Data Mapping Using Machine Learning

SAS Institute Japan Ltd.
Toru Tsunoda (Toru.Tsunoda@sas.com)
Data Mapping

Problem Definition

- Source Data mapped to Destination Data based on Standards
  - Time-consuming process
- Standards and Implementation Guides leave room for individual interpretation
  - Inconsistent Mappings – Not what a “Standard” should be
- Mapping is generally done in individual SAS programs
  - Lacks collection of central metadata/mappings
  - Only way to re-use previous mappings is copying programs

Standards / CTs ➔ Source Data ➔ SAS Programs / Mappings ➔ Destination Data
History of Clinical Data Mapping tools

Various SAS tools

Life Science Analytics Framework
(ex SAS Drug Development)

Data Mapper
(ex LSAF Extension)

Clinical Data Integration

Analytics Pro (PC-SAS)

Programing tool using GUI

Programing tool using Machine Learning

CDR (Clinical Data Repository) plus MDR (Meta Data Repository)
Data Mapping using “Data Mapper”

Solution

1. Standards / CTs
2. Source Data
3. Study Mappings
4. Mapping Collections
5. SAS Programs / Mappings
6. Destination Data
Data Mapper

- User interface to mapping process
- **Libraries** - Collection of mapping rules in central database
- Collect data and standards, and allow user to define rules for mapping
- Ability to re-use mapping rules in future studies
- **Auto Mapping** - One-to-One mapping Rule between Source Data and its variable, to Destination Data and its variable
  - e.g. adverse.patid → AE.USUBJID
  - Next study, adverse.patid is automatically mapped
- **Smart Mapping** - Similar Variables mapped in past provide guidelines to map new variables
  - e.g. adverse.ptid → AE.USUBJID
  - Suggested mapping based on previous mapping, adverse.Patid → AE.USUBJID
- Generate the SAS programs based on defined mapping
**Data Mapper**

- **Department**
  - Associate Data Sources
  - Metadata Review
  - Data Review

- **Data**
  - Study Setup
  - Metrics Reports

- **Mappings**
  - Tables Mapping/Transformation
  - Variables Mapping Transformation
  - Auto/Smart Mapping

- **SAS Files**
  - Generate SAS Program(s)/Dataset(s)

- **Standards/CTs**
  - Add Standards
  - Add CTs
  - Metrics Reports and Comparison

- **Study**
  - Study Setup
  - Metrics Reports

- **Mapping Collections**
  - Publish Mappings
  - Publish Transformations

* Version 2
Data Mapper
Administration - Users

- Users – Ability to register user in system
- Authentication Type: Application / Active Directory
Data Mapper
Administration - Roles

• Roles – Pre-defined set of Roles with different privileges
• Ability to add /remove privileges
Data Mapper
Administration – Data Sources

- Data Sources – define data locations

<table>
<thead>
<tr>
<th>Name</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>custom_formats</td>
<td>/tl/warehouse/oracle_content/data_sources/custom_formats</td>
</tr>
<tr>
<td>dataSourceA</td>
<td>/tl/warehouse/oracle_content/dataSources/dataSourceA</td>
</tr>
<tr>
<td>dataSourceB</td>
<td>/tl/warehouse/oracle_content/data_sources/dataSourceB</td>
</tr>
<tr>
<td>dataSourceC</td>
<td>/tl/warehouse/oracle_content/data_sources/dataSourceC</td>
</tr>
<tr>
<td>dupe</td>
<td>/tl/warehouse/oracle_content/data_sources/dupe</td>
</tr>
<tr>
<td>empty</td>
<td>/tl/warehouse/oracle_content/data_sources/empty</td>
</tr>
<tr>
<td>FSDDataSource1</td>
<td>/tl/warehouse/oracle_content/data_sources/FSDDataSource1</td>
</tr>
<tr>
<td>FSDDataSource2</td>
<td>/tl/warehouse/oracle_content/data_sources/FSDDataSource2</td>
</tr>
<tr>
<td>NICS AH_DATA</td>
<td>/tl/warehouse/oracle_content/data_sources/NICS AH_DATA</td>
</tr>
</tbody>
</table>
Data Mapper
Administration - Departments

- Departments – Group similar studies with associated data sources and members
- Provides better permission and data access control

<table>
<thead>
<tr>
<th>Name</th>
<th>Users</th>
<th>Studies</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>CardioVascular</td>
<td>34</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>CNS</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Dermatology</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Endocrinology</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Data Mapper
Administration - Departments

• Details / Studies / Users / Data Sources
## Data Mapper Knowledge Base - Standards

- Register Standard by Importing
- View list of Registered Standards

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Last Modified</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataStandard1</td>
<td>This is a description of data standard #1</td>
<td>August 1, 2018, 7:01 PM</td>
<td>ADaM</td>
<td>In Review</td>
</tr>
<tr>
<td>DataStandard2</td>
<td>This is a shorter description</td>
<td>August 1, 2018, 7:01 PM</td>
<td>SDTM</td>
<td>Active</td>
</tr>
<tr>
<td>DataStandard3</td>
<td>This is the longest description of the three data standards</td>
<td>August 1, 2018, 7:01 PM</td>
<td>Other</td>
<td>Inactive</td>
</tr>
<tr>
<td>DataStandard4</td>
<td></td>
<td>August 1, 2018, 7:01 PM</td>
<td>SDTM</td>
<td>Active</td>
</tr>
<tr>
<td>DataStandard5</td>
<td></td>
<td>August 1, 2018, 7:01 PM</td>
<td>SDTM</td>
<td>Inactive</td>
</tr>
<tr>
<td>QA_SDTM_std</td>
<td>For QA</td>
<td>August 1, 2018, 7:01 PM</td>
<td>SDTM</td>
<td>In Review</td>
</tr>
<tr>
<td>RWE 4.3</td>
<td>Real World Evidence 4.3 Data Dictionary</td>
<td>August 1, 2018, 7:01 PM</td>
<td>Other</td>
<td>Active</td>
</tr>
</tbody>
</table>
Data Mapper
Knowledge Base - Standards

• View Domains Information
• View Domain Variables Details
Data Mapper
Knowledge Base – Controlled Terminology (CT)

• Register Controlled Terminology by Importing
• View list of Registered Controlled Terminologies
## Data Mapper

### Knowledge Base – Controlled Terminology (CT)

- View Codelists Information
- View Codelist Values

![Data Mapper Screen](image)

<table>
<thead>
<tr>
<th>Code</th>
<th>Codelists Code</th>
<th>Codelist Name</th>
<th>CDISC Submission Value</th>
<th>CDISC Synonyms</th>
<th>CDISC Definition</th>
<th>NCI Preferred Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGERANGE-1 QA2</td>
<td>AgeBuckets</td>
<td>&lt;2</td>
<td></td>
<td></td>
<td>AGERANGE-1-CDISC DEFINITION</td>
<td>AGERANGE-1-NCI PREFERRED TERM</td>
</tr>
<tr>
<td>AGERANGE-2 QA2</td>
<td>AgeBuckets</td>
<td>2-5</td>
<td></td>
<td></td>
<td>AGERANGE-2-CDISC DEFINITION</td>
<td>AGERANGE-2-NCI PREFERRED TERM</td>
</tr>
<tr>
<td>AGERANGE-3 QA2</td>
<td>AgeBuckets</td>
<td>6-12</td>
<td></td>
<td></td>
<td>AGERANGE-3-CDISC DEFINITION</td>
<td>AGERANGE-3-NCI PREFERRED TERM</td>
</tr>
<tr>
<td>AGERANGE-4 QA2</td>
<td>AgeBuckets</td>
<td>13-20</td>
<td></td>
<td></td>
<td>AGERANGE-4-CDISC DEFINITION</td>
<td>AGERANGE-4-NCI PREFERRED TERM</td>
</tr>
<tr>
<td>AGERANGE-5 QA2</td>
<td>AgeBuckets</td>
<td>21-64</td>
<td></td>
<td></td>
<td>AGERANGE-5-CDISC DEFINITION</td>
<td>AGERANGE-5-NCI PREFERRED TERM</td>
</tr>
<tr>
<td>AGERANGE-6 QA2</td>
<td>AgeBuckets</td>
<td>&gt;64</td>
<td></td>
<td></td>
<td>AGERANGE-6-CDISC DEFINITION</td>
<td>AGERANGE-6-NCI PREFERRED TERM</td>
</tr>
</tbody>
</table>
Data Mapper

Knowledge Base – Mapping Collection (Under Development)

• Define Mapping Collections by Importing
• View list of defined Mapping Collections
Data Mapper
Knowledge Base – Mapping Collections (Under Development)

• View Mapping Collection Information
Data Mapper

Study Flow

- **Department**
  - **Data**
    - Associate Data Sources
    - Metadata Review
    - Data Review

- **Stds / CTs**
  - Add Standards
  - Add CTs
  - Metrics Reports and Comparison

- **Study**
  - Study Setup
  - Metrics Reports

- **Mappings**
  - Tables Mapping / Transformation
  - Variables Mapping Transformation
  - Auto / Smart Mapping

- **SAS Files**
  - Generate SAS Program(s) / Dataset(s)

- **Mapping Collections**
  - Publish Mappings
  - Publish Transformations

* Version 2

Knowledge Base

Study
Study Mapping Process

1. Metadata –
   • Define study name and description
   • Mapping Library
   • Associate Standards
   • Associate Controlled Terminologies
2. Data Sources –

- Associate Data Source(s) defined at Departments with study
Study Mapping Process

3. Data Transform –
   • Provide ability to read non-SAS datasets and convert them into SAS datasets.
   • Live SAS Session with ability to view log and output datasets

```
LIBNAME DS1 ’/tlawarehouse/content/data_sources/datasourceA’;
LIBNAME DS5 ’/tlawarehouse/content/data_sources/NICSAH_DATA’;
LIBNAME DERIVED ’/tlawarehouse/content/studies/study-1/derived’;

proc export
data=ds5.fube
dbms=csv
    outfile=’/tlawarehouse/content/data_sources/NICSAH_DATA/labs.csv’
replace;
run;
PROC IMPORT OUT= DERIVED.labs
    DATAFILE= ’/tlawarehouse/content/data_sources/NICSAH_DATA/labs.csv’
    DBMS=CSV REPLACE ;
    GETNAMES=YES;
run;
```
4. **Tables** –

- Provide ability to map Source datasets to Destination Domains.
- List all associated data sources
- Mapping count is maintained.
5. Tables Transform –

- Provide ability to realign source datasets and derive datasets for better alignment for variable mapping.
- Live SAS Session with ability to view log and output datasets
6. Variables –

- Provide ability to map Source variables to Destination Domain variables.
- List all associated data source tables variables per domain
- Mapping count is maintained
7. Variables Transform –

- Provide ability to add any derivations/macro calls for Domain variables.
- Live SAS Session with ability to view log and final output Domain
8. **Export** –

- Ability to export mapped domains.
- Ability to publish mapping to Library
## REST APIs

### sasSubmissions

<table>
<thead>
<tr>
<th>Method</th>
<th>Endpoint</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET</td>
<td>/sas/submissions/{submissionId}</td>
<td>Get a SAS submission based on ID</td>
</tr>
<tr>
<td>DELETE</td>
<td>/sas/submissions/{submissionId}</td>
<td>Cancel a SAS submission based on ID</td>
</tr>
<tr>
<td>POST</td>
<td>/sas/submissions/executeCode</td>
<td>Execute provided SAS code and create a SAS submission</td>
</tr>
</tbody>
</table>

### studyMappings

<table>
<thead>
<tr>
<th>Method</th>
<th>Endpoint</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET</td>
<td>/departments/[departmentId]/studies/[studyId]/mappings</td>
<td>Get the study mappings</td>
</tr>
<tr>
<td>PUT</td>
<td>/departments/[departmentId]/studies/[studyId]/mappings</td>
<td>Update the study mapping by ID</td>
</tr>
</tbody>
</table>

### Models

- **Credentials**
  
  ```json
  "Credentials": {
    "username": "string",
    "password": "string(password)"
  }
  ```

- **TokenResponse**
  
  ```json
  "TokenResponse": {
    "version": "integer",
    "The API resource version": "The API resource version"
  }
  ```
4. **Tables AutoMap** –

- Provide ability to Automap Source datasets to Destination Domains based on previous mapping information available in published libraries.
4. **Tables AutoMap** –

- Smart Mapping function is under development for improvement & enhancement.
- The following screen shot was from previous version of Data Mapper.
<table>
<thead>
<tr>
<th>MAPPING TYPE</th>
<th>DESCRIPTION</th>
<th>SOURCE</th>
<th>DESTINATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLES</td>
<td>1-1 MATCH (Top Similarity Value)</td>
<td>DATASET</td>
<td>DOMAIN</td>
</tr>
<tr>
<td>VARIABLE (AUTOMAP)</td>
<td>1-1 MATCH (Top Similarity Value)</td>
<td>DATASET</td>
<td>VARIABLE</td>
</tr>
<tr>
<td>VARIABLE (SMART MAP)</td>
<td>1-3 MATCH (&gt;0.25 Similarity)</td>
<td>DATASET</td>
<td>VARIABLE</td>
</tr>
<tr>
<td>CONTROLLED TERMS</td>
<td>1-1 MATCH (Top Similarity Value)</td>
<td>DATASET</td>
<td>VARIABLE</td>
</tr>
</tbody>
</table>
# Data Mapper

## Design Flow

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Modules</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Read Data</td>
<td>Pandas Dataframe</td>
<td>Read data into Pandas Dataframe – A tabular data structure in Python</td>
</tr>
<tr>
<td>2</td>
<td>Cleanup Data</td>
<td>Pandas Dataframe</td>
<td>Drop NaN Records, convert to lower case, combine columns, filter categories based on MIN count.</td>
</tr>
<tr>
<td>3</td>
<td>n-gram Data</td>
<td>NGRAM</td>
<td>n-gram data using character n-gram with limit of 2. This gives us more data to work with when comparing similarity in variable names.</td>
</tr>
</tbody>
</table>
Pickle n-gram

Step 4: Pickle n-gram

- **Pickled n-gram pair of (Source, Destination) value and pickle it.** The pickle file is the Python data object after serialization, and is an efficient way to store and use any Python object data.

String to Numeric Value conversion

Step 5: String to Numeric Value conversion

- **Generate Dictionary with TF-IDF values and re-scale it using MinMaxScaler.** For machine learning to work, the values must be numeric between 0 and 1.

Balance Data

Step 6: Balance Data

- **Balance data to handle under-sampling and over-sampling.** If the data we have is skewing our results, we want to normalize closer to 0.5.
### Step 7: Evaluate Models

**Modules:** sklearn

**Process:** Plenty of different models to evaluate:
- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- GaussianNB
- MultinomialNB
- OneVsRestClassifier (LinearSVC)
- OneVsRestClassifier (SGDClassifier)

### Step 8: Compute Model Params

**Modules:** Sklearn, GridSearchCV

**Process:** GridSearchCV attempts multiple combinations on the model parameters, tweaking them to find the best parameter values to use.

### Step 9: Train Model

**Modules:** sklearn

**Process:** Fit model.
### Data Mapper

#### Design Flow

<table>
<thead>
<tr>
<th>STEP</th>
<th>DESCRIPTION</th>
<th>MODULES</th>
<th>PROCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Score Model</td>
<td>sklearn.metrics</td>
<td>Classification_report – Precision, Recall, F1 Score. This tells us how much data was relevant to the suggested mappings, and gives a value on how accurate the model thinks the mapping is.</td>
</tr>
<tr>
<td>11</td>
<td>Pickle Model</td>
<td>Pickle</td>
<td>Save the trained model in pickle file.</td>
</tr>
<tr>
<td>12</td>
<td>Predict</td>
<td>sklearn</td>
<td>Predict based on trained model. With user interaction with the predicted mapping (Supervised Machine Learning), the model will become smarter and more accurate over time.</td>
</tr>
</tbody>
</table>