PHUSE EU Connect 2019

Paper TT04

Deconstructing SDTM – starting again

Johannes Ulander, S-Cubed, Copenhagen, Denmark

ABSTRACT

The CDISC SDTM standard has been with us for many years and thousands of SDTM datasets, aCRF's and define-xml's are created and submitted every year. Since producing SDTM deliverables today is an everyday task the maturity and experience in the industry must have grown, but according to FDA there are quality issues and variability between submissions which makes the Clinical Study Data Reviewers Guide (cSDRG) the most important delivery, which is there to explain when we are not able to conform to the standard! So what is the problem with SDTM? You would expect quality to improve and cSDRG's to become smaller as we get better at adopting the standard, yet this does not seem to happen. This presentation will deconstruct SDTM using new technology, increase your understanding of the standard and tell you how to produce a valid SDTM dataset, aCRF and define-xml with the click of a button.

INTRODUCTION

For making a complete submission to FDA or PMDA you will need to create a number of standardized deliverables for them to be able to consume your data, because a data point itself does not have any meaning unless it is accompanied with other data points and especially the metadata that tells you what it is.

It is now 20 years since the CDISC organization was formed and the CDISC Submission Data Standards team (SDS) started meeting to work on what we now call the Study Data Tabulation Model (SDTM). This was a reaction to updated FDA eCTD guidance requiring that data listings, patient profiles, data tabulations, and analysis datasets should be submitted. As the first three in essence are different views on the same data, the thinking was that they could be replaced with a single model which would allow review tools to produce the views for the different use cases instead.

Since 2008 and the release of SDTM v.1.2 the model has been stable, which is also true for the accompanying SDTM Implementation Guide (SDTMIG) v3.1.2 for human clinical trials. The SDTMIG though has had some changes (e.g. adding or removing domains, moving coding variables from supplemental qualifiers to the parent domain) that has had quite big impact when creating systems that support the creation of coherent domains, but the general concepts and how you "think SDTM" as observations with a topic and qualifiers remain intact.

So why are people still struggling with implementing it as it has been stable for 10 years?

COMPLICATIONS

One reason for this is that SDTMIG has a flexible way to incorporate data not defined in any SDTMIG using custom domains and supplemental qualifiers. When something is not defined you can create conformant domains and variables but with the inherent risk that your implementation will risk being obsolete at some point in the future when the appropriate domain or variable is added. This is not so much of a problem when transforming a single study, but if you are building a tool to support the transformations this is going to create headache.

Another reason is that the SDTMIG is open for interpretation, which is has to be, to be able to support the different ways of how data is collected, at what level of detail it is collected and how the study is designed. Which is why FDA also wants a clinical Study Data Reviewers Guide written (cSDRG) to explain how you have interpreted the SDTMIG and where you are deviating from the standard.

This means that a submission contains a lot more than just data, it also has a large amount of metadata:

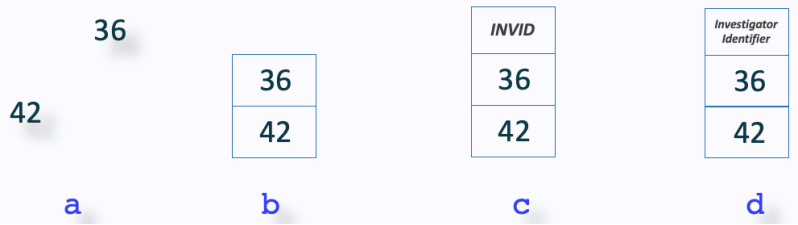
- Data+metadata: SDTM domains (datasets)
- Metadata: *annotated CRFs*, *define-xml* and *cSDRG*

All of the deliverables need to be glued together to create traceability, and e.g. the possible values need to be represented in aCRF and define-xml, whereas the data in the domains will only contain what has actually been collected. As the process of creating these deliverables usually happen after the study has started and from multiple different sources (e.g. eCRF, ePRO, dictionaries, lab transfers etc.) it is not an easy task make sure that the submission deliverables make sense and are under control. It usually creates a maintenance nightmare where the easiest way out is to write in the cSDRG why you cannot comply with the standard, rather than making the necessary change to actually comply.

To some extent I think the problem is that SDTM is misunderstood, and that our perception of data might be wrong.

IS THE PROBLEM IN OUR HEADS?

If you are presented with two numbers, e.g. 36 and 42, on a blank sheet of paper you would probably not directly identify them as datapoints. They would just be random numbers. (See *a* in the below picture.)



If the numbers were arranged directly above each other and a rectangle was added around each of them, you might suspect that there is something more to them (*b*). They seem to represent something, although you are not sure what. If a rectangle is added above them with some alphanumeric text (*c*), most would recognize that it contains the letters I and D that possibly form the word ID, and someone familiar with the SDTM might even draw the conclusion that the numbers in fact represents identifiers in the variable with label *Investigator Identifier* in the SDTM standard (*d*).
Which in this case would be absolutely correct.

INVID	Investigator Identifier	Char		Record Qualifier	An identifier to describe the Investigator for the study. May be used in addition to SITEID. Not needed if SITEID is equivalent to INVID.	Perm

Excerpt from SDTMIG v.3.2 DM domain, variable INVID

As soon as someone mentions the word data, it is very possible that you are seeing rectangular structures, like spreadsheets or relational databases, and perhaps also some metadata like variable names and labels, as it is probably the most common way of representing data. By adding metadata to the data we can make sense of what datasets and variables represent. In this case above that we have two investigators which have been assigned (somewhat unusal) identifiers. So we need both data and metadata to understand what a datapoint represents, unfortunately we have very limited options to actually have them together in one single database so we will need to create something like the FDA submission deliverables to complete the information and make it obvious what we are storing and how the variables are used.
So by constructing datapoints with structure and metadata we could make meaning of the individual datapoints. But what happens if we go the other way around and start deconstructing our SDTM datasets?

DATA AND METADATA

We will start with a simple example with only four variables from the Demographics (DM) domain in SDTM.

DM – Specification for the Demographics Domain Model

dm.xpt, Demographics — Version 3.2. One record per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product. This must be a unique number, and could be a compound identifier formed by concatenating STUDYID-SITEID-SUBJID.	Req
INVID	Investigator Identifier	Char		Record Qualifier	An identifier to describe the Investigator for the study. May be used in addition to SITEID. Not needed if SITEID is equivalent to INVID.	Perm
AGE	Age	Num		Record Qualifier	Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy concerns).	Exp
AGEU	Age Units	Char	(AGEU)	Variable	Units associated with AGE.	Exp

Excerpt from SDTMIG v.3.2 DM domain, variables USUBJID, INVID, AGE and AGEU

These are very simple variables and are very easy to implement. You are free to define how the subject's identifier (USUBJID) and the investigator identifier (INVID) are created, decide how to derive AGE and follow controlled terminology for AGEU.
Below is an example of how these variables can look when viewed in a dataset.

USUBJID	INVID	AGE	AGEU
S01	I01	30	30

Can you tell me if it is the *Subject* or the *Investigator* who is 30 years old?

Well actually it is not stated anywhere, it is only implied because of the note suggesting how it may be derived, using RFSTDTC and BRTHDTC (which both refer to the *subject* in the note.) And that BRTHDTC may not be available due to *subject* privacy concerns. The note for BRTHDTC specifically states that it is for the *subject*.

DM – Specification for the Demographics Domain Model

dm.xpt, Demographics — Version 3.2. One record per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
RFSTDTC	Subject Reference Start Date/Time	Char	ISO 8601	Record Qualifier	Reference Start Date/time for the subject in ISO 8601 character format. Usually equivalent to date/time when subject was first exposed to study treatment. Required for all randomized subjects; will be null for all subjects who did not meet the milestone the date requires, such as screen failures or unassigned subjects.	Exp
BRTHDTC	Date/Time of Birth	Char	ISO 8601	Record Qualifier	Date/time of birth of the subject.	Perm
AGE	Age	Num		Record Qualifier	Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC, but BRTHDTC may not be available in all cases (due to subject privacy concerns).	Exp

Excerpt from SDTMIG v.3.2 DM domain. Variables RFSTDTC, BRTHDTC and AGE

The order of the variables in example above are correct and if you didn't know SDTM or are unfamiliar with clinical data, chances are that you would guess that the AGE is more related to INVID than USUBJID.
Is seems that our data is somewhat incomplete even for simple datapoints and requires a lot of extra knowledge about our business and specific business rules for our standards (and I would say that the same is true for most sponsor systems/standards as well.)
Let's deconstruct the above DM variables and see what we find.

DECONSTRUCTION SDTM

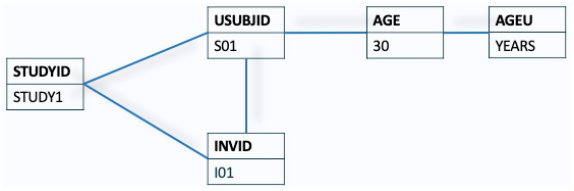
The below image is just a repeated from above, where the STUDYID variable has been added.

STUDYID	USUBJID	INVID	AGE	AGEU
STUDY1	S01	I01	30	YEARS

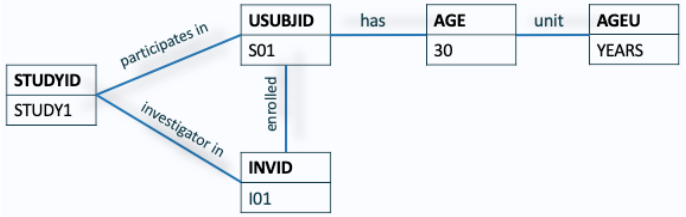
So what we have here are:

- Two distinct persons
- One person is an investigator in the trial
- The other person is a subject in the same trial
- The subject is (probably) enrolled in the study by this specific investigator
- The subject in the trial is of age 30 years.

Deconstructing the dataset it will look something like this, where lines express that the metadata implies that there is a relationships between the individual variables:

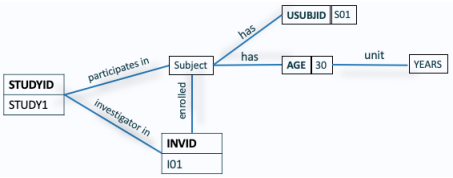


Nothing has really changed, but the lines make it clear e.g. that it is the subject, and not the investigator, that is 30 years old.
Let's put some of the implied knowledge onto the lines.

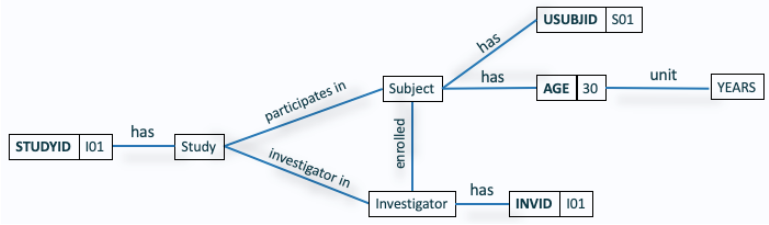


It makes it fairly easy to read, and I think that the risk for misunderstanding this representation is less than that single row. If we want to, it is very easy to go back to single row representation if we wanted to. It is just a matter of removing the lines and the added metadata.

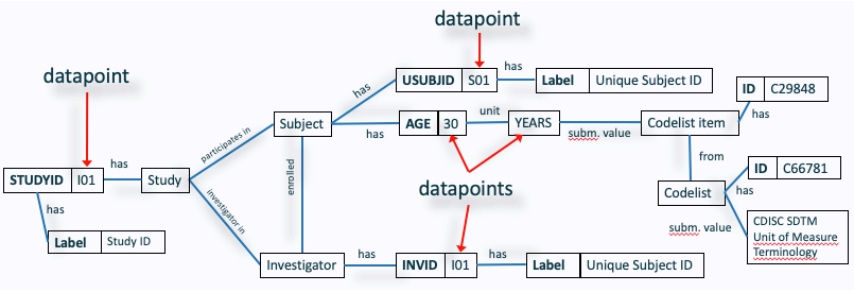
But I think we need to deconstruct some more, because we still have missing implied knowledge in our representation. We'll start by deconstructing the subject.



The subject has now been deconstructed, resulting in a subject participating in the study with a USUBJID=S01 and AGE=30 in the unit of years. Let's deconstruct the study and the investigator as well.

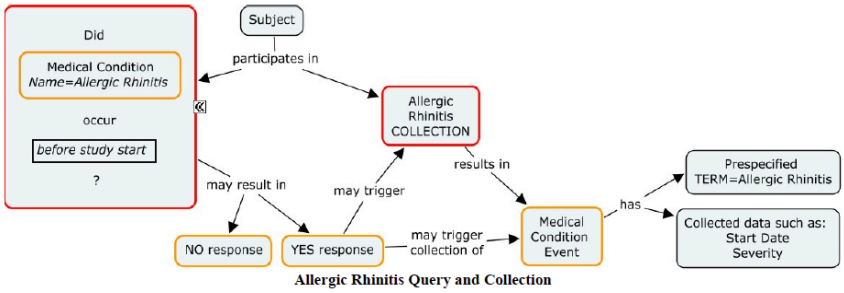


Now we have three different "things" in our visualization. A *Study*, a *Subject* and an *Investigator*, which have their own distinct datapoints and metadata attached to them. Still it is not different from our original example, it contains the same information plus some additional metadata explaining the relationship between our original variables.
Next step is to add some more metadata from SDTM controlled terminology and the labels from the standard, which we use when we create our define-xml.



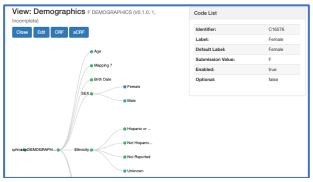
In this picture we now have all the datapoints together with their metadata. The variables have their labels and also the relationship from YEARS to the correct submission value (shortened to *subm. value*) with its codelist item ID and the codelist ID. By deconstructing the example DM row for one subject, it revealed a pattern of business logic that is representative of all the subjects in the DM domain. Instead of having a single row representation, I think that this representation can be more helpful when explaining these variables in the DM domain to anyone beginning with SDTM.
And if we store this information together like this, instead of having multiple documents, the standard would be easier to learn, maintain and use, as it is possible to express information more specific.

If you've ever opened a CDISC Therapeutic Area Guide (TAUG) you might have encountered something similar as the deconstructed SDTM record above, i.e. the concept maps like the one below from the Asthma TAUG.

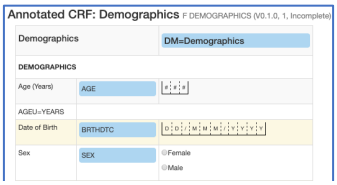


Concept map from CDISC Asthma TAUG v1.0

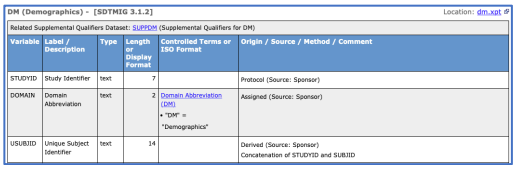
The deconstructed picture of the record from SDTM DM domain and the concept map may not look like data, but it should, as by using graph databases you are able to store data and metadata together as it is represented in the pictures above. The graph database can then be queried to retrieve the necessary information, just like a normal SQL statement does in a relational database.
By implementing it into tools using graph technology a push of a button triggers the correct query that will retrieve the necessary information and present it as a CRF, aCRF, SDTM dataset or define-xml.



Graph DB view



aCRF view



define-xml view

CONCLUSION

By deconstructing CDISC SDTM/SDTMIG it reveals patterns that are useful for implementing into graph databases, which enables automation of submission deliverables.
It also can be used to improve the quality of the standard itself as one can find inconsistencies and where clarifications can be needed.

REFERENCES

1. The CDISC Study Data Tabulation Model (SDTM): History, Perspective, and Basics, PharmaSug 2008, Fred Wood
2. SDTM, https://www.cdisc.org/standards/foundational/sdtm
3. SDTMIG, https://www.cdisc.org/standards/foundational/sdtmig
4. CDISC Asthma Therapeutic Area User Guide v1.0 https://www.cdisc.org/standards/therapeutic-areas/asthma
5. Graph databases: https://en.wikipedia.org/wiki/Linked_data and https://en.wikipedia.org/wiki/Graph_database
6. PhUSE Annual Conference 2017. TT09 (2017) SDTM Domains by Query - is it Possible? Link to: Paper Link to: Presentation

CONTACT INFORMATION

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

Johannes Ulander
S-Cubed ApS,
Lille Strandstraede 20C, 5
Copenhagen / 1254
Web: http://www.s-cubed-global.com/