Abstract Mappathon – Team MDRCupid

„Using a Similarity Toolbox for Matching Metadata“

Noemi Deppenwiese[2], Hannes Ulrich[1,2]

IT for Clinical Research, Universität zu Lübeck, Lübeck, Deutschland[1],
Universität zu Lübeck, Lübeck, Deutschland[2]

The secondary use of EHR data for clinical trial is an emerging field of interest in the section of medical computer science. Due to the rising production of healthcare data an automatic method to manage and classify the information is needed to utilize the rich information set. A promising approach is the use of metadata to describe and categorize the input data. In order to use this information, the data itself has to be analyzed to understand and find correspondences in the dataset. In our approach, we designed and implemented a toolbox, called *MDRCupid*, to find corresponding metadata elements in from clinical trials, across different datasets. It utilizes different lexical string-matching algorithms as well as the statistical bag-of-words approach. The combination and weights of these methods can be chosen at will. In addition, an optimization module was developed to calculate the best configuration for a given set of training data. The results are presented to the user via a graphical user interface, which shows a ranked list of possibly corresponding data elements from which the user may select one. These choices are saved in a HL7 FHIR ConceptMap. These manually confirmed matches may be used as new training data for the optimizer to further improve the matching parameters. We trained the toolbox on various dataset including the provided Mappathon datasets, as well sets from different research fields. Due to organizational constrains, we modified the tool to fit the challenge. The FHIR concept maps are transformed into the needed notation and contain only the first matching prediction. Our mapping pipeline consists of four steps:

1.) Transforming the CDISC ODM files into Samply.MDR import file
2.) Import the transformed dataset into a Samply.MDR instance
3.) Using the UI of MDRCupid to select and auto-match the namespaces
4.) Export the matching result in the desired notation

The first results are promising and we looking forward to apply our method on the challenge datasets.