Paper TT01

Using Graph Technology for Metadata Driven Clinical Development

Frederik Malfait, Nurocor, Austin, USA

ABSTRACT

Data standards driven architectures are taking a more pronounced and center stage role in establishing end-to-end clinical development processes. The information architecture required to achieve this is facing increasing demands to express high numbers of relationships and to interconnect many different sources of information. A graph based approach addresses both challenges. We explain the basic principles of graph technology and demonstrate how this technology can be applied to design an information architecture that covers domain models, API endpoints, and data federation. We highlight the role of the Resource Description Framework (RDF), linked data principles, graph databases, and GraphQL based API design. We demonstrate the application of graph-based principles, in particular the use of RDF as the foundation of the CDISC Library, the use of RDF and ISO 11179 as the foundation of metadata registries, and the use of GraphQL and ArangoDB for a federated data layer in a microservice architecture.

INTRODUCTION

The use and application of data standards for the collection, tabulation, analysis, and submission of clinical trial data have gained increased importance over the last decade, in particular with the mandated use of CDISC data standards by the FDA. However, the evolution from paper based data standards to digital data standards has generally been slow within the pharmaceutical industry, primarily because of challenges related to the development of sound metadata models for clinical trial standards and the implementation of such models in enterprise grade software systems. Earlier work by the author in this area includes the design of a data standards metadata model for a large pharma using the Resource Description Framework (RDF) alongside the implementation of a metadata registry (MDR) based on linked data principles (2011-2014), the initiation of the Semantic Technology working group within PhUSE, and the development of the CDISC Standards in RDF v 1.0, published by CDISC in 2015. This work was followed in 2016-2018 by the design and implementation of a cloud-based platform for the CDISC Library, including an RDF based metamodel, a configurable hypermedia driven REST API, and an easy to use data standards browser. Current work extends into several directions. From a domain perspective, the scope of the metadata driven method now spans the full life cycle from clinical development plan and digital protocol to submission. From a systems perspective, an enterprise grade platform is being developed based on a modular microservices framework with an API first approach built on a GraphQL backbone.

GRAPH BASED METADATA MODELS FOR CLINICAL DATA STANDARDS

Until recently, CDISC clinical data standards have largely been available in paper form only. Transitioning to a digital format poses several challenges. One challenge is related to the organization of the material; in the case of CDSIC, data standards are organized by life cycle stage (data collection, data tabulation, data analysis), with separate publications covering therapeutic areas and indications with different degrees of completeness. This results in a silo based approach to data standards with considerable loss of continuity and consistency. The CDSIC 360 project is an attempt to address this issue by using a more conceptual and unified approach across the current silos. Other challenges concern the nature of the information. Clinical trial data standards define metadata about clinical trial data, rather than the data itself, which results in relatively small volumes of information, but with a high degree of connectedness between information elements. In addition, versioning of metadata and sponsor extensions play an important role in the usability of the standards. Taken together, it turns out that a representation of data standards is more easily expressed by a graph of information elements rather than by a set of relational tables. The justification of this is twofold. A large degree of connectedness is naturally expressed as links between nodes in a graph. In addition, a graph based approach accommodates more easily for new information to be added or integrated. Just expand existing graphs with additional nodes and links, or create new links between nodes in different graphs. Such changes are usually harder to achieve in a relational model because relationships are less granular (row based rather than element based) and can break more easily when extending the schema.

The following example illustrates the connectedness and the graph based approach as the natural representation of the CDISC data standards.

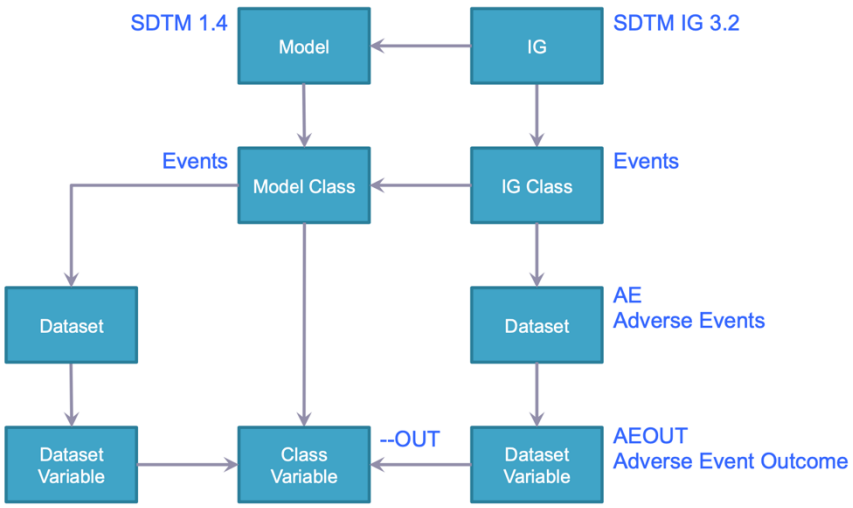

Figure 1. Relationships between SDTM and SDTMIG Elements

The degree of connectedness increases further when we take versioning into account.

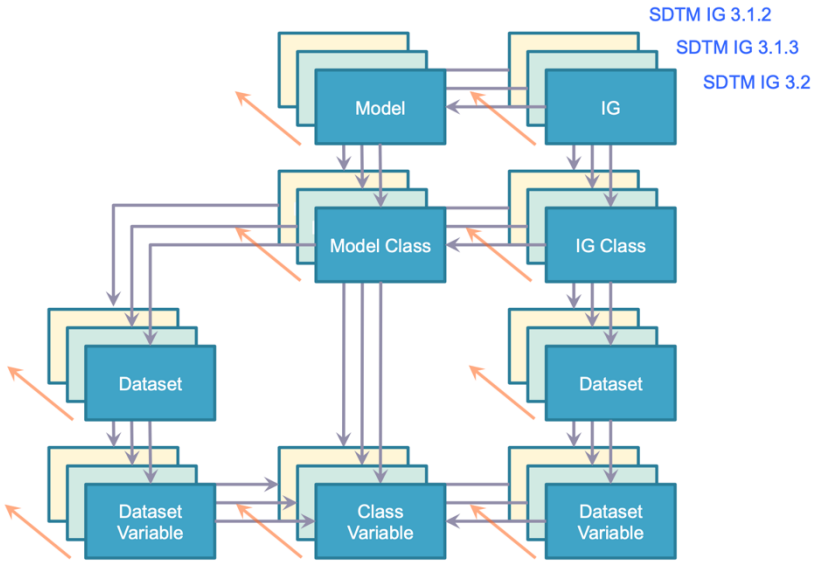

Figure 2. Relationships between Versions

USING STANDARDS TO DESCRIBE STANDARDS

If clinical trial data is subject to data standards expressed by metadata models, then it makes sense to also use standards for writing those metadata models. The leading formal standard for linked data is the Resource Description Framework (RDF), the cornerstone of a set of semantic technology standards published by the W3C. The RDF standard defines a mathematically precise formal language, based on first-order predicate logic and model theory, together with a graph query language (SPARQL). Information is expressed as statements consisting of a subject, predicate, and object (called a triple). Subjects and objects represent the nodes of a graph; predicates represent directed links between nodes in the graph. Subjects are resources in the graph about one can make statements. Objects can be either resources or literal values. Predicates are also considered resources; statements about predicates define information about the relationships between resources, and hence define the schema of the model. This means that both the information and the schema about the information are both expressed in RDF. Finally, each resource is uniquely identified by a Uniform Resource Identifier (URI). When used as a network location (URL), it is

possible to publish RDF resources on the web as true hypermedia linked data, thereby providing seamless API access to the RDF graph.

This has been the approach taken to design the content of the CDISC Library as a set of RDF graphs and to expose that information through the CDISC Library API, which is a hypermedia linked data API that allows navigation of the underlying RDF data. The RDF schema for the CDISC Library has been guided by the ISO 11179 standard for metadata registries (MDR), resulting in the following architecture of the CDISC Library model.

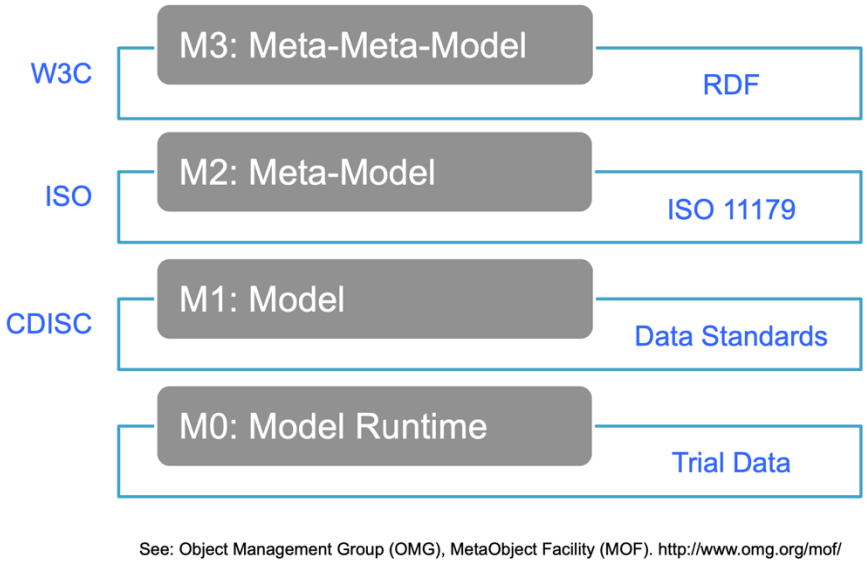

Figure 3. Metamodel Structure of the CDISC Library

The architecture fits within the 4-level metamodel approach promoted by the Object Management Group (OMG). The model runtime represents data -- in this case clinical trial data -- constrained by a model that is defined by the CDISC data standards. The structure of the model is described by a meta-model and implements the principles of ISO 11179. The meta-model itself and its instantiation expressing the data standards are both written in RDF. Level M3 closes out the hierarchy since RDF itself is written in RDF, i.e. RDF bootstraps itself. Levels M1, M2, M3 all live in the RDF world and are deployed on a single cloud based platform. One could make the logical argument to also express M0 in RDF. It would eliminate many problems currently faced with clinical trial data by eliminating silos, having built-in traceability, improving semantics, defining uniform exchange formats, and making it easy to pool data, but for obvious reasons there is a huge barrier to entry to execute this at scale.

EXAMPLE

Figure 4 shows an example of CDISC standards content expressed as part of an RDF graph, which illustrates the connectedness of the metadata.

Each version of an SDTMIG class, dataset, and variable is represented by an RDF resource. Resources are connected to represent the class, dataset, variable hierarchy, but also to connect the different versions of the classes, datasets, and variables. Variables subject to controlled terminology are linked to their code lists and each code list links to the terms it contains. Different versions of code lists and terms are linked across the terminology packages. Not shown in the graph are further nodes to represent the resource properties such as name and label.

All the content of the CDISC Library is expressed this way in RDF. This content is then exposed through the CDISC Library API, which typically exposes an API endpoint for each available RDF resource. Thus the API design maps consistently to the RDF graph structure. Following the hypermedia links exposed by the API corresponds to navigating the RDF graph.

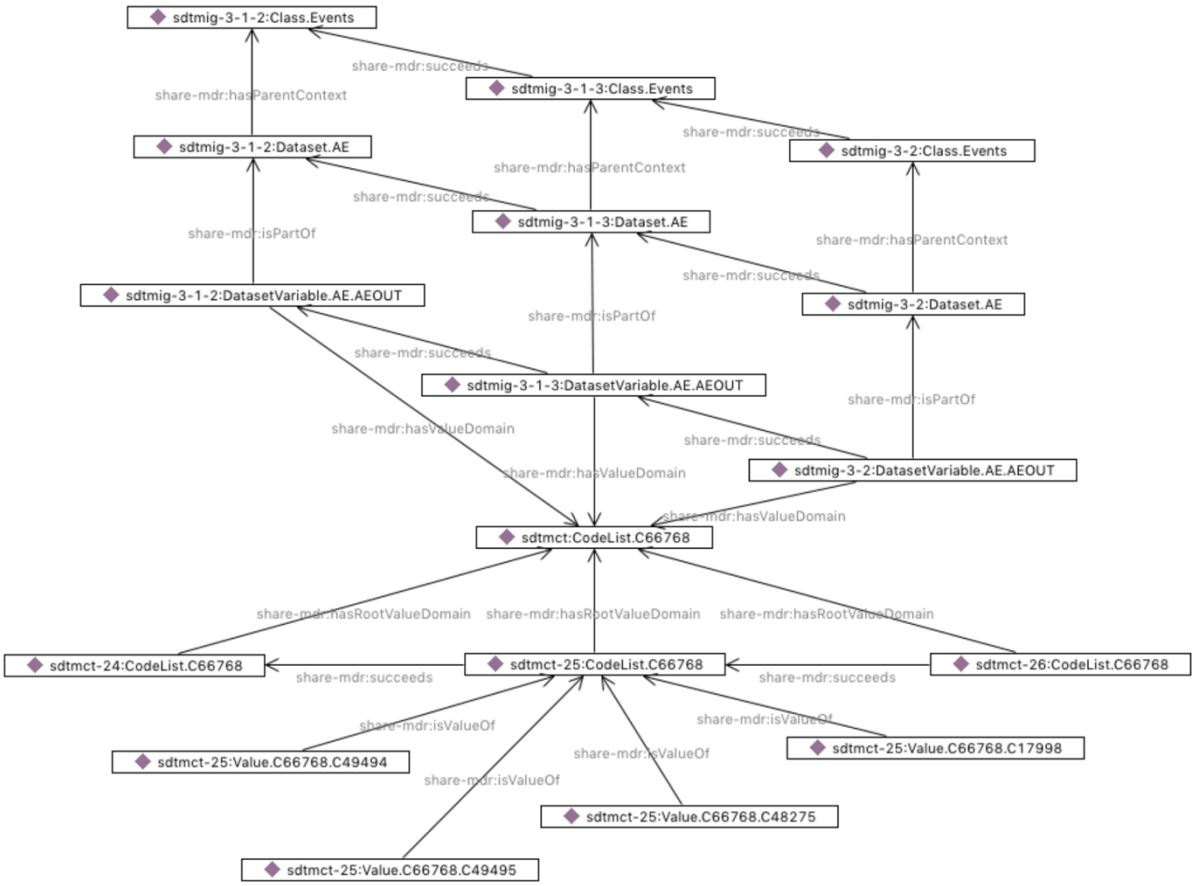

Figure 4. RDF Representation of SDTM Standards

EXTENDING THE SCOPE OF METADATA DRIVEN CLINICAL DEVELOPMENT

So far, we have mostly looked at clinical trial data standards from a standards development organization (SDO) perspective. From a sponsor perspective, there are some important additional issues to consider, such as the need for a more sophisticated versioning mechanism, the extension of data standards by sponsor specific content, and the provisioning of a more elaborate application infrastructure to support curators and users of data standards within the organization. From a technology approach, the goal then can be described to make clinical trial data standards available in a digital format, preferably in a graph based representation, integrated with sponsor defined extensions, implemented and deployed within a metadata registry, and accessible by easy to use interfaces (API, web UI). Although challenging in its own right and despite the significant need for further progress, this is less of a technology issue, but rather requires improvements on the standards development and management side, in particular a better approach to consistent end-to-end standards and the establishment of more robust governance and workflow processes.

Looking beyond the curation and compliance aspects of data standards, there are also gains to be made from the application of metadata driven process automation. The usual examples in this area are the automation of the Electronic Data Capture (EDC) build and the data transformation task from collected to tabulated data, both driven by extended metadata attributes on the data collection and data tabulation standards. Point solutions have been developed in these areas, but they fall short of a more comprehensive solution due to a lack of a more holistic approach that should take the nature of protocol driven research into account.

PROTOCOL DRIVEN RESEARCH

The planning and execution of clinical trials is by definition protocol driven. The current processes of protocol driven research suffer from a number of serious limitations that block a comprehensive and consistent metadata driven approach.

- Protocols are written as text documents, so that essential protocol information is not available in a structured and digital format. Protocol information is therefore not actionable for automation of downstream processes.
- Unstructured documents may be incomplete, open for interpretation, and harder to review.
- Protocol information is repeated across multiple documents.
- Protocols do not contain sufficient detail to provision downstream processes and systems.
- Protocols are insufficiently based on metadata and protocol standards.

The following schema illustrates the notion of a standards driven digital protocol, supported by a clinical development plan, and actionable for the automation of downstream processes that are part of study execution.

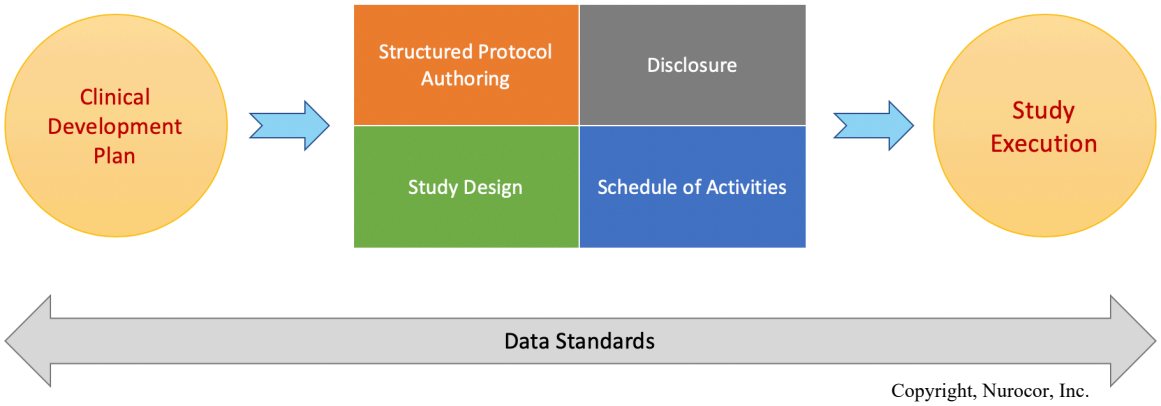

Figure 5. Standards Driven Digital Protocol Development

DIGITAL PROTOCOL DRIVEN RESEARCH

To enable digitized protocol driven research, one can break the protocol and supporting information down into four main application components.

- **Study Design** covering the structural definition of the trial design, including trial elements, epochs, arms, branching rules, and study schematics.
- **Schedule of Activities** where activities are fully governed by data standards. Activities can be decomposed into item groups and items (data elements), and where applicable code list terms. Item definitions can be based on current data standard models or future models based on biomedical concepts as currently explored in the CDISC 360 project. Once the schedule is defined, all the detailed information needed for automated study build and data transformations are known.
- **Disclosure and structured protocol elements** cover all the data elements from a protocol that can be arranged in a proper protocol model, including all the data elements needed for disclosure and trial registration. An extensive analysis of available protocol standards and requirements for disclosure shows that there is currently no single industry standard that covers all the required data elements.
- **Structured Protocol Authoring** in a multi-user cloud based authoring application, driven by a library of protocol templates (such as TransCelerate's Common Protocol Template) and backed by reusable component model for structured protocol elements.

The vision for digital protocol driven research is to enable a Lean Protocol ™ process, following principles of lean process management.

INFRASTRUCTURE FOR DIGITAL PROTOCOL DRIVEN RESEARCH

Graph based technology still plays a role in architecting the application infrastructure, e.g. for defining a data model for a schedule of activities or the structured protocol elements, but it is not the leading design factor for creating a robust infrastructure that can support digital protocol driven research. The reason for this is that the required infrastructure must support a wide array of application components, each implemented with fit-for purpose technology, and must provide integration points across the existing system landscape for each sponsor.

The recommended approach for building out such an infrastructure is based on a microservices architecture, e.g. running in a cloud based Kubernetes backplane. Each microservice can run in its own Docker container and use an application stack best suited for that application. Common services such as authentication, request handling, workflow, notification, audit, and logging are available to all application components. A first version of a Schedule of Activities app has been developed as a microservice running in such a platform, and work for a full Study Designer, Disclosure, and Protocol Authoring apps has been started. We don't go into further detail about the architecture here, except for the following two areas that touch on graph-based technology.

GRAPHQL

Any robust microservices platform should be designed from the ground up from "API First" principles. This is reflected in the use of an API Gateway to handle all incoming requests, but more interestingly the API technology is built for GraphQL first and REST second. GraphQL is strictly speaking not a graph based technology (as its name would suggest), but it rather originated as an API technology at Facebook to federate data that is distributed across shards that make up the Facebook data graph. As such it is both an API and a data federation technology that provides uniform API access to all data across all services.

ARANGODB

The data underlying the Schedule of Activities is fairly complex and highly connected following in part the domain model from BRIDG (defined activity, study activity, planned activity), but with additional drill down capabilities from activities to item groups to items to terms as modeled in the metadata registry, and additional relationships from objectives to endpoints to activities. The data is accessed from a microservice hosting an ArangoDB data store. ArangoDB implements a hybrid data model; it is a multi-model database for graph, document, and key/value based data.

ACKNOWLEDGMENTS

The author acknowledges all the hard work by the development and professional services teams at Nurocor for creating the vision and the platform to support digital protocol driven research.

CONTACT INFORMATION

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

Frederik Malfait
Nurocor, Inc.
100 Congress Avenue
STE 2000
Austin, TX 78701
email: frederik.malfait@nurocor.com
web: nurocor.com