MOMIS: automate the harmonisation process through Semantic Integration and Machine Learning

*The **MOMIS** platform: semantic integration and machine learning to improve the OMOP CDM harmonisation process.*

Background

Data harmonisation constitutes an important basis for the future of research and have proven to be an effective way to for promoting greater awareness among those involved in the process of designing, collecting, managing and analysing clinical data. The harmonization process carried out together with the **physicians**, the **domain experts** and the **IT teams** of each data partner on the ***heterogeneous disease data*** made it possible to highlight some critical points in the process:

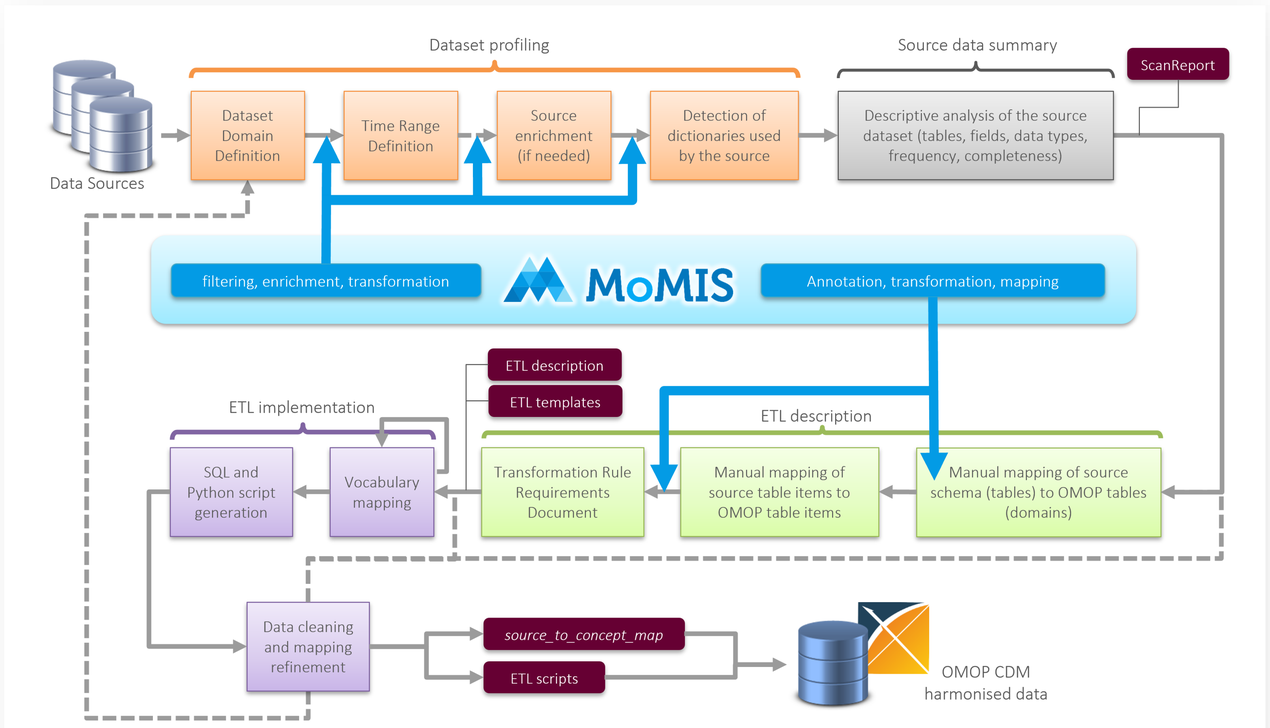
- Schema and source mapping requires a **collaborative effort** and a lot of time from physicians, domain experts and the IT team;
- Data extraction often **loses information about data types and the relationships between them**.;
- Refreshing extracts introduces **unknowns** about changes to the data and its structure.

Methods

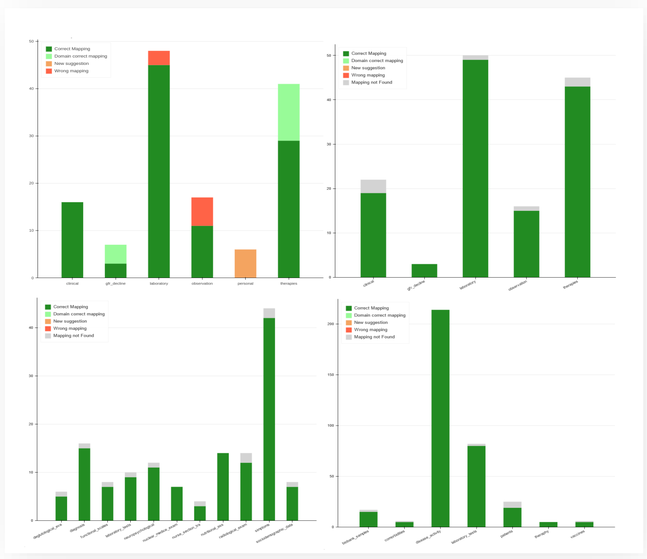
DataRiver's virtual semantic integration system **MOMIS** is designed to **preserve data privacy** and has been used in several clinical data projects to **exploit the semantics within the source data use to find relationships among the information source schemas**. MOMIS has been used to leverage semantic from ***specific vocabularies*** and exploit machine learning techniques to improve the quality of mapping by:

- Reducing the large number of options from less relevant vocabularies;
- exploiting the potential of the **multilanguage capability** to avoid large amounts of data to be translated in English;
- **Valuing and reusing the knowledge** contributed by the domain experts to *continuously improve the service* by providing automatic annotations and mappings to identify the correct domains and contexts;
- **Minimising the effort and the time required** for repetitive annotation and mapping tasks.

The MOMIS approach focused on integrating and reusing the annotations to automate the process.



Results



The charts show the progress in automatic annotation with respect to realised projects: information is already in the data!

- Tests of automatic annotation and mapping conducted on several projects have shown very positive results, **reducing the time** required for manual annotation and mapping operations by **at least 10 times**.
- The promising results from the first tests of mapping projects encourage the extension of the approach to a large number of sources to **increase the accuracy of the service**, facilitate the work of data partners and domain experts, and foster a wider dissemination of the OMOP CDM.
- The **collaborative MOMIS approach**, by using semantic and machine learning techniques, also makes it possible to focus and improve the quality of the harmonization by reducing the excessive number of less relevant concepts that prevent more accurate mapping.
- The potential of the **multilingual capability** proved to be very effective in harmonising data collected in languages other than English, **reducing the number of steps and therefore the human error factor**.

Limitation: Tests performed on a limited amount of datasets by applying semantic similarity criteria using the main standard vocabularies, even in the absence of specific ontologies. Validation of mappings by experts is a key value for improving performance and can be a significant indicator to guide developments and harmonisation of projects.

OHDSI

Enrico Calanchi, Laura Delsante, Mirko Orsini, Andrea Livaldi, Riccardo Martoglia, Luca Magnotta, Domenico Beneventano, Sonia Bergamaschi

DataRiver contract research organization