PhUSE 2017

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).

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. The CDISC mission statement emphasizes the development of data standards for medical research:

> ***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." (2)

By working cooperatively with agencies like the Food and Drug Administration (FDA), CDISC efforts led to implementation of numerous standards that try to accommodate data producer and consumer alike. The Study Data Tabulation Model (SDTM) was one of the first standards developed (3), supporting the submission of data to the FDA in standard domains, variables, terminology, and rule sets.

As standards continued to develop in support of the clinical trial lifecycle, so did their number, scope, and complexity. Examples include the Operational Data Model (ODM), Clinical Data Acquisition Standards Harmonization (CDASH), the Analysis Dataset Model (ADaM), and Define.XML. Inconsistent implementation across sponsors is widespread. A recent survey (1) showed 26% of CDER SDTM applications had at least one error.

Limitations in the CDISC models lead 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 DM 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, which would decrease redundancy. Similar arguments can be made for each domain in SDTM, especially when considering the supplemental domains in the earlier SDTM versions.

Additional challenges 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

When real-world, multi-dimensional clinical data are modeled to rigid two-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 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.

Linked Data as the solution

Linked Data as Resource Description Framework (RDF) can remedy many of the limitations of the CDISC standards. RDF ontologies facilitate the modeling and representation of real-world clinical trial concepts, entities, and relationships. Meaning (semantics) becomes integral to the data itself, which includes code lists, terminology, and metadata - all intimately connected with results data. When validation rules are employed on top of this data, the result is high quality, valid data for submissions and use within organizations.

Linked Data also addresses the shortcomings of the antiquated V5 SAS Transport Format (4), 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. Use of RDF is a paradigm shift from the SDTM as SAS XPT. How do we get there from where we are now?

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

The CTDasRDF project was initiated at the PhUSE CSS conference in Silver Spring Maryland on 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 (3). 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 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 (5). 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. Future implementation 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 CTDasRDF 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.xml document and a supporting ontology will also be delivered by the conclusion of the project, scheduled for the March 2018 PhUSE 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 an earlier PhUSE project CDISC to RDF, 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 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 for validation.
- Separation of the results data from the standards data and metadata, resulting in a version-free graph data structure for clinical trial results.
- 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 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. 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 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 concepts present in the pilot study demographics (DM) and vital signs (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 (Reference Information Model) concepts when necessary (e.g. Activities, Entities).

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. The R package "rrdf" (https://github.com/egonw/rrdf) was chosen for its intuitive approach to RDF triple creation and ease of querying both TTL files and triplestores. The "redland" CRAN package (https://cran.r-project.org/web/packages/redland/index.html) is a viable alternative. A series of R scripts read in the source XPT, massage the data as needed, map it to ontological concepts, then create RDF files in Turtle (.ttl). TTL files may be uploaded into a triplestore or consumed by other applications.

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. 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 blank case report forms (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 Study "Mini Ontology"

The first step was to create a study “Mini-Ontology” using Web Ontology Language (OWL). We chose the concept of a “mini” 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, but 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” we mean these variables have standard definitions and derivations across studies so that their derivation can be expressed in a machine-readable expression using SPIN (SPARQL Inference Notation), 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. It treats a study as a collection of Activities that are performed on Study Participants (i.e. HumanStudySubject). Participants may be afflicted by one or more Medical Conditions. It also recognizes that studies contain different types of activities: administrative activities (e.g. obtain informed consent, randomization), Interventions (e.g. product administration, surgery), Observations, Analyses. It further recognizes that all Activities have Outcomes, 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 input to generate useful analysis results. 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 (i.e. resolves to “true”) or is not met (resolves to “false”). The core mini-ontology therefore has the following class structure:

- Activity
  - Observation
  - Analysis
    - Rule
- Entity
  - HumanStudySubject
  - Medical Condition

A more detailed concept map is shown in **Figure 1**. 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.

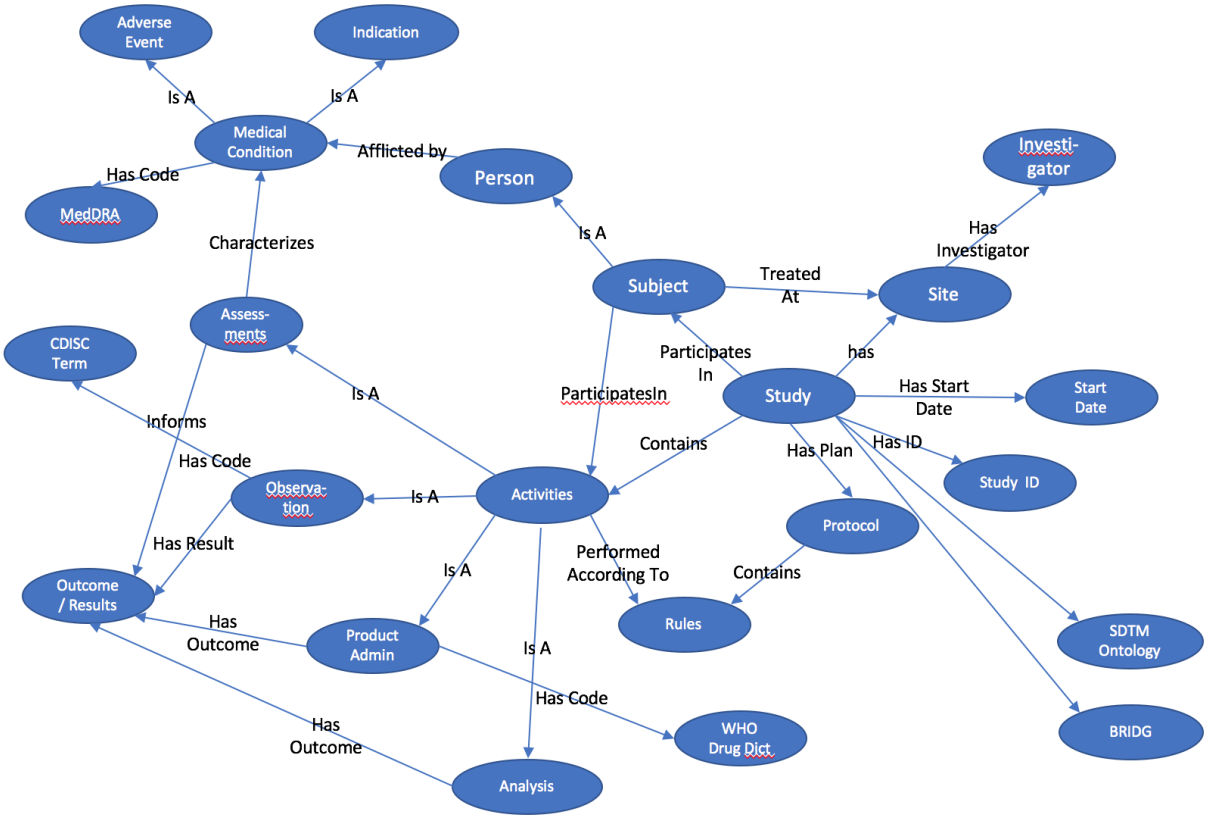

Figure 1 Minimal Study Ontology

BRINGING DATA TOGETHER

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

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 National Cancer Institute 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 below.

1. **code.ttl** – this file contains or links to resources representing defined concepts such as controlled terminologies. It includes Defined Activities. It currently provides links to SDTM terminology and the 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. 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 may 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 this file will also 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 case report form 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.
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 2 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.

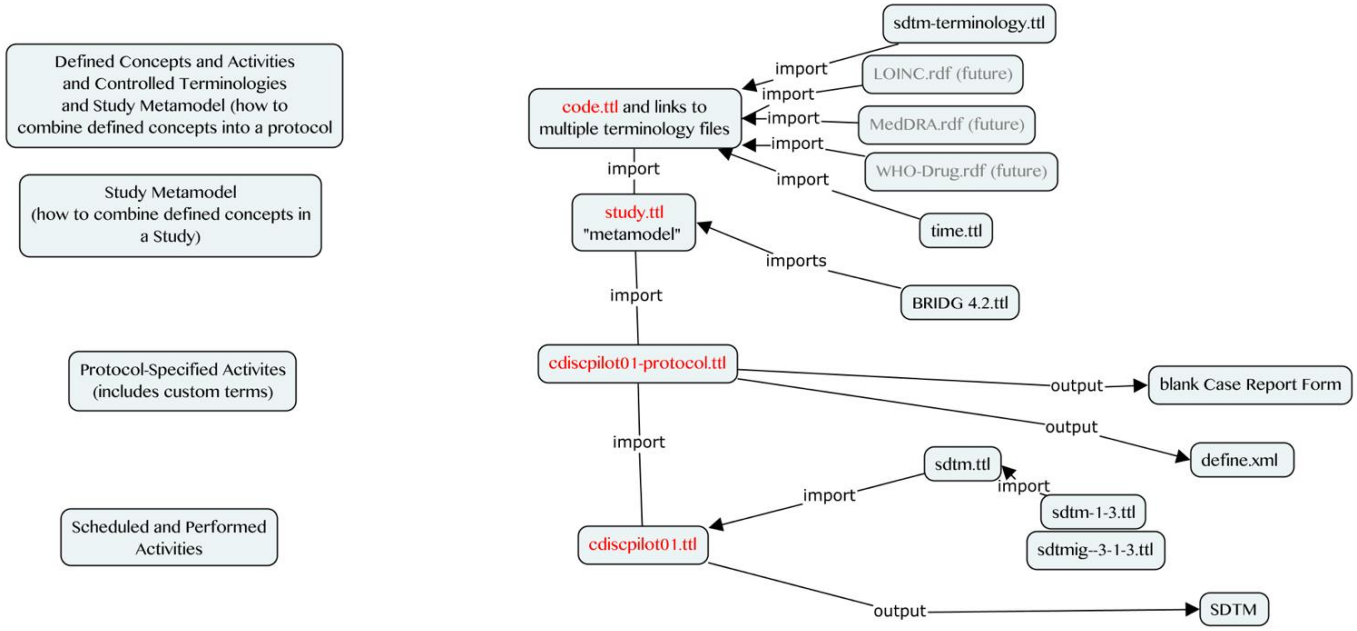

Figure 2 Importing Existing Data and Ontologies

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. 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 they describe, have inaccuracies, and are costly to produce and maintain.

CREATING DEFINE

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 intend to demonstrate 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.

There is a substantial disconnect between the data and supporting metadata when the two are not stored together[1], 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 while decreasing time and effort.

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, just add more nodes to the graph.

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 (4). Powerful mapping constructs like `owl:equivalentClass` and `owl:sameAs` facilitate compatibility with legacy data or other standards. The CTDasRDF project provides such a stepping stone for compatibility with CDISC and other standards like HL7 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 but one 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.

It is foreseeable that in the future, companies could provide 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 for secure delivery of 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 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

1. *Technical Rejection Criteria for Study Data - Preliminary Findings.* **Allard, Crystal.** Boston : PhUSE Single Day Event (SDE), 2017.
2. **CDISC.** About CDISC. *CDISC Website.* [Online] [Cited: 05 01, 2017.] https://www.cdisc.org/about.
3. *State of the Union: The Crossroads of CDISC Standards and SAS' Supporting Role.* **Decker, Chris.** Las Vegas, Nevada : SAS Global Forum 2011, 2011.
4. **PhUSE Emerging Trends and Technologies.** *Transport for the Next Generation.* s.l. : PhUSE, 2017.
5. *Managing Study Workflow Using the Resource Description Framework (RDF).* **Oliva, Armando.** Endinburgh : PhUSE Annual Conference, 2017.

ACKNOWLEDGEMENTS

The authors are indebted to the "CTDasRDF" PhUSE project team members, the R Project, Egon Willighagen for rrdf, Frederik Malfait, and 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 authors at:

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

[1] "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".