

From EDC to SDTM – faster & better!

Mor Meyerovich, Bioforum, Ness-Ziona, Israel

Lena Hazanov, Bioforum, Ness-Ziona, Israel

ABSTRACT

CDISC standards are set to support a broad array of clinical data elements and are therefore extensible and often open for interpretation. In addition, CRFs in EDC are usually 'custom-made' for each specific study. Together, these factors create many inconsistencies at both the CRF level and the SDTM level.

Here, we will present a methodology which replaces the conventional conversion process by assigning concise SDTM mapping attributes that we call 'tags', to EDC data elements, essentially directly connecting the source element to its SDTM destination. A set of specialized programs can then parse these attributes and automatically convert the captured data into high quality and standardized SDTM datasets. Furthermore, the use of forms with standardized structure and attributes results in uniform SDTM datasets across different studies making the downstream process of cross-study data integration much more efficient and easy to handle.

INTRODUCTION

THE APPLICATION OF CDISC STANDARDS IS CHALLENGING AND TIME CONSUMING

Due to regulatory requirements, the use of CDISC standards is gradually becoming more prevalent and is therefore a leading candidate as the data standard for presentation of clinical data in the industry. However, the use of CDISC standards poses many challenges, primarily due to the fact that CDISC standards were created to support a broad array of clinical data elements. This makes the standards extensible, often open for interpretation and therefore, easily misused. As an example, sending the same raw study data for conversion into SDTM by two different CROs will most probably result in different SDTM interpretation and outputs. In these kinds of scenarios, the Sponsor is left with the task of consolidating the different SDTM flavors into one cohesive approach, which makes the process of cross-data analysis as well as getting ready for submission long, cumbersome and painful.

Moreover, the introduction of data standards into the clinical study workflow, presents a major challenge to the working teams, trying to balance regulatory requirements for data standardization with tight study setup timelines. Often, the teams decide to push SDTM to the very end of the study conduct due to timelines. This, in turn, means that activities such as risk based monitoring, analysis and biostatistics as well as safety monitoring are all performed on raw-EDC data which needs to be configured and setup anew for each study. Beyond time and resources being spent on reinventing the wheel for every study, many times, repeating the analyses after the data went through the SDTM conversion may result in significant quality issues.

WE SHOULD AIM FOR A CAPABLE, LOW MAINTANACE AND REUSEABLE METHOD

Although there are multiple source-to-SDTM solutions, most handle simple mapping use cases and often come short when dealing with more complex data scenarios. Since mapping always gets complex, manual intervention is needed and therefore, the value added by those solutions is small and maintainability becomes challenging.

A potential solution has to be integrated as early in the process as possible, to minimize any back-tracking and patch work down the line. It would also have to be efficient and simple to use, so it could be incorporated into current processes without adding much overhead and burden to the existing infrastructure and support staff. Other features, such as reusability across studies and the ability to use the solution across EDC systems, can also greatly enhance its productivity and appeal.

CONVERSION THROUGH TAGGING

Here, we present a methodology which replaces the conventional conversion process of manually writing SAS® programs for each study. The methodology relies on two main components - “Clinical Tagging Language” (“CTL”) and a “Global Forms Library”. The CTL enables the association of a simple set of conversion instructions with each EDC field, and thus facilitates the transfer of information from the source EDC data to their SDTM target. The role of the “Global forms library” is to ensure reusability, consistency and time efficiency during the process of study building. Finally, an engine has to parse the tagged study data and automatically create the SDTM datasets on a pre-determined schedule or on-demand.

It is important to note that using the CTL does not require the use of the Global Library – the use of the library simply allows keeping ‘Global’ tags which can then be reused when initiating a new study. If no global library exists, tagging is done directly on a study by study level.

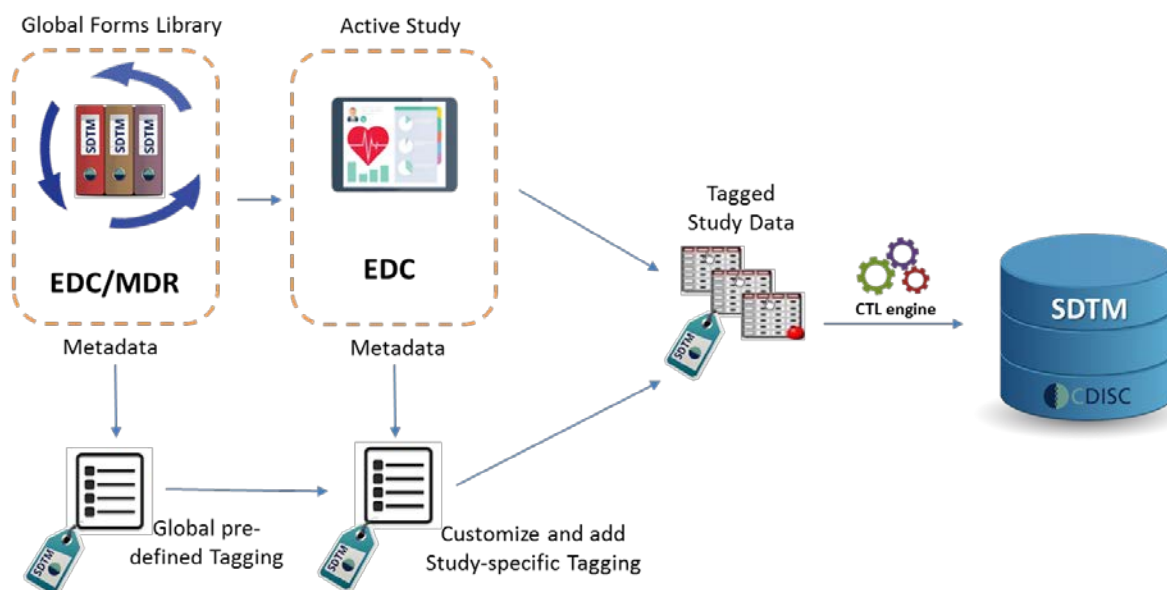


Figure 1: A typical “Data flow” using tagging.

GLOBAL FORMS LIBRARY

A Global forms library includes generic forms that are used when a new study build starts. The design of forms should support SDTM structure, naming conventions, controlled terminology and other considerations, to ensure the consistent handling of similar data elements. Moreover, in order to accelerate the SDTM conversion process while ensuring consistent SDTM mapping, the global forms should also be accompanied by pre-defined tagging.

Defining a set of global forms may prove to be a challenge, as studies vary considerably and very often include study-specific elements. To overcome this, the global library should allow for flexibility in the study building process, both by including more than one flavor of common forms and also by supporting a consistent methodology for creation study-specific forms that could then be reused in subsequent studies. There is no real limit to define only one type of a specific form, so forms such as “Vital Signs”, “Adverse Events”, and others could have a few templates as long as they are ready for use, and are well-maintained. The study-specific forms can be added as needed, but without building them from scratch. These forms are based on generic Findings, Events and Interventions forms as a baseline which can then be quickly customized to accommodate specific types of information.

CLINICAL TAGGING LANGUAGE

A Clinical Tagging Language (CTL) is a set of simple, human- and machine-readable instructions, which define the source-to-target mapping logic. In order to be effective, the CTL should support most possible mapping scenarios in a simple and concise fashion, especially in the common data scenarios encountered in clinical studies. Focusing

PhUSE US Connect 2018

purely on the mapping instructions and not on to the mode of implementation, long manual SAS® programs turn into a single line of tagging. At a later stage, a set of specialized programs can parse these attributes and automatically convert the source data into high-quality standardized SDTM datasets.

In order to facilitate the mapping instructions, we need to control three main components for each EDC element we tag -

- **Source** – What is it exactly that we want to map? As it would be too simplistic to assume that we want to target all the values in any given CRF field, we should have the ability to either tag all values, or just a subset of values. Examples for source subsets would be all values in a given visit, first/last value, highest/lowest value, highest/lowest value in a given visit, etc.
- **Target Key** – Which SDTM record/s are the target of the tagging? The exact domain and record/s in our target SDTM need to be defined, in order to eliminate ambiguities regarding placement of source values in the SDTM. Examples for target keys would be a Demographic (DM) record, a systolic blood pressure record in Vital Signs (VS), a Disposition (DS) milestone of randomization, etc. Because we are in a clinical study setup, it is safe to assume that subject and visit (if available) are always part of our target key and thus do not need to be mentioned explicitly.
- **Mapping logic** – How, and in which variable would the individual values in the source data be represented in the SDTM record? We should clearly define how we interpret the value/s and what we do with them. This should go beyond simple one-to-one mapping and encompass conditional population and assignment of values. Examples for mapping logic would be, one-to-one mapping of values to an SDTM variable, conditional mapping of values to one or more SDTM variable/s, assignment or conditional assignment of values to one or more SDTM variable/s, etc.

CLINICAL TAGGING LANGUAGE – EXAMPLES

As an example, consider heart rate information being captured on a study form. The following illustrates the information needed in order to tag each data element into SDTM. Please note the color coding to distinguish between the different tagging components mentioned above -

- o Subject → All values should be captured in "Heart Rate" VS (Vital Signs) records (identified by Subject, Visit and Time-point) and be populated in USUBJID (derivation specified in a configuration file).
- o Visit → All values should be captured in "Heart Rate" VS (Vital Signs) records (identified by Subject, Visit and Time-point) and be populated in VISIT without any modifications.
- o Time-point → All values should be captured in "Heart Rate" VS (Vital Signs) records (identified by Subject, Visit and Time-point) and be populated in VSTPT without any modifications.
- o Heart-rate-result → All values should be captured in "Heart Rate" VS (Vital Signs) records (identified by Subject, Visit and Time-point) and be populated in VSORRES without modifications, in VSSTRESC/N after conversion to standard unit, and assign "Standing" in position (VSPOS)
- o Heart-rate-unit → All values should be captured in "Heart Rate" VS (Vital Signs) records (identified by Subject, Visit and Time-point) and be populated in VSORRESU after matching to control terminology and in VSSTRESU after conversion to standard unit.

Vital Sign results can, also, be captured in one variable "All-results", while the test name is captured in a separate variable "Test-name". The difference between this form and the one above, results in quite different SAS® conversion programs while done traditionally, whereas tagging remains almost similar (see differences in bold)

- o Subject → All values should be captured in VS records (identified by Subject, Visit Time-point and **Test-name**) and be populated in USUBJID (derivation specified in a configuration file).
- o Visit → All values should be captured in VS records (identified by Subject, Visit Time-point and **Test-name**) and be populated in VISIT without any modifications.

PhUSE US Connect 2018

- o Time-point → All values should be captured in VS records (identified by Subject, Visit Time-point and Test-name) and be populated in VSTPT without any modifications.
- o All-results (instead of Heart-rate-result) → All values should be captured in VS records (identified by Subject, Visit Time-point and Test-name) and be populated in VSORRES without any modifications, in VSSTRESC/N after conversion to standard unit.
- o All-units (instead of Heart-rate-unit) → All values should be captured in VS records (identified by Subject, Visit Time-point and Test-name) and be populated in VSORRESU after matching to control terminology and in VSSTRESU after conversion to standard unit.
- o Test-name → All values should be captured in VS records (identified by subject, visit, time-point and test-name) and be populated in VSTESTCD and VSTEST after matching to control terminology, and if test-name is "HR" assign "Standing" in position (VSPOS).

Another, more elaborate example would be to use one source data element in a couple of SDTM targets. This could be accomplished by tagging the element twice, once for each target. Let us consider a log form which captures drug administration, one possible target is the Exposure domain, in which case each entry populates one SDTM record, another target is a Disposition record, where the first administration is interpreted as a randomization milestone.

Exposure -

- o Subject → All values should be captured in EX records (identified by Subject, Treatment and Time-of-administration) and be populated in USUBJID (derivation specified in a configuration file).
- o Time-of-administration → All values should be captured in EX records (identified by Subject, Treatment and Time-of-administration) and be populated in EXSTDTC (convert to ISO8601).
- o Dose → All values should be captured in EX records (identified by Subject, Treatment and Time-of-administration) and be populated in EXDOSE without any modifications and EXDOSU is assigned with appropriate unit.

Disposition -

- o Subject → First value should be captured in a randomization milestone DS record (identified by subject) and be populated in USUBJID (derivation specified in a configuration file).
- o Time of administration → First value should be captured in a randomization milestone DS record (identified by subject) and be populated in DSSTDTC (convert to ISO8601).

CTL ENGINE AUTOMATICALLY CONVERTS TAGGED EDC DATA TO SDTM

Parsing the tagging and manipulating the data accordingly could then be achieved by using an engine that scans the tagged data and translates it into a set of commands which copy, assign and derive values into SDTM datasets. The commands could be translated into any programming language, whether it is SAS® scripts or other, and could then be run on any occasion throughout the lifecycle of study.

To fully support an automated approach, one needs to take into account a few other considerations such as the definition of trial design datasets, configuration of study visits and schedule, correct handling of EDC specific variables, handling of multiple values in many-to-one mappings, handling exceptions and others. Most of those could be defined alongside the study building setup by using a configuration file which, in turn, will be utilized by the CTL engine either as a pre/post process task or by augmenting the conversion process itself.

THE USE OF TAGGING OFFERS QUALITY AND CONSISTANCY IN SHORTER TIMELINES

The benefits of using a solution such as a CTL and its implementation during study building are numerous, this approach could be easily integrated into current workflows without compromising timelines or quality. On the

PhUSE US Connect 2018

contrary, the ability to standardize data during study conduct, rather than as a post-hoc activity, opens the possibility for seamless integration of ongoing study data in advanced visualization, analytics and monitoring tools without the need for study-specific configuration and at a fraction of the cost and effort. Moreover, coupled with a global library approach, the timelines are further shortened and the quality and consistency of the data are further enhanced, ensuring quality visualization and analytics, both during study conduct and as part of post-hoc analysis across multiple studies.

CONCLUSION

In this paper we introduced a methodology in which conventional conversion of EDC data into standardized SDTM data is replaced with a set of instructions using a CTL. The mechanism and components of the tagging language have been illustrated and it is postulated that, done early enough in the process with the aid of a global forms library, the added benefits are numerous. Not only are time and effort substantially reduced during study building and data conversion, but the subsequent production of high-quality and consistent data could further be leveraged with cutting-edge analytics and visualization tools.

CONTACT INFORMATION

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

Mor Meyerovich

Bioforum
3 Golda Meir St., Weizmann Science Park
Ness Ziona / 70400
Israel
Email: mor.meyerovich@bioforumgroup.com
Web: www.bioforumgroup.com

Lena Hazanov

Bioforum
3 Golda Meir St., Weizmann Science Park
Ness Ziona / 70400
Israel
Email: lena.hazanov@bioforumgroup.com
Web: www.bioforumgroup.com

Brand and product names are trademarks of their respective companies.