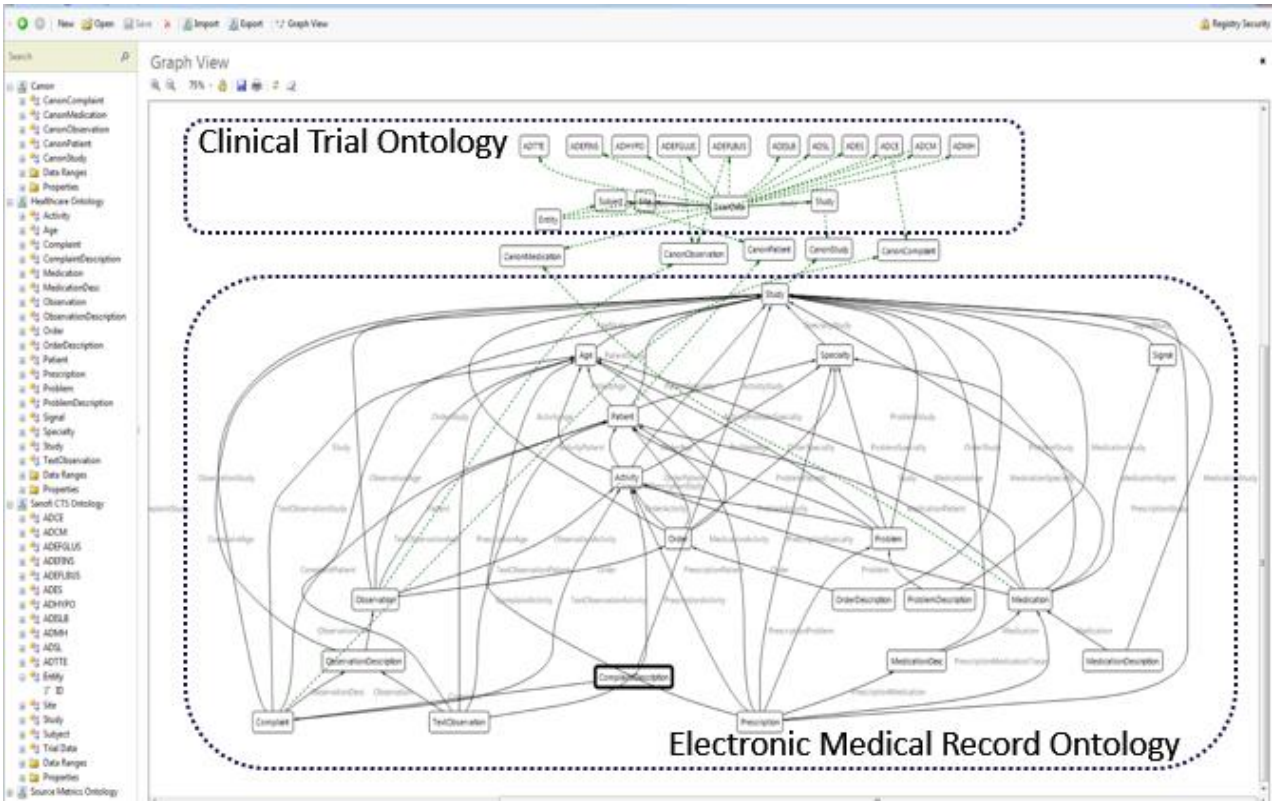

Figure 6 Real World Data (RWD) Canonical Data Model

This Anzo-based RWD solution was able to render sophisticated cross data source analytics to identify key characteristics of the healthcare population using clinical trial terms and filtering techniques. The resulting set of dashboard analytics were used to create a series of interesting health population answer sets for further simulation activities. A high-performance data conduit was developed to pipe answer sets directly to a ShinyR® application for clinical simulations. However, the close coupling of the Anzo RWD solution and the ShinyR simulation required too much processing power and memory to be sustainable. Instead, a loosely coupled mechanism was developed to land interesting answer sets in an AWS S3 bucket for access by the ShinyR application. This technique permitted the reuse of AWS nodes and enabled self-contained rapid iterations within the ShinyR application. The results of this innovative project were diverse data source analyses used to significantly increase the Probability of Success (PoS) of end-point attainment for several next generation investigative products.

CONCLUSION

Clinical trial data management is crucial to the ongoing success of all BioPharmceutical companies and is continually asked to do more with less resources. The complexity and rapid changes to the clinical data management landscape requires a new set of tools to keep pace with business initiatives. Many of us believe those tools exist but need wider adoption.

Next generation clinical data management tools are already being developed based on;

- RDF-based graph structures and OWL ontologies managed as IT industry standards by the W3C, not proprietary vendor technology stacks
- Life sciences industry ontologies from CDISC and HL7 which capture the shared experiences and requirements of many sponsors could greatly accelerate future ontology-driven solutions
- Sharing experiences of prior graph-based, ontology-driven use cases and solutions which demonstrate the possibility of future implementations and help capitalize on prior lessons learned

Cambridge Semantics is a leading proponent and developer of graph-based, ontology-driven clinical data management solutions. We have successfully demonstrated the use of these technologies, via our Anzo platform, that difficult clinical data management challenges can be addressed with this technology on a use case – by – use case basis. We will continue to share our experiences in utilizing these technologies to promote wider understanding and use.

We'd like to recommend the follow two calls-to-action as possible next steps toward next generation clinical data management solutions;

- First, the possibly of linking each unique use case ontology to one another is inherently available within the RDF-OWL technology framework to create a full lifecycle clinical data ontology. This ontology could be managed as a next generation data standard by either CDISC or HL7. Once done, next generation clinical data management platforms would become possible that addresses both current and future clinical data requirements.
- Second, we urge data management solution vendors to adopt these graph-based data structures to speed implements and promote interoperability amongst different vendors.

Once next-generation clinical data management solutions are being developed, it will become possible to enable rapid analyses of clinical trial patient data both within individual trials and across multiple clinical trials for meta-analyses. The RDF-OWL defined clinical data management ontology, based on a combination of CDISC / HL7 foundational standards and company-unique data standards, enable the rapid integration and consistent conformance of patient data from EDC to submission-ready RDF-based graph analytics in a few hours instead of weeks or months. Data from different trials and RWD may be rapidly combined, compared and contrasted via the ontology to help guide future clinical trial designs, to better understand overall compound performance and to develop better patient health outcomes.

REFERENCES

1. CDISC Foundational Standards for CDASH, SDTM, ADaM, and PRM
2. CDISC Standards in RDF Reference Guide. CDISC
3. CDISC Data Exchange Standards for ODM-XML and RDF
4. W3C Standards for RDF
5. W3C Standards for OWL
6. W3C Standards for SPARQL

ACKNOWLEDGMENTS

The author wishes to thank the following leading clinical data standards practitioners and RDF technology early-adopters for sharing their use case insights;

1. Patrick Jackson, Cambridge Semantics, Inc ®
2. Brooke Hinkson, Merck®
3. Mike Hamidi, previously of Merck® and now at CDISC™
4. Jeff Barrett, previously of Sanofi® & now at the Bill & Melinda Gates Foundation
5. Dan Boisvert, Biogen®

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jim LaPointe
Cambridge Semantics, Inc.
One Beacon Street, 15th Floor
Boston, MA 02108
Mobile Phone: +1 401-524-0297
Email: jtlapointe@cambridgesemantics.com
Web: www.cambridgesemantics.com