How to Prepare High-quality Metadata for Submission

Varun Debbeti, Cytel Inc., Waltham, MA, USA

Paper SI12
Contents

• What is Metadata?
• How to Prepare High-quality Metadata
• Conclusion
• References
What is Metadata?

- Definition: “structured data about data”
- “Metadata is descriptive information about an object or resource, whether it is physical or electronic”.
- Description could be about content, origin etc.
How to Prepare High-quality Metadata

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues
# Knowledge on Standards and Latest Requirements

## FDA Data Standards Catalog v4.10 (10-24-2017) - Supported and Required Standards

This table contains a listing of the data exchange, file formats and terminology standards supported at FDA. These standards have gone through all the steps necessary to make this part of the regulatory review process, including posting of regulatory guidance documents and associated implementation guidelines and technical specifications. The submission of standardized data using any standard not listed, or to an FDA Center not listed, should be discussed with the Agency in advance. This catalog is incorporated by reference in the guidance to industry. Providing Regulatory Submissions in Electronic Format: Standardized Study Data (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/UCM49334.pdf).

<table>
<thead>
<tr>
<th>Use</th>
<th>Data Exchange Standard</th>
<th>Exchange Format</th>
<th>Standards Development Organization (SDO)</th>
<th>Supported Version</th>
<th>Implementation Guide Version</th>
<th>FDA Center(s)</th>
<th>Date Support Begins (MM/DD/YYYY)</th>
<th>Date Support Ends (MM/DD/YYYY)</th>
<th>Date Requirement Begins (MM/DD/YYYY)</th>
<th>Date Requirement Ends (MM/DD/YYYY)</th>
<th>Regulatory Reference and Information Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical study datasets</td>
<td>Study Data Tabulation Model (SDTM)</td>
<td>XPT</td>
<td>Clinical Data Interchange Standards Consortium (CDISC)</td>
<td>1.1</td>
<td>2.1</td>
<td>CDER, CBER</td>
<td>Ongoing</td>
<td>01/28/2015</td>
<td>01/28/2016</td>
<td></td>
<td>CDISC org - SDTM See Technical Conformance Guide</td>
</tr>
<tr>
<td>Clinical study datasets</td>
<td>SDTM</td>
<td>XPT</td>
<td>CDISC</td>
<td>1.2</td>
<td>Version 3.1.2 Amendment 1</td>
<td>CDER, CBER</td>
<td>08/07/2013</td>
<td>03/15/2019</td>
<td>03/15/2020</td>
<td>03/15/2020</td>
<td>CDISC org - SDTM</td>
</tr>
<tr>
<td>Clinical study datasets</td>
<td>SDTM</td>
<td>XPT</td>
<td>CDISC</td>
<td>1.2</td>
<td>3.1.2</td>
<td>CDER, CBER</td>
<td>10/30/2009</td>
<td>03/15/2019</td>
<td>03/15/2020</td>
<td>03/15/2020</td>
<td>CDISC org - SDTM</td>
</tr>
<tr>
<td>Clinical study datasets</td>
<td>SDTM</td>
<td>XPT</td>
<td>CDISC</td>
<td>1.3</td>
<td>3.1.3</td>
<td>CDER, CBER</td>
<td>12/01/2012</td>
<td></td>
<td></td>
<td></td>
<td>CDISC org - SDTM</td>
</tr>
<tr>
<td>Clinical study datasets</td>
<td>SDTM</td>
<td>XPT</td>
<td>CDISC</td>
<td>1.4</td>
<td>3</td>
<td>CDER, CBER</td>
<td>08/17/2015</td>
<td>03/15/2019</td>
<td>03/15/2020</td>
<td></td>
<td>CDISC org - SDTM</td>
</tr>
<tr>
<td>Clinical study datasets</td>
<td>Analysis Data Model (ADaM)</td>
<td>XPT</td>
<td>CDISC</td>
<td>2.1</td>
<td>1.0</td>
<td>CDER, CBER</td>
<td>Ongoing</td>
<td>03/15/2019</td>
<td>03/15/2020</td>
<td></td>
<td>CDISC org - ADaM</td>
</tr>
<tr>
<td>Clinical study datasets</td>
<td>Analysis Data Model (ADaM)</td>
<td>XPT</td>
<td>CDISC</td>
<td>2.1</td>
<td>1.1</td>
<td>CDER, CBER</td>
<td>03/15/2018</td>
<td>03/15/2019</td>
<td>03/15/2020</td>
<td></td>
<td>CDISC org - ADaM</td>
</tr>
<tr>
<td>Study data definition</td>
<td>Define</td>
<td>XML</td>
<td>CDISC</td>
<td>1.0</td>
<td>N/A</td>
<td>CDER, CBER, CORH</td>
<td>Ongoing</td>
<td>03/15/2018</td>
<td>03/15/2019</td>
<td></td>
<td>CDISC org - Define-XML</td>
</tr>
<tr>
<td>Study data definition</td>
<td>Define</td>
<td>XML</td>
<td>CDISC</td>
<td>2.0</td>
<td>N/A</td>
<td>CDER, CBER, CORH</td>
<td>08/07/2013</td>
<td>12/17/2016</td>
<td>12/17/2017</td>
<td></td>
<td>CDISC org - Define-XML</td>
</tr>
</tbody>
</table>
Knowledge on Standards and Latest Requirements

The Catalog lists
• currently supported and/or required standards
• the FDA support begin/end date
• the requirement begin/end date

Any study starting after 17DEC2016 for NDA, BLA and ANDAs should follow the set standards in order to avoid refuse to file (RTF).

https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM340684.xlsx
How to Prepare High-quality Metadata

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues
Metadata Preparation/Validation

Guidance documents that help in preparation and validation

- Define.xml schema validation
- Define.xml Pinnacle 21 validation
- FDA Business run/Validator rules
- SDTM-Metadata Submission Guidelines
- Define-XML-2-0-Specification
- eCTD Technical Conformance Guide
- Technical Rejection Criteria
- Study Data - Technical Conformance guide (TCG)
- Other manual checks
Define.xml schema validation

- First step is to validate the structure of the define.xml
- CDISC provides schema files to be used by validator tools [http://www.cdisc.org/define-xml](http://www.cdisc.org/define-xml)
- External websites for some XML tools evaluated by CDISC E.g.,
  - Oxygen XML Editor [http://www.oxygenxml.com](http://www.oxygenxml.com)
  - Cladonia XMLExchanger [http://www.exchangerxml.com](http://www.exchangerxml.com)
  - Validome [http://www.validome.org](http://www.validome.org)
Define.xml Pinnacle 21 validation

- It is recommended to validate datasets in relation to define.xml
- Pinnacle 21 community can validate
  - Datasets alone
  - Define.xml alone
  - Datasets in relation to define.xml

Study-data-reviewers-guide template from PhUSE (released as SDRG Package v1.2)
Define.xml Pinnacle 21 validation

Pinnacle 21 Community tool
FDA Business rules/Validator rules

• FDA Business rules: Published Mar, 2017
  • Developed to reinforce internal business processes.
  • Supersede previously published validation rules.

• FDA validation rules: Published Nov, 2014
  • The validator rules are technical version of business rules
  • The validator rules were published for transparency purposes.

• One may not have to read or remember all these rules. Tools like P21 cover these issues under its rules set.
# FDA Business rules/Validator rules

<table>
<thead>
<tr>
<th>FDA Business Rule ID</th>
<th>FDA Business Rule</th>
<th>FDA Validator Rule</th>
<th>Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDAB035</td>
<td>The definition of datasets, variables, and codelists in define.xml must reflect the actual study data.</td>
<td>Datasets included in study data must be described in the data definition document (define.xml).</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Domains referenced in data definition document (define.xml) should be included in the submission.</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variables listed in the data definition document (define.xml) should be included in the dataset.</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variables included in the dataset must be described in the data definition document (define.xml).</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variable Data Types in the dataset must match the variable data types described in the data definition document (define.xml).</td>
<td>ALL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variable values should be populated with terms found in the user-defined codelist associated with the variable in define.xml.</td>
<td>ALL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pinnacle 21 ID</th>
<th>Publisher ID</th>
<th>Message</th>
<th>Description</th>
<th>Category</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD1053</td>
<td>FDAC023</td>
<td>Dataset is not present in define.xml</td>
<td>Datasets included in study data must be described in the data definition document (define.xml).</td>
<td>Metadata</td>
<td>Error</td>
</tr>
<tr>
<td>SD0061</td>
<td>FDAC024</td>
<td>Domain referenced in define.xml but dataset is missing</td>
<td>Domains referenced in data definition document (define.xml) should be included in the submission.</td>
<td>Metadata</td>
<td>Warning</td>
</tr>
<tr>
<td>SD0054</td>
<td>FDAC025</td>
<td>Variable in define.xml is not present in the dataset</td>
<td>Variables listed in the data definition document (define.xml) should be included in the dataset.</td>
<td>Metadata</td>
<td>Warning</td>
</tr>
<tr>
<td>SD0056</td>
<td>FDAC026</td>
<td>Variable in dataset is not present in define.xml</td>
<td>Variables included in the dataset must be described in the data definition document (define.xml).</td>
<td>Metadata</td>
<td>Error</td>
</tr>
<tr>
<td>SD0059</td>
<td>FDAC034</td>
<td>Define.xml/dataset variable type mismatch</td>
<td>Variable Data Types in the dataset must match the variable data types described in the data definition document (define.xml).</td>
<td>Metadata</td>
<td>Error</td>
</tr>
<tr>
<td>SD0037</td>
<td>FDAC037</td>
<td>Value for variable not found in user-defined codelist</td>
<td>Variable values should be populated with terms found in the user-defined codelist associated with the variable in define.xml.</td>
<td>Terminology</td>
<td>Error</td>
</tr>
</tbody>
</table>
FDA Business rules/Validator rules

• While most of the business rules/validator rules are covered by P21, there are a few issues that are not covered

• Issue: **Sequence of model permissible variables** in data which are **not matching** in define.xml are not found by P21 community

• e.g., When QSMETHOD is added to QS domain, and the order of QSMETHOD in define is different in define and dataset, it is not identified by P21 community tool.
SDTM-Metadata Submission Guidelines

• Purpose: To provide guidance for compiling the eCTD module 5 “sdtm” folder.

• Domain Name vs Dataset name
  • A dataset is just a collection of data with rows and columns.
  • A domain is defined as per the SDTM model.

• As per SDTM IG 3.2 “A domain may be comprised of more than one physical dataset” For e.g., QS domain = QS and SUPPQS datasets.

• Split domains
  • Annotated CRF: The annotated CRF refers to the domain name (QS) as opposed to the dataset name (QSCG, QSCS, or QSMM).
  • DOMAIN variable value should be QS (i.e., DOMAIN = “QS”) in all split datasets e.g., QSCG, QSCS or QSMM.
  • the define.xml controlled terminology-code list for Domains to include 4 letter split dataset names.

• Advise: Sponsor should check with their review agency regarding exactly what needs to be included in the submission, i.e. the split datasets or both the split datasets and the un-split datasets.
Define-XML-2-0-Specification

CDISC Define-XML Specification
Version 2.0

Prepared by
CDISC Define-XML Team

• Purpose: The specification describes the Define-XML model that is used to describe SDTM/ADaM datasets for submission to FDA
Define-XML-2-0-Specification

• Dataset level: Natural keys describe the structure of each dataset. “A natural key is a piece of data that uniquely identify and distinguish observation from any other observation in the dataset.”

• Natural keys have business meaning, natural keys are often coupled to the business. Example:
  • Sponsor-A natural keys: STUDYID, USUBJID, VISTNUM, PETESTCD
  • Sponsor-B natural keys: STUDYID, USUBJID, PEDTC, PETESTCD, PELOC, PEMETHOD
Define-XML-2-0-Specification

• Natural keys may not always match standard keys used

• Types of duplicate records due to fixed standard keys
  • True double entry: Two or more records with same keys (USUBJID, VISTNUM, PETESTCD) and only PEORRES is different
  • Insufficient standard Keys: The above example of PE falls under this category, and should be added in define.xml for easy identification by reviewers.
  • Incorrect standard Keys: Example: Sponsor Key variables USUBJID FACAT FASCAT FATESTCD VISITNUM, standard keys are USUBJID FATESTCD FAOBJ –should be fixed before submission
  • Missing values in keys: Example: Standard keys are STUDYID USUBJID DATESTCD and DADTC in DA domain. If the data has missing values of DADTC, multiple records will have same DADTC (i.e., missing) as per Pinnacle tool and are identified as duplicates.
Study Data - Technical Conformance guide (TCG)

• **Traceability Issues:** Traceability permits an understanding of the relationships between the analysis results and datasets.

• Report → Define.xml → ADaM → SDTM → CRF.

• Traceability has been a proven issue for Legacy data conversions.

• **Legacy Data:** Study data in a format that is not supported in FDA review process, and not ever listed in the FDA data standards catalog.

• Sponsors should use processes for legacy data conversion that account for traceability.

• As per TCG, there are issues like “Limited traceable path from SDTM to the ADaM datasets” and/or “Limited ability to replicate ADaM datasets using SDTM datasets” when legacy study data and legacy analysis data are independently converted to SDTM and ADaM formats, respectively.
Study Data - Technical Conformance guide (TCG)

• A tool to check if all raw data is mapped to SDTM was developed.
• It reads all existing SDTM logs to find input data library references.
• An excel file output will be generated with columns for ‘Raw FILE NAME’ and ‘SDTM DOMAIN’.
• Refer to paper for code.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FILENAME</td>
<td>NAME</td>
<td>TYPE</td>
<td>LABEL</td>
<td>FORMAT</td>
</tr>
<tr>
<td>1</td>
<td>AB</td>
<td>ASSAYNO</td>
<td>char</td>
<td>Assay Number</td>
<td>$20.</td>
</tr>
<tr>
<td>2</td>
<td>AE</td>
<td>ACNOTSP</td>
<td>char</td>
<td>Other specify</td>
<td>$600.</td>
</tr>
<tr>
<td>3</td>
<td>AETARG</td>
<td>ACNOTSP</td>
<td>char</td>
<td>Other specify</td>
<td>$600.</td>
</tr>
<tr>
<td>4</td>
<td>ANTIBSA</td>
<td>ASSAYNO</td>
<td>char</td>
<td>Assay Number</td>
<td>$28.</td>
</tr>
<tr>
<td>5</td>
<td>ANITITRT</td>
<td>ASCT</td>
<td>char</td>
<td>Conditioning chemotherapy for ASCT</td>
<td>$9.</td>
</tr>
<tr>
<td>6</td>
<td>BSYPMT</td>
<td>DATAPAGEID</td>
<td>num</td>
<td>Internal id for data page</td>
<td>10.</td>
</tr>
<tr>
<td>7</td>
<td>BSYPMT2</td>
<td>DATAPAGEID</td>
<td>num</td>
<td>Internal id for data page</td>
<td>10.</td>
</tr>
<tr>
<td>8</td>
<td>CHEMLABS</td>
<td>BASEHIGH</td>
<td>num</td>
<td>Baseline Grade for Positive Shift</td>
<td>LB</td>
</tr>
</tbody>
</table>
How to Prepare High-quality Metadata

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues
Interpret Validation Results

- P21 Interpretation of severity (like ‘Error’, ‘Warning’, and ‘Notice’).
  - errors as issues reported with 100% confidence (e.g., AE start date is after the latest Disposition date)
  - warnings as potential issues requiring manual review (e.g., EXDOSU value not found in 'Unit' extensible codelist).

- Review agencies like FDA interprets
  - Error/Warning as issues with severe impact on review process
  - Reject as critical issues that prevent review and automation processes.

- Example: FDA Business Rules uses words like “must” and “should”, P21 reports these issues with a severity as ‘Warning’ but not as an ‘Error’ or ‘Reject’.

- Despite having differences in interpretation, sponsors are encouraged to fix as many issues that are fixable

<table>
<thead>
<tr>
<th>FDA Business Rule ID</th>
<th>FDA Business Rule</th>
<th>FDA Validator Rule</th>
<th>Domains</th>
<th>SDTM 3.1.2</th>
<th>SDTM 3.1.3</th>
<th>SDTM 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOAB001</td>
<td>A treatment-emergent flag must be submitted.</td>
<td>A treatment-emergent flag should be included in SUPPAE according to SDTMIG v3.1.2 #8.4.3</td>
<td>SUPPAE</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FOAB022</td>
<td>EPOCH should be included for clinical subject-level observations (e.g., adverse events, laboratory, concomitant medications, exposure, and vital signs).</td>
<td>Variables requested by FDA in policy documents should be included in the dataset. E.g., EPOCH and ELEMENT.</td>
<td>ALL</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Example:
- SD1077 FDAC021: FDA Expected variable EPOCH not found
  - Review: Warning
  - Score: 1
- SD1097 FDAC022: No Treatment Emergent info for Adverse Event
  - Review: Warning
  - Score: 6794
How to Prepare High-quality Metadata

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues
Evaluate Impact of Issues on Data

• Meetings with review divisions (e.g., CDER and CBER) play a crucial role on approvals

• Meetings can be organized at different stages of drug development process.
  • Pre-IND meetings
  • End-of-phase2 meetings
  • Pre-submission meetings
  • Post-submission meetings
  • Adhoc meetings to resolve uncertainty about how some issues are going to affect their approval
  • Type C meeting to discuss substantive data standardization issues for NDAs and BLAs. Example: a sponsor’s desire to use a standard (e.g., therapeutic area standard in SDTM format) that is not currently supported by FDA

• “Sponsors and applicants may submit technical questions related to data standards at any time to the technical support team identified by each Center” - Providing Regulatory Submissions In Electronic Format — Standardized Study Data
How to Prepare High-quality Metadata

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues
Fix Critical Issues

• Critical issues if not fixed leads to rejection
• Study data standards catalog must be followed for studies that start after December 17, 2016.
• Technical rejection criteria is additional to the existing validation criteria
• There are two rules so far.
  • Rule #1734 – Trial Summary (TS) dataset must be present for each study in Module 4 and 5
  • Rule #1736 – Demographic (DM) dataset, Subject level analysis dataset (ADSL) and define.xml must be submitted in Module 5 for clinical data
• The above 2 rules are covered by P21 tool set as below.

<table>
<thead>
<tr>
<th>Pinnacle 21 ID</th>
<th>Message</th>
<th>Description</th>
<th>Category</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD1020</td>
<td>Missing DM dataset</td>
<td>Demographics (DM) dataset must be included in every submission.</td>
<td>Presence</td>
<td>Error</td>
</tr>
<tr>
<td>AD0001</td>
<td>Missing ADSL dataset</td>
<td>ADaM Subject level (ADSL) dataset should be included in every submission.</td>
<td>Presence</td>
<td>Error</td>
</tr>
<tr>
<td>SD1115</td>
<td>Missing TS dataset</td>
<td>Trial Summary (TS) dataset must be included in every submission.</td>
<td>Presence</td>
<td>Error</td>
</tr>
<tr>
<td>DD0101</td>
<td>Missing Define.xml</td>
<td>FDA eCTD submissions must include a define.xml file for each study in Module 4 (nonclinical) and Module 5 (clinical)</td>
<td>Presence</td>
<td>Error</td>
</tr>
</tbody>
</table>
How to Prepare High-quality Metadata

1. Knowledge on standards and latest requirements
2. Efficient preparation and Extensive Validation
3. Interpretation of issues
4. Evaluation of Impact of Issues
5. Fixing Critical Issues
6. Explaining non-fixable issues
Explain Non-Fixable Issues

- Explain non-fixable issues in reviewers guide (RG) (optional, but recommended)
- RG helps understanding of complex issues associated with data review and validation.

Issue types

- CRF design or data collection issues - should be explained Section 3.3 in clinical SDRG or section 5.2 in ADRG.

- Mapping issues could be mentioned in section 4.1 of SDRG.
  - Example 1: Compound adverse events (that are combination of simple adverse events) are mapped to a Custom Events domain (e.g., XC), and it is used as input to ADAE along with AE, then P21 report on ADAE may throw errors due to the domain variable having values other than AE.
  - Example 2: Study day 0 is not ideal as per SDTM model, but some sponsors may prefer to have study day zero for treatment onset day.

Sponsors are encouraged to include and explain issues generated on Define xml alone, and data + define.xml issues to be added to reviewer’s guides.
Conclusion

• It is important to have a balance between use of standard guidance from FDA, and automatic tools like Pinnacle 21 for creation of high quality metadata.

• All sponsors should educate their Standards implementation teams about the available resources and how they could be used to have successful approvals for submissions for regulatory review.
References

Questions