PhUSE US Connect 2019

Paper SI13

Removing Silos: Placing Data at the Centre

Dave Iberson-Hurst, A3 Informatics

ABSTRACT

The industry has long talked about 'end-to-end' as a way of improving processes within research, but the industry is bedevilled by silos and standards that are hard to use. This presentation will describe progress made in moving towards a world where silos are broken down, standards are easier to deploy, where data is at the centre and move towards a world where automation becomes a reality.

The presentation will quickly outline the vision and then describe the tangible progress made including:

- Improved version control;
- The visibility of the impact of change;
- The creation of CRFs;
- The ability to build studies;
- The automated creation of annotated CRFs;
- The automated creation of a consistent and aligned define.xml.

The longer-term vision will then be discussed including analysis, ADaM and TLFs as well as the link to EHRs as iterations in capability take us towards the end-to-end goal.

INTRODUCTION

As part of last year's inaugural PhUSE US Connect event the author presented a paper about 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." [1]

The purpose of the paper was to argue the case for improving our world of clinical research. Through standardisation we have failed to reach our goals. Some would argue that the goal of standardisation is a regulatory submission and the subsequent faster regulatory approval of products. However, the author would argue that the ultimate goal is useful data, data that does not decay – data decay – once the approval has been achieved, data that is useful for more than one purpose. We should be able to readily merge data, create pools and lakes of data and easily share our data with others for the common good.

It should be noted that while this paper may seem critical of standards development to date, it is not intended as such. We have come a long way since CDISC [7] started in the late 1990s. But we have learnt from our experiences, we can do better, we need to do better.

The key aspect in [1] was the use of Biomedical Concepts (BCs) as a core component of moving away from a variable based world, see [2, 3], into one where variables are related, bound to the appropriate terminology and built into meaningful units of knowledge that can be defined once and reused many times. A major aim of the work is to move away from mapping.

Mapping is seen as a necessary evil in our industry, but mapping can be viewed as the replacement of relationships that were never considered, were never thought about or we never bothered to create – the implicit links. When a human opens MS Excel and a worksheet is displayed the eyes naturally fix on row one. The eyes move to the column headers and the brain goes into "create relationship mode" relating columns and cells within the data. –ORRES is related to –ORRESU, rows are related, and tens of other relationships are imagined. We then craft these relationships in our SAS code. The simple ones we do consistently but the more complex ones vary across sponsors,

interpretation creeps in. As a result, variation occurs, we don't achieve standardisation and our efforts are somewhat wasted.

CONSISTENCY

There is a lack of consistency in other areas as well. Consider the term 'Study' as in Clinical Trial or Clinical Study. Ask several people what a 'Study' is, and you will receive a range of different answers depending on the respondent's perspective of the industry, a perspective that may well reflect where they work within an organisation: Clinical Science, Clinical Trial Management, Data Management, Biostatistics or Regulatory.

To find out how a study should be defined within a machine, I could look at BRIDG, I could look at the SDTM Trial Design Domains, ODM variants such as the Study Design Model, and Define.xml may have some useful information. It might be worth an examination of some of the trial registry standards from Clinical Trials.gov, the CDISC Protocol Representation Model (PRM) or the HL7 Clinical Trial Registration and Results (CTR&R) standard. The nice thing about standards is there are so many to choose from.

What we cannot do is find a single definitive statement of a Clinical Trial Study. Where can I find the necessary attributes for a study such that I could implement some code once and use it for all application that I may wish to create? There is no single model. So, when I communicate and exchange information with others, I am faced with an integration issue and conversations are at cross purposes. Scope is also important here. Is a study its identifier, its design, the entire protocol? Scope and size are important; we need to create reusable models which comes with the implication that we keep them small.

We need better and more consistent models. It might be that we cannot agree on what a study is – consensus is an elusive beast - but we should have some core understanding of these concepts and their constituent parts. We are left with a Venn Diagram of standards, each taking a view, each overlapping with the others but we have, as an industry and as data standards, no notion of the super set. Each of the models / standards are created in a silo and, as a result, we have mismatches.

A more specific example is the manner in which we define variables in each of our standards thus creating silos. The specific example of AGE and AGEU is described later in the paper.

SINGLE VIEW

TODAY

We need a single view, a single model. Now we have the BRIDG model, but it is large, complex, difficult to get to grips with and difficult consume in machine readable forms. Awareness is growing in this area and sponsors are beginning to look at aligning models and trying to understand what is a complex landscape; both in terms of the standards available and the information contained therein. The author is undertaking some work in trying to align models in certain spaces using some early tooling but, unfortunately, it is too early in that work to report on it.

SILOS

So how do we proceed? The purpose of this paper is to suggest a route by re-visiting the way we cut up the information we use in our day to day work. As described in [1] we divided our world using the organisation and process boundaries that we are all familiar with. Protocol, Study Creation, Data Capture, Tabulation, Analysis. But we do not have the links across these silos, we have cut them as a consequence of the silos in which we created the standards, see Figure 1.

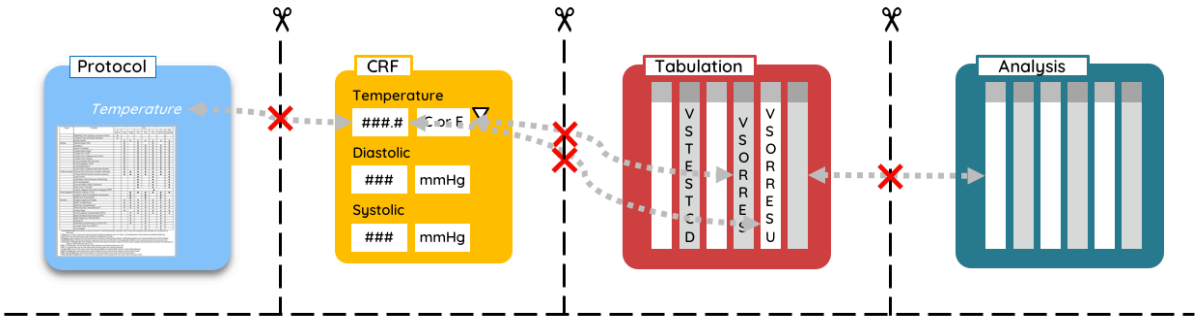

Figure 1 - Operational Silos

CAPTURE & REPORTING

As detailed in [1] we can start to break down those silos by creating BCs and building our operational structures based on the BCs, see Figure 2.

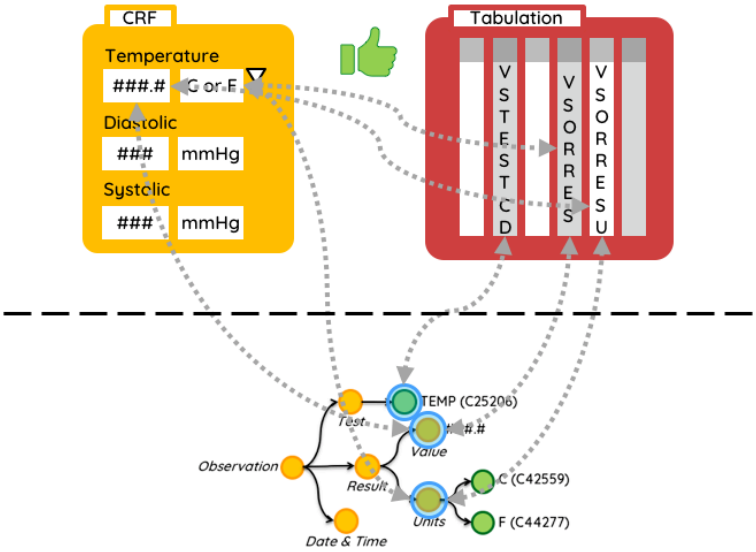

Figure 2 - Linking Forms and Tabulations

Here we are linking the CRF with a tabulation via the BC, thus linking the two worlds and removing the silos. As a result, we now have put in place the relationships between our data, the operational mechanism by which we wish to capture that data (the CRF) and the operational mechanism we wish to use to report the data (the tabulation). An important aspect to note at this point is that we see our current standards as views of the data not the data themselves.

We can redraw Figure 2 in the manner shown in Figure 3.

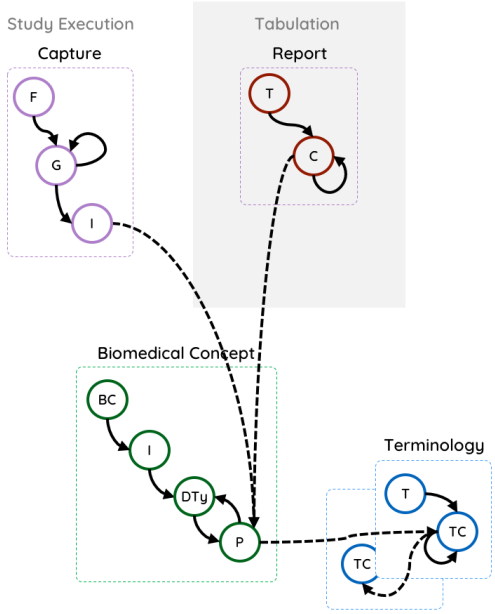

Figure 3 - Linked Models

Rather than showing the CRF as a traditional form and the domain as a tabulation we represent each one in a small model. The model is there for a single focused purpose, to capture data or to report on it. The models have knowledge of the BC model. However, it is important to note that the BC model has no knowledge of the capture or reporting models. The BC model is concerned only with the correct representation of a real-world entity as noted in [1]. A physician can measure your pulse, your weight, perform a lab test without knowledge of SDTM. SDTM cannot report these observations without knowledge of the content.

There is also a model for the terminology. The BC model has knowledge of the terminology, but the opposite is not true. We may wish to link to multiple terminologies to provide equivalence across terminologies. Again, as with BCs and the forms and domains, the terminology model should have no knowledge of the BC model. We should also be cognizant of the many terminologies out there; we need to stop reinventing the wheel and re-use knowledge that already exists.

An interesting aspect of the CDISC terminology is that, while we wish to bind our terminology to the BCs, the terminology should not assume how a term is to be used. Currently the CDISC test codes are organised into code lists linked to a target domain, VSTESTCD, EGTESTCD, LBTESTCD etc. However, there is no need for a BC to be fixed to a given domain. We should allow for flexibility that a finding test could be placed into any Findings domain for example.

STUDY DESIGN

We can add further models to expand our world. A single model for the study protocol and the study design contained therein has proved elusive. We have an opportunity to move on that, linking the study design with the BCs. This then brings an opportunity to examine the methods by which we capture our data within studies and make it easier to integrate the multitude of data sources the modern study now incorporates such as forms, lab datasets, wearables etc.

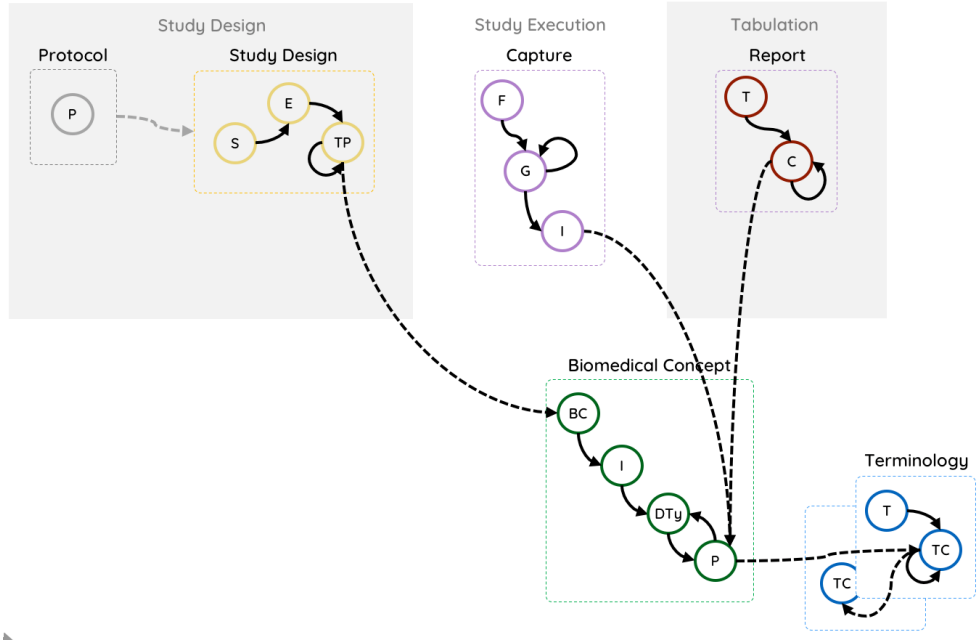

Figure 4 - Protocol and Study Design

With the Study Design linked to the BCs we have the opportunity to link the data to our definitions to create a single unified environment, see Figure 5. Now we need a model that defines our Subjects and all the properties of a subject. Again, like the notion of study discussed above, we do not have a single industry notion of what a subject is. We also are in need of better relationships between humans such as investigators, those who perform evaluations, the results of those evaluations as well as subjects and those related to them such as children, siblings etc.

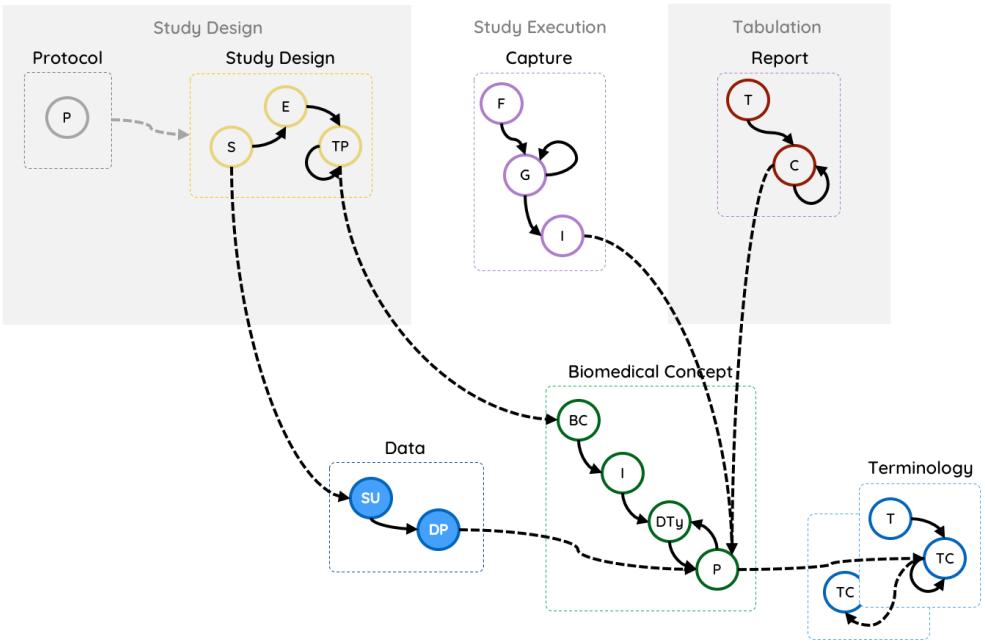

Figure 5 - Data

ANALYSIS

We can then complete our 'model' with the analysis 'silo' and the notion of an analysis concept. As stated in [1]

> *"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."*

This results in Figure 6

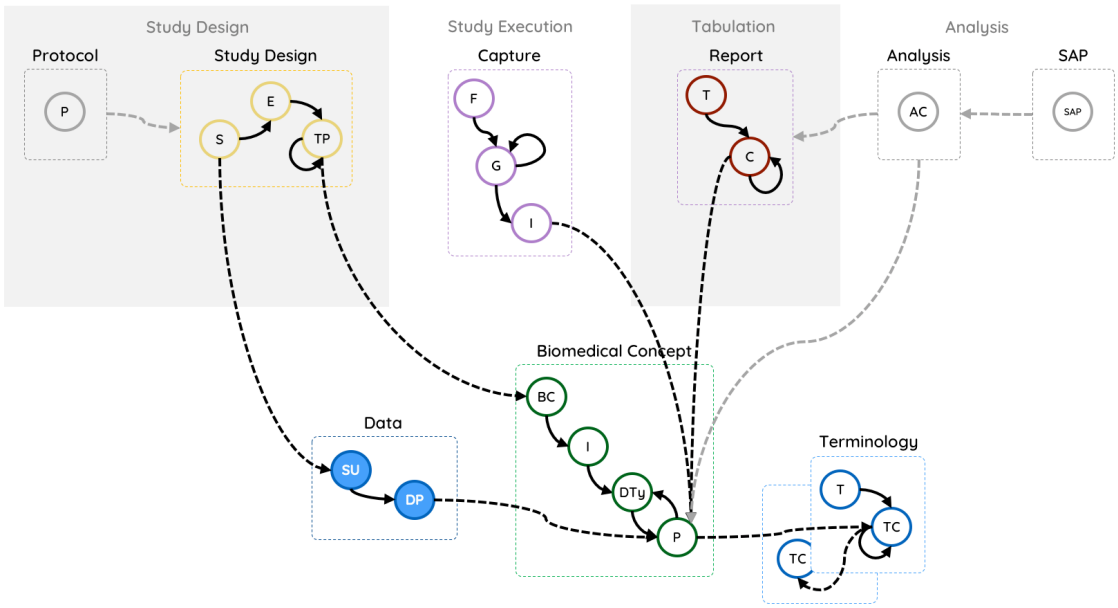

Figure 6 - Full Lifecycle

This gives a full lifecycle with the silos removed:

- The figures have been simplified to show the overall concept / notion; small discrete models / ontologies performing a given role linked together.
- Terminology at the base. We need to accommodate different terminologies and link them.
- BCs provide the key standardized, version managed, definitions.
- Capture and reporting mechanisms built upon BCs
- Study Design built as a timeline of BCs providing a complete, unambiguous and machine-readable definition of the clinical study.
- Data linked to the study design that can be queried for whatever purpose be it reporting, submission, data monitoring etc or pooled readily with other study data.
- The data and definitions can be queried for what is present and what is not; have qualifiers been used for example.
- No silos, no need for mapping, just querying of the result graph.
- Analysis will be added as another ontology to link in. We will probably require the 'analysis BC' equivalent.
- While the SAP is shown on the right of Figure 6, it is recognized that it would be better placed on the left with the protocol.

ALTERNATIVE VIEW

The logic in Figure 6 can be drawn in a different manner as shown in Figure 7.

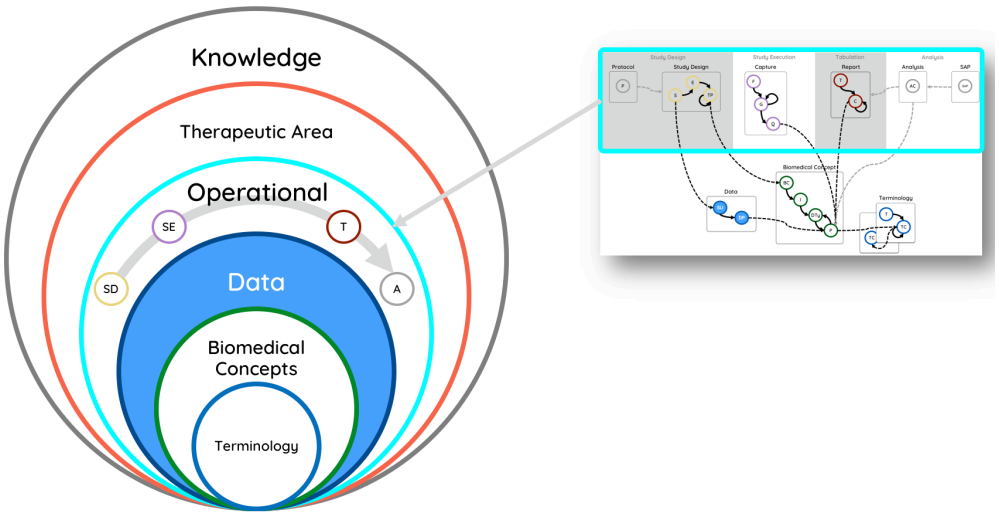

Figure 7 - Data at the Centre

This view places the terminology, the BCs and the data at the centre or base of our world. Wrapping these core definitions are our operational models that cover the lifecycle, with Figure 7 showing the link to the previous left-to-right view. This is then wrapped by a Therapeutic Area (TA) layer. This has been done to note the TAs significance and to not forget that TAs should be built from BCs to make them more useable. The final layer is a knowledge layer.

Knowledge graphs are beginning to come to prominence. They provide a model and put structure to some area of knowledge such as a disease. Their view is different, the operational layer is about conducting our study. The knowledge graph is about understanding the disease. Both need data. If the data are structured consistently such that both worlds can understand and employ those data, we will be able to reuse our hard-earned data for greater benefit.

SILOS REVISITED

When looking at the layered view, it is possible to illustrate the current world of variables in our standards and how BCs help us. Figure 8 reflects our current siloed standards. We reference the notion or concept of Age in our protocol and study design, we capture the data within CDASH and define it there, we repeat again in the SDTM Implementation Guide definitions, and then again within ADaM and ADSL. This is obviously a simple case, but we do

this time and again. We can link the definitions, but a better way is to move them into a single definition and share it; this is all the BC is.
What the BC also does – as already stated – is to relate variables in the correct manner such as result and units and also bind in the terminology. This is illustrated in Figure 9, where a simple notion of a BC is shown with an incomplete set of YEARS and MONTHS. The terminology has been restricted to these two values simple to ease the complexity of the figure, it is recognized that WEEKS, HOURS and DAYS are also normally used.

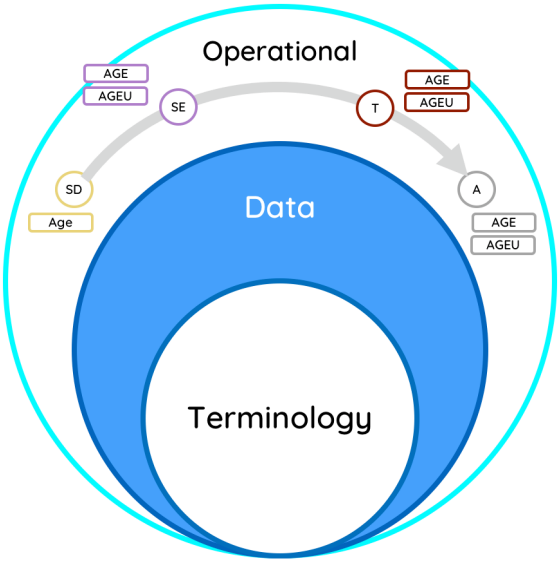

Figure 8 – Repeated Definition

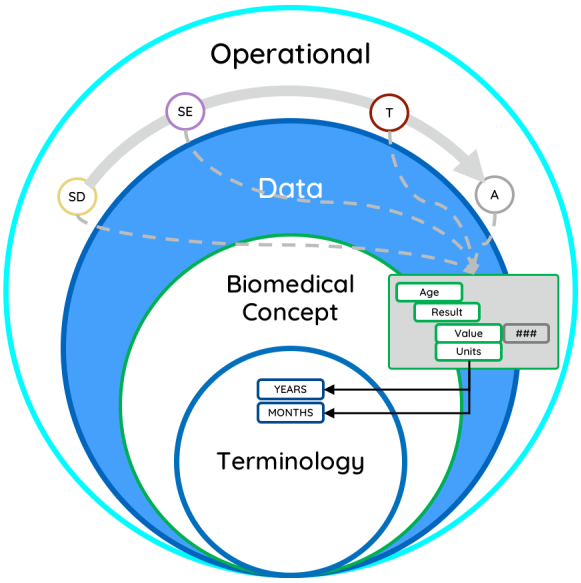

Figure 9 - Single Definition

Another aspect illustrated by Figure 8 is our current way of working. We need something new, usually something in SDTM. We design the solution within that SDTM silo, rarely considering the other parts of the process. This is improving but the author feels we have some way to go. Building common definitions would be a better way. We first model the real world, what is the nature of the item we are dealing with. The second step is to determine how we use it operationally across the lifecycle. This can be made easier by having templates and standard mechanisms for the BCs.

SUMMARY

There is a need to organise our world better, organise around the data that we collect and not see our world purely as the ways in which we wish to report it. To this end, we need to develop a series of models, loosely coupled but with well-defined connectivity that allow us to build better, more rigorous standards. The models must be shared, common across the industry allowing sponsors, CROs, vendors and others to work with common understanding that allow tools to collaborate without the need to perform expensive integrations and mapping. The models must be non-proprietary. It should be remembered we are guardians of the standards; no-one person or organization owns them – they were created by the community and are owned by them.

NEXT STEPS

References [1], [3] and [6] have shown some of these models already implemented and working. One significant effort is the PhUSE Clinical Trials as RDF project [5] that contains some interesting ideas, in particular the approach to modelling observations, adverse events and other such concepts. The project is an example of trying to understand our world better and developing models that the Author believes can make a significant contribution to the models that are described within this paper.

As noted earlier, there are so many standards. We need to align, understand what each offer and build that model that enables us to take the next step. The author has already taken some early steps in this direction and there are Initiatives within sponsor organisations. It is early days but there is promise.

REFERENCES

[1] Iberson-Hurst, PhUSE US Connect 2018, It's Time To Change.
[accessed 2019 Jan 10]
https://phusewiki.org/docs/2018_US%20Connect18/RG%20STREAM/rg06%20final%20.pdf

[2] Iberson-Hurst, PhUSE 2015, CDISC Standards and the Semantic Web
[accessed 2019 Jan 10]
https://www.lexjansen.com/phuse/2015/tt/TT09.pdf

[3] Langendorf. PhUSE US Connect 2018, SI19. Easing Your Pain with Biomedical Concepts
[accessed 2019 Jan 10]
https://phusewiki.org/docs/2018_US%20Connect18/SI%20STREAM%202/si19%20final%20.pdf

[4] BRIDG. Biomedical Research Integrated Domain Group
[accessed 2019 Jan 10]
https://bridgmodel.nci.nih.gov

[5] PhUSE White Paper. Clinical Trials Data as RDF
[accessed 2019 Jan 12]
https://www.phusewiki.org/docs/Deliverables/Clinical%20Trial%20Data%20as%20RDF%20White%20Paper%20Final%20.pdf

[6] Iberson-Hurst, PhUSE EU Connect 2018, SI12, Into the Fire, Linking CDISC & FHIR
[accessed 2019 Jan 12]
https://www.phusewiki.org/docs/Frankfut%20Connect%202018/SI/Papers/SI24-si12-v1-19314.pdf

[7] CDISC. Clinical Data Interchange Standards Consortium
[accessed 2019 Jan 12]
https://www.cdisc.org

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