A Vision For Better Standards

Over the last four or five years, I have had the pleasure of working with the CDISC standards in an operational environment, managing a library of definitions and deploying those standards for use within clinical studies. In the summer of last year, I had the opportunity to reflect on what I had experienced when working with paper specifications, updates, new versions and emerging Therapeutic Area (TA) specifications. Combined with my continuing involvement within CDISC and PhUSE it is apparent that:

- Using standards is hard work;
- The management of standards is not as easy as may be it should be;
- Sponsors are being overwhelmed with updates;
- The variability of standards is causing a wide range of problems;
- The TA standards are hard to consume, are inconsistent in approach and not all have appropriate metadata. Additionally, they are published in PDF form and thus are not machine-readable.

My frustration peaked in the summer of 2015.

I have long believed that we need to move from a variable-based world to one based on observations. The observations are a number of variables combined in conformance with a standard framework; the CDISC Biomedical Concept (BC). We see such units of knowledge in healthcare: openEHR archetypes, CIMI clinical models, Detailed Clinical Models and others. Each represents the same set of knowledge but in different machine representations.

The vision and road-map I see is:

- Our efforts should be centered and focused on the creation of BCs.
- We should ensure that the existing standards are enhanced such that they are based on BCs; CDASH and SDTM would, in essence, be split into two with a higher-level 'presentation' - traditional CRFs and Tabulations - with a lower-level rooted in BCs based on the inherent structure of the data and not an artificial tabular structure constrained by the conflicting demands of transport, storage and presentation.
- It is not necessary that everything be done in a 'boil the ocean' fashion. We obviously cannot do this but we can do this almost BC-by-BC, our world can improve day-to-day with the introduction of individual BCs. TAs can be asked to produced the core set of BCs in a consistent and

electronic form from a library of well-constructed patterns / templates that are in turn based on a common framework (the BRIDG model). Once a core set is available additional BCs can be added.

- The desire would then be to capture study data in a BC format. Again we cannot boil the ocean. Forms today are variable-based. But we can replace a small set of variables with the equivalent BC as the relevant BC becomes defined. Again we improve our world BC-by-BC, form-by-form, domain-by-domain, day-by-day.
- Captured data could then be kept in a structure based on the BC metadata. Tabular structures can be derived from the raw data.
- This would then enable a move towards linked data and easier data integration.
- As we progress we start to see more consistent data across sponsors as consistent metadata is deployed.
- We could then place such data and metadata into different transport formats. We will always need to be able to place such data into a rectangular form but BCs also allows us to 'upgrade' to formats such as HL7 FHIR. I have had some preliminary discussions on this topic with others in the industry.
- This would facilitate and ease the integration research and healthcare
- Such BCs would also allow for a different approach to the protocol. I have long considered that BCs would allow us to think of the protocol more as a subject timeline of assessments (each assessment being a BC) rather than the current rectangular schedule of assessments approach. Again, I have had preliminary discussions on this topic.

With this vision in mind, in late July 2015 I had the opportunity to do something about it. I decided to fund a project to develop a metadata repository (MDR) capable of handling BCs. My logic was that if we decide to use BCs, we need something to manage them and the objects based upon them. To me a MDR and associated tools seem the logical place to start. I also wanted something to actually show BCs working rather than just having yet another power-point slide deck with the promise of jam tomorrow. We also need to be realistic; I know there will be issues along a rocky road.

Longer term, I wanted an open source solution based on semantic technologies that people could deploy that brought the benefits of improved metadata to all not just to the larger sponsor companies that can afford the risk of big metadata projects. The desire is to lower the risk for mid-size

sponsors and do so by working with established service providers to bring scale and support.

The system already has interfaces to the CDISC SHARE environment (it imports terminology and SDTM semantic exports) and the aim is to keep pace with SHARE API developments. It is worth noting that SHARE is designed to develop standards and not to provide the breadth of functions needed by a sponsor.

Not all of this note may make sense; I have tried to keep it short. This is one reason why I constructed the MDR and the associated tools so as to be able to demonstrate the ideas in a tangible form and allow for an informed discussion.

You will notice no mention of ADaM and other work in the latter part of the life cycle. I have ideas but my expertise begins to run out. But I do believe it is possible to bring ideas similar to those expressed above into play such that the whole lifecycle from collection to submission can be handled in a linked data fashion. I am keeping in touch with those in PhUSE working in the analysis area using semantic technology and have demonstrated the tools to them. Hopefully we can slowly bring the pieces together.

D Iberson-Hurst
Assero Limited
22nd April 2016