PhUSE US Connect 2018

Paper DS05

Keeping control in a changing world

Johannes Ulander, S-Cubed & A3 Informatics, Copenhagen, Denmark

ABSTRACT

Due to the complexity of clinical research data, it is hard to get an overview of what you have collected during all phases of a research project. At the same time, new versions of standards and controlled terminology are released, which make it essential to keep control of exactly which version has been used for any given study. As the lifecycle of the data extends for several years, we need better ways of understanding how these changes impact the data we already have, but also prepare for how to move from one version to another. This presentation will show how this process with all its moving parts can be controlled and managed with the assistance of automated tools using linked data and graph databases designed to put the user back in control.

INTRODUCTION

Keeping control in the world of standards has always been a delicate task, especially in larger companies. Whether you are talking about Medical Dictionaries, laboratory tests or in which format you store your data, it will always involve decision making and require robust processes for how, when and why you make changes to it or upgrade to a newer version. With increasing requirements from regulatory authorities, where traceability and openness of the data from start to finish is crucial, the task of managing standards has not become easier.
One standard that exemplifies this complexity is CDISC SDTM Controlled Terminology, as the effect of a change in a submission value or a definition might have side effects, which are not immediately obvious.

UNDERSTANDING THE IMPACT OF CHANGES

It has become a de-facto standard when creating CDISC SDTM to call this process “mapping”, but the word hides the complexity of the great many things that it might consist of. It can be everything from re-labelling of synonyms, spitting information in source variables to a program that involves many derivations. But a “mapping” does not have to be complex to create unexpected side effects.

THEN BUN MYSTERY

The BUN mystery started in March 2016. A Clinical Project with an upcoming submission had been ongoing for more than a year and was in a stable situation. Programs for SDTM creation where running smoothly, CRF’s annotated, processes for QC and validation in place and only one more study to include in the submission package. Due to the stable situation, the projects decision to always use the latest version of SDTM Controlled Terminology was not controversial. On Friday 25’th of March a new version of CDISC SDTM terminology was released and the validation tool updated to use the latest version, and it reports of an incorrect submission value. An error that was not present the day before.
In the SDTM Terminology Changes 2016-03-25 document you can find the cause of the error, that codelist item C61019 with Submission Value “BUN” had been removed and replaced with another submission value, codelist item C125949 with Submission Value “UREAN”. (See figure 1)

Release Date	Change Type	NCI Code	CDISC Term Type	CDISC Codelist (Short Nar	CDISC Codelist (Long Nar	Change Summary	Original	New	Change Implementation Instructions
2016-03-25	Add	C125949	Term	LBTESTCD	Laboratory Test Code	Add new term to existing codelist	- - -	UREAN	Urea Nitrogen can be measured in substances other than Blood, Serum, or Plasma. This concept will replace C61019.
2016-03-25	Remove	C61019	Term	LBTESTCD	Laboratory Test Code	Remove term entirely from codelist	BUN	- - -	Urea Nitrogen can be measured in substances other than Blood, Serum, or Plasma. Therefore please use the generic code for Urea Nitrogen (C125949), which has been published with the 2016-03-25 publication.

Figure 1. From SDTM Terminology Changes 2016-03-25 document

But the interesting part of this change is not the submission value itself, it is the change in definition and that the term from the beginning contained information that didn’t fit directly to the SDTM model. “BUN” is Urea Nitrogen measured in blood, whereas “UREAN” is only the lab test Urea Nitrogen without reference to the specimen in which it is collected. Which meant that the SDTM LB datasets produced also needed to change, as specimen had not been added to the dataset as it already existed in the definition of the item itself. (See Figure 2. Of course, in the perfect world, the specimen should have been added from the start, but that is another discussion.)

SDTM CT 2015-12-18

Code	CDISC Submission Value	CDISC Synonym(s)
C111142	BUCHE	Acylcholine Acylhydrolase; Butyrylcho
C61019	BUN	Blood Urea Nitrogen
C75352	BUPREN	Buprenorphine

LBTESTCD	LBTEST	LBORRES	LBORRESU
BUN	Blood Urea Nitrogen	7,1	mmol/L

SDTM CT 2016-03-25

Code	CDISC Submission Value	CDISC Synonym(s)
C111142	BUCHE	Acylcholine Acylhydrolase; Butyrylcholin
~~C61019~~	~~BUN~~	~~Blood Urea Nitrogen~~
C125949	UREAN	Urea Nitrogen
C75352	BUPREN	Buprenorphine

LBTESTCD	LBTEST	LBORRES	LBORRESU	LBSPEC
UREA	Urea Nitrogen	7,1	mmol/L	BLOOD

Figure 2. SDTM Controlled Terminology and excerpt of produced of SDTM LB datasets

The take home message from the previous example is not that specific change itself, it is that the world of standards is in constant motion and that understanding the impact of implementing something new is not always straight forward. As we extend the reach of our existing standards and find new use cases for which they do not fit the need, new standards are introduced to handle them (e.g. the addition of Associated Persons and Therapeutic Areas Guidelines, TAUGs, for SDTM and ADaM). At the same time the rate of changes to existing standards is not likely going to decrease. Even if one would want to keep internal tools and processes up to date with the quarterly release of CDISC SDTM Controlled Terminology, it is not really possible (at least not for larger companies.) Partly because the risk is quite small for a substantial impact, but also due to the resource cost of making the impact assessment, for which you will need someone with extensive knowledge in the end to end process of producing submission deliverables.

In the long run this is not sustainable, and we need help from computers to make life easier for people managing standards.

AUTOMATED IMPACT ASSEMENT

In order to automate the process, the tools involved need to share their definitions and be machine understandable (as opposed to machine readable, which you could argue that excel or comma separated files are as well.) Linked data (see reference 1.) provides the necessary features for making this possible.

The BUN mystery will serve as an example for how automated impact assessment can be achieved.

LINKING CRF TO TERMINOLOGY

All CDISC SDTM Controlled Terminology releases are made available within the tool, so when creating the CRF form named BUNFORM, the question on the CRF is linked to the BUN codelist item from the last version before it was removed (i.e. 2015-12-18.)

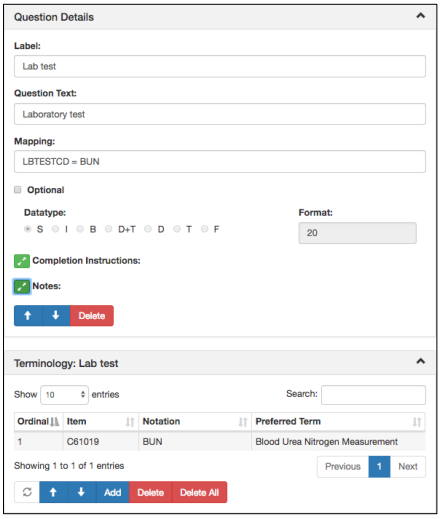


Figure 3. Linking the CRF question to terminology

PERFORMING THE INITIAL IMPACT ANALYSIS

Within the tool, I can now select CDISC SDTM Controlled Terminology release 2016-03-25, where BUN does not exist, and select impact analysis. The result is listed in Figure 4.

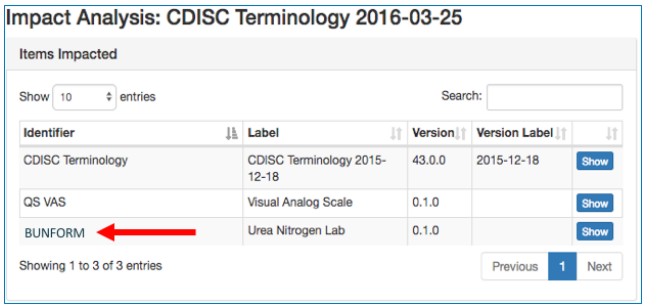


Figure 4. Result of impact analysis

As you can see, the BUN_FORM is listed as being impacted (along with other things within the tool that are also linked to the terminology that is changed.) But it still does not tell me what I learned from the SDTM Terminology Changes 2016-03-25 document, i.e. that the BUN has been replaced by UREAN.

IMPROVED IMPACT ANALYSIS

To improve the impact analysis, the information about the changes needed to be included as well. As the change implementation instructions are not machine readable some additional columns where added (as shown in Figure 5 below.)

		BUN		UREAN
Change Implementation Instructions	Prev Codelist	Prev Term	New Codelists	New Terms
Urea Nitrogen can be measured in substances other than Blood, Serum, or Plasma. This concept will replace C61019.	C67154	C61019	C67154	C125949

Figure 5. Machine readable change instruction

With this extra information incorporated into the tool as links, it was possible to also show that BUN is replaced by a new term. (See Figure 6.)

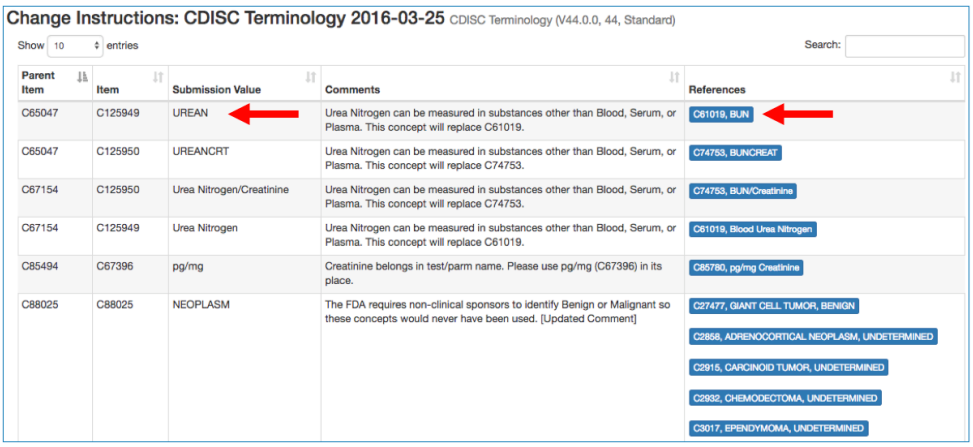


Figure 6. Excerpt of submission values changed from SDTM CT 2015-12-18 to SDTM CT 2016-03-25

NEXT STEPS

The next obvious step would be to add the full change instruction including the changes needed to the SDTM target, in the case of the BUN mystery that would include adding the variable LBSPEC with terminology ”BLOOD”.

WHY USE LINKED DATA?

The impact assessment demonstrated above can be implemented using other technologies than linked data. However, it is important to understand that there are some key features of linked data that removes some of the obstacles that we as an industry are struggling with and which are common to all companies and all relate to how we store our data, i.e. as tabular structures.

Even within a company it is hard to maintain a single structure to store demographic information, and more variations are introduced by TAUGs and by the regulators which are starting to define their own rules for how the standard is to be used. This is why an alternative way of expressing the standards, rather than the 2-dimensional structures that are being used today, is urgently needed. (See reference 2 for examples of storing demographic and medical history data as linked data.)

To implement the impact assessment as linked data, no new structures or variables have been introduced, only links between the already existing components, i.e. the CRF form with the result for BUN, the codelist item C61019 (BUN) and the CDISC Controlled Terminology 2015-12-18 release (see figure 7.)

Figure 7. Simple visualization of links between Form, result, terminology and version of terminology

If one wants to update the BUNFORM to use CDISC Controlled Terminology 2016-03-25 with the new term C125949 (UREAN), it is possible to change the link from the old to the new terminology as the new terminology also exist within the same database. (See figure 8.)

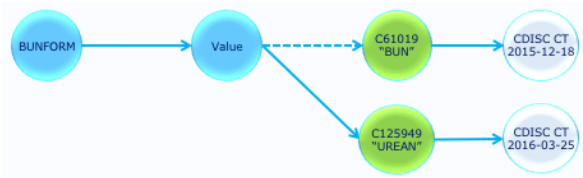


Figure 8. Simple visualization of link added to new version of terminology

If desirable, both links can exist at the same time as different versions. The next step would then be to also add the changed mapping to SDTM.

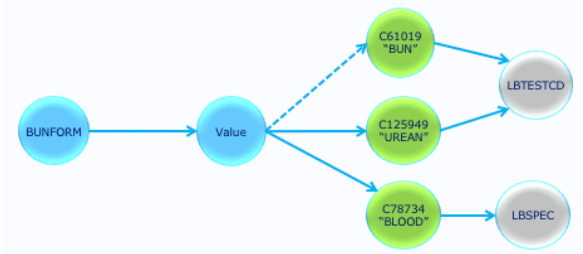
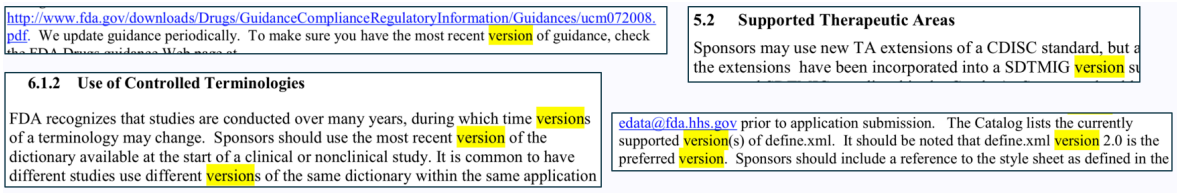


Figure 9. Simple visualization of two different "mappings" from collected value to SDTM LB

VERSION MANAGEMENT

Documenting your data properly is key to obtain control and without a version to refer to it is impossible to make sure that someone else can follow the traceability of data. Versions are therefore important to regulatory authorities (see Figure 7) and there is even a specific document to tell which versions that are acceptable. Which is why it being necessary to document it whenever it is used, e.g. in Clinical Study Data Reviewers guides, Analysis Reviewers Guides and define.xml. And if you are using multiple versions of the same standard, you might even need to add the version to individual observations within the data itself. (N.B. The next version of define.xml will have the ability to reference multiple versions of the same standard.)


http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072008.pdf. We update guidance periodically. To make sure you have the most recent version of guidance, check

5.2 Supported Therapeutic Areas

Sponsors may use new TA extensions of a CDISC standard, but a the extensions have been incorporated into a SDTMIG version st

6.1.2 Use of Controlled Terminologies

FDA recognizes that studies are conducted over many years, during which time versions of a terminology may change. Sponsors should use the most recent version of the dictionary available at the start of a clinical or nonclinical study. It is common to have different studies use different versions of the same dictionary within the same application

edata@fda.hhs.gov prior to application submission. The Catalog lists the currently supported version(s) of define.xml. It should be noted that define.xml version 2.0 is the preferred version. Sponsors should include a reference to the style sheet as defined in the

Figure 10. Some mentions of version in FDA Technical Conformance Guide, version: March 2018

RELEASE VS. VERSION

Currently versioning of standards is done by adding a date or a version number to a complete document. Therefore, there exist e.g. a 2015-12-18 and a 2016-03-25 release of SDTM Controlled Terminology. By looking at the overview of changes for different codelists it becomes clear that the they do not follow the same pattern of change. (See figure 11.)

Identifier	Label	Submission Value	2015-03-27	2015-06-26	2015-09-25	2015-12-18	2016-03-25	2016-06-24	2016-09-30	2016-12-13	2017-03-31	2017-06-30	2017-09-29	2017-12-22
C117743	CDISC SDTM Ophthalmic Exam Test Code Terminology	OETESTCD	✎	✎	–	–	–	–	✎	–	✎	–	✎	✎
C117746	CDISC SDTM West Haven Hepatic Encephalopathy Grade Terminology	HEPENCGR	–	–	–	–	–	–	–	–	–	–	–	–
C117986	CDISC Questionnaire CHART-SF Interview Version Test Name Terminology	CHSF1TN	–	–	–	✎	✎	–	–	–	–	–	–	–
C117987	CDISC Questionnaire CHART-SF Interview Version Test Code Terminology	CHSF1TC	–	–	–	✎	✎	–	–	–	–	–	–	–
C117988	CDISC Questionnaire C-SSRS Lifetime/Recent Version Test Name Terminology	CSS11TN	–	–	✎	✎	✎	–	–	–	–	–	–	–

✎ Changes – No changes

Figure 11. 5 different codelist and their changes between 2015-03-27 and 2017-12-22.

And that some codelist haven't been changed for many years, if at all, e.g. codelist C66742 - No Yes Response. (See figure 12.)

Code List Item Changes

Show 10 entries Search:

Identifier	Label	Submission Value	2015-03-27	2015-06-26	2015-09-25	2015-12-18	2016-03-25	2016-06-24	2016-09-30	2016-12-13	2017-03-31	2017-06-30	2017-09-29	2017-12-22	
C17998	Unknown	U	–	–	–	–	–	–	–	–	–	–	–	–	Changes
C48660	Not Applicable	NA	–	–	–	–	–	–	–	–	–	–	–	–	Changes
C49487	No	N	–	–	–	–	–	–	–	–	–	–	–	–	Changes
C49488	Yes	Y	–	–	–	–	–	–	–	–	–	–	–	–	Changes

Figure 12. Codelist item changes in codelist C66742 (No Yes Response)

Please note the codelist item C17998 (Submission Value = U) in the upper row. When making linked data of that specific item, we notice an issue. As C17998 exists in multiple codelists, in our linked data we only need one item to represent it, with links to the codelists where it exists. But looking at the CDISC SDTM Controlled Terminology released 2017-12-22 and filtering on C17998 it reveals that it has two different submission values. To be able to reproduce the content in controlled terminology for a specific release, we will need to support both submission values as we would otherwise get an error when validating SDTM datasets. (See figure 13.)

Code	Codelist Name	CDISC Submission Value
C17998	Action Taken with Study Treatment	UNKNOWN
C17998	Cardiac Procedure Indication	UNKNOWN
C17998	Coronary Vessel Disease Extent	UNKNOWN
C17998	Ethnic Group	UNKNOWN
C17998	Frequency	UNKNOWN
C17998	Microbiology Susceptibility Testing Result Category	UNKNOWN
C17998	No Yes Response	U
C17998	Normal Abnormal Response	UNKNOWN
C17998	Outcome of Event	UNKNOWN
C17998	Relation to Reference Period	UNKNOWN
C17998	Risk Assessment	UNKNOWN
C17998	Sex	U
C17998	Subject Status Response	UNKNOWN

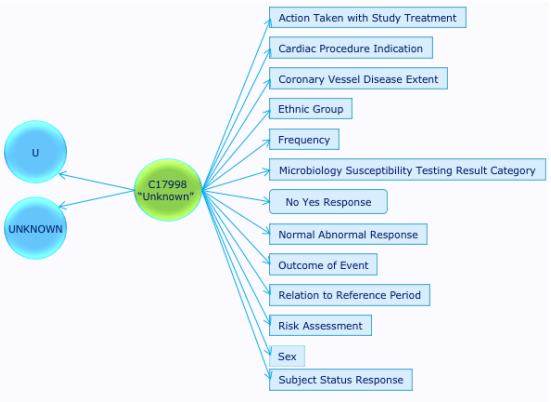


Figure 12. Codelist item C17998 – Unknown in Excel and as linked data

In linked data this issue is immediately spotted, whereas it is easy to miss in the current format.
With our current way of versioning and satisfying regulatory needs in documents, a complete new release is needed to correct issues, and it becomes a blunt instrument for managing changes.
This change does not require linked data, it is enough to adapt a new way of versioning. But linked data showed that it has a positive impact on the quality of the data.

CONCLUSION

Keeping control in a moving world is a complex task and we need computers to help us maintain control. Linked Data and graph technology is no longer a novel technology and can express standards and their relationships with greater precision, improve data quality and be adapted to fit our future needs. So even if we don't get it right from the start, we can iterate and automate moving between versions.

REFERENCES

1. Visit: https://en.wikipedia.org/wiki/Linked_data and https://en.wikipedia.org/wiki/Graph_database
2. PhUSE Annual Conference 2017. TT09(2017) SDTM Domains by Query - is it Possible?
   Link to: Paper Link to: Presentation

CONTACT INFORMATION

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

Johannes Ulander
S-Cubed ApS,
Lille Strandstraede 20C, 5
Copenhagen / 1254
Web: http://www.s-cubed-global.com/