What is high quality study metadata?

Sergiy Sirichenko

PhUSE Annual Conference
Barcelona, 2016

PINNACLE 21

Topics

- What is study metadata?
- Trial Design domains
- Reviewer's Guides
- aCRF
- Define.xml
- Conclusion

What is study metadata?

- "*data about data*"
- "*physical data and knowledge-containing info about business, tech processes, and data, used by corporation*" [1]
- 2 types of metadata
  - "*physical data*" that is stored in software and other machine-readable media
  - "*knowledge*" retained by employees and contained in other media

Study metadata in regulatory submission

- Trial Design domains
- Annotated Case Report Forms (aCRF)
- Reviewers Guides
- Define.xml
- Additional documents

- Study metadata made available to reviewers is limited to what included into submission, while highly utilized company internal knowledge is often not documented

CDER Technical Conformance Guide [2]

- *"The data definition file describes the metadata of the submitted electronic datasets, and is considered arguably the most important part of the electronic dataset submission for regulatory review".*

- At the same time, *"An insufficiently documented data definition file is a common deficiency that reviewers have noted".*

FDA Janus CTR case

- Since 2014 studies received FDA Jump Start service were uploaded into Janus CTR [3]
- 77% of all studies failed to load on first attempt
- There are many different reasons for the various load failures
- A missing or issue-laden Define.xml files were a big contributor

Metadata domains

- While most study metadata is represented by define.xml file and PDF documents, there are special standard Trial Design domains

Domain	Description
TA	Trial Arms
TD	Trial Disease Assessments*
TE	Trial Elements
TV	Trial Visit
TI	Trial Inclusion/Exclusion Criteria
TS	Trial Summary

- * Introduced in SDTM IG 3.2

Protocol info

- TA, TD, TE, TV, TI store information about study Protocol visits, treatment and disease assessment schedules, and subject screening criteria
- TS domain contains a short, high-level representation of study Protocol
  - TS is especially important for automation
  - It's the only machine-readable source for
    - Trial Indication, Diagnosis Group, Trial Phase Classification, Trial Title, Trial Type, Pharmacological Class of Investigational Therapy, Clinical Study Sponsor, and other key protocol characteristics

Reviewers Guides

- Relatively new type of study metadata developed by PhUSE
- Rapidly adopted by industry
- Valued by reviewers
  - 30 pages of high level “executive summary” of study metadata
- Study Data Reviewer Guide (SDRG) [4] - 2013
  - high-level summary and additional context for submission data package
  - purposefully duplicates information found in other submission documentation
  - single point of orientation for reviewers to the submission data

SDRG

- Additional information about
  - mapping decisions
  - sponsor-defined domains
  - study specific implementation
  - sponsor extensions to CDISC controlled terminology
- Sponsor's explanations of data validation issues
  - specifically the reason why those issues were not addressed during study conduct, mapping, and submission preparation

ADRG

- Analysis Data Reviewer Guide (ADRG) [5] – 2014
- A structure and expected content of this document are specific to analysis ADaM data
  - list of CORE variables
  - description of SAS® programs

- Overall, quality of Reviewer's Guides have been improving, however a number of common issues are still observed

Issues with Reviewers Guides

- Not following the recommended structure
  - Missing expected sections reduce value for reviewers
- Missing or meaningless explanations for data conformance issues
  - Outdated versions of OpenCDISC / P21 Validator
  - Examples of invalid explanations
    - *"Expected result"*
    - *"This is our common practice"*
    - *"As received from our vendor"*
    - *"Sponsor decided not to fix"*
    - *"We did not collect nor derive this data element"*
    - *"We do it differently than the standard"*

Generic invalid explanations

- Issue: "*Duplicate records*" in PP domain
- Sponsor explanation: "*The validation rule does not include PPORRES when determining the uniqueness of records. Accordingly, we consider these to be false positive warnings*"
- PPORRES is not a Key Variable in PP domain according to Sponsor's define.xml file
- An actual reason for duplicate records validation warnings is that PP structure in this pre-clinical study based on POOLID, while P21 check relies only on USUBJID

Generic invalid explanations

- Similar issue: "*Duplicate records*" in FW domain
- Sponsor explanation: "*The validation rule does not include FWORRES when determining the uniqueness of records…*"
- FWORRES is not a Key Variable in FW domain according to Sponsor's define.xml file
- A reason for these false-positive validation messages is that in this study FW domain utilized FWDY variable for Timing info, while P21 Validator uses other generic Timing variables to duplicate records (FWDTC, VISITNUM, FWTPTNUM)
- Explanation must be study specific and real!

Document formatting issues

- The following format issues are an immediate indication of lack of attention for this document by sponsor
  - inconsistent fonts or their size
  - missing or incorrectly working hyperlinks
  - different formats used across tables
  - unnecessary text brakes in table cells across pages
  - invisible or odd special characters copied from other documents, etc.
- Poor format almost always correlates with poor content

New documents from PhUSE

- Study Data Standardization Plan (SDSP) [6]
- Legacy Data Conversion Plan & Report (LDCP) [7]
- Driven by FDA need defined in TCG [2]
- The initial versions of these documents are expected in 2016

Annotated CRFs

- Represent data collection and SDTM mapping processes
- Metadata provided in aCRF is quite reliable, however there are few issues that sponsors should be aware of and fix before submission
  - Misspelling in variable name
  - Missing annotations
  - Mostly in SUPPQUAL domains due to “last-minute” modification in mapping specs
  - ~10-15 in a study

aCRFs

- Invalid mapping to EDC variables
- Missing annotations

- A year ago, FDA guidance documents changed the requested name for aCRFs from "*blankcrf.pdf*" to "*acrf.pdf*"
- Nevertheless, about 50% of submissions to FDA currently still use old name

Define.xml

- Describes datasets
- Based on Define-XML standardized format
  - This standardized machine-readable format allows the detailed study metadata to support automation
- Low quality of define.xml file makes it unusable by computers and by people
- Today define file is the *most overlooked part* of submission data package
- There are still many technical errors in define.xml files
- However, the most severe problem is inadequate content

V1.0 must die

- Define-XML v1.0 is outdated standard
  - Created as “last-minute” metadata fix for SDTM IG 3.1.1
  - Cannot handle Value Level
    - Important in Analysis data!
  - Lack of specific requirements for the capture of data origins resulted in common errors like:
    - Missing Origin
    - Origin=”CRF”, but no reference to particular page(s)
    - Inconsistency between origin and derivation (ex: Origin=”CRF Page” and ComputationMethod populated)
    - Origin=“Derived” without detailed derivation algorithm

Define-XML v2.0

- Released in 2013
- Resolved most limitation of v1.0
- More robust and is better suited to support current reviewer's needs (e.g., ARM)
- However, the industry has been very slow to implement Define-XML v2.0
- New FDA TCG recommend use of v2.0 as "preferred version"
- Recently FDA announced that the support for version 1.0 will end for studies that starts 12 months after March 15, 2017 [8]

Technical Issues

- Inconsistency in Character Case and use of special characters breaks XML, which is case-sensitive
  - For example, “*NO*”, “*No*”, and “*No* “ are three different values in XML

Duplicate order of Items

- For example, two different CodeList terms have the same OrderNumber:

```
<CodeList OID="CL.SEX" Name="Sex" DataType="text">
    <EnumeratedItem CodedValue="F" OrderNumber="1">
       <Alias Name="C16576" Context="nci:ExtCodeID"/>
    </EnumeratedItem>
    <EnumeratedItem CodedValue="M" OrderNumber="1">
       <Alias Name="C20197" Context="nci:ExtCodeID"/>
    </EnumeratedItem>
    <Alias Name="C66731" Context="nci:ExtCodeID"/>
</CodeList>
```

Inconsistent use of Decode attribute

- for some items within the same CodeList
  - results in ignoring items (terms) with missing Decode attribute
  - for example, the second term *"SAMPLE"* will be ignored by most tools including browsers and P21 Validator

```
<CodeList OID="CL.LBTESTCD" Name="Laboratory Test Code"
DataType="text">
  <CodeListItem CodedValue="ALB" OrderNumber="1">
    <Decode> <TranslatedText xml:lang="en">Albumin</TranslatedText>
    </Decode>
    <Alias Name="C64431" Context="nci:ExtCodeID"/>
  </CodeListItem>
  <CodeListItem CodedValue="SAMPLE" OrderNumber="2"
    def:ExtendedValue="Yes"/>
</CodeList>
```

Technical Issues

- Usage of CodeList or any other object (variable, comment, method, etc.) without defining it
- Opposite case when CodeList (or other object) is defined, but not used
- Improper utilization of dedicated elements for particular type of metadata
  - Comments are used instead of
    - Codelists
    - Computational Methods for Derived variables
    - ExternalCodelist for providing info about coding dictionary (MedDRA)

Recommendations

- Always refer to Standards documentation
- Use specialized tools for Define.xml
  - friendly interface for business users instead of direct editing of XML text
- Remember, that FDA requires
  - validation of Define.xml file
  - all technical issues must be fixed before submission

Missing Codelists

- While technical issues are critical for reading Define.xml files, it's the content deficiencies that are most commonly observed problems
- Missing Codelists for study specific data elements
  - sponsors populate Codelists only for variables that have standard CDISC Control Terminology (AEACN), but do not create study specific Codelists
    - For example, for Category (--CAT), Subcategory (--SCAT), Severity for Clinical Events (CESEV) or EPOCH variables

Missing or incorrect codelists

- Missing Codelists for Value Level metadata
  - SUPPQUAL domains are typically described using value level metadata, but sponsors often leave out Codelists for supplemental qualifiers that have controlled terminology
- Codelists created for variables collected as a free text
  - Codelists in Define.xml should describe data collection process
  - We recommend creating Codelists only for variables where data was collected, derived or assigned based on a list of pre-specified terms
  - In most cases study data Codelists with more than 30 terms are impractical and are never used. Exceptions are QNAM, --TESTCD, PARAMCD variables

Collapsed Codelists

- Collapsed Codelists for multiple variables across domains
  - For example, a single (UNIT) Codelist for all --ORRESU, --STRESU and --DOSU variables within a study
  - In some studies, such collapsed (UNIT) Codelist can result in >500 terms assigned to EXDOSU variable, while in reality EXDOSU variable only used one term “mg”
- We strongly recommend creating a separate Codelist for each variable
  - For example, (EXDOSU), (LBSTRESU), etc.
  - Exception is when Codelists for variables are identical

Missing, unclear or invalid Computational Algorithms

- All "*Derived*" variables must have clear and detailed description of Computational Algorithms
  - so reviewers can understand how values were derived and can independently reproduce them if needed
- However, majority of submissions still have missing or poorly documented Computational Algorithms
  - Quite often sponsors provide "generic" algorithms for Study Day and Baseline Flag variables, but do not provide any information for important study specific derivations like EPOCH, SESTDTC, RFPENDTC, etc.

Missing descriptions for study and sponsor specific variables

- --SPID (Sponsor ID), --GRPID (Group ID), etc.
  - Often these sponsor-specific variables are part of the dataset Key Variables
  - However, if sponsor did not fully describe these variables (e.g., meaning, source, computational algorithms, etc.), then there is no way to understand the submitted data
- The biggest value of Define file is to provide descriptions for study specific data elements
  - But unfortunately some sponsors just copy CDISC notes from SDTM IG in place of providing the important study specific metadata

A need for high quality define.xml

- Unfortunately, current level of industry compliance and quality of define.xml is very low
- Define.xml file is not ready to be used as a source of reliable machine-readable metadata
  - For example, P21 Validator cannot rely on define.xml. It has switched to manual entry of MedDRA info and uses “generic” Key Variables in datasets for duplicate records checks

Invalid Key Variables

- Usage of --SEQ variables, which are surrogate key representing artificial identifier
  - "*USUBJID, AESEQ*" – invalid metadata
  - "*USUBJID, AETERM, AESTDTC*" – expected metadata
- Usage of too many variables as Key Variables in dataset
  - "*USUBJID, AETERM, AEDECOD, AELLT, AEHLT, AESOC, AESEV, AESER, AEREL, AESHOSP, AESTDTC, AEENDT, VISIT*"
- Usage of --REFID, --SPID variables without any details about them in define.xml file

Artificial Keys

- Usage of --SPID variable as artificial surrogate key
  - Such approach does not explain what is a source for duplicate records and how to analyze data. For example,
    - --SPID is a Key Variable
    - Comment/derivation in define.xml: *"--SPID variable was populate to ensure uniqueness of Key Variables"*
  - This metadata is not much different from missing one

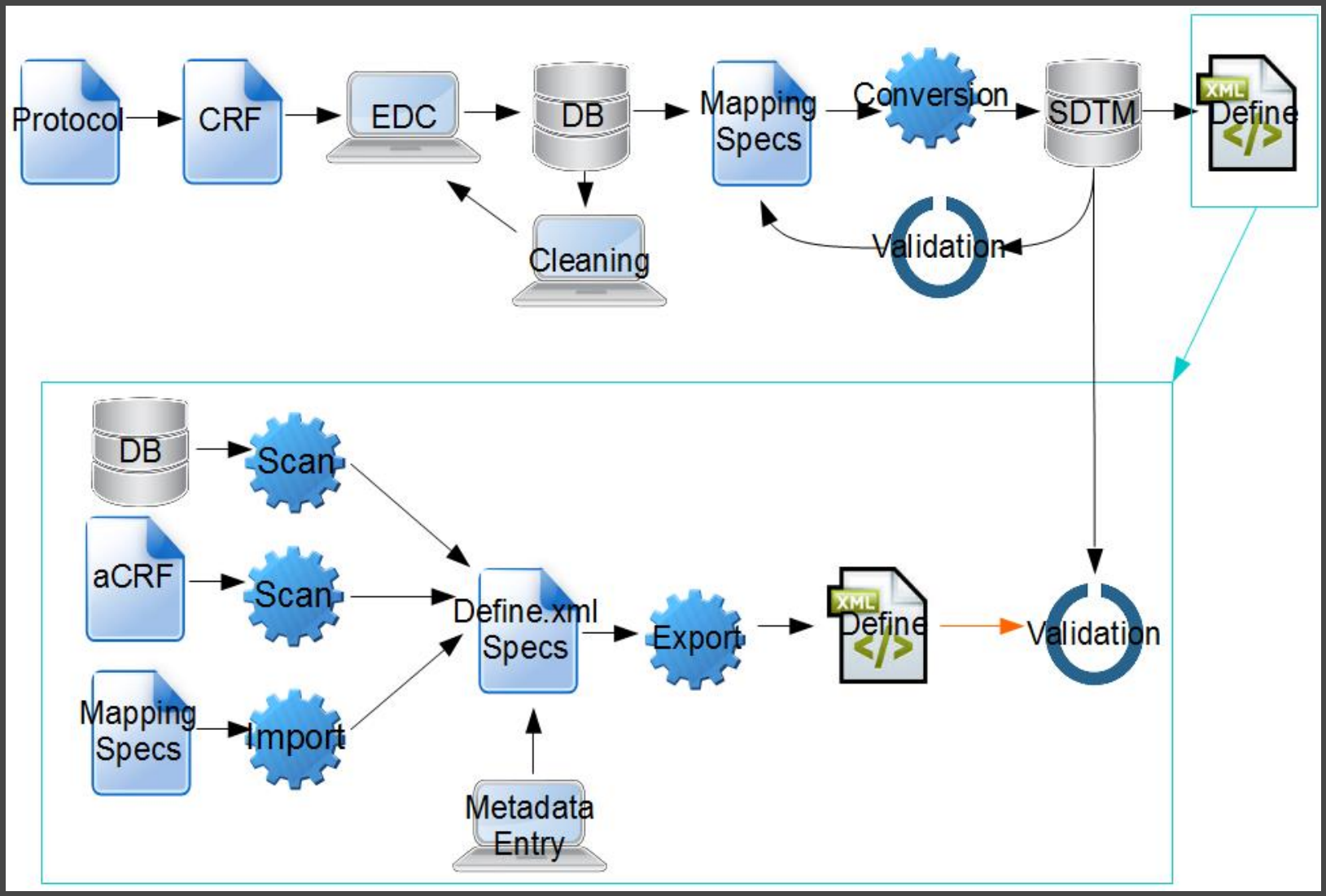
Quality of study metadata

- Today, quality of different types of study metadata varies significantly
- Usually the quality of aCRFs and SDRGs are much better than quality of Define files
- We believe the major reason for this discrepancy is due to the low utilization of Define file by the industry

Low utilization of Define

- The aCRFs are used internally for mapping and SDTM programming
- SDRGs are prepared to improve communication with reviewers
- Define files, on the other hand, are typically only created descriptively at the very last moment before submission
- Define file is not actually utilized by programmers or other users within a company

Process for descriptive define.xml



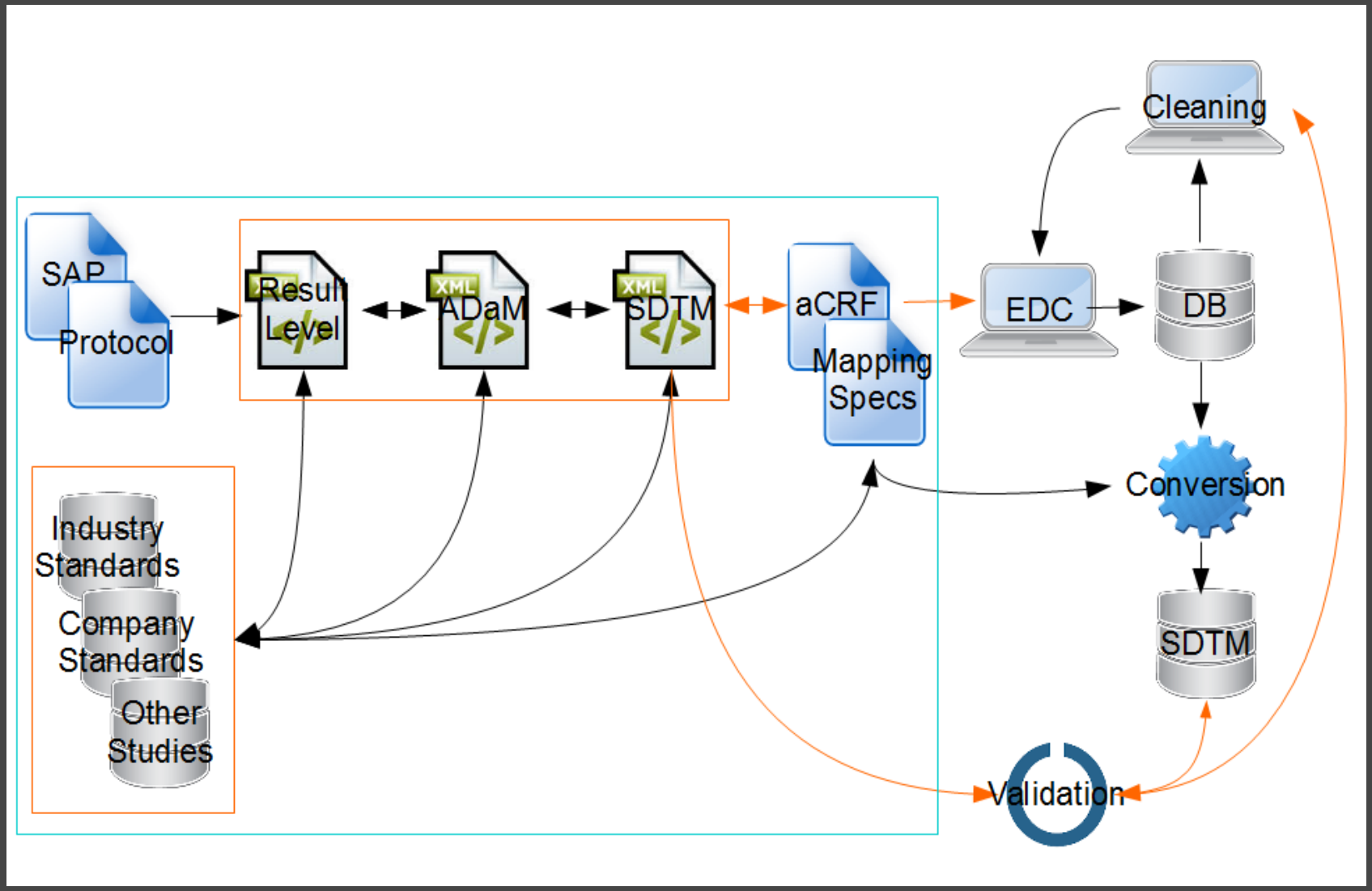
Solution for improving quality of define.xml

- Define.xml should be used actively, thus creating demand for higher quality
- We recommend exploring options to create Define file in advance and use it as a source of specifications for study data (*prescriptive approach*)
- There are many potential benefits to utilize Define-XML as a foundation for company specific metadata

Define-XML as foundation for internal metadata standard

- Define-XML was developed as a standard for study metadata
- Adding new Elements and Attributes (Define-XML+) allows simple customization for company specific needs, but still keep all standard structure for automatic creation of define.xml file and metadata exchange across companies
- It may be easier to start with ADaM prescriptive Define.xml as specifications for Analysis data

Process for prescriptive define.xml



Define-XML Implementation Guide

- Obvious reason for low quality Define.xml file is a lack of knowledge about expected content in Define files
  - Many observed issues are due to lack of experience
- Industry needs “Define-XML 2.0 Implementation Guide”
  - similar to SDTM or ADaM Implementation Guides that already exist and are used as a primary reference in addition to SDTM and ADaM Models
- PhUSE started a new working group to develop Define-XML 2.0 Implementation Guide

Summary

- High quality study metadata is extremely important for regulatory review process
  - It allows reviewers to better understand study data. It also allows tools to rely on this metadata to automate review and analysis.
- Today, quality is different for Define.xml, aCRF, and Reviewer's Guide
  - with Define.xml being less compliant with regulatory expectations and requires special attention during submission preparation
- To ensure high quality study metadata a company should have a team of experts, the right tools, and a robust process

References

1. Marco, David. 2000. Building and Managing the Meta Data Repository: A Full Lifecycle Guide. New York: John Wiley and Sons
2. "Study Data Technical Conformance Guide". CDER. March 2016. Available at http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf
3. Allard, Crystal. "Common Errors in Loading SDTM Data to the Clinical Trials Repository. Why Getting it Right Matters" PhUSE SDE. December 2015. Available at http://www.phusewiki.org/docs/2015_California_SDE/3._CommonErrorsLoadingSDTMData_CAllard.pdf#page=4
4. "Study Data Reviewer's Guide Completion Guidelines v1.2". PhUSE. January 2015. Available at http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide
5. "ADRG Package v1.1". PhUSE. January 2015. Available at http://www.phusewiki.org/wiki/index.php?title=Analysis_Data_Reviewer%27s_Guide
6. "Study Data Standardization Plan". PhUSE. Available at http://www.phusewiki.org/wiki/index.php?title=Study_Data_Standardization_Plan_%28SDSP%29
7. Legacy Data Conversion Plan & Report". PhUSE. Available at http://www.phusewiki.org/wiki/index.php?title=Legacy_Data_Conversion_Plan_%26_Report
8. "Electronic Study Data Submission; Data Standards; Support End Date for Case Report Tabulation Data Definition Specification Version 1.0". Federal Register. March 2016. Available at https://www.federalregister.gov/articles/2016/03/17/2016-05958/electronic-study-data-submission-data-standards-support-end-date-for-case-report-tabulation-data
9. Pinnacle 21 Community. Available at www.pinnacle21.net/download
10. Pinnacle 21 Enterprise. Available at www.pinnacle21.net

Contact info:

Sergiy Sirichenko
ssirichenko@pinnacle21.net