

E2E data standards, the need for a new generation of metadata repositories

Isabelle de Zegher, PAREXEL® Informatics, Belgium
Alan Cantrell, PAREXEL®, United Kingdom
Julie James, PAREXEL® Informatics, United Kingdom

ABSTRACT

This paper describes the implementation of E2E data standards within PAREXEL. It first introduces the rationale and the vision for E2E data standards (from protocol to submission), to ensure compliance in a cost effective way. And then it focuses on the core component of this implementation, i.e. the development of a metadata repository and the related data standards governance framework.

The paper concentrates on the key functionalities needed for the metadata repository (MDR) we are deploying within PAREXEL and explains why we need more than CDISC data standards management in an MDR. And more particularly we explain

- How to author and maintain data standards content that ensures not only compliance to CDISC standards with data lineage between CDASH and SDTM, but also manages semantic enrichment implicit in the definition of SDTM and needed to be made explicit for (secondary) data analysis
- How to integrate such a metadata repository within operational reality, mainly at study set-up, through the definition and management of study instance metadata, as a machine readable form of parts of the protocol.

In parallel we will touch on the impact on processes, from definition of data standards to enforcement within data operations.

Introduction

Data management today is focused on providing high quality data at the end of the trial, for analysis and reporting. With the increasing demand for near real life data visualization and data surveillance, as well as the need for comparative data across trials to support personalized medicine and data mining within Big Data, data management is now being asked to:

- provide high quality data throughout the conduct of the trial;
- ensure that clinical trial data structures are consistent across trials; and
- ensure compliance with enterprise and SDTM data standards from data collection onwards (as FDA is mandating CDISC SDTM – with traceability to collected data – as of December 2016, and we expect Japanese authorities to start requiring SDTM as of July 2016).

This triple challenge can only be met through E2E implementation of data standards, from protocol to submission. This requires new approaches and new tools for metadata management.

Rationale and vision for E2E data standards: the need for a MDR

In many organizations, standards are being maintained in silos to support individual specific processes:

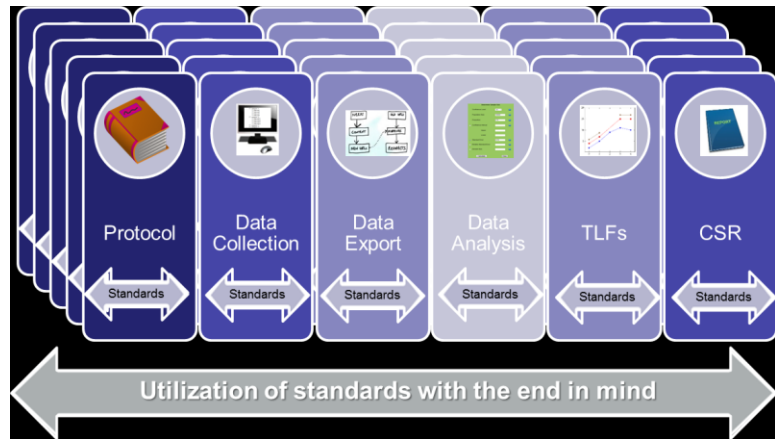
- data management maintain eCRF libraries with data collection standards/ CDASH to support data collection;
- clinical programmers focus on SDTM maintenance to ensure correct generation of SDTM; and
- statistical programmers focus on ADaM and TLFs to be included in the Clinical Study Report.

However, all these standards are linked: data collection/CDASH is the source of SDTM which in turn is the source of ADaM and TLFs. So it is critical that these standards are managed from the protocol onwards to ADaM and TLFs in a collaborative way, not only within a therapeutic area – or a franchise – but across the organization.

PhUSE 2015

E2E management of data standards brings several benefits:

- Increase data quality and efficiency for cross study reporting. Usage of similar standards across studies ease the production of safety reports (ISS, ISE) and any other cross study analytics reports; there is no need for mapping from each study to one common standard.
- Increase efficiency for study set-up. When data collection standards are properly governed, it is possible to implement libraries of eCRF forms (or equivalent for other data collection instruments) which can be validated and then re-used for each study, increasing speed and efficiency of study set-up.
- Increase efficiency for SDTM mapping, defining the SDTM mapping in tandem with the data collection standards means that the majority of the mapping is pre-defined and consistently re-useable. Full traceability from source to output. If an error is identified at any point, a root cause analysis, impact analysis and potential remediation activities are considerably simplified, if data collection, transformation and analysis are based on standards that are all mapped together.

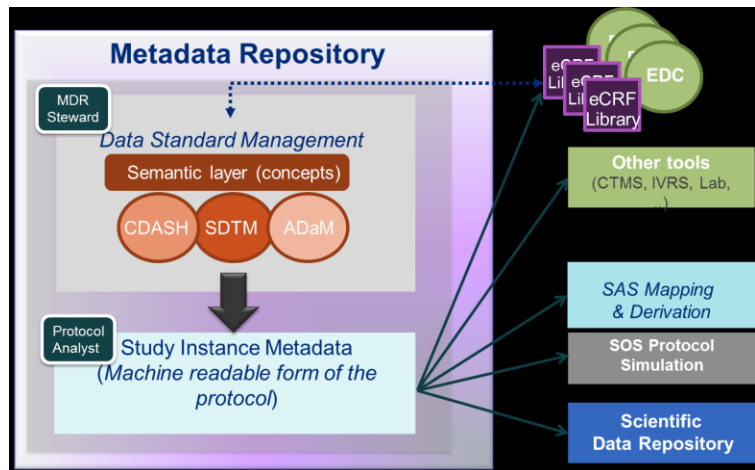


To support true E2E management of data standards, we need tools to support effective mapping, use and traceability of these data standards for data collection, to SDTM to ADaM. Following market analysis, in partnership with a software product company, we at PAREXEL decided to develop and deploy our own metadata repository,

Core functionalities of a semantic MDR

While refining our requirements for E2E data standards management, we decided that we needed two main sets of functionalities within an MDR

1. Authoring and maintenance of data standards – grouped together in a meaningful way through concepts to which the variables from data collection, SDTM and ADaM standards are mapped. Other important artefacts, such as generic eCRF forms specification or data transformation formulae are also defined as part of the data standards.
2. Use of the data standards for specific studies, with export of the relevant metadata for the specific study for use in downstream tools. This also includes descriptive metadata (trial design datasets) and the visit schedule. – This is the purpose of the Study Instance Metadata (SIM).



Concept for maintaining data standard content

At the core of any MDR is variables i.e.

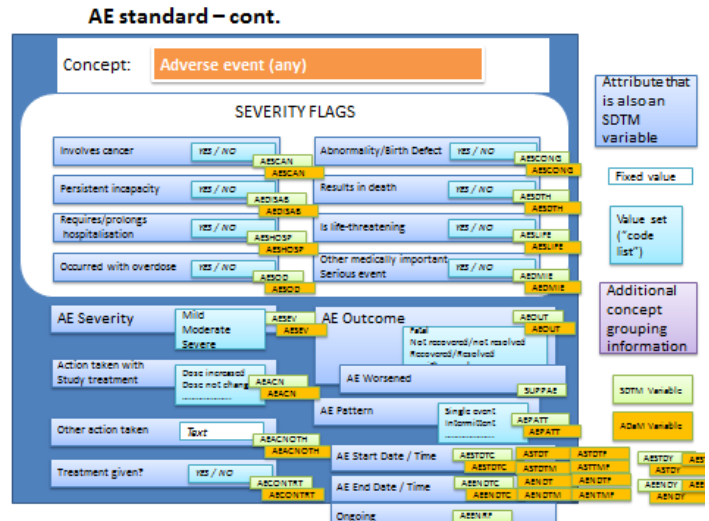
- Data collection (CDASH – linked to SDTM - or sponsor specific)
- Data output (SDTM or sponsor specific)
- Derived Data for Analysis & reporting (ADaM – linked to SDTM, or sponsor specific)

Management of standalone variables is not sufficient for correct management of E2E data standards; we need a mechanism to “glue” these variables together. Also, while CDISC SDTM is appropriate for data transfer to authorities, it is incomplete to support analysis & reporting and cross study analytics, as there is room for interpretation and many variables are optional.

There is a need to add a semantic layer on top of variables; this semantic layer is composed by concepts aka a group of - potentially ordered -variables or group of variables, that must be managed together to have meaning or semantic. For instance, a variable containing a value does not have any meaning without being linked to a variable

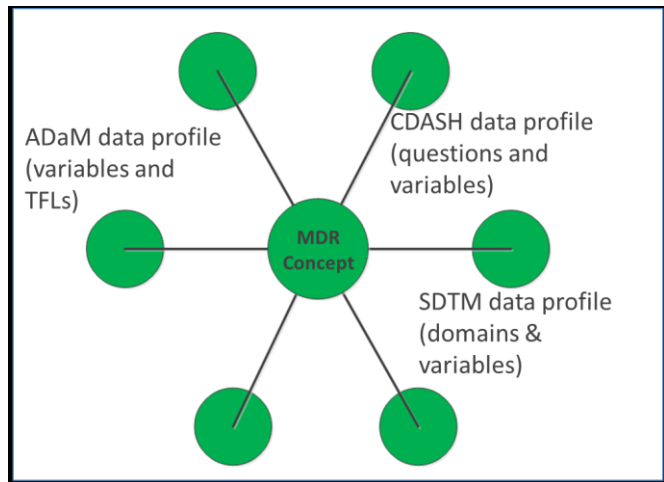
with a unit (ISO21090 example). A more complex example is the Na+ test that can be performed in urine or whole blood; the specimen information is often implicit in the protocol and not explicitly collected. The SDTM lab domain includes specimen specification as optional variables, and it is possible to deliver a compliant SDTM dataset with Na+ test, without specimen information. It is only at analysis and reporting time that the question is raised and the SDTM dataset completed.

The PAREXEL MDR supports management of an additional semantic layer of concepts to ensure all needed data – implicit or collected – are included as variables in the concepts and therefore in the underlying study dataset. Concepts are based on a hub & spoke model where the “hub” contains the concept definition, derived from ISO21090 and BRIDG, ensuring semantic consistency in these concepts grounded on industry standards. Each spoke is the standard representation of the concept, mapped to the concept defined within the hub.



There are several benefits in managing standards through concepts in a hub and spoke approach

- Concepts are more intuitive and easier to manage than variables. For instance, in a protocol we speak about an AE or demographics, represented by a whole set of variables, rather than each individual variable. In addition, we expect that around 200.000 variables may be needed to cover all therapeutic areas (there are 2 Millions in the exiting UK NHS Medical Record System); this cannot not managed efficiently by a normal human being without such additional / grouping construct.
- Concepts can be used with the data collection standards to define groups of variables to be collected together aka the eCRF forms specification for EDC or the variable groups for other data collection instrument
- Concepts are used to support mapping between each “spoke” in an efficient way. If another standard or another version of an existing standard needs to be integrated it is mapped through the “hub” – the mapping between the “hub” and the other spokes remain valid. Also the additional semantic included within a concept supports easier mapping between standards for which a variable to variable mapping is not possible (particularly in the analysis datasets).
- Concepts can and should be used to generate – in a semantically sound way – additional therapeutic area standards. Within PAREXEL we are currently testing the use of Concept templates, derived from the BRIDG model and stored within Enterprise Architect. Concept templates can support the generation of TA specific concepts by indicating which type of variable and/or terminology must be defined to generate a semantically consistent concept



Study Instance metadata for integration into operations

To ensure full value, the MDR needs to support study specifications by having a machine readable form of the relevant parts of the protocol. We call this Study-specific Instance Metadata (SIM); it includes

- The study design – based on SDTM Trial Data Models including all the parameters defined in TSPARM
- The list of forms – the data collection “spoke” described above potentially augmented for the study
- The visit schedule linking visit and forms as appropriate for the study.

The SIM is used to support EDC set-up but also other data collection instruments (as EDC covers between 30 to 70% of the data of a study). The SIM also supports other downstream systems such as Reporting including mapping and transformations and for Analysis, both of which are used in a Statistical Computing Environment. It can therefore be exported in different forms to serve these different purposes.

Deployment challenges

We are in final stages of development of the PAREXEL MDR; we are working in an AGILE approach with rolling acceptance which, in addition to the well-known benefits related to software development, ensures continuous involvement from the business team and is a great tool to support organization mobilization and change management.

Deployment of an MDR includes 2 main aspects:

1. Process and technology alignment with process update and change management.
 - The approach for change management are well known, including process updates/redesign, communication, training, potentially definition of new roles and adapted incentives.
 - However, what is particularly challenging in the context of an MDR deployment is the scope of change. It includes processes related to managing the standards themselves as well as, in the spirit of E2E support, processes related to study set-up, SDTM mapping and ADaM generation. It also includes potential organization changes to deploy/consolidate a data standards stewardship group and to define and develop new skills (see point 2 below).
 - In the context of a CRO like PAREXEL changes are even more complicated than for a sponsor; we estimate 50% of our customers would accept the use of PAREXEL defined standards but some sponsors will want to use their internally defined standards; so we need to take into account these two different operational paradigms.
2. Data standards content curation. While CDISC is providing CDASH, SDTM and ADaM with mapping, these standards are incomplete: there is no underlying formal concepts based on proven standards like BRIDG and ISO210190, and the standards do not cover all therapeutic areas'. To build the semantic layer needed to support E2E data standards we need to "curate" data standards by
 - Extracting/defining concept templates – as the basis of a hub concept– from BRIDG and ISO210190
 - Defining domain specific business concepts based on CDISC standards and from the emerging CFAST c-MAP concepts
 - Use the concept template to introduce the business concepts and generate formal concepts within the MDRWe have started Data Standards content curation within PAREXEL for the safety domain of CDISC as they are quite stable. We will be expanding on the efficacy domains with sponsors on an as needed basis, building on CSHARE whenever the "thick layer concept" will be delivered.

CONCLUSION

Big data analytics in clinical trials requires new ways to manage data in what we call Big Data Management; and this starts with managing data standards E2E from protocol to submission.

In this paper we described our approach to E2E data standards management and the development of an MDR to support definition and use of these standards within operations.

At the core of our MDR are the concepts, supporting semantically sound definition of standards and easier mapping across the different implementation levels, and different standard versions. We believe that this approach will help to build a solid set of integrated standards, which is critical for the data integration and data quality needed for Big Data Analytics, while supporting compliance. It would be of high interest to have other companies testing this hub and spoke concept approach and help develop faster and better set of standards to be managed at the industry level within CSHARE.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:

Dr. Isabelle de Zegher
VP, Integrated Solutions
PAREXEL Informatics, Belgium
T +32 2 767 16 48
M +32 478 48 28 54

Brand and product names are trademarks of their respective companies