Demystifying Define-XML Codelist Handling for Nonclinical Studies

Project Outcomes

The team has identified 5 priority questions about codelists that the team will research and provide 'best practice' recommendations; those questions are listed below.

1. **What terms should be included in a codelist? What are the criteria for determining what terms are included in a codelist?**
2. What variables, other than those that have associated CDISC codelists in the SENDIG, must/should have user-defined codelists in define.xml? What are the criteria for determining this?
3. How is the define.xml file currently being used in nonclinical (by FDA, Sponsors, CROs, system vendors)? What are possible uses (and thus valid reasons for providing codelists in define.xml)? What guidance for industry for their use?
4. **When should a Display Value (Decode) be included with each codelist term? What should be in the Decode?**
5. **When and how can you use a codelist for multiple variables?**

The questions with recommendations are shown above in bold text.

Project Overview

The goal of this project is to provide recommendations for specific problems/questions encountered, when fulfilling the define.xml codelist section for a nonclinical study.

The project will be initiated by defining a specific list of issues/concerns/questions/challenges related to nonclinical codelist in define.xml to be explored.

It will then include an exploration of published general information (clinical or nonclinical) about implementing codelists in define.xml and relating this to specific challenges expressed by nonclinical data stakeholders. It will also include exploration of how codelists are used by the FDA as a consumer of the nonclinical define.xml file.

The result of this project is expected to be recommendations to the public on best practices for fulfilling codelist section of the define.xml for SEND study submissions.

Deliverables are expected to be poster(s) for the 2018 CSS; published recommendation on the SEND Implementation Wiki and a white paper may be warranted as well.

This project will not include education in the define-xml standard, generally how to create define.xml files, or any survey or analysis of define-xml tools. It will focus solely on codelist content. Participants should have a working knowledge of the define-xml standard.

Project Updates

The following questions and recommendations have been posted to the Nonclinical Implementation FAQ, Define.XML Section

In a define.xml CodeList, when should a Display Value (Decode) be included with a term?
Define.xml 2.0 codelists can include either EnumeratedItem or CodeListItem. Use of CodeListItem allows both a term and its display value (decode) to be included. All terms in a CodeList must use either EnumeratedItem or CodeListItem.

- Use EnumeratedItem elements in a CodeList when the terms themselves are sufficient for data interpretation.
- Use CodeListItem elements with a Decode in a CodeList when a decode facilitates data interpretation - when the code value is an abbreviation, acronym or short code that represents a word or phrase.

In a define.xml Codelist, what should be in the Display Value (Decode) entry for a term?
Decode for a CodeListItem element should contain the following: When the coded value has a definition (decode) in a paired variable in the data (whether or not the variables used CDISC Controlled Terminology), use the value of the paired variable in the decode. For example, value of LBTESTCD has its decode in LBTEST. While this is not a recommendation on which variables should have codelists, the instances of paired code /decode variables in SENDIG 3.0 are ARMCD/ARM, CLTESTCD/CLTEST, ETCD/ELEMENT, PCTESTCD/PCTEST, QNAM/QLABEL, SETCD /SET, --TOXGR/--TOX, --TPTNUM/--TPT.

When the coded value does not have a decode in a paired variable in the data:

- if the coded value is CDISC Controlled Terminology, use the value in the "NCI Preferred Term" column of the published Controlled Terminology version in the associated datasets
- if the coded value is NOT CDISC Controlled Terminology and the coded value is in an associated study report table, there is likely a key or footnote or other explanatory text in the study report tables that include the code value. The decode should match the information explaining the value in the study report table. If this information is not in the study report, discuss this with the report author.
- if the coded value is NOT CDISC Controlled Terminology and the coded value is NOT in an associated study report table (for example, if it is metadata collected but not reported, or if it is from SOP information associated with the study), the decode should contain a short unambiguous word or phrase that explains the coded value.

See SENDIG 3.1 Section 4.3.4 regarding use of coded result values in SEND datasets.

Note that some sponsors have received a comment from the FDA that the decode should not contain the full definition of a term.

Published Resources

CDISC Define-XML Specification Version 2.0

CDISC Define.xml Implementation Wiki - Working with Controlled Terminology

FDA Study Data Business Rules
Note specifically FDAB035 The definition of datasets, variables, and codelists in define.xml should reflect the actual study data.

Understanding the define.xml File and Converting It to a Relational Database
Lex Jansen, Octagon Research Solutions, Wayne, PA
SAS Global Forum 2010

PharmaSUG 2016 - Paper DS16 Codelists Here, Versions There, Controlled Terminology Everywhere
Shelley Dunn, Regulus Therapeutics, San Diego, California

From the November 2017 Study Data Technical Conformance Guide: "For variables for which no standard terms exists, or if the available terminology is insufficient, the sponsor should propose its own terms. The sponsor should provide this information in the define.xml file and in the SDRG."

Definitions

From CDISC Define-XML Specification Version 2.0, Section 4.3

The term “Controlled Terminology” in the context of a study refers to the set of all allowable values across all variables that have finite sets of allowable values in the study. A “Codelist” is a unique subset of the controlled terminology to which one or more variables are subject. Beginning with SDTM Version 1.2, the SDTM-IG requires controlled terminology for many SDTM variables. For some variables, sponsor-specific controlled terminology is recommended. All controlled terminology used in a study must be provided within the Define-XML document. Each codelist referenced by a study item shall be represented in the Define-XML document using a CodeList element.

Questions about Define.xml Codelist

Member Questions

A. General Questions to Address

1. How is the define.xml file currently being used in nonclinical (by FDA, Sponsors, CROs, system vendors)? What are possible uses (and thus valid reasons for providing codelists in define.xml)?

2. What variables, other than those that have associated CDISC codelists in the SENDIG, must/should have user-defined codelists in define.xml? What are the criteria for determining this?

3. What are the criteria for determining what terms are included in a codelist? Is the criteria different when a published codelist is referenced?

- When are all possible terms included? Consider scoring scales such as are used in FOBs, dermal and ocular observations; pH scale; semi-quantitative urinalysis results; severities that use CDISC CT like MISEV; severities that do not use CDISC CT such as CLSEV; subsets of the NY codelist.
- When are only terms used on the study included?
- If some codelists contain all possible terms (rather than just the terms used on the study) should the rationale be documented somewhere (define.xml comments and/or nSDRG)?
- Is there a case where only user-defined extensions of CDISC terminology are included, and published CDISC terms are excluded?
- Should the codelist included in define.xml align with the menu selections available during data collection (which would be the "allowed" terms?
- Is a free-text term added during collection represented any differently than a selected from a menu during collection?
- Can you include user-define terms that are used on the associated study, but may be used on a different study, in a codelist? It may be that it is more convenient when creating the define.xml to always include a specific list of terms whether or not they were used on the study.
- When should user defined codelists (nonCDISC codelists) be included in the define file?
- Are the criteria "shoulds" or "musts"?

4. How should published CDISC CT codelists associated with multiple variables be referenced in define.xml (UNIT, NY for example)? More specifically, when can a codelist be shared across multiple variables and when should a unique codelist be referenced?

- For example, if body weights, body weight gains and food consumption are all in grams, can BWORRESU, BWSTRESU, BGORRESU, BGSTRESU and FWORRESU all reference a codelist UNIT_G that contains only "g"?
- If the protocol list of required tissues for gross and micro are different, should there be two different codelists referenced, each with only the list of required tissues for the type of observation? If so, how are required tissues related to a specific sex and death status represented? How are tissues added to the required tissue list (due to findings present) represented? Since the CDISC SPEC codelist contains tissues for any species and sex, presumably the entire SPEC codelist can never be used on a stud?
- When a variable entry contains more than one value (separated by a semi-colon) do you include two entries in the codelist? or one entry with both terms?

5. Is there any specific naming convention that should be used for codelists? Are there any restrictions on naming (special characters)?

- for codelists that contains a subset of published controlled terminology?
- for user-defined codelists

6. When should a Display Value (Decode) be included with each codelist term? Codelist entries that have a decode are included in the define.xml with *CodeListItem* rather than *EnumeratedItem*

- When should CodelistItem be used and when should EnumeratedItem be used?

- What should be in the Decode field?

7. If a term is entered via free text during data collection (i.e., it was not an "allowable" term predefined for selection) is it represented in the codelist any different than terms that were in an allowable list?

8. Should user defined codes use the CDISC code if it exists in a future CT package, or should it use a nonstandard code? It is understood that this would still be an extended term to the CT package used to generate the dataset.

9. If abbreviations are included as findings in a report, should be abbreviations be in ORRES, then included in a codelist with a definition (decode), which would be similar to how the report is presented (abbreviations in the body of the report with a separate explanations page/key)? Or should the abbreviations be replaced with the meanings in the ORRES? What about abbreviations in modifiers that do not have controlled terminology or in comments?

10. Is there a list of know issues or help with troubleshooting Pinnacle21 checker results?

11. What is the appropriate way to indicate that no external codelists were used?

12. What guidance for industry can we give for use of information in the define.xml file that is associated with a SEND dataset?

13. Can an original term that has been translated to controlled terminology, be included in the define file to aid in traceability to the study report? Is that a decode or something else?

Use Cases for Codelists in a Nonclinical Define.xml

1. To see what codes, terms and acronyms mean.

2. To see the full scale associated with qualitative or semi-quantitative results

3. To be able to look at a unique list of user-defined terms for a variable to ensure that they have been properly harmonized (not different just based on case, not more than one with the same meaning, etc)

4. To identify extensions to published terminology

5. To provide definitions for user-defined terms (in user-defined codelists and extensions to published codelists) in human and computer-readable form rather than in an nSDRG document.

6. To check conformance, comparing the terms in a define.xml codelist with terms in a published codelist to identify errors such as failure to use a required codelist, systematic capitalization errors, or extending a nonextensible list (Elaine)

7. When codelist information is in computer-readable form, it can be integrated with views of the study data. For example, when looking at one result on a scale, it could be possible to pop-up a list of values on the scale and their meanings. Another example is that when an abbreviation is shown, it would be possible to pop-up the definition.

8. A codelist for VISITDY/--NOMDY could be used to show the schedule of data collection for a particular test or collection of tests. It would then be easier to tell if the days in the dataset matched the study plan.