PhUSE 2010

Paper DH10

Converting Data to CDISC Models: Research Project on Metadata Extraction, Exploration and Pooling

Ronald Steinhau and Dimitri Kutsenko, Entimo AG, Berlin, Germany

Table of Content

INTRODUCTION

The starting point of an innovative research project conducted by Entimo AG under the patronage of the German Federal Ministry of Economy and Technology is the CDISC SHARE Pilot Report and its conclusions that the currently technology available in the pharmaceutical sector is not sufficient for “viewing foundational terminologies“ and shall be improved to “facilitate a robust search of terms and provide users with views that make comparison of the data elements easy and instinctive” (CDISC, 2010).

The process of transforming clinical data to CDISC SDTM and/or enterprise pooling structures might be very challenging. The difficulties start already with the creation of the mapping specification. In the worst case, structure and content are not exactly known, proper definitions are embedded in documents stored elsewhere in paper form and knowledgeable experts have already left the organization - every data item has to be tagged with metadata from the scratch in this case. Old specifications, if retrievable at all, might contain textual descriptions of the mapping logic. However, one logical step requires often several transformation steps where source data even might be located in multiple datasets. Blind usage of available mappings would provide no big help, even with tool support.

Another aspect of the mapping process is related to codelists: Codelists are extremely helpful for understanding data elements, especially if comprehensive metadata is missing. The rationale is strong for matching codelist values (from older codelists) to standard terminology or current codelist versions. However, manual matching of codelists with many items is time consuming and error-prone.

Metadata (or data element) definitions are needed in order to help understand the meaning of datasets. The effort of creating metadata for existing studies can be very high, especially considering the large number of existing (ongoing and legacy) studies which wait for their transformation to CDISC SDTM (like) formats. In addition, with takeovers and M&As, companies face similar problems: The formats of studies from merging companies might deviate from each other and from the established standards. The urgent need to accelerate the mapping process is evident.

GOALS AND CONCEPTS

In the project we have extended the research scope to include not only terminology, but also other metadata types involved into the mapping process.

Looking at the reality, the first step – analysis of existing data sets and assigning meanings to the data columns - is only the beginning of the exploration process. More complex relations between data and the hidden rules of their interpretation (e.g. “when was the first visit and how relates the age of a patient to it” or imputation rules for missing values) can not be easily caught by visual data exploration.

The project goal is to investigate mechanisms for automatic recognition and extraction of metadata based on available clinical data. As a project outcome, an intelligent solution shall be developed to explore data structures and their relations and re-use metadata and mappings from comparable studies managed in a central information pool. User shall receive support in specifying all necessary information to simplify the creation of mappings.

As full automation is hardly possible, ways to create and present open issues to the user for resolution in an interactive way shall be created. Mechanisms to enhance rule taxonomies with resolved issues shall be investigated ensuring non-ambiguous, non-duplicate metadata definition.

The key is the ability to support users in finding necessary information: the mechanism shall detect similarities or let users query for such, related not only to whole items and lists, but also to metadata parts. Relations and dependencies between models (e.g. BRIDG and SDTM) shall be involved into the analysis. Innovative principles such as fuzzy logic, statistical analysis and adaptive learning shall be utilized to develop domain specific knowledge systems (see also sections Architecture and Technical Base). By specifying queries and/or rules, metadata shall be found which are useful in new situations, even if modifications are required. The described principles shall make also recognition of synonyms possible. Such search tactics as “best match per domain” or “best match for all data structures” shall be used to define user priorities and optimize search results.

After finding information the next smart step would be to apply this information to the new situation – e.g. to make suggestions to the user how to adjust an existing mapping to a different structure. This would bring significant time advantages in the process to develop consistent and high-quality data definitions in comparison with manual creation of study metadata and support their harmonization. Extracted metadata of various types will be pooled and used for mapping of clinical studies to CDISC models and beyond. Different versions of metadata shall be managed and be accessible simultaneously.

A major project requirement is the cross-model capability with model tagging: Metadata of different models such as BRIDG, SDTM and company specific models shall be involved into the analysis process and allow the user to move between the models and hierarchy levels.

At any time the user shall be able to relate metadata with each other - by assigning contexts and relate metadata from different contexts. (A context is a tagged domain of metadata such as BRIDG, SDTM, a clinical study, a domain in the clinical study pool, a set of studies, a mapping process, for example.) Context switch allows the user to focus on the filtered information instead of being disturbed and overloaded by all available information.

Creation of flexible mechanisms for visualizing and reporting metadata of different types (such as codelists, domain definitions, mappings) is another key element of the research project.

The metadata explorer shall contain intense reporting capabilities which allow users to browse:

- Metadata of a selected entity (e.g. dataset or data column) in one or more contexts

- Relations between metadata in different contexts
- Available mapping specifications in different layouts

Information shall be exportable in a table oriented form (Excel), in an XML oriented format and as a domain specific language (DSL), whereby the first is optimized for human readability and the third one for automatic processing (e.g. import into other tools).

ARCHITECTURE

The solution aims to make use of existing mathematical and statistical methods: Novel pattern recognition principles such as fuzzy logic, statistical data analysis and adaptive learning shall be applied to clinical trial data and metadata in an innovative way. A domain-specific expert system shall be created. The following figure displays top-level parts of the system:

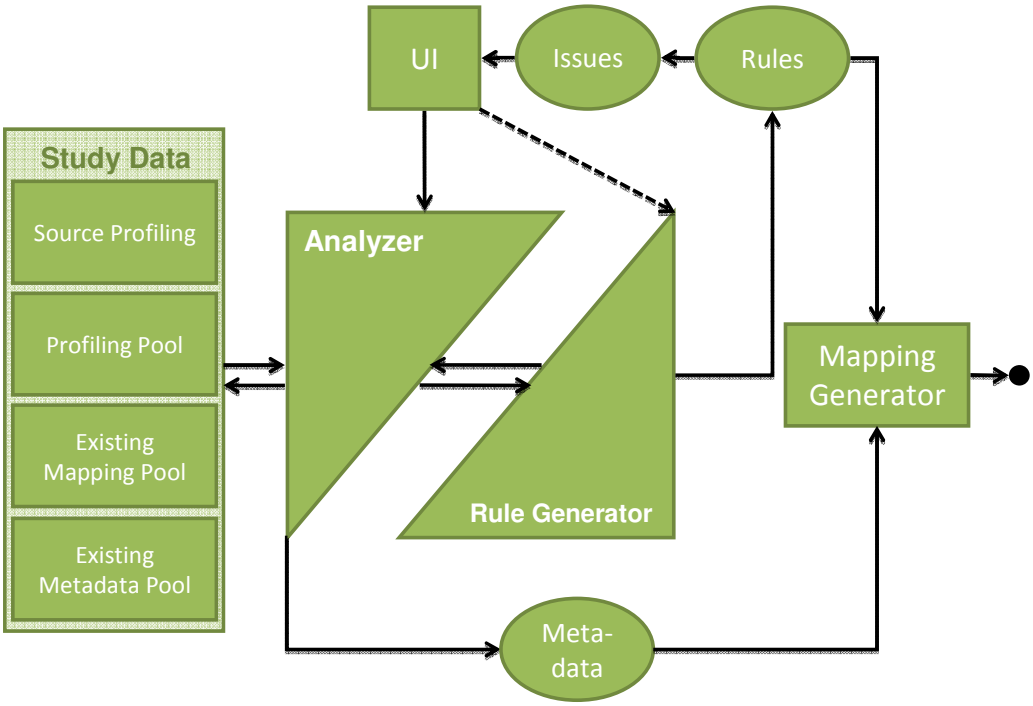

Illustration 1: Top level architecture

The module architecture distinguishes between the following categories:

1. Information collection / source profiling (e.g. according to name patterns and statistical analysis of values)
2. Information processing / analyzer (e.g. based on profiling and metadata pools from different studies and rules from the rule engine)
3. Rule definitions in the rule engine / rule generator (e.g. domain or cross-domain rule catalogs)
4. Mapping generator (e.g. mapping specifications)
5. User interface (UI) for presentation and resolution of open issue in the mapping definition

The system will be able to automatically derive issues from older, similar studies (if properly maintained by users). This allows to create a "minimum set of questions to a study" which are presented in a user interface to specialists. The answers flow back into the rule generator and

will be directly used in the mapping process. The final step is the generation of a mapping sequence in different formats (also for post-processing purposes).

The great advantage of the mechanisms is massive time saving compared to manual study analysis and mapping definition.

TECHNICAL BASE

The model-driven approach - supported by the Eclipse Modeling Framework (EMF) - is extensively used in the project as a technical environment. For the development of the DSL the software tool XText is used. XText supports grammatical formulation of a language and its model-based analysis.

SAS datasets, CSV, Excel and XML files as well as direct database queries (JDBC) are supported among many other file formats in order to import information and/or directly access external data and metadata sources.

A broad range of mathematical and statistical methods is already implemented in the Java community and available as modules. Usability of such rule systems as JTP (Java Theorem Prover, object oriented "Reasoning System") or JFuzzyLogic (library for the implementation of fuzzy logic aspects and Bayesian classification) is being evaluated in the project context as every system has its advantages and disadvantages.

CONCLUSIONS

While moving forward phase after phase in the project, we obtain very promising results of pilots and feasibility studies related to the described mechanisms. Though there is still a way ahead to a rule-based expert system, our vision turns out to be a steady source of new learning experiences, helps us become more effective in many aspects and brings us closer to the technology- and metadata-driven future.

REFERENCES

CDISC 2010, CDISC SHARE Pilot Report, http://www.cdisc.org, viewed on 2 September 2010.

CONTACT INFORMATION

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

Ronald Steinhau / 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

Brand and product names are trademarks of their respective companies.