PhUSE 2009

Paper CD08

From raw data to submission: A metadata-driven, repository-based process of data conversion to CDISC models

Dimitri Kutsenko, Entimo AG, Berlin, Germany

ABSTRACT

The paper presents a visionary framework to solve selected practical problems when converting data to CDISC models. It will guide the reader through the steps of a data conversion process from raw data to SDTM conformant data and show through practical examples how efficient metadata management in a repository based environment supported by smart tools helps speed up the whole workflow.

The examples will touch on generation of mapping specifications and mapping programs. Special focus will be placed on such issues as quality control, validation and reusability of work results – all supported by metadata-driven mechanisms.

The article is intended to be a top-level guideline for organizations which are just starting or have been involved in data conversion to CDISC models and are looking for innovative ways to design a flexible data conversion workflow developed with a metadata-based vision.

INTRODUCTION

Good news - with every new version, the CDISC SDTM model becomes more stable and robust. Bad news - numerous mapping projects set up to transform data to SDTM-conformant structures are still suffering from significant inefficiencies along the entire data conversion workflow: A typical mapping process separates distinct phases of mapping specification and mapping program development. Mapping programs are manually coded based on the mapping specification which is provided as Excel sheets. Due to the fact that models are dynamic and change over time, mapping programs as well as specifications are difficult to maintain due to the need of interlinked information and are error-prone if model adjustments are necessary. The degree of reusability for structure definitions and program code is minimal: For every study the structure definitions and code need to be created from scratch or are “re-used” by copy and paste “automation” in manual replacements. Moreover, the phases are often carried out by different organizational roles. Communication-related friction at the interfaces between organizational roles and units is “pre-programmed” in such a process.
Without doubt, projects with a few team members are easily manageable in terms of coordination efforts and communication flows. However, even small teams would benefit from a regulatory compliant and traceable process. With increasing project team size and number of studies to be processed, especially when teams are distributed in space and time zones, the management of such projects becomes a challenge.

Imagine a world where everything is possible. What would the mapping process look like? What smart tools could make the process more efficient and release you from routine tasks by replacing them with automation?

As an agile software R&D company and technology provider, Entimo has been able to absorb enormous know-how in different areas over the years and has stayed at the forefront of radical changes in technology, adapting them in development processes and implementing them in its products.

In this paper, I would like to share with you our vision of the data conversion process. We will use CDISC SDTM as an example model for transformations (CDISC 2005b). However, SDTM is just a metadata model in a general sense. If metadata handling is flexible enough, the same process and tools can be used to support any target model – be it specific or standardized.

Certainly, organizational contexts differ from company to company as well as in terms of SOPs, experience of people involved, and their technical background. For this reason, concrete implementations of the mapping process including steps and roles may take various forms with different names. However,

looking at the process, we can crystallize out key elements and tasks which are generic in their nature. To analyze them, we will break down the whole process into parts, try to detect possible inefficiencies in the separate process steps, look for means of eliminating these inefficiencies and, finally, reintegrate the parts into a continuous process. Our additional goal is to find out how organization-wide initiatives may help to increase efficiency even further in the long run.

DELIVERABLES

Following the trend, and for a number of other reasons, big pharmaceutical companies are on the way to partly or completely outsourcing study conversion to SDTM, especially with regards to the large stock of legacy studies available in every house. (The reasons for this are outside the scope of this paper and shall not be investigated here). On the other side, to cope with a growing workload and enjoy globalization advantages, CROs and conversion specialists – who typically take over the task – have experienced a rapid and large-scale concentration wave through numerous M&As and have gone global.

Serving the customers at their best, many CROs often find themselves in a situation where they have to maintain several EDC systems simultaneously – differing not only in current technological capabilities with regards to SDTM support, but also in terms of future potential. In such a situation CROs face a dilemma: to wait until all EDC are upgraded and natively support SDTM, a process which might last years (if realistic at all, as some of the EDC systems are not being actively developed and will hardly survive). On the other hand, mastering all in-house EDC systems in order to be time and cost efficient in comparison with the competition, is expensive due to significant training needs among other things. In addition to the ongoing studies, numerous legacy studies need to be brought into a standardized form (SDTM or SDTM+, for example) for pooling or other reasons.

Is there an alternative option to waiting or working with many processes simultaneously?

One answer is to find an optimum point where all the EDC systems cross and set up from this point a unified, company-wide mapping process for which training is required only once and which makes the task cost and time efficient through use of intelligent tools. Such a process would allow the user to be competitive and fast, which is especially important considering the tough project timelines shortly before submissions.

We begin our journey from raw to SDTM conformant data with a minimal set of deliverables including raw datasets and CRF and will put all other information sources aside for simplification reasons (however, for legacy studies, even such minimal set is possible).

Already the procedure of deliverables transfer reveals some potential for improvement. Due to the fact that source data does not need to be changed, the question arises if it is necessary to migrate the data from one system to another? In the best case, the data will be kept in one place, but be accessible for transformation to SDTM from other tools. Further, data might come from customers or older legacy studies stored in SAS® datasets or other formats. So the process should provide means to access external data sources on the one hand, and on the other, to import and store datasets, metadata descriptions and specifications ready to be used for transformations.

MAPPING SPECIFICATION

The first stop on our journey is the creation of the mapping specification – a crucial part of the whole process. Classically, it begins with annotation of the CRF according to target SDTM structures. The data manager who is typically responsible for this task uses PDF commenting tools and a solid knowledge of CDISC models and raw data. This process is time consuming, so innovation is advancing into this area: Next generation EDC systems can natively generate target annotated CRF - which is obvious as they support the study from the very beginning - which will simplify this step in the future. However, older legacy studies are excluded from this convenient situation due to the fact that CRFs are stored in different formats, either scanned or as searchable content. For this reason, a certain amount of manual work will remain here in spite of the technical progress, even if some technical aid can be imagined.

Once the CRF is target-annotated, it is passed along the workflow to the role creating the mapping specification (let us call this role “mapper” for convenience reasons).

Let us have a look at the source side first. Classical tools (e.g. SAS PROC CONTENTS) certainly deliver an exact description of the raw datasets. Though very useful, this information cannot be directly used for

mapping. With every dataset, the mapper needs to create source structures manually from scratch as they are different from study to study, from dataset to dataset. Simple multiplication of the number of studies with the number of datasets in a study shows that already at this stage some automation would save a significant amount of time and routine work. A more user-friendly solution would provide the possibility of selecting the attributes from a list, thus at least eliminating the need to type in attribute names manually and reducing the number of errors.

In our ideal world, a tool is available which allows us to automatically extract metadata definitions from raw datasets and use them as data definition elements in mapping specifications and programs (we will see later that both outcomes don't need to be considered separately).

In addition the source metadata definitions can be used for quality checks before carrying out the transformation: while appending several datasets, for example, incoming datasets can be checked for conformance with the metadata definitions to eliminate such basic errors as deviations in attribute types or in the number of attributes.

The target side looks slightly differently. The structure definitions for CDISC SDTM are publicly available on the CDISC web page (www.cdisc.org). As a harmonization model, the SDTM structures do not completely reflect the clinical reality of particular studies by default. Typically, sponsors collect more information than required for submission. When converting studies to SDTM, this information should not get lost and it might be necessary in the future. Some studies might not contain all attributes available in SDTM. The term SDTM+ incorporates company- or study-specific adjustments of the standard SDTM structures - be it extension or reduction. The SDTM model allows model-specific modifications with regards to domains and attributes as long as certain rules apply.

Tracing the usage of SDTM structures on the target side, we can see that they are used or referenced in several outcomes of the process: in the target-annotated CRF, mapping specification, mapping programs and define.xml. In addition, the submission information spans several interconnected dimensions (or multi level metadata, in other words, which are linked together). With increasing complexity, its consistent management becomes a challenge. Considering the fact that the core structures do not change very often, the management of target metadata reveals a great deal of potential for reusability within organizations and studies. However, for this reason, the metadata management mechanism must remain flexible to address tailoring and future model changes.

Imagine how comfortable and time-saving it would be to define the core target metadata only once and then reuse it in different mapping projects and studies to create the required outcomes!

This is exactly the approach our metadata driven world is built on: to begin the mapping process with the definition of metadata templates for target structures. Such an approach brings enormous synergies: Following its internal SOPs, an organization can create a set of predefined templates with company-wide standards, possibly going beyond pure SDTM and containing all relevant attribute definitions.

The templates contain all core information such as domain and attribute level metadata as well as links to controlled terminology and/or dictionaries already pre-configured. Promulgated to the study level or other intermediate organizational levels, the templates can be adjusted to concrete needs: not available (and optional) attributes can be removed from the target metadata definitions; additional study specific elements can be added. This approach reduces the work in a mapping project to small adjustments, minimal in comparison with setting up the structures from scratch. Certainly it means more work initially, but allows users to benefit later from reusable templates. The templates can be used in many ways: in mapping specifications and programs, for SDTM compliance checks as well as for creation of define.xml. These four advantages are a powerful reason to invest more time in the beginning in order to have time and resource savings at the end when time is usually critical.

Certainly, deploying such an approach requires a central metadata repository (for an overview of general requirements to the metadata management see Kutsenko 2009). If implemented in an organization, the repository would provide controlled access to templates for the roles involved in the mapping process and would support means to work flexibly with reusable templates at the project and study level with version control and configurable workflows. CDISC has also been actively working on its Metadata Repository (CMDR) to present a pilot in the coming winter (CDISC 2008).

Once the target structures are fixed, they can be related to the available sources, with conversion functions, if necessary.

What sounds easy might turn out to be an extremely time consuming task which might even involve different organizational roles and knowledge domains in resolving ambiguities. In the best case an attribute required or expected in the target structure is directly available in the source definition (even if with a different name). In this case a simple 1:1 assignment can be done (as practice shows, simple assignments account for the largest part of all assignments and transformations to SDTM). If you consider, for example, the LB domain with 44 attributes in the standard CDISC structure (CDISC 2005c), a little help for the mapping definition (which might consist of several transformation steps built on each other and include transpositions) would save a great deal of time for this routine job and reduce the number of typing errors.

In our environment, automations are available for such tasks starting with a drag-and-drop interface for attributes with different names (unfortunately even the smartest tools can't interpret the data - the human brain still is and will be required) up to automatic assignment of attributes with the same names.

More difficult cases (e.g. unclear or unavailable sources which happen often in older legacy studies or specific transformation functions) may require the involvement of data managers and programmers who might be globally distributed at several company locations. Imagine a situation where a mapper has to resolve some ambiguous source-target relation and to create consequently a complex conversion algorithm. The mapper would need to specify the problem to solve (unclear source for the attribute A), contact a data manager or another person who would help with the source side (let's assume, for example, that the attribute A requires a complex transformation based on several source datasets). This person would need to store comments on the attribute which would help the programmer to develop an efficient algorithm. Then all the information is passed to the programmer who can develop an algorithm for this particular conversion (in our example, the mapper can not program SAS very well).

In order to make the process more efficient, communication between the team mates should be supported by tools which would help them find their way in complex structures and focus on problem zones, and to work according to specified workflows and processes.

In our visionary world, communication between all roles involved would happen in an environment which provides effective collaboration tools and automatically documents communication flows. In such an encapsulated environment it is an easy task: you can send a role-bound notification (via email or as an internal system message) to your colleagues pointing to specific problems in a mapping definition. They can review the specification online, provide additional information as attribute comments and forward the issue back and forth for resolution according to the defined workflow. You can even start a discussion in a linked forum or discussion group. All communication flows are stored related to the mapping definition, so the communication and decisions are transparent even later for new personnel.

For numerous standard transformation algorithms, it is even not necessary to involve programmers – drag-and-drop GUIs are available (e.g. for selection of available attributes and formats) and can be assigned even by users who are not very familiar with SAS. To deal with terminology mapping, a GUI-driven profiling tool helps us to compare several datasets and to detect differences in terminology used and consequent mapping needs. The mapping rules are stored in the mapping definition (also supported by drag-and-drop GUIs) and used later on.

The principle of visual definition helps to reduce the SAS skills required for mapping to an absolute minimum.

At the end of this key process step, a well-defined mapping specification is available. It is stored in the repository as a metadata-based mapping definition of sources, targets and their relations including conversion algorithms, is accessible in a dedicated view to all users with access rights online, and is up-to-date and exportable in different formats (as an overview or full mapping report, for example).

MAPPING PROGRAM DEVELOPMENT

Now we have reached the stage with the highest inefficiency in the whole mapping process: In a classical process, after the mapping specification is finalized, programmers have to implement the conversions in mapping programs based on the developed mapping specification.

Looking at the mapping specification from the abstract perspective, you can ask yourself the following question: if the mapping specification is supposed to be a 1:1 documentation of the sources, targets and algorithms, why can't I automate this task? At this point, we have the exact descriptions of the sources

(derived from the raw datasets) and targets (adjusted standard templates); we defined all assignments and standard conversion algorithms (supported by GUIs - so standardized). The programmers were involved during the specification phase to develop specific conversions algorithms. These algorithms were stored using the syntax of the mapping program's target language directly in the mapping specification. It would be obvious to generate mapping programs from the specification!

You're absolutely right! In our visionary process, programs don't need to be manually developed. All inputs described above are sufficient to automatically create the mapping programs (e.g. SAS) based on the mapping specification. Just one button click is required when we are supported by a smart program generator! This frees your programming resources for more complex tasks than programming of 1:1 assignments and standard algorithms.

DATA CONVERSION

As you noticed above, we haven't mentioned real data yet (except for the first step to create source metadata from raw data). Real data only needs to come into play at this step – when the generated mapping programs are executed with real data to be converted to the target SDTM structures.

To keep the generated mapping programs flexible and reusable with different datasets, no hard-coded information about the location and names of the datasets to be transformed is stored in programs. They contain only parameters for all objects involved. While running the mapping program, the user simply assigns real objects to the program parameters (drag-and-drop support is also available in the system for this purpose, together with the storage of used parameter values).

By default, generated mapping programs contain built-in checks for source data: existence of the attributes and their conformance with the metadata definitions used in the specification. These checks are the first step of the quality control. If the source checks have been passed and there are no logical errors in the definition ("human error"), datasets compliant with the mapping specification are available after the mapping program execution.

An important question arises at this step: Are my converted datasets conformant with SDTM? How could this process be made efficient? One solution is to use a SAS check program and to integrate this program into the conversion workflow. This brings several advantages: not only can I launch the check program directly after the conversion and so see the results of the quality checks on the converted datasets immediately; I can also, with a stand alone check program, obtain the flexibility to enhance the checks with my specific ones. (Many organizations have already implemented own checks. In addition, CROs may need to execute specific checks required by their customers.) Integration of standard and specific checks into the workflow makes the whole process flexible, convenient and fast.

DEFINE.XML

We are approaching the last stop of the mapping process – define.xml (CDISC 2005a). In order to get the submission bundle ready, just a few more actions are required - time consuming in the technology-conservative world, a few mouse clicks away in the technology-driven one.

There are certainly many ways to create define.xml. At this stage in the process, we already have target metadata of SDTM domains which we also used for the creation of the mapping specification and generation of the mapping programs. Controlled terminology and value level metadata need to be created - be it manually or with a mapping program (this program follows the same principles as described above for the SDTM target datasets). Transport files are created from target datasets with dedicated format converters.

The newest version of the official CDISC style sheet for define.xml supports page references to annotated CRFs and allows navigation from define.xml directly to the pages of the annotated CRF were the attributes appear. This means that the full annotated CRF needs to be scanned for attributes! For this task, we developed a tool which automatically scans annotated CRFs for domain attributes used, extracts page numbers and writes them into the attribute level metadata. It's as easy as this – define.xml requires only a few mouse clicks! With such automation, define.xml can be created within the shortest time - from less than an hour up to several hours. It means a significant improvement in comparison with manual programming of define.xml. In addition, the whole process can be easily repeated to update define.xml with changed content.

FINAL NOTE

By the way, the process described above is not a vision; it is a reality with Entimo's tools! entimICE DARE (Data Analysis and Reporting Environment) provides a collaboration environment with a central repository, configurable workflows, notifications, versioning support and many other useful features. A smart mapping tool (entimICE Mapping) creates mapping specifications and programs from metadata definitions stored in the repository. Define.xml Generator bundles define.xml and SDTM Checker validates the conformance of datasets with SDTM.

Finally, the illustration below depicts the described process as an overview with the stages of our journey from raw to SDTM conformant data.

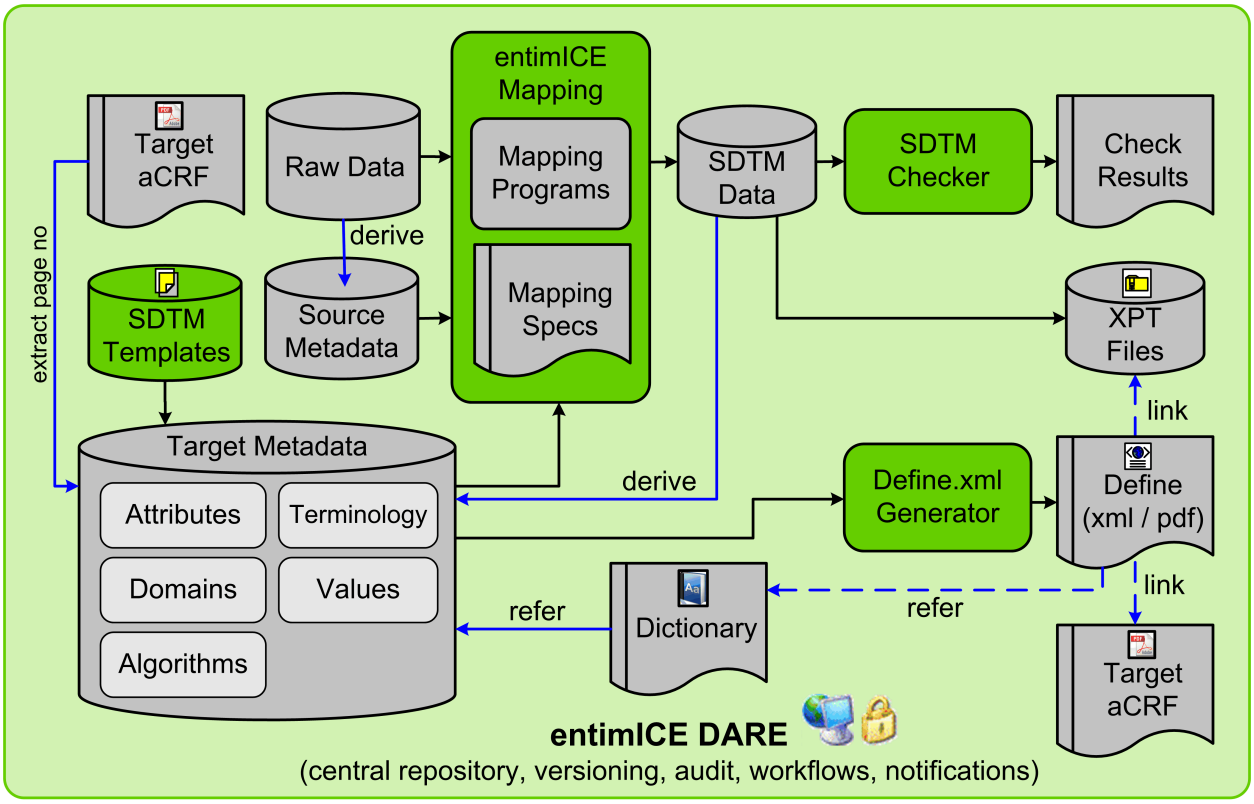

Illustration 1: From raw data to submission with entimICE tools

REFERENCES

CDISC 2005a, Case Report Tabulation Data Definition Specification (define.xml) Version 1.0, http://cdisc.org/models/def/v1.0/index.html, viewed 2 June 2009.

CDISC 2008, CDISC Meta data Repository. Enrichment & integration of CDISC Data Standards toward semantic interoperability. Business Requirements, http://iis2.imise.uni-leipzig.de/beanformer/search/CMDR/, viewed 2 June 2009.

CDISC 2009, CDISC Meta data Repository. Enrichment & integration of CDISC Data Standards toward semantic interoperability. Pilot Specification: Business Scenarios (working document), http://iis2.imise.uni-leipzig.de/beanformer/search/CMDR/CDISC_MDR_-_PilotSpecification_v0.4.doc, viewed 10 July 2009.

CDISC 2005b, Study Data Tabulation Model (SDTM) Final Version 1.1, http://www.cdisc.org/models/sdtm/v1.1/index.html, viewed 2 June 2009.

CDISC 2005c, Study Data Tabulation Model (SDTM) Implementation Guide Final Version 3.1.1, http://www.cdisc.org/models/sdtm/v1.1/index.html, viewed 2 June 2009.

Kutsenko, D 2009, An analysis of requirements for metadata in metadata driven mapping projects and organizations, Data Basics (SCDM), vol.15, no. 2, pp. 12-13.

CONTACT INFORMATION

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

Dimitri Kutsenko
Entimo AG
Stralauer Platz 33-34
Berlin / 10243
Germany
Work Phone: +49 30 520 024 100
Fax: +49 30 520 024 101
Web: www.entimo.com