PhUSE 2010

Paper DH03

How to improve quality control in a data conversion process? By extended usage of metadata!

Dimitri Kutsenko, Entimo AG, Berlin, Germany

ABSTRACT

This paper introduces a new, integrated methodology of quality control (QC) in data mapping projects: Following a mapping process, we will first look at the points in the process where the QC can be generally applied. With a metadata based vision in mind we will analyze different means of improving the QC by extended usage of metadata. It will be discussed how this critical activity can be made efficient and flexible, can eliminate many time consuming, manual tasks and reduce later costs of errors by shifting significant parts of the QC work to an earlier stage in the process. This brief discussion will be conducted at an abstract level, allowing the transfer of the concepts to different processes and target models. The particular examples taken from the CDISC SDTM world are used only for comprehension reasons – the principles are generic and model independent in their nature.

INTRODUCTION

When analyzing different processes for data conversion to arbitrary data models including SDTM, one process aspect appears especially critical and shall be considered separately - Quality Control (QC). Due to the fact that data quality control can be understood very broadly (in the area of data management, for example), the scope of this paper is narrowed solely to the QC activities during data conversion.

It is obvious that studies which were set up following any standard – be it company specific and/or industry common - bring up a different level of task complexity and efforts with regards to data conversion than legacy studies. In legacy studies, it might even happen that a significant share of the information describing data structures is not available in electronic form, and in the worst case hardly retrievable from paper documentation due to the time lag. In addition to the information gap, experts knowledgeable about the data structures might have already left the organization. The requirements will also vary depending on the concrete goal of data conversion: study submission or data pooling.

The perspective of a pharmaceutical company on data conversion would also differ from a CRO's world: CROs have to master many different models of their numerous clients, often equipped with incomplete deliverables, under strict time and budget restrictions – which makes the whole task more challenging.

What should the QC function look like to offer maximum efficiency combined with process flexibility and to reduce tedious manual tasks to a necessary minimum?

The primary goal of this paper is to suggest a high-level vision of the QC task, even running danger not to consider some specificities obvious for an experienced practitioner. The concepts and practices presented in the paper shall be kept generally applicable and thought-provoking. They shall be adapted to concrete processes and requirements later on "in the field".

STANDARDS MANAGEMENT

Let us move to the very beginning of a generic mapping process.

A general approach of creating company-wide standard models would offer significant advantages also with regards to data quality control. (The process of negotiating, developing and managing the standards within an organization is beyond the scope of this paper.)

CDISC SDTM is one example of an open, standardized data model (CDISC, 2008). On one side SDTM can be used for study submissions. Beyond that the model allows company specific extensions as long as certain rules are adhered to. For these reasons it has been widely used as a model foundation to create organization specific standards.

A bold challenge related to standards is the fact that standards are typically arranged upon several interlinked hierarchy levels (e.g. global, project, trial). The standards from higher levels are adjusted at lower levels to reflect particular requirements of every level. Let us consider an example where specifications and datasets are created at the study level based on the study-level standard domain metadata which is related to the global standard. (The global standard contains structural domains containing dataset descriptions based on standard attributes.)

Now, the global standard shall be changed. Is it an easy task to keep track of all structural changes to standards across all hierarchies? No way! **Impact analysis** is required in this case which would determine the impact of the changes to the global standard across all hierarchy levels, all instances of the global standard and include the detection of effects on the datasets at any level.

Already this phase may contain large potential for optimization. However, the impact analysis task is hardly feasible in a manual way over longer time periods and change iterations.

Let us look at another possible issue: A global standard shall be adjusted at the study level. How to avoid that standard attributes contained in the global standard and required later for submission don't get lost during the study-level standard definition phase? If not (or wrongly) taken over, the attributes will not be implemented in mapping specifications and conversion programs and will not be available in datasets later on. At the latest during the submission, the problems will be uncovered and sent back for resolution, meaning project delays and involvement of additional, unplanned resources.

Descriptions of dataset structures (henceforward "domain metadata") themselves can and should be subject to comprehensive, cross-level quality control. This sounds trivial and obvious, but has wide-reaching consequences, especially given the number of conversion-related outcomes where domain metadata are used: annotated CRFs, mapping specifications, conversion programs, datasets and, finally, the model content overview (e.g. define.xml in the CDISC context).

Every error or ambiguity in the target domain definition will result in multiple added costs if passed down the workflow! For this reason, an organization must make every effort to discover such issues as early in the process as possible and make control of domain metadata an integral part of the quality control function!

Two quality control mechanisms can be thought of with regards to domain metadata:

On one side, domain metadata shall be compared across hierarchies to detect deviations (e.g. at the study and global levels). This helps eliminate many potential issues very early in the process. If deviations are detected they must become subject to review. However, manual comparison of dataset attributes is laborious (especially considering their significant number in a submission model and repeated character of the task). An automated **report of metadata deviations** is required which would allow for comparison of selected dataset definitions and detection of structural differences in them. Such a report could be also a sound discussion basis for following reviews.

Another visionary approach (which is already a reality, at least for Entimo!) is to implement **higher level metadata definition rules** (metadata's metadata to misuse the metadata term). Such metadata-related rules automatically apply to every metadata domain and include metadata checks: Column existence checks, uniqueness checks, value type checks, keys, patterns, ranges are only a few of them. In this case, if somebody tries to remove an attribute specified as required from the domain metadata or to enter an attribute key which is not unique, but is expected to be unique, he or she is warned about the conflicts.

Certainly, in order to make these mechanisms feasible, an encapsulated environment and adequate tools are required where domain metadata can be stored in a structured format supporting automated evaluation (e.g. in a repository).

MAPPING SPECIFICATION AND CONVERSION PROGRAMS

Moving further in the mapping process toward actual data conversion, we stop and glance at the mapping specification. Similar to the domain definition problems, errors in a mapping specification are critical if propagated down the process: A mapping program which might and often does include many transformation steps built upon each other would be developed on the wrong specification and will need to be corrected later on.

The simplest source of errors might be the widely used “copy-paste” mechanism. Ideally, it should be possible to directly **(re-)use the domain metadata** mentioned above. If it were possible to automatically use standard domain metadata and to derive and use source side dataset descriptions, the control of the two mapping sides would make the whole step less error prone. In this case, you need only to take care of conversion algorithms to guarantee conformance with the target dataset structure!

It is definitely worth investing time into the development, validation and introduction of **standard conversion algorithms** in your organization. In this case mapping specialists can depict the transformation logic by configuring available (validated!) algorithms from the library. This would minimize QC measures at this step to the review of the transformation logic and specific conversion algorithms not included into the standards library.

In our technology driven world, we go even one step further and use an innovative “single-source” strategy: Its key idea is to merge two classically separate steps - creation of the mapping specification and development of mappings programs. The unified, metadata based mapping definition carries all the necessary information to generate both products with a single mouse click and allows to re-use standard algorithms and domain metadata. Such mapping definition enhanced with textual comments is also a solid base for issue clarifications during mapping development and review activities.

Another means of quality control are **data checks against metadata**. This is especially reasonable when several datasets shall be pooled with the same mapping program. Whether to design a separate check facility or to integrate checks into conversion programs is more a question of philosophy. In any case checks of data against used structure metadata would be helpful to avoid errors caused by wrongly applied conversion algorithms and shall be executed prior to data conversion. The results of such checks should be reportable for documentation and correction purposes.

Once datasets have been converted, the next reasonable question arises: Is their content compliant to the target model (think of SDTM 3.1.1 or 3.1.2 rules, for example)? Usually the **content validation** is done at the end of the mapping process. Why shouldn’t the content checks be carried out directly after data conversion? Certainly, not all standard checks are applicable in this case, but even a subset of checks can save a lot of time and effort later on. One by one, new converted datasets can be added to the check runs until the domain list is complete and all cross-domain checks can be executed.

At Entimo, we use the same domain metadata also as a basis for the content checks. This helps us support organization-specific models and integrate additional dataset attributes into the check workflow. Such flexibility guarantees that the formal aspects and requirements defined by an organization are validated, thus leading to improved data quality.

CONCLUSIONS

In the course of this brief paper, we stopped by and looked at the following generic mechanisms which can improve QC during data conversion:

- Higher-level metadata definition rules
- Definition of standards (especially standard domain metadata and algorithms)
- Impact analysis and domain comparison across hierarchies
- Direct re-use of domain metadata in mapping specifications and programs
- Data checks against structure metadata
- Content checks of target datasets based on domain metadata

The key message of this paper is that data quality control activities carried out as early in the mapping process as possible combined with the integration of reusable metadata definitions, standard algorithms and metadata-based tools significantly shortens the process duration, supports more efficient data QC and consequently reduces the costs of errors and projects in total.

Already now, the clinical world provides powerful examples of metadata based process implementations with off-the-self tools (Entimo’s tool set entimICE® is one of them).

Boost your processes with metadata based data quality control and enjoy the benefits of being quick off the mark!

REFERENCES

CDISC 2008, Study Data Tabulation Model Implementation Guide: Human Clinical Trials, 3.1.2 Final, CDISC Submission Data Standards Team, http://www.cdisc.org/content1209, viewed 5 September 2010.

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