Have Data Standards – Now What?

Next Steps Required to Put Them to Work to Fuel the Furnace of Data Innovation

GREG STEFFENS

NOUMENA SOLUTIONS

Case Study Summary

- Process Steps in Paul's Project
  - Create metadata from the study data
  - Compare metadata content across studies to identify differences
  - Create the initial define files – version 1.0 2.0 and html
  - Study programmer populates update metadata with primary key flags and comments describing derivation descriptions
  - Compare the updated metadata to the initial metadata
  - Document SAS program files in an html file
  - Perform quality check of the SAS log files
  - Validate the xml files with Xerces and Pinnacle21
- Results were very positive, thanks to a collaborative working environment, a great study programmer (Sunil Pusarla) and effective technology

Define File Improvements in Version 2.0

- Both versions of define.xml were created
- Version 2.0 was used in the submission. Version 1.0 should never be used
- A define file in html format was created to provide more information and an easier navigation to find information. This was converted to pdf format for the submission
- Version 2.0 fixes a flaw in the 1.0 version that prevented VLM
- Version 2.0 also adds a number of variable attributes and is over twice as large as 1.0
- Any software that does not support VLM and these attributes does not really support 2.0. VLM should be added for SDTM findings domains and ADaM data sets with PARAMCD

Higher Objectives

- The Omeros project is a good example of one kind of process which demonstrates the value of retrospectively built metadata
- The topic, today, is what else should we be doing to maximize the value of standards
- CDISC objectives focus on the submission
- Our objectives should go beyond that to operational efficiency
  - A higher and more consistent quality
  - Done faster with less time
-

Creating Metadata Earlier in the Process

- Creating metadata at the start of study planning has several advantages.
  - The define file would also be created from the metadata earlier.
  - The define file can function as the detailed programming specification
  - As issues are discovered the metadata content is improved
  - Production and validation programmer using the define as a specification
  - The completeness and correctness is much more likely
  - The metadata can be a single storage of information repeated in protocol, SAP, and programming specification
  - Identification of standards nonconformance can be programmatically done
  - Inconsistencies across studies can be programmatically identified and fixed

How to Create Metadata Earlier

- The study data doesn't exist
- So, metadata can be created from the standards that are stored in the same metadata design. Providing studies with metadata content is a big advantage of having data standards
- To do this, the same metadata design should be used everywhere
- Not adopting a standard metadata causes problems
  - CDISC had this problem when they didn't define standard metadata (excel worksheet columns) for the subteams defining SDTM, ADaM, CDASH, etc. We, in pharma, needed to write different programs to import these inconsistent worksheets. When CDISC implemented and MDR, there was a great deal of work to do to bring all the standards to a single metadata

Add More Information in Metadata

- More metadata, than CDISC has in the define file, is required to attain operational efficiency
- Metadata needs to describe to flow from data state to data state and not only the data states themselves
- The industry needs a standard relational metadata design that supports operational efficiency and not just esubmissions
- A robust metadata that describes the data flow, can be used to create study metadata from standards metadata based on the raw data metadata. If you know the variables you are collections, you can programmatically determine SDTM and ADaM metadata

Dimensions of Data

- Many people think 2-dimensionally about data – variables and data sets
- But data is 4-dimensional
  - Variables, data sets, row-level or VLM, and flow from state to state
- Some people claim problems result from a square, relational, 2-dimensional data design
- But the problems are a result of 2-dimensional thinking
- A 4-dimensional design can resolve our problems, even over a relational physical data design
- Don't be mislead and enticed by new kinds of technology to solve the problems,
  - Recall xml was going to solve all our problems?
  - Now its RDF/XMl, or AI or big data or …

Metaprogramming Component

- Having data standards stored in standard metadata design goes a long way.
- But the real value comes when we add metaprogramming
- Operational efficiency appears at the convergence of good process, robust metadata design, metadata-resident standards and metaprogramming
- Take one of more of these away and the value diminishes

Conclusion So Far

- We can all benefit from an industry standard, relational metadata design
- One metadata design shared by all
  - CDISC published data standards in this format
  - Data specifications between pharma and CROs and third-party data providers
  - Company data standards
- Easily accessed from SAS programs
- Much more information from the CDISC MDR in this metadata
- Same metadata design across all CDISC standards

Other Study Information Data

- Other information about studies, stored in standard data sets, will have similar benefits as data metadata
- Trial design
- Schedule of events
- Treatment arms
- Visit definitions
- Data deidentification
- SAS log scan results
- Etc.