PharmaSUG 2016 - Paper SS11

What is high quality study metadata?

Sergiy Sirichenko, Pinnacle 21, Plymouth Meeting, Pennsylvania
Max Kanevsky, Pinnacle 21, Plymouth Meeting, Pennsylvania

ABSTRACT

High quality study metadata is an important part of regulatory submission since it allows reviewers to interpret and understand submitted data, which means your submission can potentially move through the process more quickly. However, poor study metadata is most often cited by reviewers to be deficient. In fact, 77% of submissions in 2015 could not be loaded into FDA Clinical Trial Repository mostly due to issues with Define.xml and Trial Summary dataset. In this presentation, we will share the most common issues with study metadata in our industry and provide recommendations how to avoid or correct them to ensure successful regulatory submission.

INTRODUCTION

There are many definitions for metadata. The short and commonly used one in our industry is "*data about data*". A more detailed definition describes metadata as *"physical data and knowledge-containing info about business, tech processes, and data, used by corporation*" [1]. Study metadata in regulatory submissions includes Define.xml for study datasets, annotated Case Report Forms (CRF), Trial Design datasets, Study Data Reviewer Guide (SDRG), Analysis Data Reviewer Guide (ADRG), and newly introduced Study Data Standardization Plan (SDSP) and Legacy Data Conversion Plan (LDCP). It may also be supplemented by additional custom documentation. Study metadata does not only describe submitted datasets, but also provides study specific details to help reviewers understand content, purpose, usage, and limitations of collected and derived data.

There are two types of metadata, "*physical data*" that is stored in software and other machine-readable media, and "*knowledge*" retained by employees and contained in other media. In regulatory submissions, study metadata made available to reviewers is limited to physical data, while highly utilized internal knowledge is often not documented. It's important to remember that FDA reviewers do not have access to the same study metadata as sponsor teams and their data vendors.

Missing or poor quality study metadata in regulatory submissions may lead to different interpretation of study data and different results. Study metadata is a major communication mechanism between sponsor's data managers or programmers and FDA reviewers. Unfortunately, quality of study metadata is still often overlooked.

CDER Technical Conformance Guide [2] says that "*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*".

With broad adoption by the industry of data and metadata standards, new automatic tools were developed to take advantage awarded by these standards. However, the tools rely on high quality input. For example, since 2014, clinical trial data submitted to FDA that received JumpStart service has been loaded into Janus Clinical Trials Repository (CTR). The Janus CTR provides reviewers with easy access to this data, which, in turn, supports efficient regulatory review. However, for this to work, the data in question has to be both compliant and useful. Alarmingly, 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.

TRIAL DESIGN DOMAINS

While most study metadata is represented by "documents" in PDF and XML formats, there are a few special standard Trial Design domains designed for study metadata.

Domain	Description
TA	Trail Arms
TE	Trial Elements
TV	Trial Visit
TI	Trial Inclusion/Exclusion Criteria
TS	Trial Summary

Table 1. Trial Design domains

Trail Arms (TA), Trial Elements (TE), Trial Visits (TV), and Trial Inclusion/Exclusion Criteria (TI) domains store information about study protocol visits, treatment schedule, and subject screening criteria. While Trial Summary (TS) domain contains a short, high-level representation of study protocol. TS is especially important for automation, as 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.

While the industry has already adopted these domains and include them in submissions on regular basis, there are currently very few analysis tools that actually utilize trial design information.

REVIWERS GUIDES

Reviewer's Guides are relatively new type of study metadata developed by Association Programming Pharmaceutical Users Software Exchange (PhUSE). Study Data Reviewer's Guide (SDRG) was introduced in 2013 to provide FDA reviewers with a high-level summary and additional context for the submission data package. It purposefully duplicates information found in other submission documentation (protocol, clinical study report, annotated CRFs, define.xml, etc.) in order to provide FDA reviewers with a single point of orientation to the submission data [4]. Reviewer's Guide communicates additional information about mapping decisions, sponsor-defined domains, and sponsor extensions to CDISC controlled terminology. It also captures sponsor's explanations of data validation issues, specifically the reason why those issues were not addressed during study conduct, mapping, and submission preparation.

There is a rapid adoption of Study Data Reviewer's Guide by the industry, primarily due to its popularity with FDA reviewers, but also for its usability. On average, a Reviewer's Guide has only about 30 pages, which is a lot less than hundreds of pages across protocol, define.xml, and other documents.

An initial version of Analysis Data Reviewer Guide (ADRG) was introduced in 2014. A structure and expected content of this document are specific to analysis ADaM data [5]. For example, it includes sections for a list of CORE variables, a description of SAS® programs, etc. Overall expectations and recommendations for ADRG are similar to SDRG. Watch PhUSE Wiki website for new versions of Reviewer's Guide documentation.

Overall, they quality of Reviewer's Guides have been improving, however a number of common issues are still observed. The following is a summary of common issues:

Not following the recommended structure

The structure of the Reviewer's Guide must follow the recommended template provided by PhUSE (http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide). However, some sponsors only fill out parts of the template, which significantly reduces the document's value for FDA reviewers.

Missing or meaningless explanations for data conformance issues

The Data Conformance Summary section of the Reviewer's Guide provides an opportunity for sponsors to identify and explain in detail why some of the data issues were not fixed. This helps reviewers navigate around the data issues during analysis and preempts the need for additional question and clarifications.

Many sponsors still use outdated versions of Pinnacle 21 (formerly OpenCDISC), which results in some issues not being identified or explained. Since FDA always uses the latest version of Pinnacle 21 (a.k.a. DataFit) and validation rules, this leads to missing explanations in Reviewer's Guides.

Even if sponsors use current or recent versions of Pinnacle 21, in many cases the explanations provided are not sufficient. Here is a list of our favorite explanations that you should NEVER use in a Reviewer's Guide:

- *"Expected result"*
- *"This is a 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"*

Issues explanations that show incorrect interpretation of CDISC standards and FDA requirements
Often data conformance issue explanations show that a sponsor does not understand or has an incorrect interpretation of CDISC standards and FDA requirements. For example

- Issue: *“Date is after RFPENDTC” (in most domains, 10 – 60% of records)*
- Sponsor explanation: *“RFPENDTC is the last date of participation for a subject for data included in a submission. RFPENDTC is set to the latest DSSTDTC in DS domain where DSCAT=’DISPOSITION EVENT’*”

As you can see, a sponsor derivation algorithm deviates from the CDISC definition of RFPENDTC variable. Also, in this particular study sponsor used DSDTC variable instead of DSSTDTC for most Disposition Events including “Study Completion” records.

Document formatting issues

Another often-overlooked problem in Reviewer’s Guides is the low quality in document formatting due to

- 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.

Such format issue is an immediate indication of lack of attention for this document by sponsor. Poor format almost always correlates with poor content.

We recommend paying special attention to Reviewer’s Guide. Consider this as an additional opportunity to communicate with your FDA reviewers. Remember that Reviewer’s Guide is a product of collaborative work and input is expected from your entire team, including statisticians, programmers, data managers, clinical team and data vendors.

In addition to Reviewer’s Guides, PhUSE is working on two new documents for regulatory submissions. Study Data Standardization Plan (SDSP) [6] and Legacy Data Conversion Plan & Report (LDCP) [7] are driven by FDA need defined in Technical Conformance Guide TCG [1]. The initial versions of these documents are expected in 2016.

ANNOTATED CASE REPORT FORMS

An Annotated Case Report Form (aCRF) is a required PDF document, which represents data collection process and mapping of form fields to SDTM structure. aCRF is a main reference for reviewers to understand how data was collected. The overall quality of annotated CRFs is usually much better than the quality of define.xml files. Metadata provided in aCRF is quite reliable, however there are a few issues that sponsors should be aware of and fix before submission.

Based on our experience almost every study has about 10-15 errors in CRF annotations. It is quite challenging to perform 100% error free QC of a few hundred annotated pages versus actual data. Only automated tools can catch inconsistencies when comparing define.xml or data versus aCRFs. Here are some examples of common errors in annotations:

- Misspelling in variable name
- Missing annotations
- Annotations for MedDRA coding done in EDC system. MedDRA coding is formally “Assigned”, not collected

Most annotation errors are in SUPPQUAL domains. This is likely caused by “last-minute” modifications in mapping specifications without matching updates to annotations in CRF. Programmers should ensure that SDTM mapping team corrects annotations according to changes in mapping specifications.

Another common issue is that sponsors submit invalid aCRFs with missing annotations or with annotations to ‘*raw’* EDC database instead of SDTM variables. Such violation of regulatory requirements may result in FDA request for resubmitting study data package.

Also, 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 “*blankcrf.pdf*” instead of “*acrf.pdf*”.

DEFINE.XML

Define file describes datasets. For standardized data, define file is also expected to follow the 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.

Issues caused by Define.xml v1.0 limitations

For example, Define.xml v1.0 cannot adequately handle Value Level metadata. CDISC models are normalized and store different test observations (height, weight, etc.) in same variables. Thus, Value Level metadata is needed to provide sufficient detail for each observation (like allowed units or controlled terminology) to support data review and analysis. This is especially important for analysis data, which is why FDA requests that sponsors utilize Define.xml v2.0 instead, which corrects the Value Level limitations.

Another major limitation of Define.xml v1.0 that causes issues for reviewers is the lack of specific requirements for the capture of data origins. Was the data collected on CRF, derived, or received from laboratory? If collected on CRF, then on what pages? If derived, then with what method? Define.xml v1.0 only provided optional fields called “Origin” and “ComputationMethod”, but no clear requirements or controlled terminology on how to use them. This has resulted in the following common issues:

- 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 was released in 2013 and has resolved most of the prior version’s limitations. It’s more robust and is better suited to support current reviewer’s needs. However, the industry has been very slow to implement Define.xml v2.0, which is surprising considering that Define.xml v1.0 is almost as old as SDTM IG 3.1.1. Do you know many companies are still using SDTM IG 3.1.1? We highly recommend that industry upgrade to Define.xml v2.0 to take advantage of the new functionality that improves description and reviewability of submission data.

New FDA Technical Conformance Guide [1] recommends the usage of version 2.0 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

Regardless of what version of define.xml sponsors use, there are a few other common technical errors we typically observe with define files.

- 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 values 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 and will not be displayed in internet browsers using standard stylesheet:

```
<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>
```

- Another severe inconsistency violation is a usage of CodeList or any other object (variable, comment, method, etc.) without defining it. Another common issue is with the opposite approach when CodeList (or other object) is defined, but not used.
- Programmers should ensure a proper utilization of dedicated elements for particular type of metadata. For example, often Comments are used instead of computational Methods for Derived variables or ExternalCodelist for providing info about coding dictionary (MedDRA).

To avoid technical errors, we recommend always refer to Standards documentation. There are many specialized tools for define.xml, which can handle XML and Define-XML technical implementation and provide friendly interface for business users instead of direct editing of XML text [9], [10]. FDA requires validation of define.xml files and all technical issues must be fixed before submission.

Incorrect or 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 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. For example, if CMDOSU is collected using values from a drop-down menu in EDC system, it should reference a codelist in define.xml file. However, if CMDOSU was collected as free text, a codelist is not necessary as it will result in codelist with several hundred unique terms. We believe that in most cases study data codelists with more than 30-40 terms are impractical and are never used.
- Collapsed codelists for multiple variables across domains – for example, a single UNIT codelists for all --ORRESU, --STRESU and --DOSU variables within a study. In some studies, such collapsed UNIT codelists 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.
- Codelist submitted as a complete CDISC Control Terminology. For example, study data uses only one term for COUNTRY variable. However, COUNTRY codelist includes all 249 terms from (Country, C66786) Codelist in CT. CDISC CT is used as a source of standard terms. A variable codelist explains how data were collected.

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.

Sometimes in computational algorithms sponsors refer to non-available information like raw data from EDC system or external look-up conversion tables, additional documentation which is not included in submission data package. Please ensure that all Derived variables and Value Level have clear, correct and detailed computational algorithms, which only use data elements and information included in the data package.

Missing descriptions for study and sponsor specific variables

Missing descriptions for study and sponsor specific variables, like --SPID (Sponsor ID), --GRPID (Group ID), etc. is another severe issue in Define content. Often these sponsor-specific variables are part of the dataset Key Variables, responsible for "duplicate" records and play other important roles. 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 many sponsor 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. We already provided an example with FDA CTR where missing or issue-laden Define.xml is a major contributor to extremely high rate of failures when uploading study data into the system.

As of today, Define.xml file is not ready to be used as a source of reliable machine-readable metadata. For example, early versions of Pinnacle 21 Validator used the MedDRA version specified in define.xml file (if provided) for validating study datasets. However, such metadata-driven approach did not work well because in most cases the MedDRA version information was not implemented correctly. In latest Pinnacle 21 Community 2.1.1, the user must instead select the MedDRA version from the drop-down box in the tool interface.

Another case where a reference to study metadata would be beneficial is to check for duplicate records. In theory, a define.xml file should be a source for study specific Key Variables for each dataset. However, incorrect or invalid Key Variables are too common an issue to utilize Define file for validation. Here are common examples

- Usage of --SEQ variables, which are *surrogate key* representing artificial identifier. Only *natural keys* (with only few exceptions) are expected to be used to define Key Variables in datasets
  - "*USUBJID, AESEQ*" – invalid metadata
  - "*USUBJID, AETERM, AESTDTC*" – expected metadata
- Usage of too many variables as Key Variables in dataset. Such approach does not correctly explain data structure. For example,
  - "*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
- 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 with comment/derivation in define.xml *"--SPID variable was populate to ensure uniqueness of Key Variables".* This metadata is not much different from missing one.

Currently, Pinnacle 21 Validator can rely only on some generic Key Variables pre-defined as a part of check algorithms. In the future, when the industry is able to produce reliably high quality Define files, automated tool can fully utilize them as a source of study specific metadata.

The right approach for creating Define.xml

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 by the industry.

The aCRFs are used internally for mapping and SDTM programming, while 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.

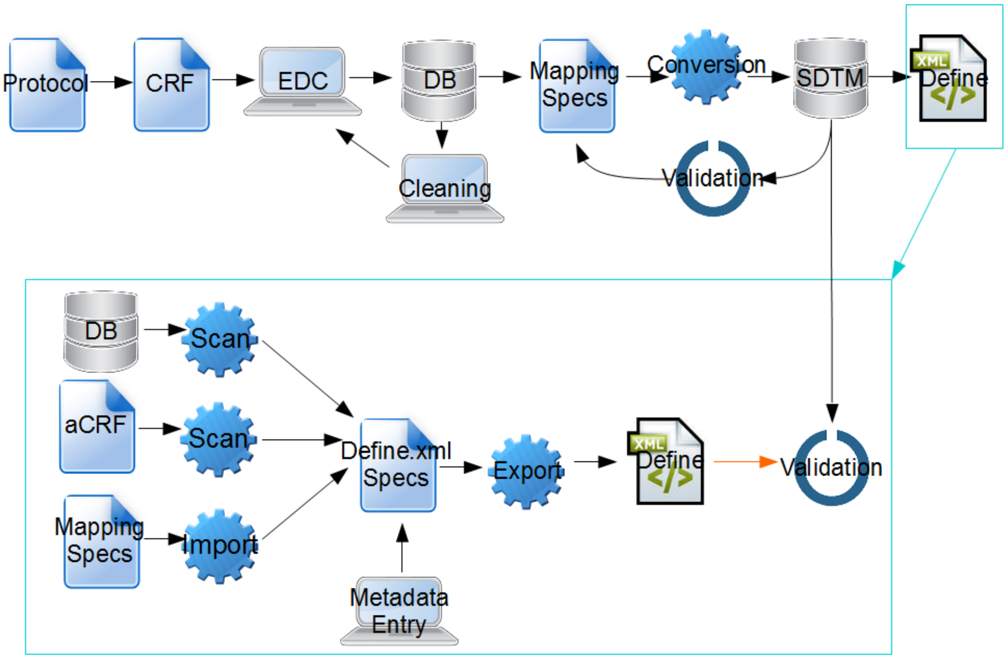

Figure 1. Process for descriptive define.xml

Therefore, we see only one solution for real improvement of Define files. They 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*). Define-XML was developed as a standard for study metadata. There are many potential benefits to utilize Define-XML as a foundation for company specific metadata. Adding new Elements and Attributed (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. Also, if your company programmers are able to use ADaM define.xml for creation of analysis data, then it's a good predictor that FDA reviewer can reproduce your programming as well.

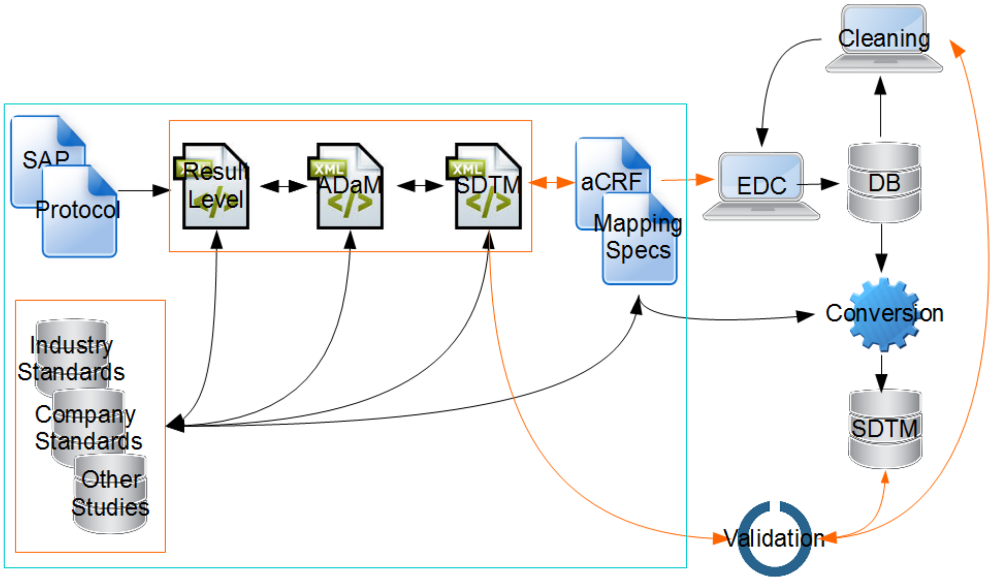

Figure 2. Process for prescriptive define.xml

Finally, another 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 Completion 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. During Computational Science Symposium 2016, PhUSE started a new working group to develop Define.xml Completion Guide. The initial document is expected be available in 2016.

CONCLUSION

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] OpenCDISC Community. Available at www.opencdisc.org

[10] Pinnacle 21 Enterprise. Available at www.pinnacle21.net

CONTACT INFORMATION

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

Name: Sergiy Sirichenko
Company: Pinnacle 21 LLC
Work Phone: 908-781-2342
E-mail: ssirichenko@pinnacle21.net

Name: Max Kanevsky
Company: Pinnacle 21 LLC
Work Phone: 267-331-4431
E-mail: mkanevsky@pinnacle21.net