

Research Project on Metadata Extraction, Exploration and Pooling: Challenges and Achievements

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

Table of Content

INTRODUCTION.....	1
WORK PACKAGES	2
Development Preparation.....	2
Specification / Modeling	2
Analysing Data.....	2
Metadata and Rules	3
Smart Recommendations.....	3
Mapping / Meta-Generator	3
Development of Modules.....	3
Testing, optimization and system documentation	3
ADVANCED WORKFLOWS	4
ACHIEVEMENTS AND OUTLOOK.....	6
Problems and Limitations	6
Look into the Future.....	6
CONTACT INFORMATION	7

INTRODUCTION

Entimo AG conducted under the abbreviation CDEE ("Clinical Data Extraction and Exploration Environment") an innovative research project - funded by the Federal Ministry of Economics and Technology - for the intelligent optimization of data and metadata use in clinical trials. The work focused on the analysis and recognition of data and metadata on the basis of similarities and profiles to facilitate mapping processes or pooling tasks. The use of adaptive learning and query of user-specific information were an additional goal of this project. A major aspect of the project was the intelligent classification of information from clinical studies based on existing metadata systems in order to accelerate the mapping process.

In the preparation phase, technical requirements were analyzed that had to be solved in the project scope in order to adequately address the technical aspects:

PhUSE 2011

- A performing database-based repository for efficient analysis and reporting of clinical data.
- An extensible interface for data import, which allows evaluation and analysis during the import process.
- Advanced metadata management based on the ISO 11179 for Metadata Repositories, which allows creating references between different metadata (with BRIDG as a universal model).
- An extended "domain specific language" (DSL), whose foundation has already been developed. The (new) commands shall support all features of the CDEE from import through analysis to mapping.
- Embedding of the DSL and the CDEE functions as an extension to the existing entimICE software system to be able to make use of its existing functionality and infrastructure.
- Additional analysis and use of a library for machine-aided learning. This library proved to be more useful than the logic-based control systems that were studied in the work package 3.

WORK PACKAGES

The work was carried out divided into several work packages which are listed and briefly explained below.

Development Preparation

This work package laid the planning ground for the project. Requirements were gathered and classified in numerous discussions and meetings.

A selection of software systems to be evaluated was made. In the following discussions about the role of "smart learning", additional software packages were added to the evaluation list, although there are no rules engines in the strict sense available on the market.

The development environment was Eclipse Helios/Scala IDE (version 3.6) established and equipped with the latest Scala IDE. Scala was chosen as a development language because of its interoperability with Java and its unique capabilities to express complex algorithms in a compact and more readable way and take advantage of parallel processing more easily.

It was determined that the entimICE platform will be used for development. As a configurable work environment for clinical studies, the platform allows configuration of actions, editors and many other aspects for each and every node and information type. CDEE is aimed as an exploration environment in the project area. Although this decision required more effort in the development than a context-free prototyping, it allows usage of existing functionalities and guarantees better migration of the software developed in later production versions.

Specification / Modeling

Analysing Data

The possibilities of statistical data analysis and the presentation of its results were examined. This resulted in the need to represent the results as interactive reports that allow to navigate from the report on details of the tree, for example, to display the actual data. The requirements for interactive reports have been finalized.

A list of statistical and quantitative analysis methods to be carried out on data sets and a specification for the flexible implementation of this analysis were created. It turned out that

PhUSE 2011

all analysis operations must be performed in streaming mode due to sometimes large amounts of data.

The work package was also used to specify an important aspect of CDEE: the comparison of data and the search for similarities at the level of data, metadata and mappings. These commands, their variants and parameters have been specified in DSL.

Metadata and Rules

Several aspects of rules based on metadata have been tested against practical usability. The work package was also used to specify the aspect of templates and enhanced metadata in the system. These include such metadata systems as BRIDG, SDTM and ADaM. A model was created for the storage of metadata in terms of metadata repositories including statistical profiles of attributes. Due to the importance of a clinical database-based repository, this WP was also used to specify its properties in detail.

Smart Recommendations

An approach for "smart recommendations" has been developed in this work package. It was specified that in places where the system possesses information on previous decisions of the user or characteristic properties of data and/or metadata collected, an interactive dialogue shall be provided displaying recommendations for decisions. The outcome of these decisions is saved for further recommendations.

The details for such interaction - that can occur at various points in the system (e.g. during import and mapping search) - have been developed.

Mapping / Meta-Generator

In this WP and based on the already existing entimICE mapping technology, the options were specified to:

- Generate an issue list for the user what adjustments are suggested based on real data (or their real metadata)
- Automatically execute mappings in the clinical repository

The work package was also used to extend the spectrum of mapping commands in DSL.

Development of Modules

The following modules were developed:

- Search module of similarities (at the level of datasets and at the level of mappings) with built-in capability for "adaptive learning".
- A streaming module for the statistical and quantitative analysis of data including automatic detection of primary key candidates.
- A comparison module for datasets (based on clinical data stored in the database)
- Interactive dialogs to query decisions whose results are included in the search module.
- Import modules for data (CSV, SAS, Excel) and metadata (BRIDG, SDTM, ADAM) with integrated streaming analysis.
- The database-based clinical repository has been partly developed and enhanced in the project.
- A module for interactive reports was designed to display analysis results.

Testing, optimization and system documentation

To verify the selected approaches test data have been manually created. For the clinical repository special test cases were developed.

The optimization was aimed mainly at the area of analysis due to large amounts of data. Constant attention was paid to the performance.

The system documentation and specification was designed in broad terms and continuously expanded in the course of the project.

ADVANCED WORKFLOWS

The entimICE CDEE support of data processing in clinical studies was extended by two major workflows:

- Import of data from different data sources
- Selection and implementation of mappings, e.g. to transform source structures to CDISC-compliant target structures.

The following figures illustrate these extensions in the workflows as well as innovations that enable these new procedures (MDR = Metadata Repository).

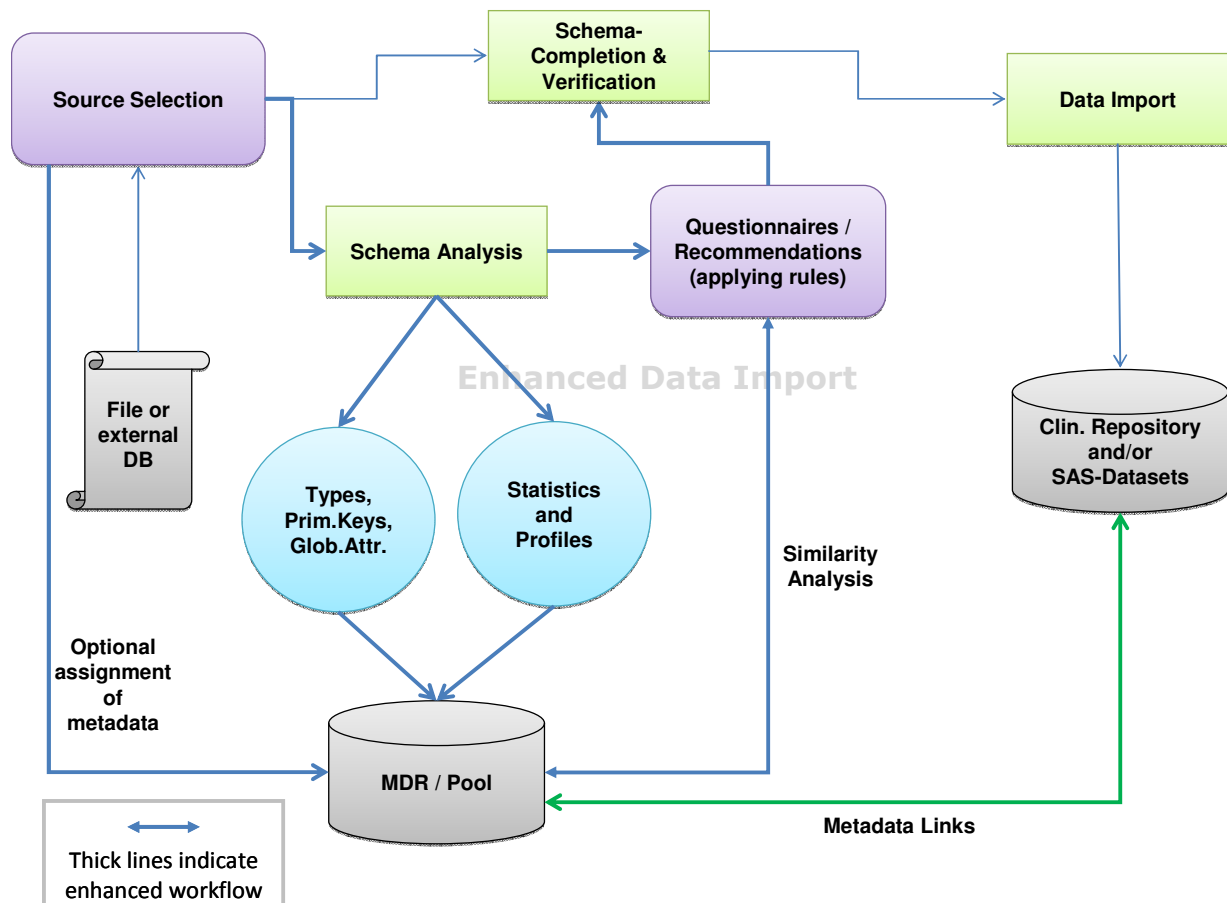


Figure 1 - Data import and data analysis workflow

The classical import functionality provided only a selection of related metadata structures, which were used to carry out schema verification. If this metadata did not exist, they had to be created manually. The new workflow (thick lines) supports an extensive preliminary analysis; appropriate metadata (including statistical profiles) is extracted from data. In case of ambiguity, the similarity analysis of existing data and related metadata is run to provide

decision help for the user. Comparative analysis based on statistics is especially important in this case to assess existing unique attributes of the MDR.

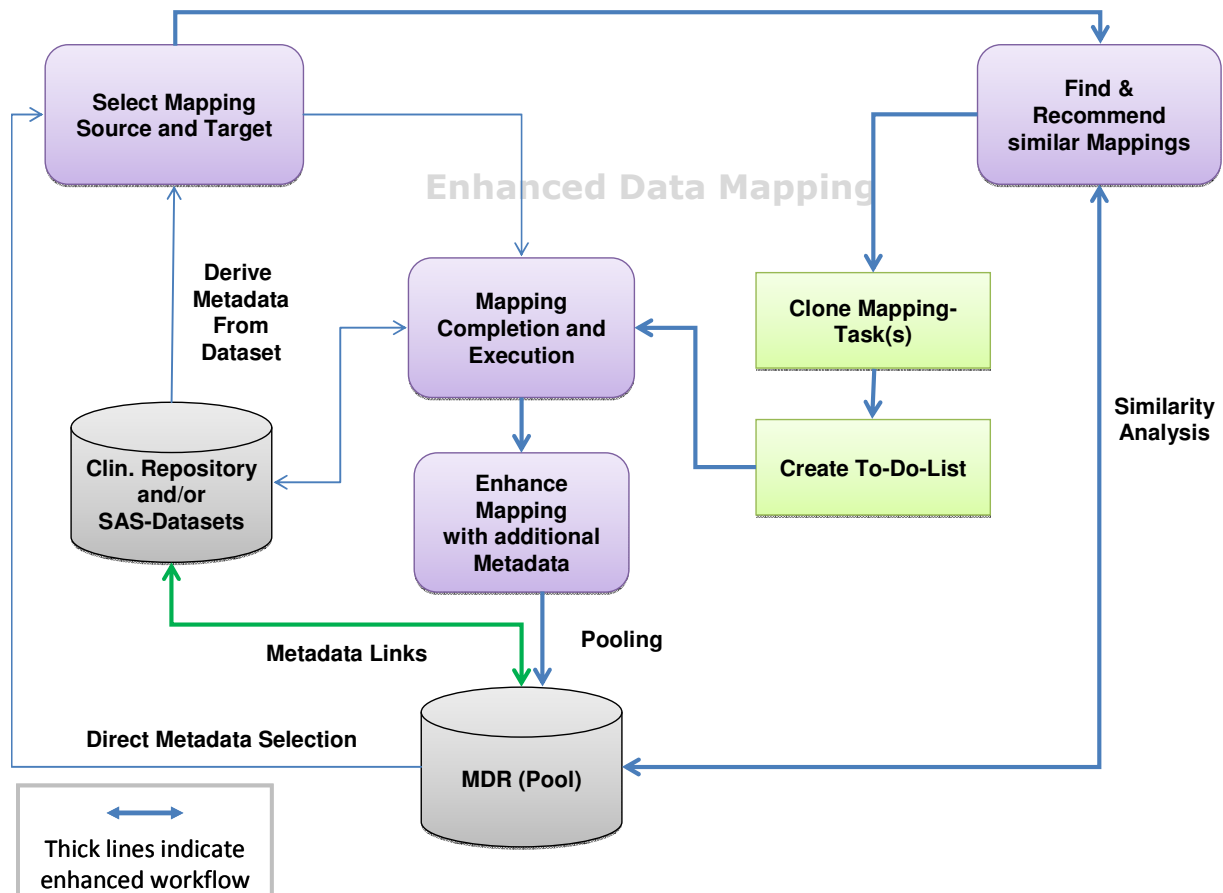


Figure 2 - Mapping Workflow Definition

A mapping is defined at the metadata level, since the actual transformation is often repeatedly executed with different datasets. Moreover, the transformation can be performed in different environments (SAS-centric, database-centric, etc.).

The new workflow allows you to search and display similar mappings that have been stored in the MDR. Statistical profiles, such as name and structure information, are also included into the search. Unique attributes of the MDR, as determined when importing the data, are also an important knowledge source. If there are several similar mappings, the candidates are listed in order of their similarity and the user can make a decision based on the shown information.

If a similar mapping was found, the new source information is used and adaptable tasks and actions are adjusted/reconfigured and used. Open actions are marked as "To Do" and the user must complete only this smaller part of the mapping.

Finally, the user can add the mapping to the available pool as a template (for future searches), refining search matches by identifying which structures are irrelevant (such as in the source-specific attributes).

In the focus of our effort, statistical methods needed to be identified that operate on the basis of values and allow assignment of attributes to existing unique attributes as well as detection of similarities. This largely corresponds to the classical discriminant analysis. In the scope of the project numerous statistical methods were investigated. The challenge was

to define a suitable similarity measures depending on the applied scale (nominal, ordinal, metric).

ACHIEVEMENTS AND OUTLOOK

The real innovation of the project can be summarized in the following key aspects:

- A novel, effective and comprehensive method of data source analysis (either beforehand or subsequently) with various statistical methods, analysis of primary keys and if possible classification of the variables found in the metadata repository. The manual selection by the user leads to a learning system.
- An implementation of a metadata repository - based on the ISO standard 11179 - with the possibility to extend statistical annotation attributes. The metadata repository has the ability to link one attribute definition to one or more attribute definitions, metadata to another domain, so that navigation through various metadata systems is possible in the future.
- Transformation of structural metadata into templates so that similarity analysis needs to consider only a subset of all columns.
- An adaptive similarity search in existing mappings, which considers all annotations stored in the metadata repository and identifies appropriate candidates.
- An implementation of a generic data repository for clinical studies which provides a balance of performance and generality and allows multiple versions of one dataset with complete audit trail of data.
- An advanced, domain-specific command language (DSL). The language can be edited in a dedicated editor with syntax checking and syntax highlighting.

The integration of these features into the existing entimICE platform posed a major challenge, but was solved due to the configurability and extensibility of the system interfaces (entimICE has been developed from the very beginning on the software architecture designed for extensibility). The overall functionality has been integrated directly into a prototype of the current entimICE platform – the basis of Entimo products.

Problems and Limitations

During the analysis of various logic-based control systems, it became clear that there are simpler and better methods for the solution of similarity detection. The actual logic, the selection of the correct analysis method based on simple properties such as the variable type can be implemented in the code.

During the analysis and implementation of statistical methods, it also became clear that some results could be improved if more contextual knowledge would be used: for example, the knowledge of the drug nature, the disease or the study. To follow this approach a workable system is missing that does not require elaborate input from the users' side. Future developments should therefore target in this direction to develop such system to extract, collect and make available necessary metadata during study creation.

Look into the Future

Management and combination of multi-level metadata is not available as a practical Metadata Explorer yet, as it would make sense from the user's perspective. No satisfactory solution to this UI was found so far. By linking on the meta-level, identification of similar mapping could be further improved and the aggregated knowledge of a company with regard to unique variables would be more visible and transparent.

An ultimate goal is still to further accelerate the mapping process. A big step was made by recommending already existing mappings and their re-use/re-configuration needs. By letting the fast database-grounded clinical repository actually perform the mappings and

PhUSE 2011

only export final results to datasets for submission, another step of acceleration is planned. Smart integration of user tasks (special transformations) – usually written in SAS – is the challenge in a non-SAS mapping workflow. Innovative ways of using SAS as a language, not a system, and the ability to use different languages (e.g. R or Java) to express user tasks can help to improve the mapping process.

The new principle “analyse/recommend/re-use” established in the research project was touched only on the surface. More and better analysis and recommendation points can be identified and used to support all tasks. This approach requires an awareness of many system dialogs for recommendations. An evolutionary change of the software has already taken place within the entimICE platform.

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.