Paper TT02

Breaking the Mold: Clinical Trials Data as RDF

Tim Williams, UCB Biosciences Inc., Raleigh, USA

Armando Oliva M.D., Semantica LLC, Fort Lauderdale, USA

# Abstract

After more than a decade since the implementation of CDISC SDTM as the standard for clinical trials data exchange, our industry continues to struggle with significant implementation challenges: [a] standards non-conformance resulting in a high incidence of rejection criteria for submissions (1). [b] Costs converting between versions. [c] Limitations of the two dimensional format and lack of intrinsic metadata. [d] Challenges linking to other standards and data.

This paper outlines the philosophy, ontology, and methods adopted by the PhUSE project “Clinical Trials Data as RDF.” By modeling *to the data* instead of to a specific standard, Resource Description Framework (RDF) supports a future-proof, multi-dimensional data store for clinical trials data while enabling strong compliance 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, standards-conformant, validated SDTM domains can be created using SPARQL rules (SPIN).

\*(1) Citation from FDA to be added.

# Introduction

The Clinical Data Interchange Standards Forum (CDISC) formed in the late 1990's to develop standards and models supporting the clinical trials data lifecycle to assist in optimizing drug development and regulatory review.

***CDISC Mission Statement***

"CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata.***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." (1)

By working cooperatively with agencies like the Food and Drug Administration (FDA), their efforts led to implementation of various standards that try to accommodate data producer and consumer alike. The Study Data Tabulation Model (SDTM) was one of the first standards developed (2), supporting the submission of data to the FDA in standard domains, variables, terminology, and rule sets. As standards developed to support the clinical trials lifecycle, so did the number and complexity of the standards themselves. Examples include the Operational Data Model (ODM), Clinical Data Acquisition Standards Harmonization (CDASH), and the Analysis Dataset Model (ADaM), and Define.XML .

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

# Limitations in current CDISC standards

Limitations in the CDISC models lead to challenges in standardized data representation and implementation. The design of each SDTM domain was chosen to represent discrete categories of information. DM is the primary source of demographics information, yet by design it must include representative 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 are best modeled independently for each type of concept they represent, with the benefit of decreasing redundancy in the data. Similar arguments can be made for each domain in SDTM, and especially the supplemental domains in the earlier SDTM versions.

Additional issues include:

* Adverse Events modeled as observations instead of medical conditions
* Multiple approaches for representing medical conditions (MH, AE, CE), leading to standardization inconsistencies
* Inconsistent approaches for representing changes in medical conditions over time
* Inconsistent approach for linking disease information (e.g. epilepsy, systemic lupus erythematosus) with the disorders associated with the disease (e.g. seizures, lupus nephritis) for any given subject
* Inconsistent representation of subjective observations/symptoms/patient reported outcomes
* No standard approach for representing assessment/adjudication information (i.e. the analysis of observations to identify and characterize medical conditions
* Sponsor defined definitions for important concepts that limit interoperability, e.g. Reference Start Date (RFSTDTC)
* Multiple locations for the same or similar information leading to data integrity issues, e.g. death information found in DM, DS, AE, others
* Data duplication and redundancy across domains
* Separation of coding and terminology from the instance data

# Consequences of the current models

As the models continue to grow in scope and complexity, inconsistent implementation across sponsors is widespread. Because real-world, multi-dimensional clinical data are modeled to the rigid 2-dimensional standard data structures, important relationships are lost, limiting 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 require new domains and variables, which significantly increase implementation challenges. The net result is that 45% of standardized data submissions to the U.S. FDA trigger rejection criteria based on only two conformance validation checks.

# Linked Data as the solution

Linked Data as (RDF) provides a solution to the limitations in the CDISC standards. RDF ontologies facilitate the modeling and representation of real-world clinical trials concepts, entities and relationships. Meaning (semantics) becomes integral to the data itself, which includes code lists, terminology and metadata intimately connected with results data. When validation rules are employed on top this data, the result is highly quality, valid clinical trials results data for submissions and use within organizations.

Linked Data is also a potential solution to the many issues facing the antiquated V5 SAS Transport Format (3), by including provenance and audit trail, flexibility and extensibility for evolving requirements, support for integration from multiple sources across the data lifecycle, and robust metadata.

Storage of clinical trials data as Linked Data represents a paradigm shift from the SDTM model. As an example, the RELREC domain becomes complete unnecessary[[1]](#endnote-1) , replaced by integral metadata and explicit linkages within the data. Linked Data provides flexible model for data with high validity. How do we get there from where we are now?

# PhUSE CSS Project: Clinical Trials Data as RDF (CTDasRDF)

The CTDasRDF project officially started at the PhUSE CSS conference in Silver Spring Maryland on March, 2017. SDTM was chosen as the staring point because it is one of the most mature and widely adopted of the CDISC models (2). It is more stable than ADaM, the implementation of which varies highly between studies and companies. 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>).

Instead of mapping the existing SDTM model and example data into directly into RDF, the project instead chose to model the *concepts* needed to support SDTM creation. Modeling the clinical trials concepts and entities means the approach can be extended past SDTM and applied with relative ease to other standards or phases in the clinical trial data lifecycle. Standards become embedded with the data and processes, meaning they can be applied earlier in the process, instead of waiting until nearer the time of the submission to then create data in the proper form (in a sense, "validating as you go"). Future implementation may propagate outward from this project in the direction of data collection, the protocol, and clinical study design, or just as easily in the other direction toward analysis datasets, results presentation, and publication.

Project deliverables extend beyond prototyping the creation of highly-valid data for select domains. The data and methods for

creating the relevant sections of the DEFINE document and a supporting ontology will also be delivered by the conclusion of the project, scheduled for March 2018 CSS conference. The value proposition for the project will be detailed in a White Paper and includes:

* Conversion of a minimum of two SDTM Domains from the CDISCPILOT01 data files. The resulting graph data will leverage preexisting work like SDTM terminology as RDF, available from CDISC.org. The project will evaluate alignment with other ontologies such as the NCI thesaurus, BRIDG, FHIR (if stable), a time ontology (for temporal concepts) and others as deemed necessary. The project will avoid SDTM domains that rely on large coding dictionaries since these would negatively impact project scope. Data will be round-tripped from SDTM source, to graph, and back to SDTM domain for validation of the results.
* Separation of the results data from the Standards data and metadata, resulting in a version-free graph data structure for clinical trials results.
* CDISC compliant data for submissions, created by mapping the standards data to the results data. A result of this approach will be a drastic reduction in the costs for recoding between CDISC versions
* Generation of highly compliant, high quality SDTM domains for study submission. Costs for data review, validation and re-work will be greatly reduced.

The project's working hypothesis is that the 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. Adverse Events are not Observations, rather Medical Conditions). If designed correctly, the 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 automatically generate high quality data in various formats including SDTM, ADaM, FHIR, etc.

The project team is approaching the problem from two directions. One sub team focuses on the creation of a mini study ontology to represent the entities 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, 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 the latter.

A second sub team converts 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 subteam. XXX The resulting RDF TTL files can be uploaded into a triplestore or directly queried from R or other applications.

The resulting queryable 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. DEFINE.XML are created on-demand for the in-scope domains.

All project files, data, and this paper are available from the project's Github repository: <https://github.com/phuse-org/CTDasRDF>

Future steps may include expanding the mini-Study ontology to accommodate data for other domains and investigating the generation of clinical summary reports.

# The study "mini ontology"

* Basing the data model on an ontological schema ensures not only the resulting (instance) data is well-formed, structurally consistent, and valid. It also ensures the data accurately captures the real-world clinical trial concepts accurately. [Could add example of AE as an Observation vs. a Condition?]
* A human Study is a set of Activities conducted on Persons (i.e. HumanStudySubject)
* Rules determine when those activities are performed
* The Outcomes of the activities (e.g. observation results) are Analyzed to make conclusions.
* These form the central study ontology ”Core Classes”
  + Activity
    - Observation
    - Analysis
  + ActivityOutcome
  + Entity
    - HumanStudySubject
  + Rule

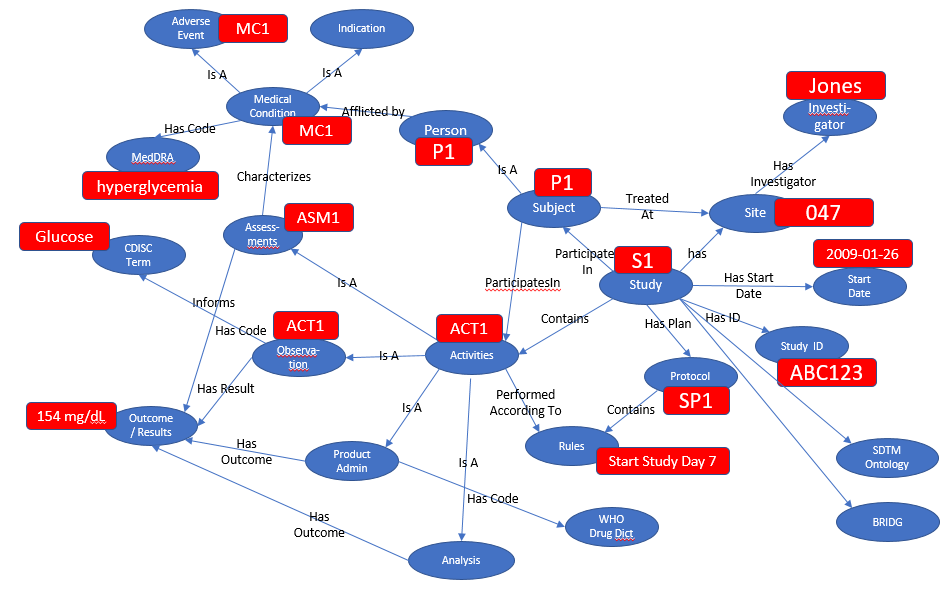


Figure 1 Minimal Study Ontology

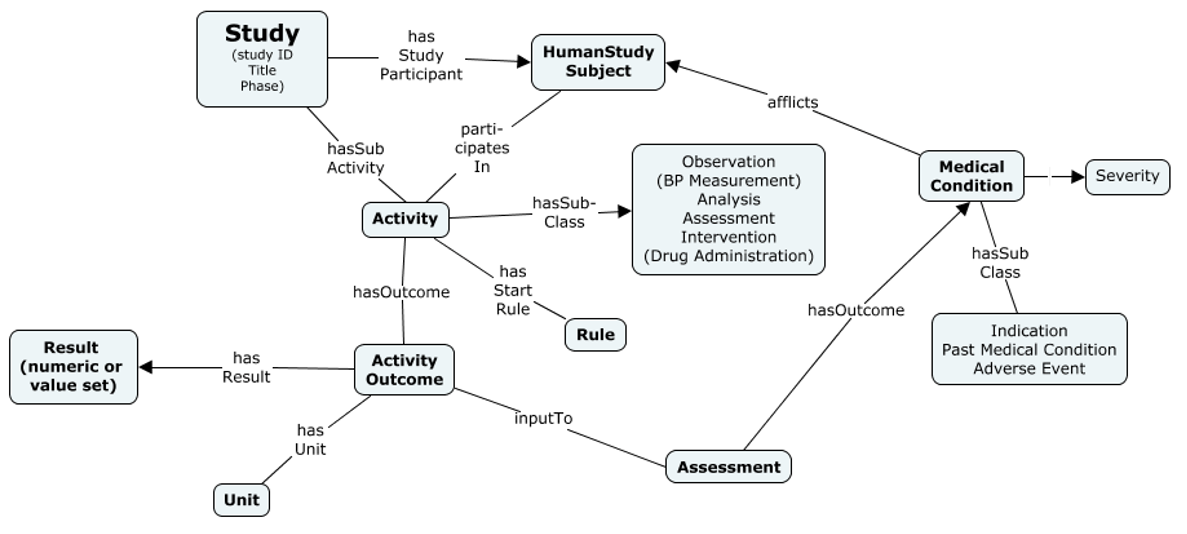


Figure 2 Minimal Study Ontology - OWL Representation

Bringing Data Together

* Leveraging pre-existing work

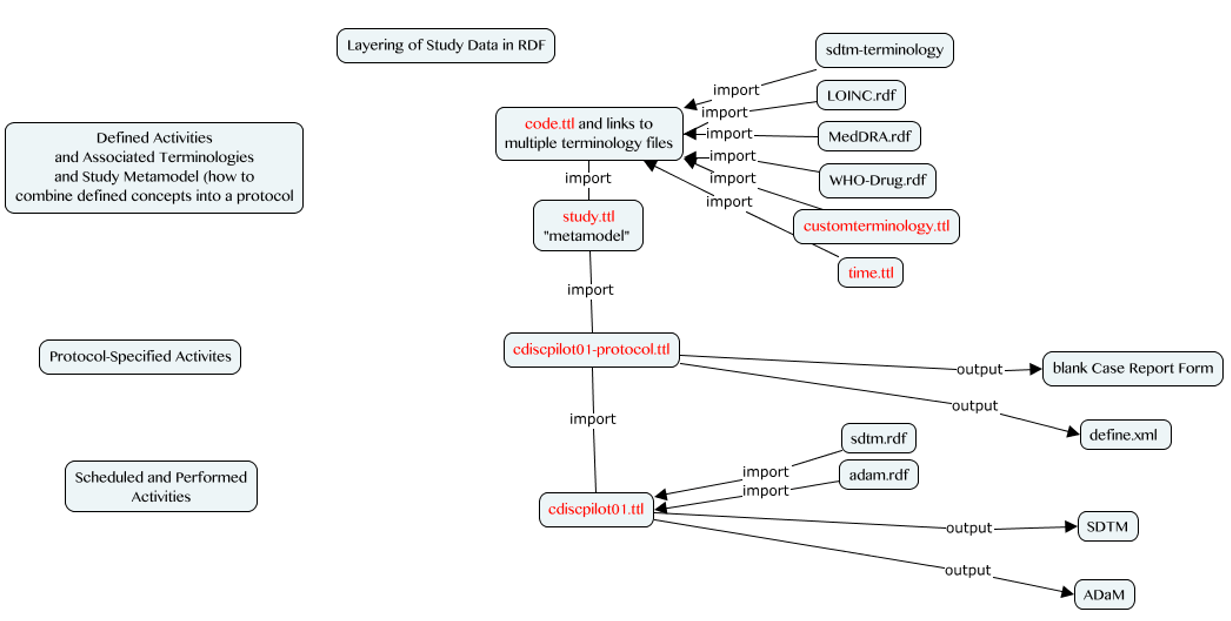


Figure 3 Importing Existing Data and Ontologies

* A core principal of Linked Data
* RDF is uniquely designed to link together multiple standards to facilitate implementation: SDTM, CDISC Terminology, WHO Drug Dictionary, MedDRA, etc.

Source data

* A human Study is a set of Activities conducted on Persons (i.e. HumanStudySubject)
* Rules determine when those activities are performed
* The Outcomes of the activities (e.g. observation results) are Analyzed to make conclusions.
* These form the central study ontology ”Core Classes”
  + Activity
    - Observation
    - Analysis
  + ActivityOutcome
  + Entity
    - HumanStudySubject
  + Rule

Creating High Quality, Valid SDTM Domains

* SDTM from RDF
* Creation of DEFINE to support SDTM
* In addition, validation rules, quality checks etc can all be expressed in the RDF....
* By storing additional metadata with the checks, the checks themselves are self-explanatory, without the need for supplemental documentation. Metadata is not limited to version and provenance information. Addition of appropriate metadata makes the values self-describing, removing any ambiguity from their interpretation and removing the need for separate files and documents to describe the data. These separate files and documents represent another point of failure in the process where documents become out-of-synch with the data the describe, have inaccuracies, and are costly to produce and maintain.

CreatinG DEFINE

Currently, the common process for creating DEFINE.XML includes the execution of SAS Macros to extract information from the SDTM domain datasets, then augment that data will user-supplied data from a number of sources. The process depends on a large number of intermediary files, manual input, and is labor intensive.

Define as a requirement to support the submission of SDTM domains

Generation becomes a query 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.

There is a substantial disconnect between the data and supporting metadata when the two are not stored together[[2]](#endnote-2), 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 creation all occur within the same environment, greatly decreasing the amount of manual input and thereby lessening the chance for errors and decreasing time and effort.

Why RDF, not LPG?

The Neo4j labeled property graph (LPG) provides an easy entry point for users new to graph data. The display of entities and their relations in an interactive web interface, coupled with the intuitive Cypher query language is attractive for new users new to Linked Data. Neo4j is used in large scale production deployment for product recommendation engines, tracing person and corporate relationships [example: Panama Papers(x)], social networks, and large number of other applications.

RDF was chosen for this project for several reasons. There is strong support in the development community, with several open source or free databases for RDF (called triplestores and quadstores) [cite]. The ability to directly use ontologies for classification and rule sets, and to a lesser extent inferencing and reasoning, was important to the project team. Rules can be applied using SPIN [x] and existing ontologies, terminologies, and data from the LOD cloud can be leveraged. Among the most important factors, RDF affords the exactness and precision needed in the pharmaceutical industry, not solely for modeling but also in the available data types by leveraging XML Schema Definition (XSD) (6).

Conclusion

The clinical research arena continues to evolve at a brisk pace. New data sources like those from wearables, ingestibles, and social media result in an increasingly diverse and complex array of data sources. Data models and structures evolve along with these technologies and flexibility of Linked Data means it is uniquely positioned to solve these challenges. When new content requirements emerge, just add more nodes to the graph. Powerful mapping constructs like owl:equivalentClass and owl:sameAs facilitate compatibility with legacy data or other standards.

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 evolution of standards should provide backward backward compatibility (3) and RDF project provides such a stepping stone for compatibility with CDISC and other standards like HL7 and FHIR.

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 SDTM data for submission is one such use case from the many that exist within the clinical trials data lifecycle. RDF provides a standards-agnostic, multi-dimensional data model that can be leveraged to extract data into various version of CDISC or in-house standards, agnostic of any specific standard.

It is foreseeable that in the future, companies could provide an a secure SPARQL endpoint to a regulatory agency for data submission. Templated, standardized queries would create the data necessary for review along with documentation, summary, and DEFINE information. Alternatively, development of Semantic Blockchain could be used as secure delivery Linked Data.

Implementation challenges remain, along with vested interests in existing data models and standards. Standards must continue to be freely available to participants to ensure their continued 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.

CUT CONCEPTS/TEXT

* Datensparsamkeit (1) , roughly translated from the German for our project as "store only what you need, link to the rest"

Leverage pre-existing work, including ontologies and terminologies. [= remove redundancy and do not recreate the wheel!]

References

1. **CDISC.** About CDISC. *CDISC WEbsite.* [Online] [Cited: 05 01, 2017.] https://www.cdisc.org/about.

2. *State of the Union: The Crossroads of CDISC Standards adn SAS' Supporting Role.* **Decker, Chris.** Las Vegas, Nevada : SAS Institute, 2011.

3. **PhUSE Emerging Trends and Technologies.** *Transport for the Next Generation.* s.l. : PhUSE, 2017.

4. **PhUSE CSS Project: Analysis Results & Metadata. *PhUSE Wiki.* [Online] [Cited: 08 23, 2016.] http://bit.ly/2bf7Pk0.**

**5. Fowler, Martin. Datensparsamkeit. *MartinFowler.com.* [Online] 12 12, 2013. [Cited: 05 01, 2017.] https://martinfowler.com/bliki/Datensparsamkeit.html.**

**6. World Wide Web Consortium. XML Schema Part 2: Datatypes Second Edition. [Online] W3C, October 28, 2004. [Cited: 05 8, 2017.] https://www.w3.org/TR/xmlschema-2/#built-in-datatypes.**

**7. Cairo, Alberto. *The Truthful Art. Data, charts, and maps for communication.* s.l. : New Riders, 2016.**

**8. Linked Open Vocabularies (LOV). [Online] Open Knowledge Foundation. [Cited: 07 26, 2016.] http://lov.okfn.org/dataset/lov/.**

**9. Uniform Resource Identifier. *Wikipedia.* [Online] https://en.wikipedia.org/wiki/Uniform\_Resource\_Identifier.**

**10. Willighagen, Egon. R Package: rrdf. *Github.* [Online] [Cited: 07 27, 2016.] http://github.com/egonw/rrdf.**

**11. Thieurmel, B. Introduction to visNetwork. *CRAN.* [Online] [Cited: 07 27, 2016.] http://cran.r-project.org/web/packages/visNetwork/vignettes/Introduction-to-visNetwork.html.**

**12. *PhUSE Workshop: Semantics 101 for Pharma.* Anderson, Marc and Williams, Tim. Barcelona : PhUSE, 2016.**

**13. CDISC. CDISC Standards in RDF. *CDISC.* [Online] [Cited: 08 23, 2016.] http://www.cdisc.org/rdf.**

**14. Robinson, Ian, Webber, Jim and Eifrem, Emil. Graph Databases. [Online] [Cited: 08 23, 2016.] http://neo4j.com/book-graph-databases/.**

**15. Resource Description Framework. [Online] WikiPedia. [Cited: 08 23, 2016.] http://en.wikipedia.org/wiki/Resource\_Description\_Framework.**

**16. W3C. SPARQL 1.1 Query Language. W3C Recommendation 21 March 2013. [Online] W3C. [Cited: 08 23, 2016.] http://www.w3.org/TR/sparql11-query/.**

**17. Bostock, Mike. D3 : Data Driven Documents. [Online] [Cited: 08 23, 2016.] https://d3js.org/.**

**18. W3C. SPARQL 1.1 Federated Query. W3C Recommendation. [Online] W3.ORG. [Cited: 08 23, 2016.] http://www.w3.org/TR/sparql11-federated-query.**

**19. PhUSE Semantic Technology Working Group Overview. [Online] PhUSE. [Cited: 07 27, 2016.] http://www.phusewiki.org/wiki/index.php?title=Semantic\_Technology.**

ACKNOWLEDGEMENTS

The authors are indebted to the following people, organizations, and companies:

* PhUSE "Emerging Trends and Technologies" working group
* "SDTM as RDF" PhUSE project team members
* R-project, Egon Willighagen for rrdf
* Frederik Malfait & CDISC

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

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

|  |  |
| --- | --- |
| Tim Williams  UCB BioSciences, Inc  Raleigh, NC  tim.williams@ucb.com (work)  NovasTaylor@gmail.com (personal)  C:\_sandbox\sas\Conferences\PhUSE-SDE-CARY2015\images\Twitter_logo_blue.png @NovasTaylor  <https://www.linkedin.com/in/timpwilliams> | Armando Oliva  Semantica LLC  Fort Lauderdale, USA  aolivamd@gmail.com  C:\_sandbox\sas\Conferences\PhUSE-SDE-CARY2015\images\Twitter_logo_blue.png @nomini  https://www.linkedin.com/in/aolivamd |

Brand and product names are trademarks of their respective companies.

1. Don't panic! RELREC can be recreated as needed just like any other SDTM domain, based on rule sets applied to the data and metadata in the triplestore. [↑](#endnote-ref-1)
2. "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". [↑](#endnote-ref-2)