Preliminary Recommendations for Traceability documentation using Define-XML 2.0

-Final Draft, December 15, 2013-

Background

This deliverable, produced by the "Define-XML 2.0 - Traceability Features Review" sub-team, falls under the "Optimizing the Use of Data Standards" working group, "Traceability and Data Flow" project kicked off at the 2013 FDA/PhUSE Computational Science Symposium.

Summary

The Define-XML 2.0 Specifications, "Metadata Submission Guidelines" (MSG) V1 and the "ADaM Implementation Guide" V1 documents, describe metadata for SDTM, Analysis Data Sets (ADS) [*1] and Analysis Results. The Define-XML specification document is expected to handle the technical implementation of what is described in less technical terms in implementation guidelines documents; however, it must be recognized that the Define-XML 2.0 Specifications document is reflecting the current version of the Metadata requirements, while the implementation guidelines are not updated yet. Nevertheless, all documents are very useful to understand what metadata should be provided in a submission package and how that metadata can support traceability.

The metadata is delivered in an XML format and the stylesheets display that metadata into a "human readable" format where content across the documents can be connected via hyperlinks (in case the information is available in the XML file). The Define-XML 2.0 Specifications production release package includes an example of such metadata for SDTM and ADaM, with the html version based on the standard stylesheet supplied in the package. Unfortunately, the specifications and, in consequence, an example for results (tables, figures) are outstanding.

The metadata attributes to be defined can be divided into 2 parts: one part to describe the content of the dataset, variable or result (table); the other part to describe the relationship between the source and the target. The latter offers great value in case of traceability.

According to the Define-XML 2.0 Specifications, the Metadata is defined in a way that the association is explicitly given in the target dataset or variable referencing the source. Therefore, the metadata documents support the *trace back* direction more so than a *trace forward* direction - see Table 1. Specialized visualization and analytical tools are needed to facilitate a *trace forward* functionality.

[*1] In general, the Define-XML 2.0 Specifications allows Metadata for legacy data also; therefore, the term ADS applies in a generic way to the analysis data sets, whether they follow the ADaM standard or not.

Table 1: Target - Source supporting attributes:

Target	Source	
Result	ADS	DATASET: link to the dataset used SELECTION CRITERIA, PARAM/PARAMCD, ANALYSIS VARIABLE: describes the sub-setting of the dataset DOCUMENTATION, PROGRAMMING STATEMENTS: describe the analysis method used.
ADS	ADS	Derivation: describes the derivation rule
ADS	SDTM	Predecessor: link to the variable in SDTM
SDTM	STDM	Origin: Derived, Derivation: describes the derivation rule
SDTM	CRF	Origin: link to the CRF page

Metadata attributes

Since the Metadata attributes are well defined, it is unnecessary to repeat it here. In the following, Table 2, the only attributes addressed are those where the team would like to provide further recommendation. In some cases, the intention is to make the metadata more user friendly without any effect on traceability.

Table 2: Metadata attributes

References	Independent if it is the reference to a CRF, a dataset or any other document (e.g. SAP) a link (relative reference) should be provided. Should be given whenever possible - References to large documents like Protocol, SAP etc, should also include the effected section. E.g. SAP, section 3.2even if the link is only to the SAP.
Derived information	- a reference to a document should be given, if applicable - a short description should be provided in addition to the reference. - Programming code should be limited to the essential algorithm. - Programming code should be avoided in the documentation attributes
Value level Metadata	- The existence of value level metadata for a variable should be indicated by a hyperlink on the variable. - Should be provided when attributes like CONTROLLED TERMS OR FORMAT, ORIGIN (?) or DERIVATION/COMMENT are depending on the content of another variable. - Consider multiple origins where the attributes are the same. - In case the content of the 3 attributes is true for several expressions of the describing variable, then these entries should be combined.

Variable level MD vs. Value level MD	- When value level metadata is provided, two scenarios are recommended to handle the content of CONTROLLED TERMS OR FORMAT, ORIGIN and DERIVATION/COMMENT - Scenario 1: the information on the variable level remains blank, all information are handled in the value level metadata. Grouping of describing variables is done as much as possible (see value-level metadata). Also “negative” expressions for the describing variable can/should be used instead of long lists of “positive” expression. (e.g. TESTCD not eq (a b c) instead of TESTCD eq (d e f g …) - Scenario 2: the information on variable level is filled and reflects the “general” information. The value-level metadata keeps only those entries which deviate from the general. - In case the second scenario is used, a note should be included in the submission document explaining the scenario followed.

Results level Metadata (RLM)	- Are seen as very useful to reproduce the results
	  - The FDA reviewers would prefer to have a complete set of MD, but are aware that it cannot be done for legacy studies.
	  - It should be considered that the request for those MD may increase for new studies.
	- Recommendation to implement a similar concept of value level MD for results. This could support e.g. repeated tables.

DATASET	The ADSL dataset should only be referenced in case that it is the primary source of the table. e.g. ADSL in case of baseline characteristic table, but not for an AE tables where the subgroup information is repeated in ADAE - Should ADSL be documented here in case the information is merged for analysis and part of the selection criteria?
SELECTION CRITERIA	The PARAMCD /PARAM selection criteria should also appear in this field. - In case multiple datasets are used then define the crititria by using the 2 level naming convention. E.g. Where=(ADLB.PARAMCD=”ALBUM” ADVS.PARAMCD=”WEIGHT”)
PROGRAMMING STATEMENTS (RLM)	- Content of this section should be limited to the Procedure used or the appropriate algorithm and NOT a copy of the complete program! - A reference to the used program is not required, but might be useful for internal process purposes.

Controlled Terms or Formats	- In case format /codelist is used, that codelist should be named. - A list of “valid values” should be provided in any case, where it is relevant for the result. A subset list is not required in case the whole codelist is “valid”. - Do not clone codelist for subsettings, but use that “valid value” approach to display it. - If the used controlled terminology is depending on a variable content, then value level metadata should be provided. Also in case that the list of “valid values” is depending on the describing variable. - Note: “valid values” means the possible values in the data (e.g. the pick list on the CRF). It should not reflect only the current values in the data. - Ensure that CDISC controlled terminology can be identified easily (in contrast to sponsor code/codelist).
Where to place CRF/Investigator instructions that can influence the collected content?	E.g. CRF question: did the patient get fever? Instruction: tick yes in case that the temperature increases more than 1°C. Put it into the comment field of the SDTM define.xml document.

Acknowledgments

Team members: Mikkel Traun <mt@novonordisk.com>; Tatyana Kovtun <tatyana.kovtun@bayer.com>; Aimee Basile <abasile@celgene.com>; Jingyee Kou <jingyee.kou@fda.hhs.gov>; Karin LaPann <lapannkarin@PRAintl.com>; Marcelina Hungria <mhungria@dicoregroup.com>; Pam Ryley <Pamela.ryley@takeda.com>; 'songhui.zhu@klserv.com'; 'ed.lombardi@syntheracthcr.com'; 'Peter Schaefer' <peter.schaefer@certara.com>; Tanja Petrowitsch <Tanja.petrowitsch@bayer.com>

Authors: Tanja Petrowitsch, Marcelina Hungria