PhUSE US Connect 2018

Paper RG06

It's Time To Change

D Iberson-Hurst, Assero Limited & A3 Informatics ApS

INTRODUCTION

It is now nearly two decades since CDISC was formed, the FDA has accepted data using the CDISC Study Data Tabulation Model (SDTM) format since 2004 and studies starting after 17 Dec 2016 must be submitted in approved electronic formats. As an industry, we have come a long way, yet we still have silos, we still map and we cannot work in a seamless manner.

There may be a better way. This paper will discuss a vision in which the data are represented in their natural form and technology leveraged to allow for a more seamless process, where mapping is no longer required, silos are removed, and our standards become views of well-structured data, generated more via query rather than via code. However, the lights need to be kept on, work on today's studies need to continue. The paper will also address how we can move forward with the vision in an iterative manner, while still supporting current study work.

It is all very well writing about a vision and laying out the theory but few are convinced by yet another PowerPoint deck or a theoretical paper. Therefore, since 2015, a series of practical steps have been taken to implement this vision and determine if the theory can be put into practice.

This paper is intended to paint the vision and allow others to comment upon that vision. The final part of the paper presents information on the practical work undertaken to date and what can be demonstrated today.

ISSUES

SILOS

One of the significant inhibitors to achieving the end-to-end vision is the silos within which we work, as an industry. You only need to look at the current CDISC standards to see this; we see CDASH, SDTM, ADaM but no one content specification that unites these standards[1]. When you consider the entirety of the clinical trial design, the data collected and then the subsequent calculations performed during analysis and presentation of results, you see a large interconnected set of data. The problem is that the amount of data and the interconnections are so large that it is impossible to visualise; the number of data points and the relationships between them are so numerous that it is difficult to comprehend. Add in the nature of studies and the variation between them and the problem increases in size. This complexity led to the CDISC data standards being developed in silos that reflected the organizational and process silos existing within industry. As a consequence, we have lost the connections across the silo boundaries. We then expend considerable effort attempting to put the connections back into the data at some later point in the process. The industry refers to this as mapping, a process that does not come without risk. Mapping is really just the replacement of relationships we have lost. Figure 1 is a simple visualisation of the data complexity with the raw data we collect on the left with the interconnections to calculations and subsequent data through various stages until we reach the answer.

Note: The following three figures are intended to convey the notions described in the text and not reflect exactly falls into each of the existing standards.

[1] BRIDG does exist but is not considered a content standard. It does provide a model that spans the silos

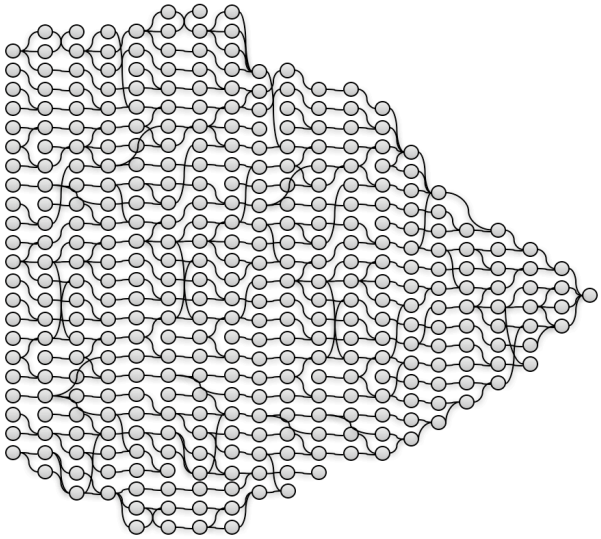


Figure 1: Clinical Data Complexity

A simple overlay in Figure 2 relates the figure to today's standards

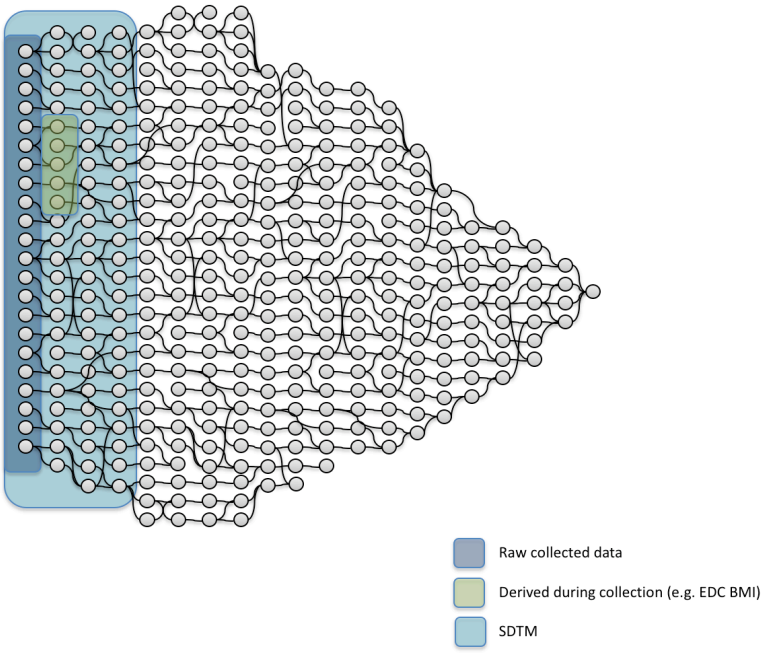


Figure 2: Clinical Data Complexity and Standards

Figure 3 illustrates how the current data standards were developed. Because of the complexity of the data we created silos by cutting it up. But in doing so everywhere we cut, we lost a relationship as well as the knowledge that the entities at either end of the relationship were related. The mapping process replaces each of these relationships but every time a sponsor does so there is the potential for doing it differently.

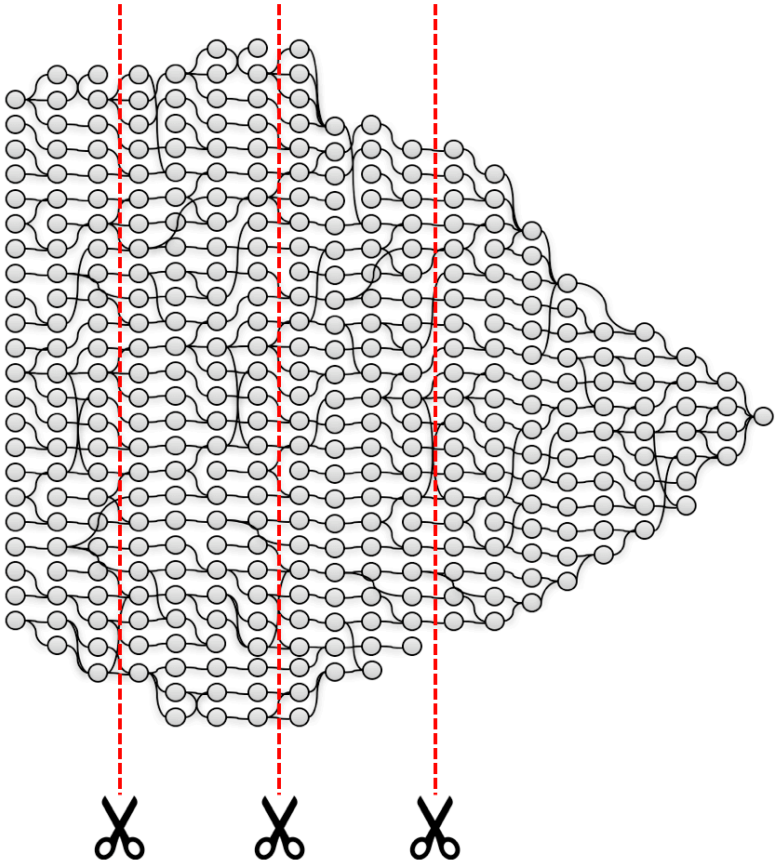


Figure 3: The Creation of Silos

THE RECTANGULAR AND THE ROUND WORLDS

Another issue is the use of rectangular structures for reporting. While rectangular structures work well for reporting, they are not optimum for storing and processing data and do not represent what occurs in the real world.

Consider Figure 4. On the left of the figure we have the structures with which we are familiar. A domain is a means of reporting one or more observations, interventions or events that have occurred within a clinical study. Define.xml is a means of stating how a given set of domains are constructed, their terminology and the variables used; the study metadata. A form is a means of capturing the clinical data. On the right-hand side, we have the world in which we live. We have people who are patients or subjects upon whom we perform observations, undertake interventions or events occur either in the context of providing healthcare or as part of a clinical study. They are related to other people who we may be interested in such as siblings, parents or children. But that real world is not rectangular, it is multi-dimensional. This world has been referred to as the round-world in the past.

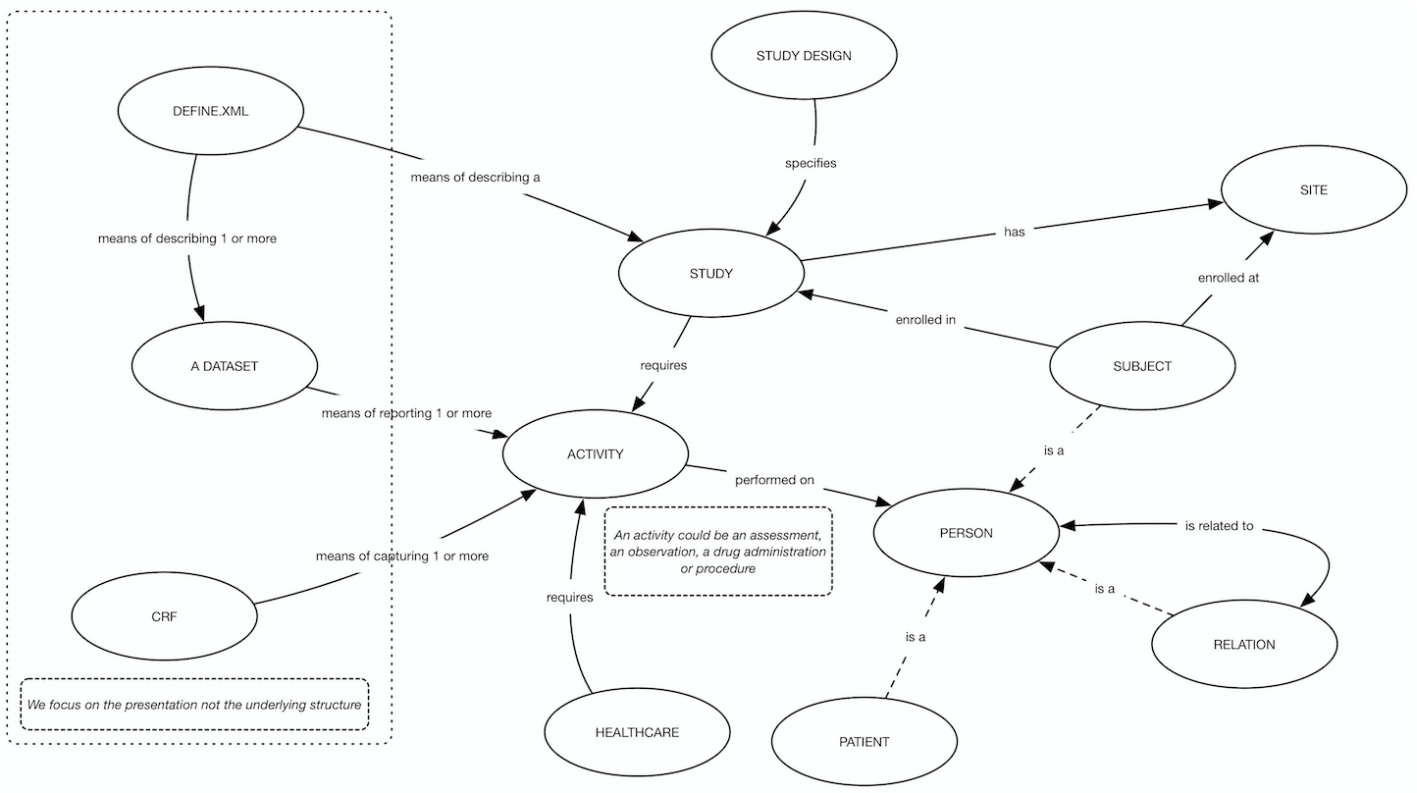


Figure 4: The Rectangular and the Round Worlds

This exhibits itself within our standards in various ways. The standards are developed in the context of the left-hand side: domains, variables and rectangular structures. As a consequence, we miss some of the nuances within the data and its true construction or form; we may miss relationships, not represent them correctly or ignore them completely.

Different teams building the standards are prone to place the same information into different variables, place different information into the same variable which is then used inconsistently or variables are used in combination in different ways.

We also see issues when users encounter situations when new types of data have not been seen before. Questions of which domain and which variables to use immediately come to mind and two sponsors can make different choices. Our data immediately becomes non-standard.

One such example is a recent creation of the Functional Test and Clinical Classification domains. Here we seem to have created two domains in addition to the Questionnaires domain so as to separate the types of QRS instruments; nothing wrong with that. But this division is a categorization. Unfortunately, the variable --CAT is already being used to hold the name of the questionnaire. So instead of the instrument name going in to a variable such as --NAME, the use of CAT was overloaded[2]; so is this variable a name or a categorization? To meet the new requirement we are forced to split by domain rather than use the correct qualifier. We have a situation where information has been placed into the wrong logical container. It is for reasons like this why we require a good reference framework along with good definitions to ensure we model our data correctly.

For further examples see Therapeutic Area standards and their impact on current SDTM implementations see the paper by Ulander and Both, (Ulander & Both, 2015)

SQUASHED

Our standards also place everything into one structure. We transport our data in datasets, we store our data as datasets, we model it as datasets and we present it as datasets. We have squashed everything into one structure but this is not what we need to do in the future; in fact we must stop doing this going forward.

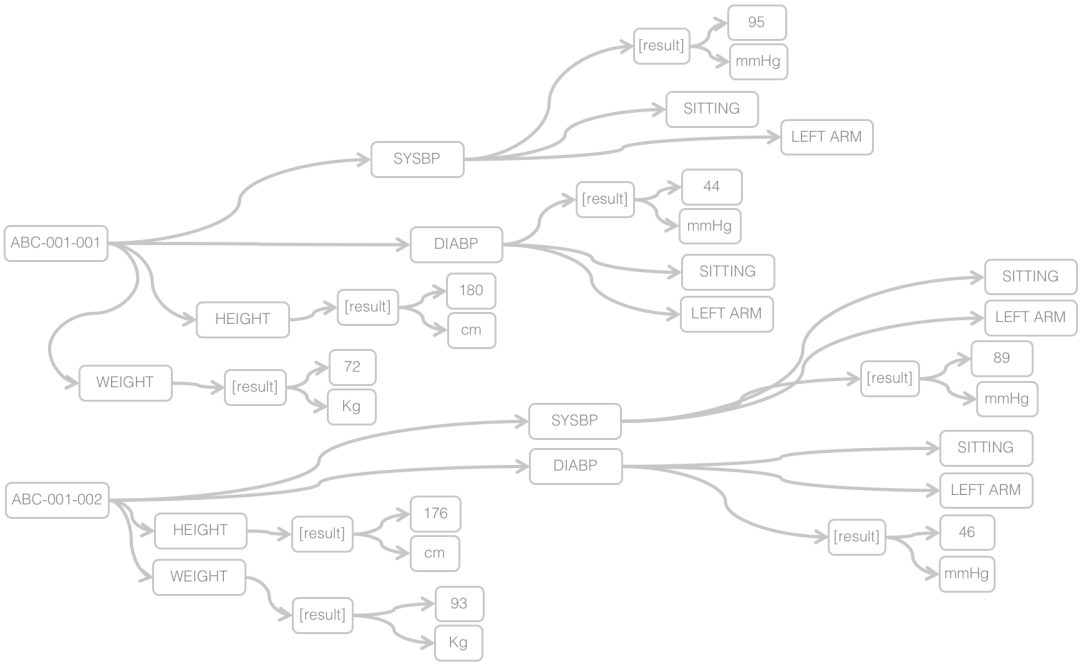
[2] Overloaded in this context is used to refer to the situation where more than one type of information is stored within a single variable.

Imagine the SDTM dataset shown in the Figure 5 below. A few rows of data with some simple findings. This is what the industry is presented with day on day with the storage, design and presentation all manifested in one structure, the SDTM standard. Within that dataset are many assumed relationships. The columns are related; the value and the unit columns for example. Rows are related if only by subject but other relationships exist such as systolic and diastolic blood pressure[3]. We place these assumptions and relationships into code when we process the datasets. Mostly this works because we have a shared understanding but as the world develops and expands the shared understanding doesn't reach every programmer. Documentation is provided in non-machine readable PDF form that is separate from the data. Unless you are already intimately familiar with the standard, a programmer will "manually" map this when trying to understand the data. Self-describing data would assist in aligning data across the industry

ABC-001-001	…	SYSBP	…	SITTING	LEFT ARM	95	mmHg
ABC-001-001	…	DIABP	…	SITTING	LEFT ARM	44	mmHg
ABC-001-001	…	HEIGHT	…			180	cm
ABC-001-001	…	WEIGHT	…			72	Kg
ABC-001-002	…	SYSBP	…	SITTING	LEFT ARM	89	mmHg
ABC-001-002	…	DIABP	…	SITTING	LEFT ARM	46	mmHg
ABC-001-002	…	HEIGHT	…			176	cm
ABC-001-002	…	WEIGHT	…			93	Kg

Figure 5: Rows and Columns

The following figures show steps in reorganizing the data into a node and relationship form. This allows us to see the actual structure of the data and its natural form. The first step, Figure 6, is to link a 'subject' node with the relevant observations and link the results and values.



[3] One solution that has been used is to utilise VSCAT set to "Vital Signs" and VSSCAT set to "Blood Pressure" but again this is overloading the use of –CAT and –SCAT as "Blood Pressure" is really a higher level of test code. The solution used is less than ideal forced upon us by the tools and structures we have available.

Figure 6: Step 1

Figure 7A second step can then be taken to form 'observations' with a more explicit recording of the actual test, the result and the qualifiers, see Figure 7

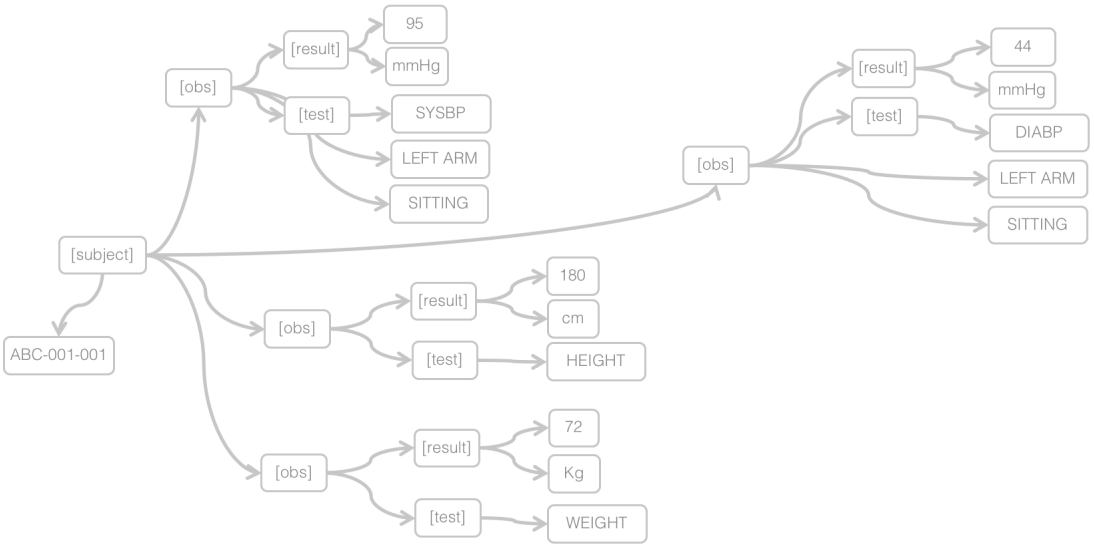


Figure 7: Step 2

A 'final' step, for the purpose of this explanation, though more could be done, is to share terminology across the 'observations'.

Note: In the following figure the green circle represents an concept within a terminology to which the data can refer.

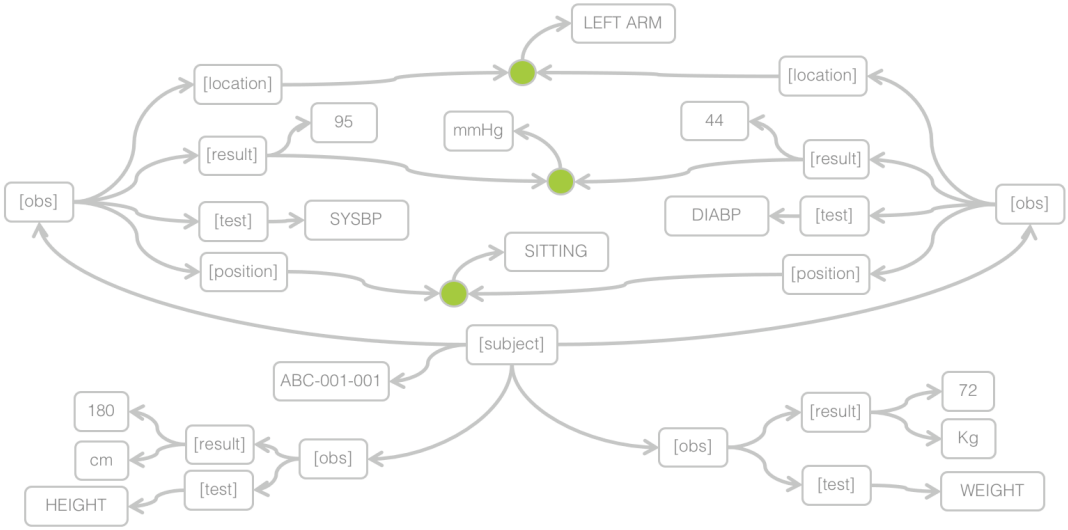


Figure 8: Step 3

Note: For the sake of clarity the timings for the observations have been left off and the final figures only include a single subject.

The first important point to note is that the two forms are equivalent, the data are the same but the graph version as the information is richer in what we can derive from it. Secondly, and more importantly, we can extract the rectangular form from the node and relationship form. A machine can 'walk' through the nodes and relationships and extract the

rows for presentation and further analysis. The storage of data and the presentation of data need not be one and the same.

The representation we have here is a graph that can be queried. This is an important point to remember.

ANOTHER PERSPECTIVE

SDTM and the tabular structure is very useful, but real-world data is not tabular in nature. Consider starting with the 'observations' that relate directly to entities we see in the real world. Clinical studies could be built from these small building blocks (a finding, the recording of an event, an intervention) with the associated timing information (the study design) and relate these 'observations' to subjects.

One of the challenges with data today is consistency across the data collected and in particular the terminology used. By building 'observations', we can be more precise up front in that, for an 'observation', we can specify precisely the terminology to be used and where it is to be used. We can set the precise units for an 'observation' and we can set the qualifiers we expect and the terminology to be used in each case. These definitions will become part of the data in that when data are collected we would link to the definition used to collect it. This is important as we merge and pool data.

These 'observations' are a small packet of knowledge that relate to real-world entities. These 'observations' will be referred to as Biomedical Concepts (BCs) for the remainder of this paper. The BC becomes a definition of a measurement or a question we ask a subject, a unique code for that measurement/question, the possible responses and the various qualifiers. Some of the information maybe optional and not always captured as part of some studies but we can define precisely all the features of our BC.

Suddenly this seems like a lot of work given that there could be 10,000 of these BCs but it isn't. We have this information today. The knowledge exists for the vast majority of these BCs in people's heads but also in a machine processable form with the multitude of define.xml files held by sponsors. There is also no need to boil the ocean, we can do this by subdividing the problem either by domain, Therapeutic Area (TA) or some other such criteria and use the BCs in parallel with current methods. The important thing is that every BC that is defined takes us one step closer to better data and better precision.

ANATOMY OF A BC

One of the key aspects of a Biomedical Concept is to ensure that the same type of information goes in the same place, that the semantics of the information are well understood and that the information be retrieved later and used consistently. To do that we need some form of reference. This role is currently fulfilled by the BRIDG model[4]. BRIDG classes, attributes and datatypes and the associated definitions can be used to build up templates against which we can build our Biomedical Concepts. It is vitally important that we have a single pigeon hole for each item of data that we wish to use and have a precise definition for it. Returning to an example from earlier, we need a slot into which place Categorization information and have a precise definition for what a Category is. We must not allow other information to be placed into that pigeon hole. This will aid users in placing the right information in the right place such that we get consistency across our data and make our data more useful.

Figure 9 shows a simplified picture of a template. It has a slot for a test code (a unique identifier for what is being measured denoted by Test in the diagram), a set of timing variables (in this case a simple date and time), the result consisting of the value and units, the units not being necessary for coded responses, a position, a location and other qualifiers

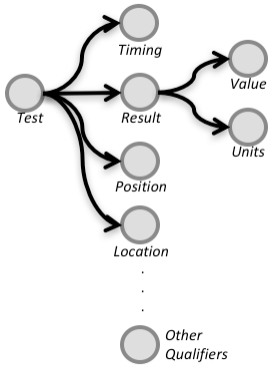


Figure 9: Biomedical Concept Template

Once a template is in place it can be filled with content. Consider a very simple example of Temperature, Figure 10. There is a result consisting of value and units. The units are only going to be Fahrenheit (F) and Centigrade (C), the value an integer of 3 digits (###).

Note: In the following figures reference is made to the NCI Thesaurus 'C-Codes' used within the CDISC terminology

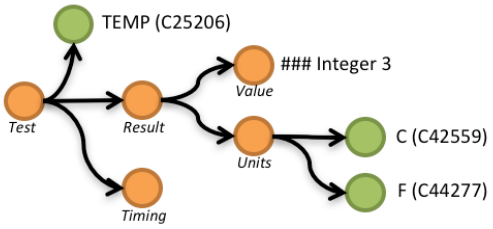


Figure 10: Simple Biomedical Concept

Is there a method of observation? Are there other qualifiers that need to be captured? Typically, such qualifiers are not currently captured for Temperature but there is no reason why not, a slot exists within the template and can be utilised if required. Consider a second example, see Figure 11, for systolic blood pressure where there is a need for a qualifier of position. Again, the terminology is bound in as part of the definition such that there can be no debate about the list of permitted values and whether or not a qualifier is required. Note that optionality can be included within the definition, do we always have to collect and store position.

[4] We might consider SDTM class variables from Findings, Events and Observations to fulfil these roles but what we don't have with these structures currently is the relationships between them. For example, a result has a unit except, of course, when coded when units are not required. BRIDG gives us this information.

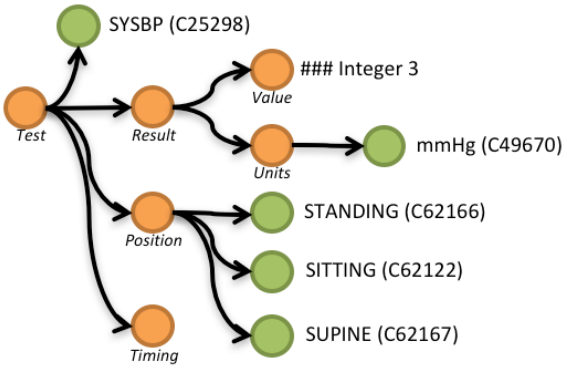


Figure 11: Biomedical Concept for SYSBP

The above discussion raises an interesting point. If an additional qualifier was required in the future (say method) then it can be added. There is already a slot in the template for method and it can be utilised. If there is no slot matching the type of data that is required, then the template will need to be updated.

This drives the need for version managing the BC definitions. In the above example of adding method the terminology for the permitted methods must be defined when method is added to the BC definition. This will mean we now have several versions of the BC but we will be able to tell. When we collect data we know against which version of the BC we used in defining what data are to be collected. This places us in a much stronger position when handling the data. For example, when handling data from two studies it would be possible to determine if the version used was the same or not and, if not, what the differences were.

Some may notice that BCs are no different than other models from such places as openEHR, FHIR, Intermountain healthcare, LOINC. They are not. It is the same knowledge and is a reusable clinical research form. A significant amount of this knowledge is available in different forms, BCs are about organizing it into useful reusable components.

An important item to note is that, other than the terminology, there is no connection to the CDISC world, there is no mention of domains or variables within the BCs. Even the terminology being used could be replaced by other terminologies if required or equivalences (mappings) noted. This would allow a route of migration away from CDISC terminologies should it be desired.

The number of templates envisaged is in the 10s of magnitude. The number of BCs is high, in the 1000s. This seems an impossible task but it should be noted that most clinical studies are using the same content time after time; the amount of new content is relatively small[5]. This existing content can be furnished from existing material; it is available in define.xml files held by sponsors. The value level metadata held in such files is not far removed from the information seen in the figures above. Simple processing could release this knowledge into the BC form and make it easily available to all.

[5] It is worth examining closely, for a single sponsor, the amount of existing content, how it could be re-engineered into BCs versus the true new content that arrives with a new clinical study. However, it needs to be remembered that, in whatever way we work, we need to have good definitions; the work still needs to be done. We also need to share this information across all sponsors.

LIFE-CYCLE

How do these BCs help the operation of trials and deliver better data? Consider first the collection of data. This is normally accomplished using forms and EDC systems, data loads, ePRO devices or wearables. These are providing a stream of 'observations' in a certain structure.

Typically, a form is a series of questions to which is attached an SDTM annotation detailing the location of the data within the database. What is not on the form are the tests (qualifiers) associated with those results. So significant effort is expended annotating forms to 'replace' the information that has been lost (assuming those building the form ever realized the relationship existed). In the simple cases this takes the form [VARIABLE] WHERE [TEST CODE] = X the WHERE piece being the relationship being manually recreated. However, a form, table or data stream can be built from BCs. Each question on a form is simply a link to the leaf of a BC definition; the result, the units; a qualifier. When forms are built in this manner the items are linked back to the test and other qualifiers such as category and subcategory by the BC definition, the relationships are not lost, they are in place the moment the form is created.

In the simple example shown below, Figure 12, the form is constructed from those items we need to collect. For temperature, we need to collect our units, F or C. However, for systolic blood pressure we have fixed units – our BC definition tells us this – it does not need to be entered, it can be displayed to aid the user, but it does not need to be collected. Of greater import is the link to the BC and from the result value or units there is, of course, a link to the remainder of the BC. Thus, the test code can be accessed as well as any qualifiers. It can be immediately seen if those qualifiers are being captured or are we ignoring them. Humans can do this but so can a machine.

Note: In the following diagrams orange nodes reflect the template structure used to define the BC while the green nodes are the CDISC NCI terminology used to create a given BC

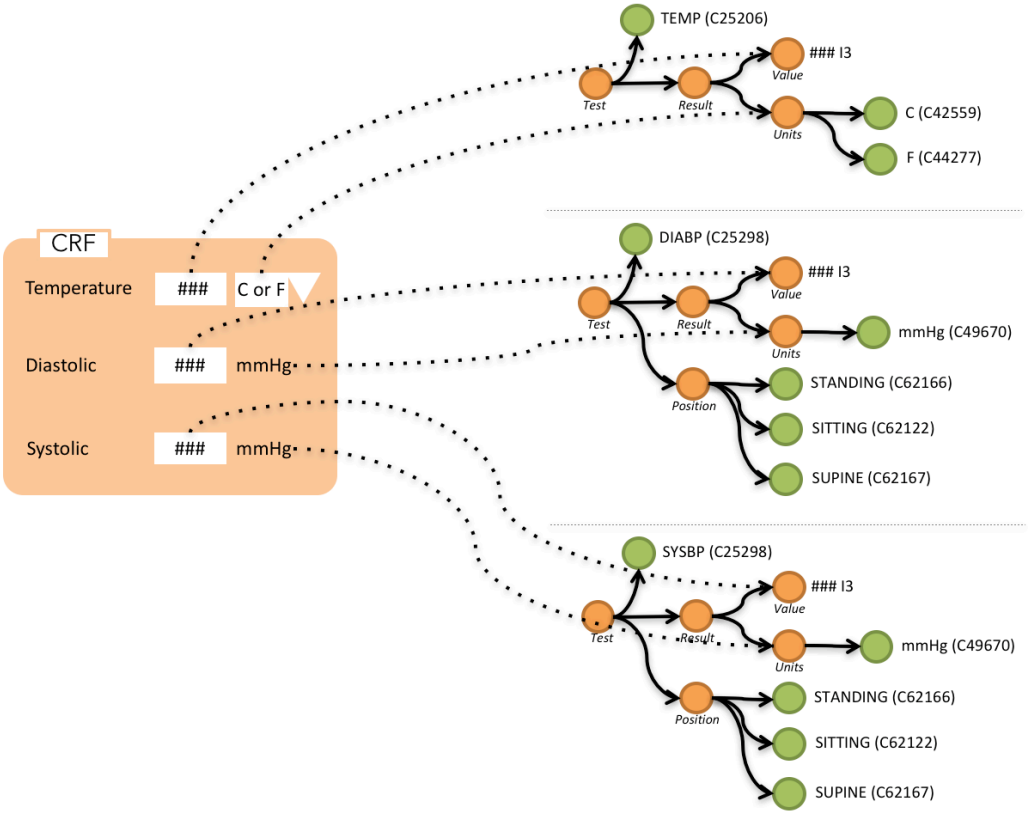


Figure 12: A Biomedical Concept based Form

One important benefit of this is that other forms using the same BCs will ensure that the data are consistent at the observation level. One implication is that standard forms become less important and the use of standard BCs becomes the aim. This also provides the foundation for alignment with non-form based data sources such as wearables.

Having used BCs for collection there is now the need to link to the tabular presentation world of reporting. The link to the desired human representation is achieved by linking the SDTM to the underlying data. Each SDTM variable is again one leaf within several BCs. BRIDG helps in that there have been SDTM BRIDG mappings available for a long time. Our templates are built from BRIDG classes and attributes. Therefore, there is a link between the collection of data and SDTM, see Figure 13.

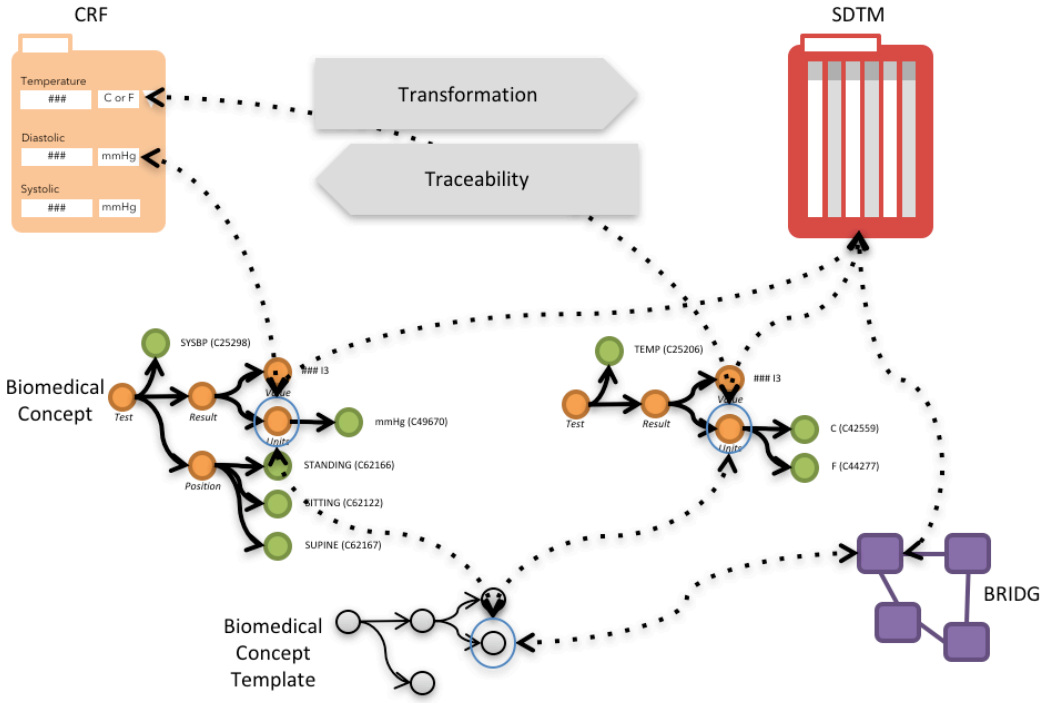


Figure 13: Link to SDTM

The world of collection is now linked with the world of reporting. It is repeatable and predictable.

As stated earlier, not all data is collected via forms. With BCs it doesn't matter. Whether the data are from a lab with a series of Lab BCs placed into tabular data for transmission or a wearable device producing an hourly average heart rate reading as a data stream. What is needed is the ability to place the data into the BC structure. It immediately can be merged with other such data that conforms.

One implication of using BCs is the removal of the restriction on where the data go in terms of domain and the domain presentation. An SDTM domain is a rectangular structure that can be defined and linked to a set of BCs (e.g. Vital Signs or ECG). A BC could equally be 'assigned' to Vital Signs or the ECG domain, the choice can be stated up front as it is today by the SDTM IG but could also be left to a later stage. This could allow greater flexibility in the future. However, flexibility can be abused so caution is recommended![6]

In essence, a BC is a common lower-level definition sitting beneath the existing SDTM & CDASH specifications. They will also extend into ADaM where raw collected data are referenced and also back towards the definition of what is to be collected within a study in places such as the protocol and statistical analysis plans. CDASH and SDTM become specifications of how we wish to collect and view our data, BCs define our data.

[6] It could be argued that we should fix our BC to Domain mappings. But even if we do, the structures allow us to change the allocation in the future if we decide we got it wrong.

DERIVED DATA

BCs cover the raw-collected data within SDTM, but it is well understood that SDTM encapsulates a lot of derived data. By defining the BCs, we start to gain much clarity about what is raw and what is derived. The derived data and the specification as to how it is generated belongs with SDTM. As an example, many domains have a --DY that is the difference between two dates (the observations --DTC value held within the domain and, generally but not always, the variable RFSTDTC held within the DM domain). This variable is not really part of the raw data, it is part of the SDTM presentation. Therefore, these derivations can be defined in a machine readable way as part of the SDTM domain definition. As a consequence of BCs clarity emerges about how data items are created and their true nature. An additional benefit is that we gain clarity on how the data are constructed and such knowledge can be shared with the machine that can perform the repetitive task of creating such data.

REAL-WORLD EVIDENCE

A great deal has been said about Real-World Evidence and the desire to link healthcare data with research data. Much work is underway within HL7 on FHIR and there is some commonality between BCs and FHIR resources. If the gap between the two can be closed, and it is believed it could be, then it would be easier to link research data in BC form to healthcare data in resource form.

One impediment to such a goal are the terminologies used and the necessary mappings between them but there might be ways in which BCs could hold definitions of other terminologies such that the necessary translations could be made.

Further progress has been made in this area and was reported on at the CDISC European Interchange.

THERAPEUTIC AREA USER GUIDES

What is the impact on Therapeutic Development and their associated User Guides? It is strongly believed that the TAUGs should define the set of relevant BCs with the necessary full definitions, in particular full terminology definitions for each BC. This allows for predictability in the content of domains and the ability to determine if additional content – that may be valid – is present.

In the first instance an 80/20 rule should be applied to TAUG development such that the most commonly used ‘observations’ are defined and then the set can be expanded in incremental steps.

Existing TAUGs should be revisited and the BCs developed. Some have the necessary information already, while others are lacking. By developing BCs, it is believed the documents could be reduced in size, be consistent in format and be readily usable. The BCs could also be delivered in an electronic form. At the most basic level, even if BCs were published in Excel spreadsheet form, this would assist industry even when working using traditional variables by having a precise definition of what were the data collection expectations.

ITERATIVE APPROACH

As with all new ideas, they need to be prototyped and lessons learned. The approach can then be adjusted, and further work undertaken. With BCs, the approach should be taking a few steps, review, learn, adjust and repeat. Semantic and graph technologies lend themselves to an iterative approach and allows to move towards our desired vision in steps learning and adjusting as we move forward.

The initial step should be taken with the existing safety domains that are well understood. Alongside this, a single TA should be tackled. Good tooling is also required to support the work to make it easy for content to be assembled and reviewed.

Data quality will be improved with every single BC that is deployed.

AND FINALLY …

Having removed the silo between collection and reporting, we obviously wish to remove the other silos. We can move back towards protocol that details the BCs (observations) that a study is to perform. Analysis uses a lot of the raw data and thus can link to BCs, but it brings into play the 'analysis concept'. It is certainly within the bounds of possibility to envisage standard analysis datasets, such as ADSL, being defined and production being automated. This is the next big step.

BUSINESS CASE

As with everything we do we need to justify the benefits that any change will bring. We have dealt with the technical benefits within the paper. It would be interested to attempt to quantify the financial benefits to sponsors and regulators alike.

REFERENCES

Ullander and Both, PhUSE, 2015. Therapeutic Area standards and their impact on current SDTM implementations [accessed 2018 May 3] https://www.lexjansen.com/phuse/2015/cd/CD03.pdf

ACKNOWLEDGMENTS

The author would like to thank the following for reviewing the paper:

- Kirsten Langendorf, S-Cubed & A3 Informatics
- Johannes Ulander, S-Cubed & A3 Informatics
- Tim Williams, UCB

CONTACT INFORMATION

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

Dave Iberson-Hurst
A3 Informatics ApS
Lille Strandstræde 20C 5.
DK-1254 Copenhagen K,
Denmark

Email: dih@a3informatics.com
Web: www.a3informatics.com