Paper ML06

Breaking down the data standards silos – a new metadata ontology for data analysis derivations

Chris Price, F. Hoffman-La Roche, UK
Nelia Lasierra, F. Hoffman-La Roche, Switzerland
Didier Clement, F. Hoffman-La Roche, Switzerland

ABSTRACT

In the last years, the FAIR principles have re-focused attention on the importance of data standards and formal vocabularies. Traditionally, these have existed in individual silos which has been a significant roadblock to end-to end automation efforts. Developing connectivity between these is essential to facilitate automation in data transformation and pooling activities.

At Roche, data standards and vocabularies are described using ontologies and knowledge graphs and stored in an RDF-based metadata repository that enables their extraction in a machine-readable format.

We conducted an experiment to evaluate the feasibility of extending our current ontologies describing SDTM and ADaM and to connect them via derivations to answer key questions related to the optimization of analysis, pooling, and curation processes. The success of this experiment led to an ongoing project focused on leveraging connected and machine-readable standards to facilitate the design of analysis dataset specifications and to address downstream challenges.

1 INTRODUCTION

The analysis dataset specification is a fundamental document for a clinical data scientist preparing to analyse data. It describes the dataset structure and associated variable metadata and the algorithms used to derive the content. It is a critical document that is written prior to the start of programming and evolves over the course of the study lifecycle. It provides the instructions for a clinical data scientist on how to create each analysis dataset as well as providing information for users and reviewers of the analysis datasets on how they were created.

The Global Data Standards Repository (GDSR) contains a set of standard specifications analysis datasets that are used as the baseline for the Study team to develop the analysis specifications. These are made available to data scientists via a spreadsheet which is then updated for each study by: Removing content that is not required; selecting the applicable option where more than one derivation is available for an individual analysis concept; and adding new non-standard content. While steps have been taken over previous years to make the contents of the study analysis dataset specification machine readable, the process of developing the content and representing them in define.xml as a final product presents a number of challenges:

GENERATION:

- **Complexity** - The writing of study specifications is a manual process requiring significant attention to detail with the accompanying risk of errors.
- **Code relationship** - The current standard specifications do not currently contain details of how these are implemented in standard code and consequently the specifications are used as a guideline for programming as opposed to an input. While previously standard code was developed at a dataset level the move to use Admiral [1], with its more granular approach to the development of functions, requires a more in depth understanding of the available standard code by users.
- **Multiple use cases** - The analysis dataset specifications have multiple different uses and audiences. These uses have varying requirements and place significant stress on the document. The analysis dataset specifications routinely need to be both functional and technical specifications as well as both human and machine readable to reduce the effort required in their development.

INTEGRATION, RE-USE AND TRACEABILITY:

- **Harmonisation** - To support activities, such as pooling and the development of reusable code, harmonisation across a molecule is desirable. Current practice of using existing study specifications as a source creates significant challenges in maintaining alignment as commonalities need to be managed at a higher level.
- **Findability** - Querying and finding implemented content across studies is not straight forward as there is no cross-molecule repository of draft or final study specifications. Additionally, while the final product (define.xml) is machine readable this is only routinely generated where the study data is being submitted to a health authority. This means that many study specifications do not have a findable, fully machine-readable version.
- **Traceability** - Understanding derivations and their provenance through the specifications currently needs to be done manually by reading individual algorithms. As the different metadata across the data flow is not connected it can be challenging to quickly assess the impact of changes in the source data on the analysis and what data is required for any individual analysis.

2 METHODS

2.1 FAIR AS A SOLUTION

Adoption of the FAIR guiding principles (published initially in Wilkinson MD et al [2]) is a recognized strategy to enhance data re-usability, findability, harmonisation, and integration. By describing our datasets according to these principles, machines can understand the data and help us to find, connect and re-use these data assets and therefore automation can be introduced for these processes. In order to address the challenges exposed before with the analysis dataset specifications and to optimise the design and delivery of ADaM datasets, we decided to explore this approach and represent the analysis dataset specifications as machine readable metadata. By having these specifications as FAIR machine-readable metadata, inheritance, and re-use of derivations across studies within a project can be facilitated (harmonisation), dependencies between datasets can be managed by an application (complexity), understanding of the specifications can be enhanced for multiple use cases, comparison of analysis dataset specifications is enabled, and findability is enhanced.

The FAIR principles have re-focused attention on the importance of data standards and formal vocabularies. Having standards and formal terminologies enhances mutual understanding and aids integration. Even though SDTM and ADaM standards are in place, for the FAIRification of the analysis dataset specifications we designed an ontology that bridges these two standards to enable end to end automation. Enabling different paths of connectivity between the derivation and the source, was our first step towards addressing the challenges that the creation of the analysis dataset specification presents today. Metadata describing the analysis specification has been then represented in the form of RDF (using unique and persistent identifiers) and planned access to it through standard APIs.

2.2 USE CASES

We approached these challenges and the development of the solution in several steps. The initial steps aimed at both assessing the feasibility of the design of the needed ontology and to gain feedback from the end users on whether the content queried would provide a business benefit. For this activity, we defined two scenarios. For each scenario we established the business objective, clearly identified the questions that we were seeking to answer and defined the expected business benefit. These were fed into the development of the ontology and population of the knowledge graph iteratively along with feedback from data scientists until it met these requirements.

SCENARIO 1: GENERATION OF ANALYSIS DATASETS

As described above, there are a number of challenges in developing and maintaining the content of analysis dataset specifications. This first scenario of assisting in the generation of analysis dataset

specifications sought to address the challenges exposed in the introduction and bring the following business value:

- All content available as machine-readable to provide the opportunity to automatically generate human and machine readable renderings of specifications (e.g., define.xml).
- Improve the data and metadata alignment to support health authority submission related processes.
- Improved relationships between specification contents to reduce manual input.
- Improved support of molecule consistency by the enablement of inheritance from molecule-level specifications as opposed to copying of content.
- Increased efficiency of data pooling activities by efficient identification of differences between individual studies.

Competency Questions

- What is the algorithm/choice of algorithm for an individual derivation?
- What are the ADaM variables and where necessary the ADaM Parameters that are created for this derivation?
- What derivations are prerequisites in order to perform another derivation?
- What source data (SDTM) is required for this derivation?
- What functions do I need to generate this dataset?

SCENARIO 2: IDENTIFYING TARGETS FOR DATA CURATION AND METADATA TRACEABILITY

This second scenario of assisting in the identification of targets for data curation to be used in secondary data analysis and metadata traceability sought to address the challenges identified previously to bring the following business value:

- Enable querying and identification of implemented content across studies.
- Improved consistency of specifications across both a molecule and therapeutic area.
- Introduce the automated ability to identify specific datasets and variables that need to be curated in order to perform a derivation to target curation efforts where they bring the most value.
- Improved capability to conduct beginning-to-end traceability for analyses.

Competency Questions

- What is the data (SDTM variables including identifiers) that impact the calculation of a derivation?
- What is the algorithm/choice of algorithm for an individual derivation?
- What derivations (and consequently endpoints) were created for a clinical study?
- What derivations could be created (based on the available source data) for a clinical study?

2.3 INITIAL STEPS

To start our FAIRification process of the analysis dataset specification we defined the so called “Experiment 0” which was focused on the design of an ontology that would facilitate answering the above exposed questions and its publication in an internal platform (GDSR) in order to gather feedback from the Data Scientists in our organisation. Our experiment consisted of the following steps.

TEAM SET UP

We formed a multidisciplinary team of information architects with expertise in semantic technologies and knowledge of CDISC, and domain experts with knowledge in both the analysis dataset specification creation process and the CDISC standards. In addition, additional support in SPARQL (SPARQL Protocol and RDF Query Language) [3] query design and feedback from those with domain knowledge was obtained during this experiment.

ONTOLOGY DEVELOPMENT AND STANDARD KNOWLEDGE GRAPH

We first designed the ontology that would explicitly connect SDTM variables and ADaM variables via derivations that was optimised to answer three key competency questions that were prioritised:

- What are all the SDTM variables that impact an ADaM derivation?
- What are the Admiral functions needed for generating X derivation?
- What are all ADaM variables and parameters required to derive and populate a derivation?

After an initial draft of our ontology, we evaluated its completeness and expressivity by populating the ontology for two derivations. These two derivations enabled us to cover a wide range of possibilities and led to minor modifications to the design of the ontology.

- Overall Survival
- Diastolic Blood Pressure Change from Baseline

PUBLICATION IN LYNX

The last step of our experiment was to publish the created ontology and knowledge graph in the GDSR and enable access to the Data Scientist through the Lynx application [4]. Lynx is an application that enables the integration of reference metadata across Roche and enables the registration of SPARQL queries in the form of a catalogue that is exposed to the end-users using natural language. Through Lynx, data scientists can explore the graphs using the full power of SPARQL queries while navigating through the data in a simpler manner by using business language. We registered the corresponding SPARQL queries for the three target competency questions.

3 RESULTS

The existing ontology in the Roche GDSR supporting analysis datasets describe the variables and associated metadata developed by CDISC as part of the Analysis Data Model (ADaM) [5] including the Subject Level Analysis Dataset (ADSL), the Basic Data Structure (BDS) and the Occurrence Data Structure (OCCDS) models. These are then extended by a linked ontology describing the implementation of these variables in individual datasets, as well as textual descriptions of each derivation. Where multiple derivations are possible for a single concept (e.g., deriving the analysis age using months or years) these are included individually and linked to the same variable. There is no explicit machine-readable link between the ontologies that describe the analysis datasets with those that describe the tabulated datasets (i.e., SDTM).

To address the competency questions defined previously it was necessary to develop an additional new ontology that would support the requirement to link the source and output, as well as the implementation of each derivation. In designing this new ontology, it was critical to ensure that we sought to reduce complexity to the minimum level to deliver on the requirements for the competency questions (i.e., we did not need to build a code generator) and that we connected to existing ontologies and other ongoing work relating to implementation (e.g., Admiral). Based on this the requirements for the ontology consisted of:

A. DEFINING THE INPUT FOR A DERIVATION

All possible inputs of a derivation needed to be considered, these included any data from SDTM domains including the dataset and variable where the data was stored, as well as any identifiers (e.g. --TESTCD) from the dataset to supported datapoint traceability. Additionally, previously derived analysis concepts may be used as a source for a derivation, so this relation was required to be defined

B. DEFINING THE OUTPUT OF A DERIVATION

The ADaM dataset and variable where the result of the derivation is stored was required. Additionally, both identifiers (i.e., PARAMCD) for BDS datasets and qualifiers (e.g., date imputation flags and units) were required to link related concepts together where they could not be separated.

C. SUPPORT OF MULTIPLE OPTIONS TO DERIVE THE SAME ANALYSIS CONCEPT

It is necessary in analysis to support and identify alternative derivations of the same concept. This may be due to the use of different input data (e.g., due different versions of SDTM as a source), alternative algorithms or alternative representations (e.g., use of different units).

D. DEFINING A LINK TO THE IMPLEMENTATION OF THE DERIVATION

A link was required to the Admiral function and the required arguments for those functions to enable simpler implementation of these tools by data scientists.

E. PROVISION OF PERSISTENT UNIQUE RESOURCE IDENTIFIERS (URIS) FOR DERIVATIONS

It is accepted that over time the structural data standards and controlled terminology will evolve. It is required that this evolution be supported while also providing an ability to track changes between different versions of the derivation.

With the new ontology (See figure 1) designed to support these competency questions, this enabled us to simplify the existing ontologies describing ADaM and to reduce the content, thus simplifying maintenance. From an ontology perspective we are now able to remove all derivation text as well as all parameter-based metadata used in ADaM datasets using BDS. From a content perspective we have removed all predecessor content as it will be possible in the future for these to be supported directly from the SDTM ontologies at a study level.

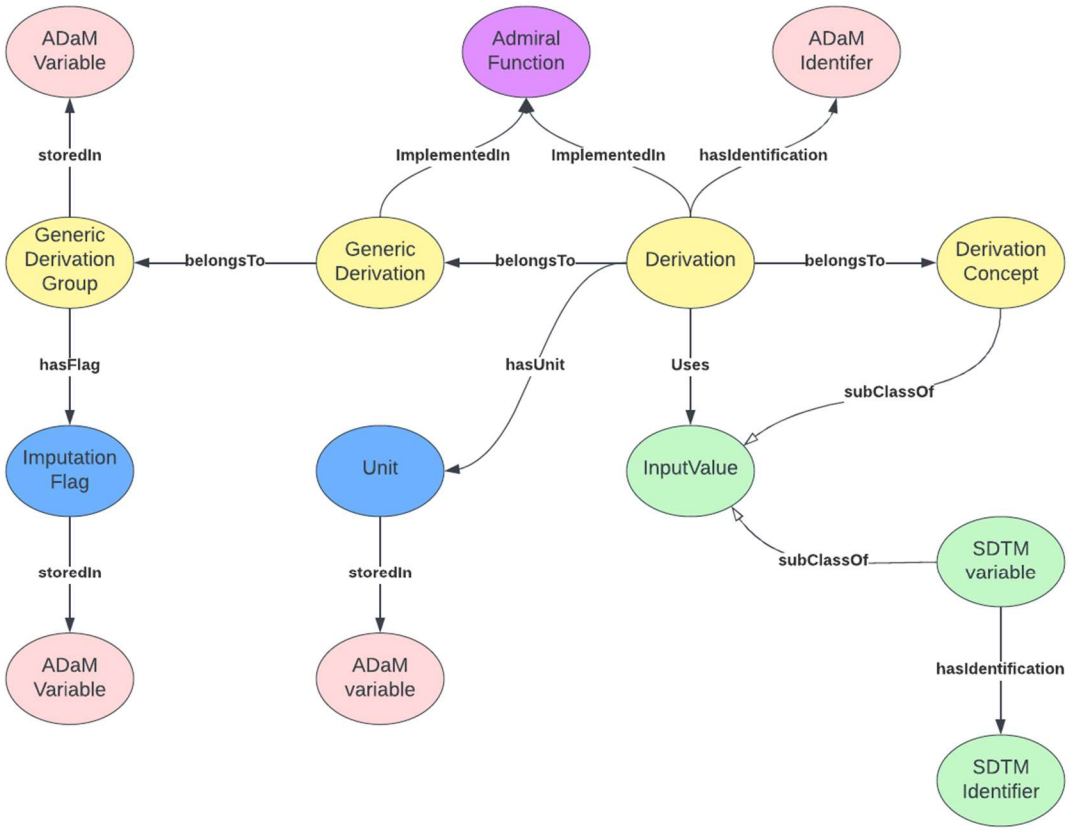

Figure 1: Simplified logical model developed for analysis derivations

4 EVALUATION AND DISCUSSION

With “Experiment 0” we demonstrated that the ontology that we have developed was sufficient to bring value to the end users as it was answering key questions they are facing when generating the analysis dataset specifications or when browsing its contents. As a continuation of this, the first step that we took was to extend the available knowledge graph by populating the set of variables and derivations needed to represent 80% of the ADSL and ADVS (vital signs) analysis datasets. This enabled us to continue to test and shape the ontology and share richer examples with the data scientists.

We gathered informal feedback from Data Scientists and conducted formal interviews with twelve of them working in distinct roles and sites within our organisation. During the interviews we analysed their current user journey for creating or using the analysis dataset specifications and obtained feedback on our vision and results of our experiment.

As a next step, we evaluated how to scale and bring the FAIRification of the analysis dataset specifications to life. Specifically, we developed a set of follow-up steps looking at:

A. Assess the feasibility of developing an engine capable of populating an analysis dataset specification as machine readable metadata using, as a baseline selected, standards (global or molecule). A Proof of Concept using GDSR technologies was developed to address this topic.
B. Evaluate the value for pooling and traceability. An R Shiny app that enabled comparison of the mock-up specifications was developed to provide a more graphical representation of already developed Lynx queries.
C. Assess the feasibility of migrating the current standard representation in GDSR to the new model

5 NEXT STEPS AND OUTLOOK

Following the successful demonstration that analysis derivations can be modelled in the ontology and subsequently queried, Roche is now exploring options on how this can be integrated into the business with the development of new applications to address the challenges identified above. This has been incorporated into the Roche wide FAIR strategy adoption efforts where we are aiming to make all data FAIR by design, moving away from the need to perform retrospective FAIRification of data. The Data Sciences aspects of this are being addressed through Project Botany. Project Botany is an ecosystem of tools born from grass-roots innovation with an aim to ensure we can efficiently and sustainably create high-quality Biomedical Data Assets that can be readily shared with others for generating actionable insights for patients. It is building practical modular automation solutions that can be embedded within our workflows, with each solution providing its own benefit.

The newly named Orchid Garden project will work with data scientists to seek to build two applications. One to enable end users to design and build their analysis dataset specifications and the second to browse and query analysis dataset specifications. The work to develop the design and build application is in progress with the intent being to create an application that supports users in creating an analysis dataset specification for both a molecule and an individual study. The basic requirements for these applications are that data scientists should be able perform the following tasks:

- Select content from the global standards and/or from previously defined molecule implementations of data standards.
- Build the study design elements (e.g., periods and visits) and associate them with the applicable analysis datasets).
- Design non-standard analysis dataset specifications based on the existing ADaM data structures.
- Store the molecule and/or study analysis specifications in a database.
- Identify derivations that are dependent on other derivations to ensure all dependencies are available.
- Generate the analysis datasets define.xml for a study.

While not currently under formal development at this stage the analysis dataset browser application envisaged to offer a number of functionalities that data scientists have identified the following as important to improve both the efficiency and the quality of their work:

- The ability to compare content selected and defined for a study against global standards as well as against other studies.
- The Admiral functions and arguments that are required for standard derivations.
- The SDTM variables and identifiers that are required to implement analysis derivations.

While the initial ontology has been defined there is significant work, some of which is already ongoing, to utilise the ontology we have defined within business processes. The experiments and the user interactions that have taken place have greatly influenced the ontology development. We believe that this will be the basis for the development of future applications that address the challenges faced by data scientists in developing analysis dataset specifications by providing an extensible and persistent link between existing data standards.

REFERENCES

[1] https://pharmaverse.org/

[2] Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018. 10.1038/sdata.2016.18 - DOI - PMC - PubMed

[3] https://www.w3.org/TR/sparql11-query/

[4] Lynx: A FAIR Knowledge Graph Engine for Reference Data Integration & Look up Service Javier D. Fernández and Nelia Lasierra, Semantics.cc 2021 (https://2021-eu.semantics.cc/lynx-fair-knowledge-graph-engine-reference-data-integration-look-service - presentation in industry track)

[5] https://www.cdisc.org/standards/foundational/adam

ACKNOWLEDGMENTS

The authors would like to acknowledge the following individuals who contributed to the development of this ontology and the ongoing work to utilise the ontology as part of the development of new applications: Selena Baset, Kathrin Flunkert, Robin Koeger, Magdalena Krochmal, Jennifer Lomax, Michal Pietrusinski, Rashad Rasul, Ivan Robinson, Javier D. Fernandez, Huw Mason, and the wider Orchid Garden team.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at
chris.price.cp1@roche.com, nelia.lasierra@roche.com and didier.clement@roche.com