Leveraging Metadata When Mapping to CDISC Standards with SAS ® Machine Learning

Sandeep Juneja, SAS Institute, Cary NC

Ben Bocchicchio, SAS Institute, Cary NC

Nathan Asselstine, SAS Institute, Cary NC

Matt Becker, SAS Institute, Cary NC

ABSTRACT

Standards define the targets to which source data needs to be mapped too. People can interpret these targets in different ways, this can lead to inconsistencies in the resulting standardized data. The ability to allow different teams working on similar studies (could be located at different off sites or offshore sites) to re-use prior knowledge gained by the team would not only save significant time mapping studies, but increase the quality in the resulting standardized data.

This paper talks about capturing source to destination data mapping as metadata into centralized libraries and applying Machine Learning algorithms to streamline and predict mapping for newer studies that have similar metadata to already mapped studies. This process could lead to consistent destination data mapping and can significantly reduce the mapping timing by re-using system suggested mappings.

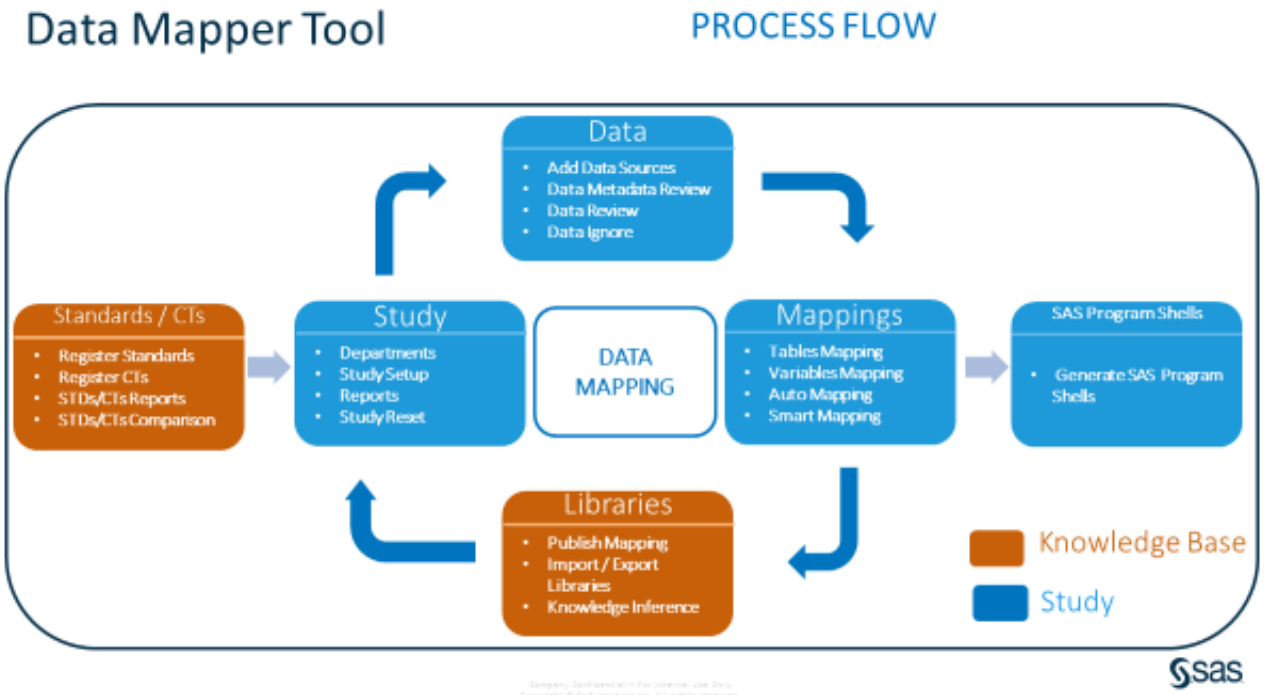
INTRODUCTION

Mapping raw data to standards is one of the most challenging process in the healthcare industry. Reusing or reapplying the information collected during mapping processes from previously mapped studies and building upon that knowledge inference is the most important part of the mapping process. Most companies struggle with building knowledge inferences and reapplying them through the data process efficiently. In addition, sometimes, there are multiple tools/programs required to go through full-cycle of data mapping. Therefore, it is difficult for the standard user to know all the different versions and tools, and use them correctly throughout the data mapping process.

What if we had a tool that provides a user-friendly interface for everything from mapping raw data to generating SDTM standards (including domain templates)? Simple User Interface (UI) and click-away concept design provides access to all the required information on a single screen. Auto-mapping and smart-mapping features in the tool, which are based on knowledge inference derived from machine learning algorithms, reduce time and effort for the user. This leads to improvements in quality, efficiency and consistency.

DATA MAPPING PROCESS

Data Mapping flow is as shown in **Error! Reference source not found.**.Error! Reference source not found.



DATA MAPPER COMPONENTS

1. Standards / Controlled Terminologies – Provides ability to register Standards like – SDTM, ADaM or company specific standards and CTs
2. Studies – Provides ability to register different studies and control permissions.
3. Data – Provides ability to capture Data for studies from different sources
4. Mapping – Provides ability to map source to destination data
5. Generate SAS Programs – Provides ability to generate SAS programs based on mapping metadata
6. Libraries – Provides ability to capture mapping metadata into different libraries.

MACHINE LEARNING ALGORITHMS

Since mapping metadata is captured into Libraries, different type of Machine Learning algorithms can be applied to learn information about existing mapping and these algorithms can be trained to help predict mapping for new source data.

Machine algorithms can be applied to different type of metadata captured at Dataset, Variable and Value level. Below screenshot represent model Similarity vs NGram similarity for Tables Mapping

```
Predictions: ['ae' 'cm' 'lb' 'fa' 'eg' 'ie'] Expected: ['AE', 'CM', 'LB', 'FA', 'EG', 'IE']
             Model_Matched_Term  Model_Similarity NGram_Matched_Term  NGram_Similarity
Search_Term
adverse2                     ae          0.717276                ae          0.583333
chemistry                    lb          0.515414                lb          1.000000
conmed                       cm          0.703553                cm          1.000000
electrocardiac               eg          0.699650                eg          0.521739
follow                       fa          0.683542                fa          1.000000
inclusion                    ie          0.537428                ie          1.000000
```