

Bimonthly newsletter of XML4Pharma,  
Katzelbachweg 18, AT-8052 Thal, Austria  
Web : [www.XML4Pharma.com](http://www.XML4Pharma.com)  
Mail: [Info@XML4Pharma.com](mailto:Info@XML4Pharma.com)  
February 2013



**XML4Pharma is a CDISC  
Registered Solutions Provider**

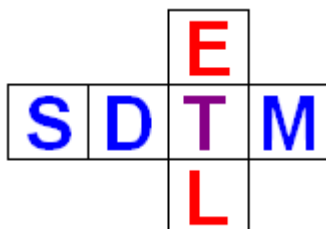


### **SDTM-ETL™ 2.2 is released**

Version 2.2 of our popular SDTM mapping tool has just been released. Most important improvements and new features include:

- full support for SDTM 1.3 - SDTM-IG 3.1.3
- Additional domain templates: medical devices, oncology, non-subject-data, draft SDTM 1.4 domains
- Suggestions for the SDTM/SEND Variable Length (on request of the FDA) - see last page of this newsletter.
- Further improved dialogs and wizards
- considerably improved performance

Additional information is available on the [SDTM-ETL website](#), where you can also download a copy of the [product flyer](#), the [product brochure](#) and the [“new features” document](#).



### **CDISC EU Interchange Bad Nauheim/Frankfurt**

The yearly European CDISC Interchange is this year taking place in Bad Nauheim, a spa town about 35km north of Frankfurt, from 22 to 26 April.

The city is also famous for music lovers, because Elvis Presley lived there during his military service in Germany.

Traditionally, the two first days of the conference will be devoted to trainings, including the famous two-day SDTM training. Also the last day (Friday) will be devoted to trainings- I have been asked to give the ODM training on that day.

The conference itself is taking place on Wednesday and Thursday. You can already find a [preliminary programm here](#).

In my role of professor in Medical Informatics at the [University of Applied Sciences FH Joanneum](#) here in Graz, I have send in two abstracts. To my surprise, both were accepted, meaning that I will give two presentations during the conference with the titles:

- *It's the semantics stupid! The last mile (or kilometer?) in interoperability between healthcare, clinical research and regulatory agencies*
- *The future of ODM*

The former presentation will be about how we can further improve semantic interoperability with healthcare, whereas in the latter I will be presenting my personal opinion on the future of the ODM standard.

Oh yes, and I will also be taking some posters from our students with me (see further)!

As usual at EU Interchanges, there will also be a social evening, with a live music show featuring Elvis.

If you have not registered yet for the European CDISC Information, [here is the link with additional information](#).

## **Preparing for define.xml 2.0**

The final specification of the “Case Report Tabulation Data Definition Specification” (CRT-DDS) v.2.0, better known as “**define.xml 2.0**” will be published in the next few days.

For us, this means ... a lot of work, which we need to prioritize.

First of all, we will be working on a new version of the “[define.xml Checker](#)”, which is a software tool for validation of define.xml instance files. This software is used by many of our customers, and as I understand it, they want to switch to the new define.xml version as soon as possible.

Updating [SDTM-ETL™](#) will obtain a somewhat lower priority. The reason is that the FDA will need some time (I expect 1-2 years) to adapt its workflows and tools and to be able to receive submissions using define.xml 2.0.

## **More about some of our students projects**

Last year, a group of 3 female bachelor students at the [University of Applied Sciences FH Joanneum](#), department of [eHealth](#), run a project where they had to construct a software program to preselect eligible subjects for an Alzheimer study from repository with electronic health records in HL7-CDA format. Some details can be found in the [previous newsletter](#).

I have further extended this software with some additional features, and will bring a poster prepared by the project team to the EU CDISC Interchange in April. I will also bring the software itself and give demonstrations to interested persons and companies.

Also, one of our bachelor students (she just passed the final exam a few weeks ago) made her thesis on the topic of XML-Schematron and its use in clinical research. The practical part of the thesis consisted of creating a Schematron for CDISC ODM 1.3(1). I am currently checking the work she did. When that is ready, I will ask my fellow CDISC volunteers in the XML Technology team to review it, so that we can later publish it as an official CDISC publication.

If you have never heard about Schematron, you might have a look [here](#). Schematron is an XML-based technology to define sets of rules that cannot be covered by XML-Schema. A simple example from the ODM standard is that all datetimestamps in a transactional file must be [earlier](#) than the creation datetime (“CreationDateTime” attribute on the ODM element) of the ODM document.

Whereas about 50% of the rules of ODM are covered by Schema, an additional 40% is covered by Schematron. The remaining 10% is mostly about content in previous ODM files in the same time series. This would require the use of XQuery within Schematron, which (as far as I know) is not possible yet.

## **Blogging**

I am a regular blogger, especially on LinkedIn.com, but also have a few blogs of my own, where I regularly publish some new ideas or comment on current issues and trends in the e-clinical world.

<http://cdisc-end-to-end.blogspot.com> discusses the use of CDISC (and other) standards in the end-to-end process: from protocol writing to electronic submissions to the regulatory authorities.

My latest contributions there have been about the use of ODM to store data points from electronic health records (such as HL7-CDA, or ISO-21090). One of the critics on ODM in the past is that it is not compatible with neither HL7-CDA nor ISO-21090. The blog entry shows that this is simply not true, and that it is a piece of cake to include data points from these standards in CDISC ODM using the extension mechanism.

Another blog entry is about things ODM (and SDTM) can learn from HL7-CDA, like the use of UCUM units of measurement, the “code” and “codeSystem” construct, and the use of “Universal Object Identifiers” (OIDs - not to be confused with the ODM OIDs). Every serious organisation in healthcare that I know has an OID, and is extending that OID to uniquely identify its standards, and its documentation and specifications. The only exception I know is ... CDISC.

The second blog that I “own” is <http://cdiscguru.blogspot.com>. Yes, I realize the naming is overambitious. During the last months, I published some short articles there about whether SDTM is a database (it isn't), on the use of supplemental qualifiers and the disability of the FDA reviewers to join them with their originating table, and about the sense or nonsense of developing CDISC controlled terminology for units of measure (why didn't they just take UCUM?).

Of course we welcome all readers of this newsletter to read the blog entries, to add comments, and even to add new articles!

## SDTM-ETL™: suggesting the ideal length for an SDTM variable

The FDA has been complaining a lot about too-large SDTM datasets due to non-optimal (or better said far-from-optimal) variable lengths in the SDTM datasets (“File size issues” in the [CDER Common Data Standards Issues Document](#)). In modern data formats like XML, there should not be an issue, but as the FDA still requires submissions to be delivered in SAS Transport 5 format, which is a fixed-length format, a non-optimal assignment of variable lengths can lead to memory problems when files are tried to be opened on the FDA computers.

Just some explanation: suppose that you have assigned a length of 200 (which is the maximal length for a variable) for LBNRIND (as I found in a dataset of the CDISC pilot), as you do not know in advance what the maximal value length will be. You now fill the cell in the table with “NORMAL”, which is 6 characters, then the remaining 194 bytes is filled with blanks. When the value is “ABNORMAL” (8 characters), the remaining 192 bytes are filled with blanks.

Don't laugh! This is happening very often!

When doing so, you easily end up with files that are 10 or more times larger than they can be when a good choice is made for the variable length.

But how do you know what the length will be in advance? People that use classic statistical software often do not care, as they just do not realize that this can lead to issues. The software does not point them to such possible issues either.

However, SDTM-ETL cares!

When doing a drag-and-drop from the ODM tree to an SDTM variable cell in SDTM-ETL, the software checks what the defined length in the ODM metadata is and compares this with the currently defined length of the target SDTM variable. For example:

| ItemDef | Value   | Domain | StudyEvent | Form        | ItemGroup |
|---------|---------|--------|------------|-------------|-----------|
| TS      | STUDYID | DOMAIN | TS.TSSEQ   | TS.TSGRPID  | TS.TSF    |
| SE      | STUDYID | DOMAIN | USUBJID    | SE.SESEQ    | SE.ETC    |
| SV      | STUDYID | DOMAIN | USUBJID    | SV.VISITNUM | SV.VISI   |
| EX      | STUDYID | DOMAIN | USUBJID    | EX.EXSEQ    | EX.EXG    |

| ItemDef        | Value   | Domain  | StudyEvent | Form   | ItemGroup |
|----------------|---------|---------|------------|--------|-----------|
| RELREC         | STUDYID | RDOMAIN | USUBJID    | IDVAR  | IDVAR     |
| SUPPQUAL       | STUDYID | RDOMAIN | USUBJID    | IDVAR  | IDVAR     |
| MyStudy:GLOBAL | RFSTDT  | RFENDTC |            |        |           |
| MyStudy:DM     | STUDYID | DOMAIN  | USUBJID    | SUBJID | DM.RF     |

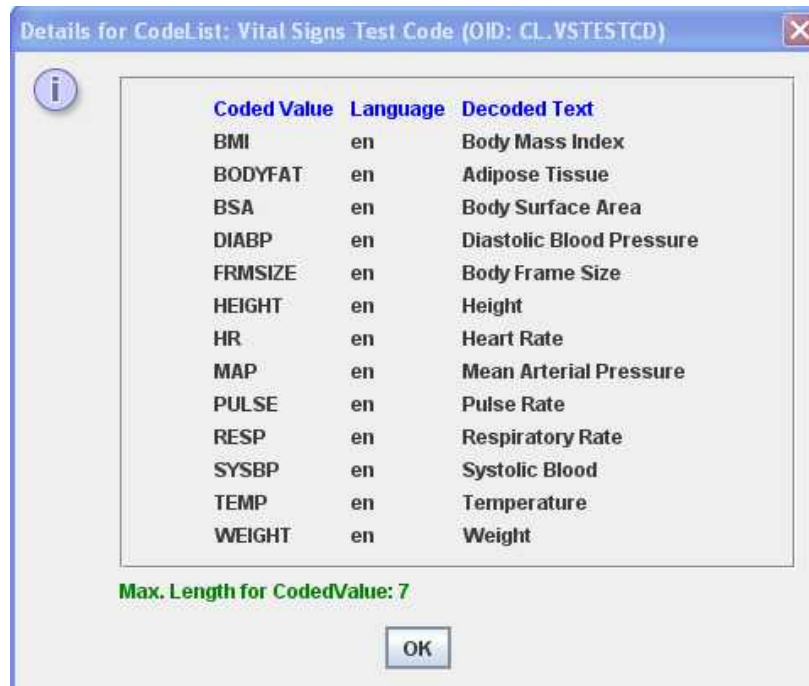
The user maps the ODM item “SubjectID” to the SDTM variable “SUBJID”. The software then reports that the maximal length in the ODM was defined as being 11 characters, and that the current maximal length for the SDTM variable is 60. It then allows to adapt the length for the SDTM variable to also 11 (using the checkbox “Set SDTM Variable Length to ODM ItemDef Length”).

The user can then later still decide otherwise and choose for another length.

Another function in SDTM-ETL is “view clinical data”. It generates a table with all the clinical data for a data point (which can be very useful to understand what the data is about), and then immediately also reports the length as defined in the ODM, and additionally the in reality maximal length found in the data. The user can then “port” that length to the corresponding SDTM variable.

Especially when the SDTM variable has an associated codelist, we have seen that in many submissions the maximal length for that variable has been set to 200. Here again, this is a waste of space, and can lead to file size issues at the FDA.

SDTM-ETL has a menu option “SDTM Variable - view associated codelist. The program then shows all the entries from the codelist and also displays the length of the longest entry in the list. This again allows to adapt the SDTM variable length to exactly the length that is needed.



| Coded Value | Language | Decoded Text             |
|-------------|----------|--------------------------|
| BMI         | en       | Body Mass Index          |
| BODYFAT     | en       | Adipose Tissue           |
| BSA         | en       | Body Surface Area        |
| DIABP       | en       | Diastolic Blood Pressure |
| FRMSIZE     | en       | Body Frame Size          |
| HEIGHT      | en       | Height                   |
| HR          | en       | Heart Rate               |
| MAP         | en       | Mean Arterial Pressure   |
| PULSE       | en       | Pulse Rate               |
| RESP        | en       | Respiratory Rate         |
| SYSBP       | en       | Systolic Blood           |
| TEMP        | en       | Temperature              |
| WEIGHT      | en       | Weight                   |

Max. Length for CodedValue: 7

OK

We added these new features to the software as the FDA requested people that do the mapping to keep the SDTM variable length at a minimum required. We so hope to contribute to less problems at the FDA when reviewers want to inspect SDTM submissions in SAS Transport 5 format.

Once we will have an XML-based format for submissions to the FDA, this kind of issues will of course not occur anymore.