Paper AANN

Breaking the Mold: Clinical Trials Data as RDF

Tim Williams, UCB Biosciences Inc., Raleigh, USA

Armando Oliva, 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, data model, and methods adopted by the PhUSE project “SDTM Data as RDF.” By modeling *to the data* instead of to a specific standard, Resource Description Framework (RDF) provides 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 a number 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. This is future can become a reality using Linked Data.

Limitations of Current models and Standards

The row-by-column problem

The standards themselves and the instance data to which they are applied are extremely limited by their rigid, two-dimensional, row-by-column representation. This results in several drawbacks, examples of which include (but are not limited to):

|  |  |
| --- | --- |
| **Issue** | **Example** |
| Duplication and redundancy of data | Study identifier values repeated in each row of DM. |
| Inconsistent structures between domains | One row per patient in DM, multiple rows per patient in VS  Structure forces one-to-many merges using identifiers across domains in multiple files, increasing changes of error and decreasing data processing efficiency |
| Instance data separated from coding | SDTM Terminology is separate from the instance data. |
| Standards, instance data, metadata, all stored in separate files | Difficult to synchronize. High chance of introducing errors or not finding errors. |
| <ADD> |  |
|  |  |

Modelling limitations in SDTM

* <things modelled incorrectly in SDTM: Provide example>
* Mixture of concepts.

The contents of each SDTM domain were 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 modelled 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.

Consequences of the current models

* Interpretation issues: Inconsistent implementation between companies
* Data is modelled to the standard and the standard's structure, making conversion between standards (and different versions within a standard) needlessly complex
* Compliance issues: 45% of SDTM data submission trigger a rejection criteria based on only two checks.
* <not sure what to call this next one…>

As a case in point, the CDISC Terminology has Codes for concepts ("Code") and Codes of value sets, which they also call code lists and value domains ("Codelist Code").

* C66741 is the code list code for Vital Sign Test Code, which are the permissible values for the SDTM variable LBTESTCD  (short name)
* C67153 is the code list code for Vital Sign Test Name, which are the permissible values for the SDTM variable LBTEST  (long name)

The same concept DiastolicBloodPressure (C25299) exists in both code lists. Given the CDISC naming convention of including both the code list code and the concept code, there are two resources for the same concept. Even worse, the two resources aren't linked to each other. I think this is a flaw in the CDISC terminology file.  So technically both are correct. Here are some options

1. For the purposes of the pilot, we can just pick one, either one, and document the need to resolve this redundancy in the CDISC terminology file as a future enhancement. I have no problem using the other resource. What do you think?

2. we can add an owl:sameAs triple to the file so it looks like the following. A search for one will automatically return both.

sdtm-terminology:C66741.C25299

        a                    owl:Class ;

        rdfs:subClassOf      code:VitalSign ;

        owl:equivalentClass  [ a                  owl:Restriction ;

                               owl:allValuesFrom  code:BloodPressureOutcome ;

                               owl:onProperty     study:hasActivityOutcome

                             ] ;

        skos:prefLabel       "DIABP"^^xsd:string ;

       owl:sameAs        sdtm-terminology:C67153.C25299 .

apparently this type of redundancy is common in the semantic web, so the owl:sameAs property was developed to avoid the need to rename resources every time a duplicate is found.

Me:

The code list code and the concept code are not even linked?  I'm going to file this under "WTF, CDISC?!!?"  This is a good example of how our graph approach is an improvement over the CDISC model. We may wish to use this example in the paper we will write for the PhUSE Annual conference?

How about we use vstestcd directly in the "data" TTL, and relate it to the value of vstest in OWL? I would chose vstestcd for the data because it is more compact than vstest and I see vstest as a (more verbose) description the thing represented by vstestcd

Linked Data as The solution

Linked Data as Resource Description Format (RDF) provides a solution to the limitations in the current models.

(an overall list here, with details to be provided in the later sections)

* the multidimensional nature
* ontologies to accurately classify the concepts present in clinical trials data
* existing terminology
* ability to provide integrated metadata
* validation rules to accompany the data (SPIN, etc.)

The result can be a highly valid and flexible data store for clinical trials data. How do we get here from where we are now?

A Linked Data approach is a potential solution to the many issues facing the antiquated V5 SAS Transport Format (3), including:

1. Provenance and audit trail
2. Flexibility and extensibility for evolving requirements
3. Support for integration from multiple sources across the data lifecycle
4. Robust metadata

Storage of clinical trials data as Linked Data represents a paradigm shift from the SDTM model. The RELREC domain becomes complete unnecessary[[1]](#endnote-1) , replaced by integral metadata explicit linkages in the data.

CSS Working Group : SDTM DATA as RDF

Why SDTM was chosen

* It is one of the most mature and widely adopted of the CDISC models (2)
* The model is well known an the implementation is more stable than ADaM, which can very highly between studies and companies.
* By fixing modelling errors and omissions, the value of the linked data approach can be demonstrated.
* Current debates of the longevity of the SAS Transport format (.XPT) provide a potential opening for Linked Data Approaches
* A previous project that focused on conversion of ADaM-like results data did not gain immediate traction (3)

Project philosophy and approach

We are not modeling the SDTM standard. Rather, we are modeling the concepts and data needed to support SDTM creation. These items are in turn an integral part of the clinical trial data lifecycle itself, allowing additional components to be added with relative ease, and for the standard to be adopted earlier in the process, instead of waiting until nearer the time of the submission to then create data in the proper form. In other words, this standard could propogate in both directions, back toward data collection, the protocol, the clinical study design, and toward the opposite end: analysis datasets, results presentation, and publication. The standards become intimately embedded within the data and the process.

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!]

Modelling to the data, not to the standard

The project will develop a prototype method for creating SDTM domain data as Resource Description Framework (RDF). Deliverables also include the data and methods for creating sections of the DEFINE and a minimal study ontology relevant to the domains selected for the prototype.  The value proposition for the project will be detailed in a White Paper and includes: 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 will be created by mapping the standards data to the results data. Costs for recoding between CDISC versions will be drastically reduced. Generation of highly compliant, high quality SDTM domains for study submission. Costs for data review, validation and re-work will be greatly reduced. A minimum of two SDTM Domains will be selected from the CDISCPILOT01 data files. The resulting graph data will leverage existing work like the 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 DDTM source, to graph, and back to SDTM domain for validation of the results.

* Modeling the data, not the standard
* Hypothesis:
  + Model is closer to how clinical study data are created and used
    - Includes explicit semantics not present in current models (e.g. Assessment, Medical Condition)
    - Corrects previous modeling constructs (e.g. Adverse Events are not Observations, rather Medical Conditions)
  + More stable over time
  + Easier to implement
  + More flexible
    - Easier to accommodate new content requirements
    - While maintaining backwards compatibility with older versions
  + Able to generate in an automated fashion, high quality data in various formats: SDTM, AdaM, FHIR, etc.
* Pilot focus is generating high quality SDTM domains: DM and VS
* Start with SDTM data: CDISCPilot01 Study
* Generate a mini-Study ontology to accommodate DM and VS data from the pilot study
* Create a Knowledgebase:
  + Convert SDTM data to RDF based on the mini-study ontology
* Query the Knowledgebase:
  + Create high quality, submission ready SDTM DM and VS domains
  + ”Round-Tripping” back to SDTM
* Dynamic DEFINE.XML creation
  + Future Steps: Expand mini-Study ontology to accommodate data for other domains
* The approach an why it was chosen: Could model from the top down : A study, a protocol…down to the individual, or from the data and domains upward. Both approaches have merit. The team chose the latter, to start modeling at the observations within DM and model upward to the higher-level concepts

Why RDF, not LPG?

Labelled property graphs like Neo4j provide an easy entry point for users new to graph data. The display of values and their relations in an interactive web interface, coupled with an intuitive query language (Cypher) is attractive for new users as a way to understand how graph databases function. Neo4j is used in large scale production deployment for product recommendation engines, tracing person and corporate relationships [Panama Papers, etc.], 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 version of the data store [cite]. The ability to directly use ontologies for classification and rule sets, and to a lesser extent inferencing and reasoning. Rules can be applied using SPIN [] and existing ontologies, terminologies, and data from the LOD cloud can be leveraged. Amoung 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).

Project Github site: <https://github.com/phuse-org/SDTMasRDF>

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

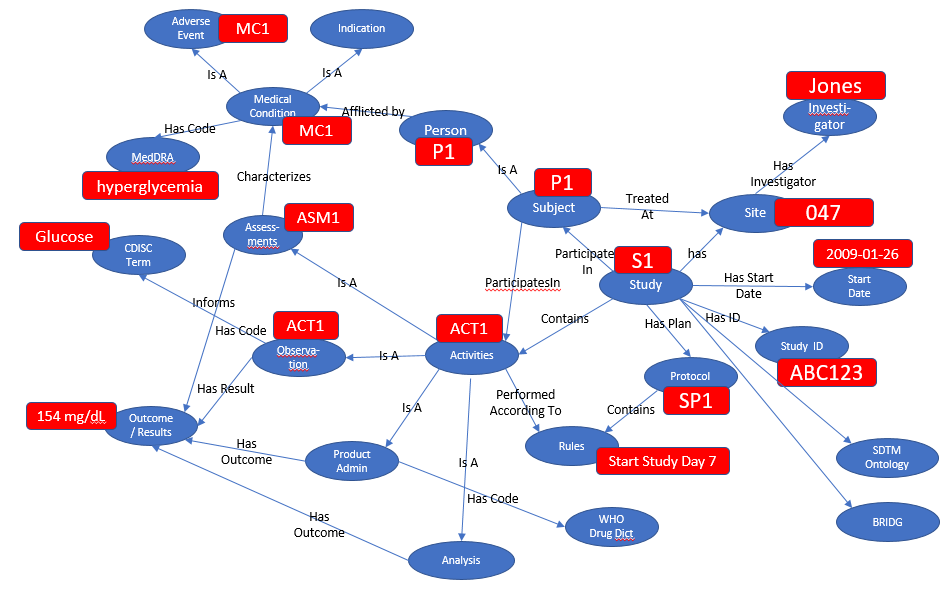


Figure 1 Minimal Study Ontology

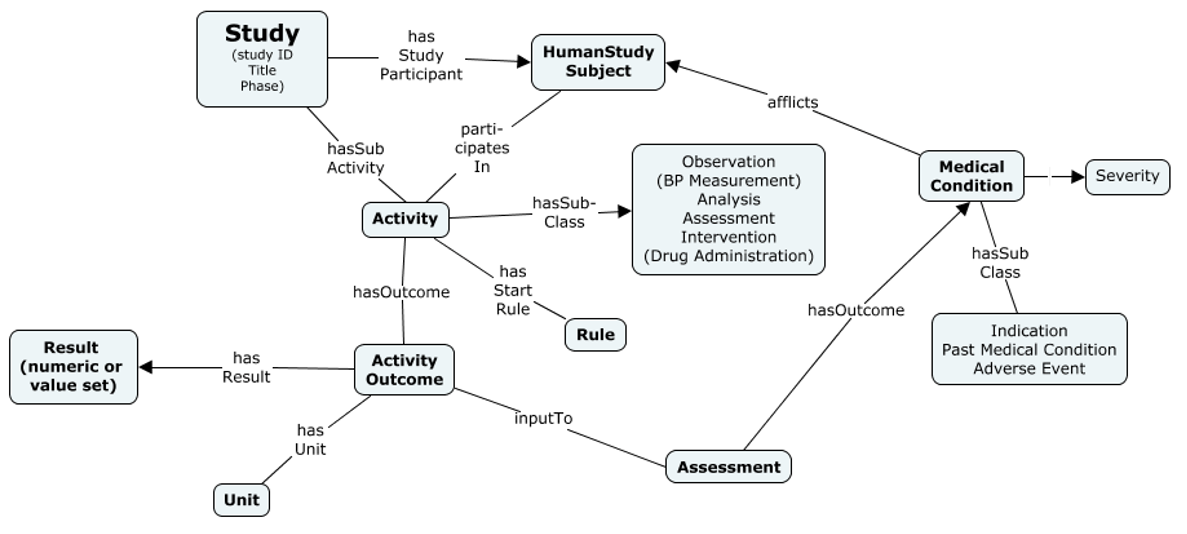


Figure 2 Minimal Study Ontology - OWL Representation

Bringing Data Together

* Leveraging pre-existing work

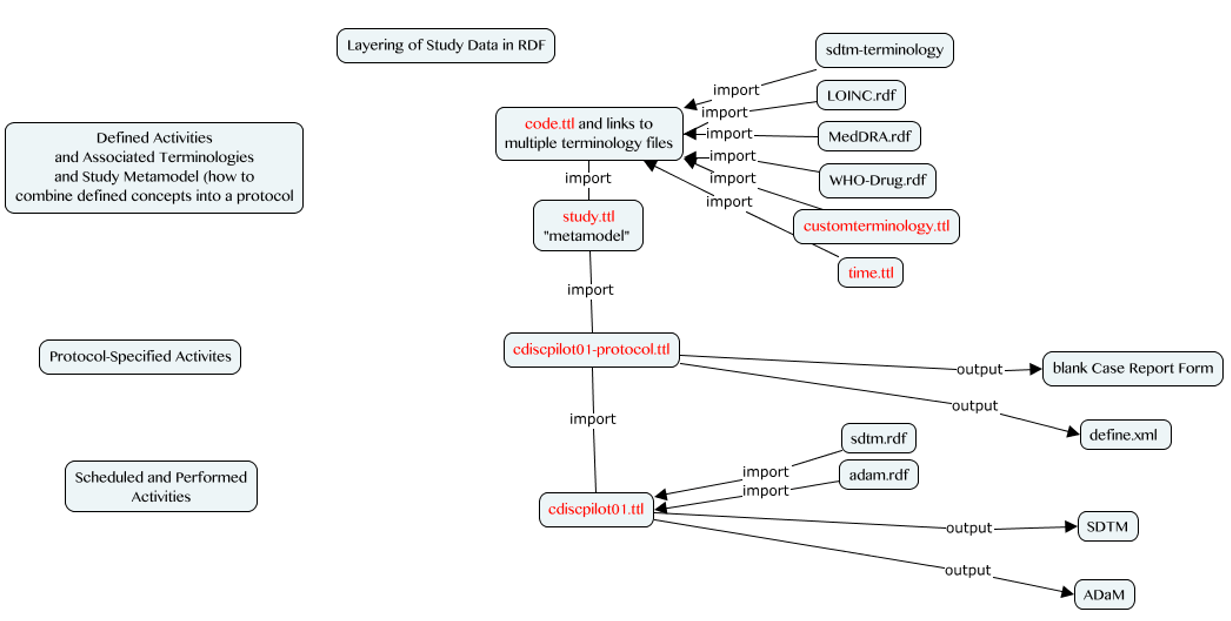


Figure 3 Importing Existing Data and Ontologies

* 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 labour 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.

Future PRoof and Flexible

* RDF can peacefully co-exist with other standards.... it can output to SDTM, ADaM, FHIR, or any other format
* Flexible: it can accommodate new content requirements easily ... just more nodes in a graph
* Powerful mapping constructs for legacy data:  (e.g. owl:equivalentClass, owl:sameAs)
* The clinical research arena continues to evolve at a brisk pace. New data sources like those from wearables, ingestables, and social media result in an increasingly diverse and complex array of data sources. Data models and structures evolve along with these technologies – Blockchain, etc. as well as the computational and statistical approaches to support complex analsyes(machine learning, Bayesian?). Robust-yet-flexible standards are needed to support this dynamic landscape. A myriad of disconnected standards and data models is not the answer: Mutlidimensional, linked data with integrated metadata can provide the solution.

By its very nature, a Linked Data solution provides relationships between values and concepts without the duplication and redundancies found in two dimensional structures. Furthermore, the relationships themselves can be further annotated with descriptive metadata, facilitating explicitly self-documented, machine-readable data.

Any new data format should provide backward compatibility (3) to existing standards, like CDISC SDTM.

**A Glimpse into the Future.**

* Companies open a secure SPARQL endpoint to the regulatory agency who could then obtain the data using standardized and templated queries, generating documentation, summary, and DEFINE information "live", as needed against the data.
* Semantic Blockchain for secure delivery, validity, ownership

Figure x foo foo foo

**CHALLENGES**

* need to coordinate efforts
* tools for visualization, API's for data usage/ease of use.
* Versioning – StarDog can version graphs.
* Concerns with the availability of information from CDISC SHARE project due to a very restrictive licensing agreement from CDISC
* Vested interests from vendors, pharmaceutical companies, and regulators provide resistance to moving to new solutions, instead preferring to band-aid and kludge existing data models and ways of thinking.
* ROI must be demonstrated: going from one SDTM version to another, fewer errors. Merely the act of converting existing data to linked data makes errors and missing data explicitly known. Add to that the validation rules (SPIN) , and highly compliant SDTM domains became routine.

Conclusion

* To be successful, Linked Data approaches mature past academic exercises to solve pertinent, practical problems and show return on investment. Until know, this has been elusive for Pharma.
* Study data represented as RDF provides richer meaning that travels with the data.
* Can use used to easily and automatically generate high quality SDTM data for submission.
* Use of "query templates" for extraction to various versions of a CDISC standard, or to any in-house custom standard.
* RDF provides a mult-dimensional data model, agnostic of any specific standard.
* Implementation challenges remain, along with vested interests in existing data models and standards. These should not limit the discussion. Rather, they should spur us into action.

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)