PhUSE White Paper
Clinical Trials Data as RDF

Contents

Overview

Historically in the clinical data field, a dataset is thought of in terms of rows and columns, variables and observations. This paper describes the motivation and utility of data as a graph, as demonstrated by the PhUSE Computational Sciences Symposium Emerging Trends and Technology working group project “Clinical Trials Data as RDF.” The project was formed to investigate the feasibility of using Linked Data as a future-proof technology for clinical trials results.

Linked data is a collection of interrelated datasets that is achieved using Semantic Web tools. The Resource Description Framework (RDF) is a general-purpose language for representing information in a graph-based data model that defines unique entities and the relationships between them.[1] Linked data also provides the ability to access different types of data from multiple sources, providing query results and analysis in real time. Although a traditional relational database may provide similar results, the resources needed to establish and maintain such a database can quickly become prohibitive.

The pharmaceutical industry deals with very large data sources of increasingly diverse origin. In addition, it must structure the data utilizing various standards (SDTM[2], ADaM[3], etc.) based on definitions that keep evolving. A common theme across the pharma industry is to increase data-driven research decisions. To facilitate this trend there is an urgent need for integration of data leading to cross-study analysis.

PhUSE initiated the Clinical Trials Data as RDF project at the annual Computational Sciences Symposium conference in Silver Spring Maryland in March, 2017 to investigate the ability of Linked Data to address the challenges inherent in the current standards. SDTM was chosen as the starting point because it is one of the most mature and widely adopted of the CDISC models (Decker, 2011). SDTM data to support the project was immediately available thanks to the previous efforts of the PhUSE Scripts project (https://github.com/phuse-org/phuse-scripts/tree/master/data/sdtm/cdiscpilot01).

In addition to supporting multi-source and multi-dimensional data, RDF represents the unique entities within a data landscape. Model development can therefore focus on these entities and their relationships instead of on a specific version of a standard. This capability in turn allows creation of a multi-dimensional data store for clinical trials data, while enabling strong alignment to past, present, and future submission standards. Linked Data is uniquely positioned to bring together multiple standards including SDTM, CDISC Terminology, WHO Drug, MedDRA, and others. High-quality, SDTM-conformant domains can be created using automated database queries instead of custom programming. A semantic reasoner[4] can be employed to infer new information based on existing data. For example, if a person is a part of the clinical trial then the person is inferred as a patient and there is no need for hardcoding this entity

Rather than map the existing SDTM model and example data directly into RDF, the project team chose to model the concepts needed to support SDTM creation. Modeling the clinical trial concepts and entities means the approach can be extended past SDTM and applied with relative ease to other aspects of the clinical trial data lifecycle (Oliva, 2017). When standards are embedded with the data and processes, they can be applied earlier to create data in the proper form (in a sense, "validating as you go"), rather than waiting until closer to the time of the submission. Because the data, standards, and rules are machine-readable, validation becomes increasingly automated. Future implementation

[1] http://linkeddata.org/faq
[2] Study Data Tabulation Model
[3] Analysis Dataset Model
[4] https://en.wikipedia.org/wiki/Semantic_reasoner

may propagate outward from this project in the direction of data collection, the protocol, and clinical study design, or in the other direction toward analysis datasets, results presentation, and publication.

The project's working hypothesis is that the Linked Data Model is closer to the data's real-world structure (natural form) rather than tabular structure industry uses currently. It includes explicit semantics not present in current models and corrects previous modeling constructs. When based on ontologies designed to represent the real-world processes in clinical trials, the model will be much more stable over time and easier to implement.

Definitions

(Clinical) Observation: a measure of the physical, physiological, or psychological state of a person or individual. Related terms: symptom, sign, subjective observation, objective observation, outcome measure, biomedical concept.

Medical Condition: a disease, injury, disorder, or transient physiological state (e.g. pregnancy) that interferes or may interfere with well-being. Medical Conditions explain abnormal clinical observations.

ADaM	Analysis Dataset Model
AE	Adverse Event, a medical condition that is temporally associated with an intervention
Assessment:	An analysis of one or more Observations(s) to identify and characterize a Medical Condition
BRIDG	Biomedical Research Integrated Domain Group
CDASH	Clinical Data Acquisition Standards Harmonization
CDER	Center for Drug Evaluation and Research
CDISC	Clinical Data Interchange Standards Consortium
CRF	Case Report Form
CSS	Computational Sciences Symposium
DM	Demographics Domain
ETT	Emerging Trends and Technology
FDA	Food and Drug Administration (United States)
FHIR	Fast Healthcare Interoperability Resources
LOINC	Logical Observation Identifiers Names and Codes
LPG	Labeled Property Graph
NCI	National Cancer Institute
Observation:	A measure of the physical, physiological, or psychological state of an individual (also known as a

	clinical observation.
ODM	Operational Data Model
OWL	Web Ontology Language
RDF	Resource Description Framework
RIM	Reference Information Model
SDTM	Study Data Tabulation Model
SPIN	SPARQL Inference Notation
TTL	Terse Triple Language (Turtle) https://www.w3.org/TeamSubmission/turtle/
URI	Unique Resource Identifier
VS	Vital Signs Domain
W3C	World Wide Web Consortium

Introduction

CDISC began working on standards in 1997, with the Submission Data Standards (SDS) team initializing development of SDTM as the standard for clinical trials data exchange in 1999. At present, our industry continues to struggle with significant implementation challenges, such as:

- Standards non-conformance resulting in a high incidence of rejection criteria for submissions (Allard, 2017).
- Multiple interpretations of the implementation guides leading to variability in standards implementation.
- Costs converting between versions.
- Limitations of the two-dimensional format and lack of intrinsic metadata.
- Challenges linking to other standards and data.

Linked Data as RDF can remedy many of the limitations of the CDISC standards. RDF ontologies[5] facilitate the modeling and representation of real-world clinical trial concepts, entities, and relationships. *Meaning* becomes integral to the data. Code lists, dictionaries, metadata, and value-level information are all intimately interconnected with the value-level results data. When validation rules are employed on top of this data, the result is high-quality, valid submissions that are reusable within organizations.

Linked Data also addresses the shortcomings of the antiquated V5 SAS Transport Format (PhUSE Emerging Trends and Technologies, 2017), by providing flexibility and extensibility for evolving requirements, support for integration from multiple sources across the data lifecycle, and robust metadata. Use of RDF is a paradigm shift from the SDTM as SAS XPT.

[5] https://www.w3.org/standards/semanticweb/ontology

RDF is not without its challenges. Graphs quickly become complex and can be difficult to navigate and understand. A lack of good user interfaces and scalable visualizations contribute to the inability to disentangle this complexity. The industry must work together to establish and implement standardized entities and semantics. As an industry, we must overcome our vested interests in antiquated technology and skill sets in order to move forward.

The Case For Linked Data and RDF

Linked Data is a way to structure and publish data using meaningful (semantic) connections. It is built upon the concept of each distinct entity having a single unique identifier, thus removing ambiguity when referring to items. When semantically interconnected data are created using robust classification schemes (ontologies), standard terminology, and rules, the data can be liberated from traditional data silos. Data can be combined from diverse sources. Standards become a view of the data instead of an imposed structure.

Introductions to Linked Data often become very technical in nature, and it is easy for readers to quickly become discouraged due to the depth of the subject. The benefits of Linked Data are best described in a relatable example.

Imagine the following scenario. You are a wine enthusiast, and you wish to find a bottle of a Cabernet Sauvignon made by Beaulieu Vineyards, Napa Valley, vintage year 2012. You want to determine where is it sold nearby for the best price, and you are willing to drive 10 miles from your house. To accomplish this, you may conduct your search as follows:

1. You go to the Google Maps website and identify the wine stores within a 10-mile radius of where you live. (Let's assume there are five stores that meet this criterion).
2. You go to Google Search, and you find each store and its website.
3. You find that three of the stores have their inventory online. You then conduct three different searches at three separate websites.
4. The other two stores do not share their inventory online, so you call each one and wait…. and wait….

After spending thirty to sixty minutes on this search, you find two wineries carry this wine. You pick the cheaper of the two, and off you go to purchase the wine.

Now imagine a different scenario where each inventory is online, and the data are linked with each other so you can conduct a single distributed search that looks something like this:

Search on	Business	=	Wine Store
	Location	=	10-mile radius from Point X (your home address)
	Item	=	Red Wine, 750 ml
	Vineyard	=	Beaulieu Vineyards
	Region	=	Napa Valley
	Variety	=	Cabernet Sauvignon
	Vintage	=	2012

Because the data are linked, the search engine has access to geographic data and wine inventory data across multiple web sites to rapidly return the answer from a single search. As an added bonus and because wine characteristics, such as taste, aroma, *etc*., are also linked to the individual wines, the search results also provide similar wines at each store that may be acceptable substitutes. If you could not find the exact wine you were looking for, the system could make recommendations based on similar characteristics, or purchasing patterns and recommendation from other customers. It is also easy to add new types of products (wine glasses, cork screws) to the store inventory without disrupting the existing database.

This is the promise of linked data: having the ability to access and query different types of data from multiple sources and being able to analyze them in real time to get the answers that one needs more quickly. Although a traditional relational database may provide comparable results, the resources needed to establish and maintain such a database can quickly become prohibitive. Federated search queries are not possible where searching for data is centralized, versus the distributed search capabilities facilitated by linked data.

The pharmaceutical industry deals with very large sources of data from a growing list of sources. In addition, the data must conform to various standard structures (SDTM, ADaM, etc.,). according to standard definitions whose specifications keep evolving. The industry is in dire need to link data from multiple sources, including the instructions for generating standardized structures, to make the data more useful. One important area is in data standards implementation. A second example serves to illustrate this point.

Let’s say you are collecting Ethnicity information on your study subjects. How do you represent this information using existing data standards when you may not know them very well? In order to accomplish this task, you may conduct your search along the following lines:

1. Go to the SDTM Implementation Guide, version 3.2 to identify the variable and domain where ethnicity information is recorded: Domain= DM (Demographics) and variable name is ETHNIC
2. Go to the latest version of the CDISC Terminology file to identify the appropriate controlled terms to use for each permissible ethnicity category
3. Go to the Define.xml v. 2.0 specification to identify how to report metadata information for ETHNIC (e.g., data type = xsd:string, CRF location, role)

As in the previous example, these important pieces of information—all of which are necessary to report ethnicity information in a standard format—exist in silos and require individual searches within each source to manually cobble together a complete picture regarding proper standardized reporting of ethnicity information.

Imagine if the data for all three sources were linked together in some way so that looking up ETHNIC in the SDTM IG automatically links you for the appropriate value set in the CDISC terminology file, and automatically links you to additional metadata around ethnicity data, for inclusion in the define.xml file.

One approach to linking data in this way uses the RDF standard, developed by World Wide Web Consortium (W3C). Each “thing” such as the ETHNIC variable in SDTM is called a “Resource” and is associated with a globally unique resource identifier so that each can be uniquely identified and linked to any other resource. In the same way that the Yahoo web page and the Google web page cannot be confused with one another because they have unique resource locators (URLs; web address), individual data resources cannot be confused either.[6]

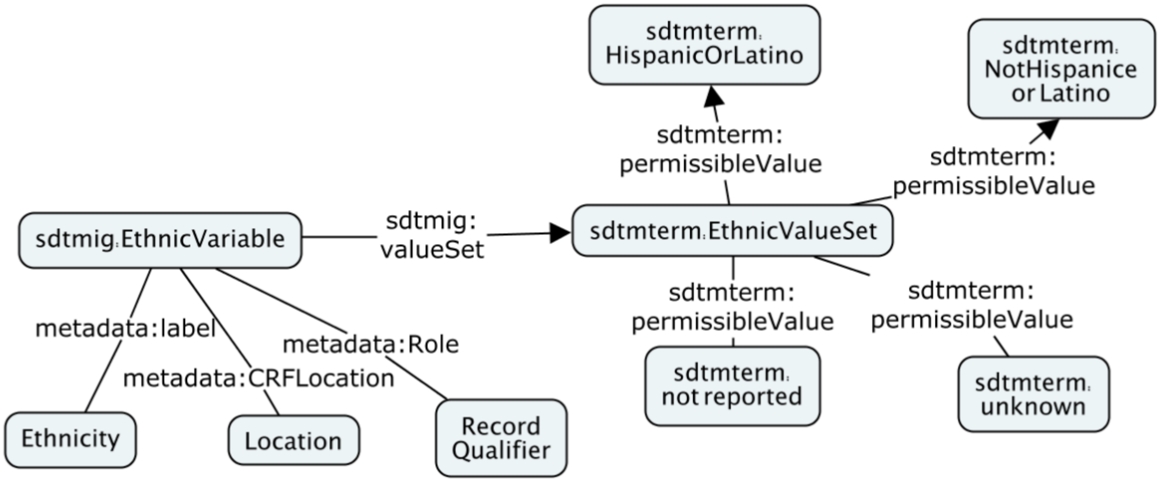
A fundamental advantage of this approach is that resources can be linked with each other, regardless of the source,

[6] One can think of a URI as a URL for data. This is not completely correct but it suffices to say that a URI allows resources to be distinguished from each other in the worldwide web or within a company’s intranet.

and the nature of the link can be described in another resource. Resources and links are considered "data as a graph" or graph data, and are stored in any one of several graph databases.

For simplicity, let's say all resources in the SDTM Implementation Guide have a prefix "sdtmig:", all resources in the CDISC terminology file have a prefix "sdtmterm:", and all resources having to do with metadata information (e.g., for define.xml) have a prefix called "metadata:". We now can create links that look like this:

Figure 1. Ethnicity Information Graph



Search engines can traverse a graph and provide useful information back to the user. For example: how many permissible values are there for ETHNIC information? What are they? The traditional approach would require manually searching through multiple, disconnected data sources. For the graph data approach, the data is linked together, facilitating a single automated search. When one considers that data standards implementation requires the use of many standards: CDASH, SDTM, ADaM, Define.xml, CDISC Terminology, MedDRA, WHO Drug Dictionary, etc., it is imperative that they all be linked in order to simplify and facilitate implementation efficiencies and inconsistencies.

"Data as a graph" is a more natural representation of real-world resources and relationships. Adding new information is easier and less disruptive than for relational databases where primary keys are required. Linked data has no such requirement, where appending data can be as simple as adding more resources and relations to both value-level data and the supporting ontologies and rules (which are themselves just more resources and relations). This encourages an incremental approach to building large databases without fear of boxing yourself in with a predetermined data model. Navigating the path of relationships in the graph is also much easier than in relational databases. RDF can also act as a translational layer between very diverse data silos, thus knocking down siloed information.

RDF Challenges

Slow uptake since its inception in the early 2000s led many to believe that Linked Data technology, and specifically RDF, is incapable of reaching mainstream adoption. However, a recent surge in interest has been aided by new players entering the space.

There are multiple reasons for the historically slow adoption. One is perceived complexity. The simple concept of Subject-Predicate-Object triples quickly builds into complex graphs, ontologies, and rules that can be difficult for newcomers to understand due to a lack of user-friendly tools for graph visualization and navigation. The complexity can be made invisible to end users by the development of proper interfaces. Unless you are interested in the technology, you do not need to know that your high-quality SDTM results data are derived from Linked Data. In fact, Linked Data sources are easily transformed into the SAS Transport file format!

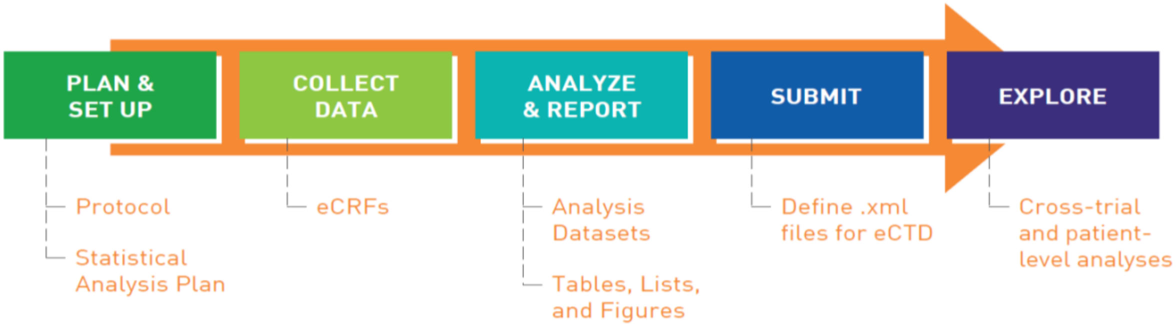
RDF graphs are viewed as complex, unwieldy, and difficult to update, maintain. Inferring values and relations using a reasoner engine is perceived as academic without practical application. When reasoning is used and the results are not what is expected, troubleshooting can be frustrating and labor-intensive. While inferencing and reasoning may initially be seen as having limited application for clinical trial results data, such capabilities may be applied more broadly within the data lifecycle, which in turn results in higher quality results data. Data quality is further enhanced when rules are integrated into the data using approaches like SPARQL Inferencing Notation (SPIN)[7], Shapes Constraint Language (SHACL)[8] and Shape Expressions (ShEx)[9].

A strength of RDF lies in its ability to leverage ontologies that are both human and machine readable. Many ontologies are available online and continue to be developed.[10] This plethora of ontologies is also a challenge, because no single ontology or approach is directly applicable to the CTD as RDF project. Clinical trials data is compartmentalized into different domains, thus distinct ontologies come into play to represent the information. Obtaining consensus is often difficult, if not impossible, to achieve. Therefore, the project will leverage existing ontologies where possible and will create new ones when necessary.

It is challenging to translate the promise of this technology into compelling business cases for managers with a solely relational database view of the world. This project is designed to serve as a basis for building and augmenting business cases using an approach designed to scale-out across the data lifecycle and broader operational ecosystem. Linked Data's capability to solve the many problems that exist in the data lifecycle make it a very disruptive technology, complicating return on investment calculations. (Williams, 2018) Companies are unlikely to replace existing systems and skill sets at great cost. Rather, both Graph and Relational systems are likely to coexist, leveraging the strengths of each.

[7] https://www.w3.org/Submission/spin-overview/
[8] https://www.w3.org/TR/shacl/
[9] https://www.w3.org/2001/sw/wiki/ShEx
[10] http://www.obofoundry.org/

Figure 2. Clinical Study Data Lifecycle

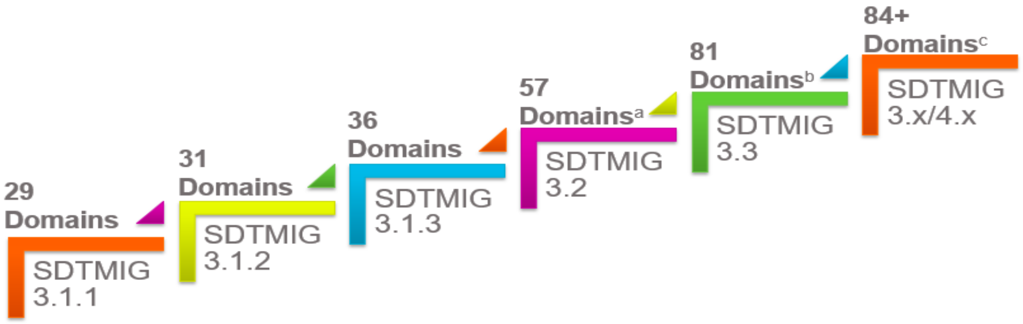

PLAN & SET UP
COLLECT DATA
ANALYZE & REPORT
SUBMIT
EXPLORE
Protocol
Statistical Analysis Plan
eCRFs
Analysis Datasets
Tables, Lists, and Figures
Define .xml files for eCTD
Cross-trial and patient-level analyses

Rationale

CDISC formed in the late 1990s to develop standards and models supporting the clinical trials data lifecycle to help optimize drug development and regulatory review. The CDISC mission statement emphasizes the development of data standards for medical research by stating **"The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare."** CDISC standards are vendor-neutral, platform-independent and freely available via the CDISC website." (CDISC, n.d.)

By working cooperatively with global health regulatory agencies, CDISC efforts led to implementation of numerous standards that try to accommodate data producer and consumer alike. SDTM was one of the first standards developed (Decker, 2011), supporting the submission of data to the FDA in standard domains, variables, terminology, and rule sets. As standards continue to develop in support of the clinical trial lifecycle, so do their number, scope, and complexity. Examples include ODM, CDASH, ADaM, and Define.XML.

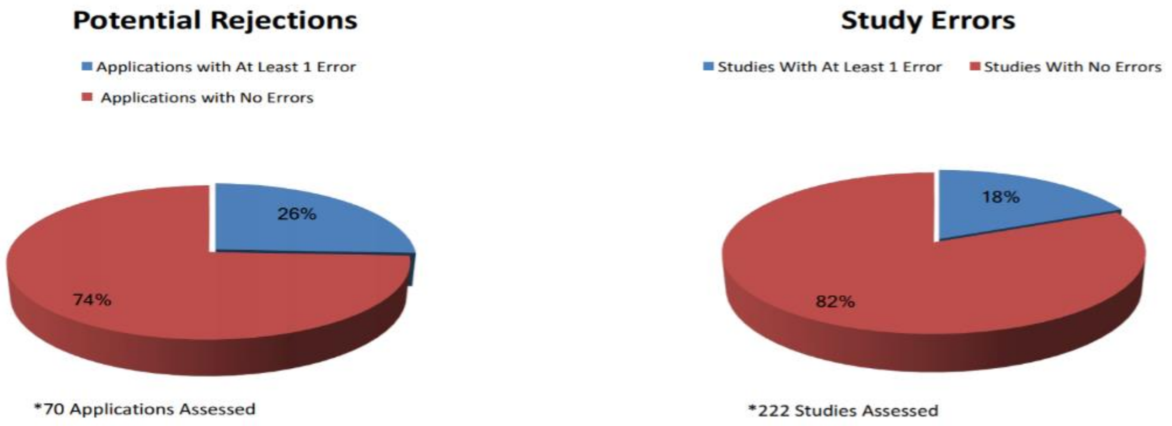
Figure 3 Evolution of the SDTIMG and Supporting Implementation Guidelines on a Domain-Level Perspective



Note: Count includes modeled domains only and does not represent SUPPQUAL or SENDIG specific domains.
[a] Includes APIGv1.0 (APDM & APRELSUB), MDIGv1.0 (7 Domains);
[b] Include items from footnote a plus PGxIG v1.0, Influenza v1.0 (VR) & Draft SDTMIG 3.3.
[c] Includes items from footnote b plus existing draft domains in CDISC development.

Additionally, inconsistent implementation across sponsors is widespread. A recent survey (Allard, 2017) showed 26% of CDER SDTM applications had at least one error. Anecdotal information from data analysts indicate that the actual incidence is much higher.

Figure 4. Errors by Application and Study[11]



Limitations in the CDISC models have led to challenges in data representation and implementation. A contributing factor is the underlying design of the SDTM structure itself, where each SDTM domain is understood to represent discrete categories of information. The DM domain serves as an example of the problems inherent in a two-dimensional, row-by-column design. While the DM domain is the primary source of demographics information, it also includes values for the study (STUDYID), treatment arm information (not just arm, but also the coded value for ARM, ARMCD), and units for the age column. These individual concepts should instead be modeled independently to decrease redundancy. Similar arguments can be made for each domain in SDTM, especially when considering the supplemental domains in the earlier SDTM versions.

When real-world, multi-dimensional clinical data are modeled to rigid two-dimensional standard data structures, important relationships are lost, which limits interoperability and reusability of the data. In addition, the tabular data structures have shown to be non-extensible, i.e., accommodating new clinical data content requirements for therapeutic areas often requires new domains and variables, which significantly increase implementation challenges.

The CDISC efforts brought much needed standardization to the industry, laying the groundwork for what needs to come next: A paradigm shift to flexible, freely available, multidimensional data models with integrated metadata and rule sets.

Project Deliverables

The project intends to deliver a prototype for the creation of high-quality, highly compliant SDTM data for select domains. The project deliverables also include data and methods for creating the relevant sections of the Define-XML

[11] Source: Allard, Crystal. " Technical Rejection Criteria for Study Data– Preliminary Findings." PhUSE 2017 Single Day Event
http://bit.ly/2HPEjha
http://www.phusewiki.org/docs/2017_SDEs/Boston%20SDE/Technical%20Rejection%20Criteria%20for%20Study%20Data%20%E2%80%93%20Preliminary%20Findings.pdf

document and a supporting ontology. The deliverables include:

- Conversion of study data from a minimum of two SDTM Domains from the CDISCPILOT01 data files. The resulting graph data leverages preexisting work, which developed RDF representation of the CDISC foundational standards (https://www.cdisc.org/standards/foundational/resource-description-framework-rdf/cdisc-standards-rdf and https://github.com/phuse-org/rdf.cdisc.org). The project explores leveraging other ontologies such as the NCI thesaurus, BRIDG, FHIR, the W3C time ontology (for temporal concepts), and others as deemed necessary and useful. The project will avoid SDTM domains that rely on large coding dictionaries, since these would negatively impact project scope. Data is round-tripped from SDTM source, to graph, and back to SDTM for validation.
- Separation of the results data values (i.e., observational or instance) from the standards data and metadata, resulting in a version-free graph data structure for clinical trial observations.
- CDISC-compliant SDTM data for submissions, created by mapping the standard to the study data. A consequence of this approach will be a drastic reduction in the costs for recoding between SDTM versions.
- Generation of highly compliant, high quality SDTM domains for study submission. Costs for data review, validation, and re-work also will be greatly reduced.
- The potential benefits of Clinical Trials Data as RDF project extend beyond the project deliverables themselves. These additional benefits may lead to future pilots to explore additional potential benefits.

Linked Data as the Solution

The project's working hypothesis is that a Linked Data model is closer to how clinical study data are created and used. It includes explicit semantics not present in current models (e.g., Assessment, Medical Condition) and corrects previous modeling constructs (e.g., SDTM models Adverse Events as Observations; whereas we believe they are best modeled as Medical Conditions). If designed correctly, the RDF model should be much more stable over time and easier to implement. Flexibility is increased since it is easier to accommodate new content requirements while maintaining backwards compatibility with older versions. When the appropriate rules are employed on top of the data, it becomes possible to generate high quality data in various formats including SDTM, ADaM, FHIR, etc.

The project team approached the problem from two directions. One sub team focused on the creation of a mini-study ontology to represent the concepts present in the pilot study DM and VS domains. The team considered the merits of a top-down modeling approach from a study, a protocol, and downward to the individual (e.g., observations), or to proceed bottom-up from observations within DM and modeling upward to the higher-level concepts, then expanding to include VS and potentially other domains. Both approaches have merit. The team chose a combined method that closely aligns with the pilot data while using a top-down approach to incorporate BRIDG and HL7 RIM concepts when necessary (e.g., Activities, Entities). A second sub team converted data from the CDISCPilot01 SAS transport files to RDF using R scripts to transform the data to match the ontology model developed by the first sub team.

The resulting query-able knowledgebase of clinical trials data includes the classification and structure of the model and its rule sets in addition to the instance data and metadata. Submission-ready SDTM domains are easily extracted and the data can be compared against the original sources in a round-trip check to ensure validity. The Define.XML are created on-demand for the in-scope domains. Future steps may include expanding the mini-Study ontology to accommodate data for other domains and investigating the automatic generation of the blank CRF. CRF generation could be based on the protocol and study ontologies along with additional metadata, impacting both the study design phase and later data validation and reporting phases.

The “Mini” Study Ontology

The first step was to create a mini study ontology using OWL (see Figure 5). We chose the concept of a mini study ontology to reflect the strategy of only modeling those concepts and relationships necessary to represent the data available in the SDTM DM and VS domains for the pilot study. Therefore, the study ontology is not complete. However, this approach minimizes complexity and with future iterations tests the hypothesis that iterative model development is not only feasible, but in fact desirable. Basing the data model on an ontological schema ensures the resulting instance data are well-formed, structurally consistent, and valid. For example, SDTM contains numerous operationally defined variables such as study day and baseline flags. By operationally defined, we mean these variables have standard definitions and derivations across studies such that their derivation can be expressed in RDF, thereby enabling their derivation “on the fly” using inferencing. This approach provides greater level of accuracy and consistency than what is currently being achieved.

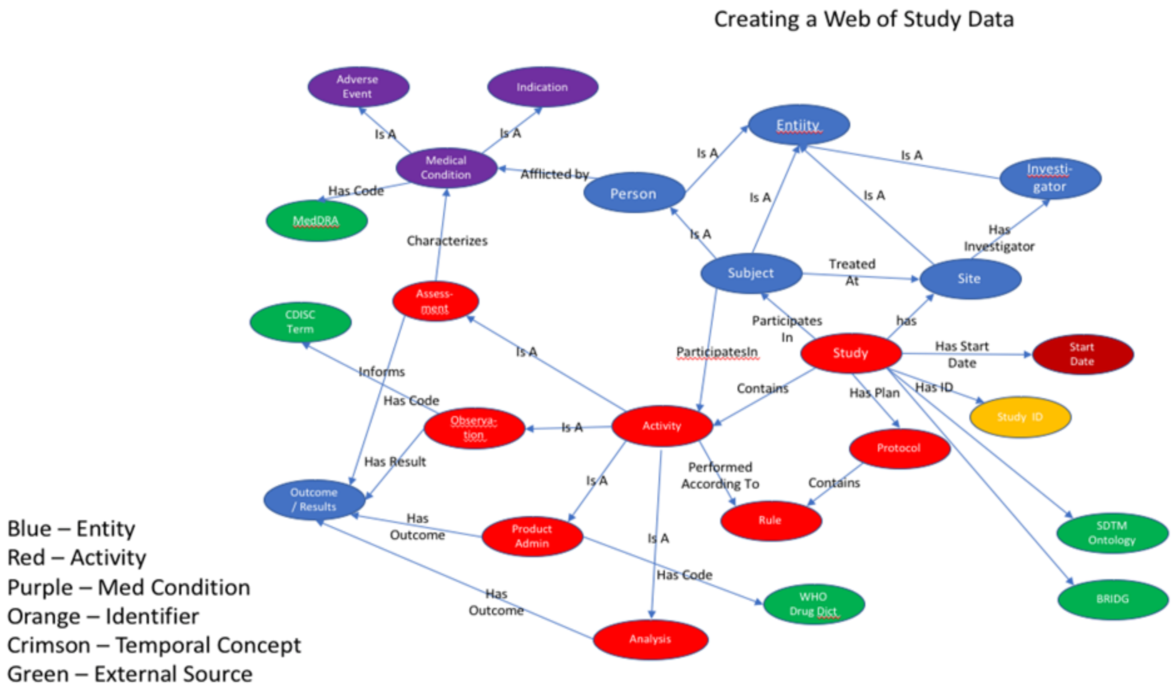
The fundamental core of the mini-ontology consists of a few classes and relationships.[12] It treats a study as a collection of Activities that are performed on Study Participants (HumanStudySubject) and their data (ActivityOutcome). HumanStudySubject may be afflicted by one or more MedicalCondition. It also recognizes that studies contain different types of activities: AdministrativeActivity (e.g., obtain informed consent, randomization), Intervention (e.g., product administration, surgery), Observation, and Analysis. It further recognizes that all Activities have Outcomes (ActivityOutcome), which in the case of Observations, are the Results. The Results can be represented using standard categorical terms from a dictionary or can be numeric data with or without associated units. Analyses are processes that take Activity Outcomes as inputs to generate useful analysis results. Analyses can be simple derivations (e.g., Age from Birthdate and ReferenceStartDate), more complex Assessments (e.g., Adjudication that a Myocardialinfarction is present based on various Observation Results), or highly detailed efficacy or safety analyses. Activities also have Rules that determine, for example, when Activities can be performed. A Rule is a type of Analysis because it takes as input the results of Observations to determine if the Rule is met (RuleOutcome resolves to “true”) or is not met (resolves to “false”). The core mini-ontology therefore has the following class structure:

- Activity
    - Observation
    - Analysis
        - Derivation
        - Assessment
        - Rule
- Entity
    - HumanStudySubject
    - Medical Condition
        - AdverseEvent

The ontology relies on the well-established principles of the clinical data lifecycle: first one observes, then one assesses before performing an intervention. A more detailed concept map is shown in Figure 5. It includes links to external data sources such as controlled terminologies and SDTM schemas allowing the extraction of instance data into highly-compliant SDTM domains.

[12] Classes are presented in CamelCase using a singular noun (e.g., StudyActivity) and relationships also in camelCase with the first letter always in lower case (e.g., HumanStudySubject participatesIn StudyActivity).

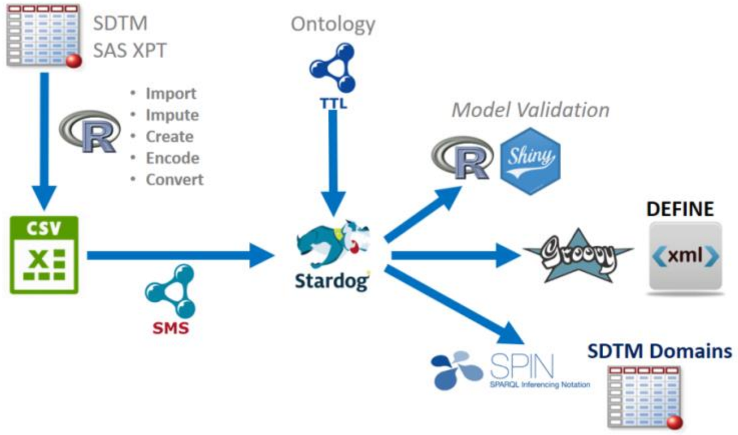
Figure 5 Mini Study Ontology



Instance Data Conversion

R was used to import the source data XPT files and perform data manipulation and imputation. Data required to validate the model that was not available in the original source was created using R or entered manually in supplemental CSV files. For example, a value for death date (`DTHDTC`) and death flag (`DTHFL`) was created in R for one patient in the DM domain even though no deaths were reported in the original study. *Investigator* and *investigatorID* information was not in the original data, so it was created in a CSV file since this information is part of most clinical trial data sets.

Figure 6 Data Conversion Process



Additional information required in the graph data was not found in the source XPT files. Examples include information such as the start rules that occur prior to making an observation or taking a measurement. Consider the case where a blood pressure is taken in the lying position (`VSPOS=SUPINE`, in the VS domain) after the subject assumed a supine position for five minutes. This corresponds to a start rule of `StartRuleLying5`[i] in the graph. These types of rules are part of the study protocol information and can be associated with the results values using OWL 2[ii] inferencing. Use of inferencing greatly reduces data redundancy. See the project's GitHub site for more details about this approach.[iii]

After the data manipulation is complete, the R data frame for each domain was saved as a comma-separated value (CSV) file that was in turn mapped to the graph database. The W3C standard R2RML "Relational Database to RDF Mapping Language"[iv] defines how to map data in a relational or row-by-column format to RDF graphs. Stardog further simplified R2RML as Stardog Mapping Syntax (SMS)[v]. SMS mappings can be converted to R2RML, allowing the project to maintain vendor neutrality for the conversion step.

Bringing Data Together

To achieve one of the major goals of the project, which was the automated generation of highly conformant SDTM data for submission, we chose to leverage previous work, including:

1. The **PhUSE CDISC to RDF** project, which modeled the CDISC standards using RDF. This work enables the derivation of SDTM datasets from the knowledgebase.
2. **SDTM Terminology in RDF**, which is published by the NCI and allows linking of important concepts in the mini-ontology to the controlled terms defined by CDISC.
3. **BRIDG 4.2 ontology**, which allows reuse of existing BRIDG concepts in the ontology as needed.
4. **W3C Time ontology**, which provides a standard representation of temporal concepts in RDF (instants, intervals, start/end dates, etc.)

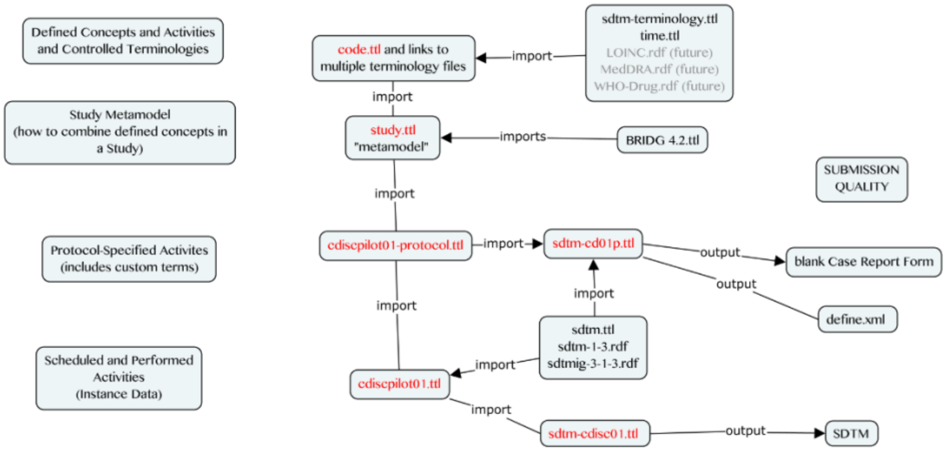
We were able to link these various external data sources to the mini-ontology to create a single seamless graph. The development process included the creation of various RDF files in Turtle format based on the type of data and how we envision the data will be managed in a production environment. A brief description of each file follows.

1. **code.ttl** – Contains or links to resources representing defined concepts such as controlled terminologies. It includes a catalog of Defined Activities. It currently provides links to SDTM terminology and the W3C time ontology. In the future, it can be expanded to link to other terminologies in RDF such as MedDRA, LOINC, and the WHO Drug Dictionary. For defined activities, it also contains links to the permissible activity outcomes (e.g. sex data collection activity outcome can only be Male, Female, Undifferentiated, or Unknown. It is anticipated that this file will reside and be maintained on a public site for all implementers to reference, although various links to proprietary terminologies would likely be restricted based on licensing agreements.
2. **study.ttl** – Contains the study metamodel in OWL. It contains the core classes and relationships previously discussed that are common to all studies. This ontology imports code.ttl. It is anticipated that the study.ttl file also will be publicly available on the web.
3. **cdiscpilot01-protocol.ttl** – Contains the concepts and relationships specific to the pilot study protocol, including the protocol-specified activities, rule sets, and controlled terms/value sets. It imports the study.ttl ontology. It is expected that this file will be the primary source to generate the blank CRF and the Define.xml contents. Since study protocols are considered proprietary, the file will likely reside behind a firewall with restricted access. It also defines a separate namespace called custom: to store protocol-specific concepts and custom terms that are not present in code.ttl.
4. **cdiscpilot01.ttl** – Contains the instance data for the study. It imports the cdiscpilot01-protocol.ttl file. This file also resides behind a firewall.
5. **sdtm.ttl** – Contains or links to the SDTM ontologies that are useful in creating valid SDTM datasets from the knowledgebase. This file is publicly available as are existing CDISC standards.

6. **sdtm-cdiscpilot01.ttl** – Links the instance data in cdiscpilot01.ttl with the SDTM ontology in sdtm.ttl from which the SDTM datasets are derived. Any protocol-specific SDTM implementation information is contained herein.

Figure 6 provides a schematic of the various files and their relationships with each other. Future links to other data sources are shown in gray. The figure illustrates a core principle of Linked Data in being able to link seamlessly to multiple external data sources; a feature missing in current SDTM implementations.

Figure 7 Importing Existing Data and Ontologies


Defined Concepts and Activities and Controlled Terminologies
Study Metamodel (how to combine defined concepts in a Study)
Protocol-Specified Activites (includes custom terms)
Scheduled and Performed Activities (Instance Data)
code.ttl and links to multiple terminology files
sdtm-terminology.ttl
time.ttl
LOINC.rdf (future)
MedDRA.rdf (future)
WHO-Drug.rdf (future)
import
study.ttl "metamodel"
imports
BRIDG 4.2.ttl
SUBMISSION QUALITY
cdiscpilot01-protocol.ttl
sdtm-cd01p.ttl
output
blank Case Report Form
define.xml
sdtm.ttl
sdtm-1-3.rdf
sdtmig-3-1-3.rdf
cdiscpilot01.ttl
sdtm-cdisc01.ttl
SDTM

Creating High-Quality, Valid SDTM Domains

Once the study ontology is completed and instance data are linked to the ontology, the implementer can use standard SPARQL queries to generate high-quality, valid SDTM domains for submission. Future enhancements allow the addition of validation rules as constraints to the data (e.g., AGE cannot be negative) to support integrated data validation.[13] There is a substantial disconnect between the data and supporting metadata when the two are not stored together[vi], which is the case in all non-graph approaches. When the data is in a graph, the data, metadata, validation checks, reporting, and domain and define.xml creation all occur within the same environment, greatly decreasing the amount of manual input and thereby lessening the chance for errors while decreasing time and effort.

Creating Define.XML

Historically, creation of define.xml required execution of SAS Macros to extract information from the SDTM domain datasets followed by augmentation from numerous sources, including intermediary files and labor-intensive manual input. The process has recently improved with new software applications but these still rely on manual addition of data and metadata that is not integral to the study data.

We demonstrated that by using a Linked Data approach, generation of define.xml becomes more automated, using SPARQL queries to extract the metadata this is now integral to the same data used to create the SDTM. In the future, this set of data+integrated metadata could be all that is needed for delivery.

[13] Machine readability is currently possible in RDF using SPARQL Inference Notation (SPIN) by defining constraint to the data (e.g., AgeOutcome) as RDF triples using the spin:constraint predicate.

Future Steps

To fully support the existing study data regulatory submission process, it is necessary to perform the following additional tasks:

1. Complete conversion of all pilot DM and VS data to RDF;
2. Extraction of complete, standards conforming SDTM datasets for DM and VS and corresponding define.xml tables
3. Expansion of the mini study ontology to support clinical data from all existing SDTM domains
4. Repeat 1 and 2 for all other SDTM domains

Additional steps beyond these depend on achieving consensus on which additional use cases will be supported next. For example, further automation of study conduct activities can benefit from:

1. Conversion of study protocol information, including all activity start rules, into RDF
2. Automated generation of the blank case report form from the study protocol

To support the data validation / data quality use case:

1. Convert SDTM conformance validation rules to RDF and use SPIN or SHACL to perform automated validation via inferencing
2. Convert FDA business rules for data quality to RDF and use same strategy as above to perform automated data quality checks
3. Convert information in FDA laws, regulations, guidances and technical specifications into RDF to automate conformance checks with existing regulatory policies.

Conclusion

The clinical research arena continues to evolve at a brisk pace. New data sources like those from wearables, ingestibles, and social media provide an increasingly diverse and complex array of data sources. Data models and structures must evolve along with these technologies. The flexibility of Linked Data means it is uniquely positioned to solve these challenges. When new content requirements emerge, all that is needed is to add more nodes to the graph. Additionally, the GO-FAIR consortium [14] endorses semantic web technology to make data findable, accessible, interoperable, and reusable.

This paper is not a proposal to replace current CDISC standards. Rather, it is a way forward to ensure their continued development. Any interim solution in the evolution of standards should provide backward compatibility (PhUSE Emerging Trends and Technologies, 2017). RDF provides powerful mapping constructs to define concepts in different standards or versions of a standard as equivalent, thus facilitating compatibility with legacy data. The Clinical Trials Data as RDF project provides such a stepping stone for compatibility with CDISC and other standards like HL7 FHIR; thus, enabling a simpler standards maintenance and implementation process is an important use case for data as RDF.

To be successful in the pharmaceutical industry, Linked Data approaches must mature past academic exercises to solve pertinent, practical problems with demonstrable return on investment. Efficient creation of high quality, valid SDTM data for submission is the second of many use cases within the clinical trial data lifecycle. RDF provides a standards-agnostic, multi-dimensional data model that can be leveraged to extract data into various versions of CDISC or in-house standards. Validation rules can be expressed using SPIN (SPARQL Inference Notation) or SHACL (Shapes Constraint Language) such that the data become self-validating via inferencing. The immediate result is more useful,

[14] https://www.go-fair.org/

higher quality data, making cross study analyses (yet a third use case) much easier to achieve, as described below. The data validation and the cross-study analysis are two use cases that are of particular interest to the U.S. FDA, as is evident by the design of their Janus Clinical Trials data warehouse. In addition, FDA continues to develop and refine analysis tools that rely on SDTM data. These highly efficient tools “break” when non-conformant data are submitted, making the validation use case absolutely critical to achieve long-term FDA goals. Furthermore, meaningful cross-study analyses depend on strict adherence to submission standards across all studies submitted to the FDA. As we have demonstrated, Linked Data (and RDF in particular) supports the auto-generation of highly standards conformant data, which is a critical “first step” towards enabling the validation and cross-study analysis use cases.

Implementation challenges remain, particularly in an industry wedded to its tabular data, existing data models, and standards. Standards must continue to be freely available to participants to ensure their evolution. We must coordinate our efforts not just between companies and regulatory agencies, but also seek solutions outside of the pharmaceutical industry. Additional tools for visualizing and working with Linked Data must be developed with a view toward lowering the bar for entry of new users.

These concerns and challenges should not limit the discussion. Rather, they should spur us into action to further develop the vast potential of Linked Data technology for the pharmaceutical industry.

References

Allard, C. (2017). Technical Rejection Criteria for Study Data - Preliminary Findings. Boston: PhUSE Single Day Event (SDE).

CDISC. (n.d.). *About CDISC*. Retrieved 05 01, 2017, from CDISC Website: https://www.cdisc.org/about

Decker, C. (2011). State of the Union: The Crossroads of CDISC Standards and SAS' Supporting Role. Las Vegas, Nevada: SAS Global Forum 2011.

Oliva, A. (2017). Managing Study Workflow Using the Resource Description Framework (RDF). Endinburgh: PhUSE Annual Conference.

PhUSE Emerging Trends and Technologies. (2017). *Transport for the Next Generation.* PhUSE.

Williams, T. (2018). Overcoming Resistance to Technology Change: A Linked Data Perspective. *PhUSE EU Connect.* Frankfurt: PhUSE.

Acknowledgements

The authors are indebted to the "Clinical Trials Data as RDF" PhUSE project team members for their contributions to the project. This paper is largely based on free, open-source software and the efforts of volunteers in PhUSE working groups. Please support those who donate their time and expertise through your own collaboration, participation, and promotion of these activities.

Contact Information

Project Co-Leads:

Tim Williams
UCB BioSciences, Inc
Raleigh, NC, USA
tim.williams@ucb.com
@NovasTaylor
https://www.linkedin.com/in/timpwilliams

Armando Oliva, M.D.
Semantica LLC
Fort Lauderdale, FL, USA
aoliva@semanticallc.com
@nomini
https://www.linkedin.com/in/aolivamd

All project files, data, and this paper are available from the project's Github repository: https://github.com/phuse-org/CTDasRDF. Study instance data: https://raw.githubusercontent.com/phuse-org/CTDasRDF/master/data/rdf/cdiscpilot01.ttl

Brand and product names are trademarks of their respective companies.

[i] This start rule states that the target activity (BP measurement) takes place only after the subject has been in the supine position for five minutes.

[ii] See https://www.w3.org/TR/owl2-overview/

[iii] See https://github.com/phuse-org/CTDasRDF/blob/master/doc/DataMappingAndConversion.md

[iv] See https://www.w3.org/TR/r2rml/

[v] See http://docs.stardog.com/#_stardog_mapping_syntax

[vi] "stored together" does not mean "in the same folder." If your data and metadata are not intimately intertwined in the same source, they are separate. This includes "a separate table in the same database".