Framework Curation & Loading For tranSMART – FC&L4tranSMART

Version 1.1 - PostgreSQL

User guide

Introduction

TranSMART curation & loading framework aim is to provide a user-friendly interface to load data into tranSMART, from already created mapping files, or by helping users to create these mapping files. For now, clinical data, gene expression data and study description can be loaded with this framework.

License

Compliance

This application has been tested with tranSMART GPL 1.0 PostgreSQL release.
Tested Java Runtime Environments are JRE 1.6 and JRE 1.7

Installation

The framework required a java environment to work.
In a Linux system, edit the file FCL4tranSMART.ini and replace the –vm option by your java path.
To launch the framework, just run the provided executable from the folder of the application.

The database has to be configured with a dot as decimal separator. If it is not the case, tm_cz stored procedures i2b2_load_clinical and i2b2_process_mrna_data have to be modified, to set the dot as decimal separator just while the procedure is running.
For that, at the beginning of the procedure (after the BEGIN), add this line:

```
EXECUTE IMMEDIATE 'alter session set NLS_NUMERIC_CHARACTERS=".,"';
```

Proxy settings

If your network used a proxy, it is required to set proxy settings with the menu item *Proxy preferences*, in the *Preferences* menu.
If your system or web browser is already configured with proxy settings, just set the *Proxy type* as *Native*, and parameters of your system or web browser will be used.
If not, you can choose the *Manual* proxy type in the list. In this case, you have to enter your proxy host and port. If your proxy required an authentification, check the *Authentification required* check box, and enter your user name and password.
In both case, click on the *OK* button to save the parameters.

Choose a workspace

At the opening of the framework, a window appears to let the user choose its workspace. This workspace will be a folder where all files will be stored. New folders will be created in it, so permissions have to be accorded. The chosen folder has to be empty.
Once the workspace has been asked once, it will not be anymore. If the user wants to modify it, he can thanks to the menu item ‘Change workspace’ of the menu.

Preferences

TranSMART curation & loading framework needs a database connection for a few steps. The database connection parameters have to be set in the window ‘Preferences’, opening from the menu. One or several database connection parameters set can be saved with a name, by filling all fields then clicking on the “Save” button. To use a database, its name has to be selected on the list of the preferences window, and then the “OK” button has to be clicked.

When a step needs a database connection, if the connection parameters are correct, the name of the database is displayed, and if it is not the case, a warning is displayed.

Warning: The database connection parameters are not secured; all usernames and passwords are written in a non encrypted file.

Add a new study

In the menu file, the menu item ‘Add a new study’ allows creating a new study. By default, this study is named ‘New_study’. This name has to be modified, because two studies can not have the same name, so no other new studies can be created before that.

Remove a study

A study can be removed from the workspace or from the database, but one does not imply the other.
To remove a study from the workspace, choose the item “Remove a study folder”. A dropdown list of all studies in the workspace is provided. Choose one of them then click on the button “OK”.
To remove a study from the database, choose the item “Remove a in database”. A dropdown list of all studies in the database is provided. Choose one of them then click on the button “OK”.

Data curation & loading

Process

The study with which the user wants to work has to be first selected in the part called “Study selection”. Then the type of data to load has to be selected too, in the part called “Data type selection”.
A list of steps will appear depending of this data type in the part called “Step selection”. If a step is selected, the part called “Work part” will be consequently updated, and a description of

this step will be displayed, with the guide to use this particular step, in the part called “Step description”.
For a given data type, the different steps has to be done in the correct order, so steps which can not be done before another step had been executed are grayed, and can not be selected.
For most of the steps, a button “OK” is present, and has to be clicked before passing to next step.
For all data types, a top node has to be defined before loading. This top node can be defined in the step “Set top node” of the data type “Study description”.

Raw data files

For clinical data loading, inputs files can be of two types:
-A tab delimited text file, with lines representing subjects and columns representing variables. A column has to contain unique subject identifiers.
-A file in soft format, which is one of the format proposed on GEO site for public data sets. In this case, a tab delimited file is created from the soft file, by getting the sample identifiers and the different characteristics under the form “tag: value”.

For gene expression data loading, the input file to provide is a file created by Affymetrix Expression Console, from a 3’ Expression array - RMA analysis.

Mapping files

For clinical data, there are two types of mapping files, the column mapping file handling the column of the raw data files, and the word mapping file handling the term of the raw data files. Both can be provided to the curation tool only if they are correctly built.

For the column mapping file, the criteria checked when a column mapping file is provided are the following:
-All the lines must have six columns
-All the raw data files listed must have been selected in the step “Select raw files”
-A line has to be set as the subject identifiers
-All column numbers have to be set, and have to be numbers
-All data labels have to be set
-All category codes corresponding to properties have to be set
-If a property has a data label source (data label is set to “\”), this data label source has to be provided

For the word mapping file, the criteria checked when a column mapping file is provided are the following:
-All the lines must have four columns
-All the raw data files listed must have been selected in the step “Select raw files”
-All column numbers have to be set, and have to be numbers
-All original values and new values have to be set

For gene expression data, there is only one type of mapping file, called the subject to sample mapping file, which handles the mapping between the samples and the subject identifiers, and the displaying of gene expression data in tranSMART dataset explorer tree. For this file, the criteria checked when a column mapping file is provided are the following:
-All the lines must have nine columns

-Category codes have to be always the same
-The following columns have to be set: Study identifiers, Subjects identifiers, Sample identifiers, Platform, Tissue type, Category codes.

For all files

The different files for a given study and a given data type are listed in the part called “Used files”. A file can be selected to be displayed in the part called “Files viewer”. For the data types “Clinical data” and “Gene expression data”, if a study has been loaded, the list of files contains a file called “*kettle.log*”, containing the error logs for the executed kettle job.

Copy, paste and map

In the *Edit* menu, three menu items allows copying from clipboard, pasting to clipboard and mapping from clipboard. These menu items are available only if the selected step contains columns of data.
The *copy columns* feature allows copying all vales contained in fields, including the non editable columns. The copied columns can then be pasted or mapped in columns from FC&L4tranSMART, or copied in an Excel spreadsheet, a text file…
The *paste from columns* feature allows pasting columns from the clipboard into the editable columns of the step. Columns can be copied from FC&L4tranSMART, an Excel spreadsheet, a text file…
The *map from clipboard* feature allows mapping columns in the clipboard with columns of the current work part. The mapping copies a line if the value of the first column of the work part fits with a value of the first column of the clipboard. This feature will so do nothing if the clipboard contains less than two columns. Columns can be copied from FC&L4tranSMART, an Excel spreadsheet, a text file…

Parts organization

The part organization is saved when the application is closed, and will be the same in the next session. To restore the default part organization, close the application if it is opened, remove the folder named “workspace” in the application folder (warning: this is a different folder than the one chosen as your workspace), and open it again.

Limitations

In the file viewer, the number of lines of a file is limited to 500. If a file is greater, it is indicated at the end of this file in the viewer that the file is too long.

Steps description

Study description steps
-Change identifier:
This step allows modifying the identifier of a study, which is a unique identifier used in database. It is specially used when a new study is created, and is named 'New_study' by default.

For now, it is not possible to modify a name of a study which has data already loaded in database, so a verification is done, and the fields is not editable if there is data for this study in the database. A database connection is needed for this step.

-Set study node:
This step allows defining the study node of the study, which is the place where the study will appear in tranSMART dataset explorer.
A tree can be built by adding free text nodes. Then the study can be added to the tree by indicating its name and click on the 'Add' button.
The study to load is indicated in orange. The other studies are indicated in grey.
This step requires a database connection.

-Load description:
This step allows loading a study description in the database.
The study node has to be defined in the last step to display description into tranSMART.
Study description is loaded as pairs of key/value. A pair can be added by clicking on the button 'Add a tag'. A pair can be removed by clicking on the button 'Remove tag' corresponding to the wanted line.
When this step is chosen, if the description has already been loaded into the database, data is retrieved to be displayed. If they are modified and loaded again, data will be replaced.
A database connection is needed for this step.

Clinical data steps
-Select raw files:
This step allows adding one or several raw files, by indicating the paths, separated with a ';' character, or by choosing them with the 'browse' button.
The format of the file, tab delimited or soft(GEO), has to be indicated in the 'Format' dropdown.
When the button 'Add files' is clicked, the format of the files is checked, and then the files are:
-For a tab delimited file: copied in the workspace
-For a soft file: information are get from the "^SAMPLE = value" line, and from the "!Sample_characteristics_ch1 = property: value" lines (under the form of a property/value couple), then a tab delimited file is created in the workspace with this information
It is also possible to remove one or several files by selecting them in the corresponding list and then click on the 'Remove selected files' button

-Set column mapping file (optional):
This step allows choosing a column mapping file, by indicating the path or by choosing it with the 'browse' button.
When the button 'Add file' is clicked, the format of the file is checked, and then the file is copied in the workspace with the extension '.columns'

-Set word mapping file (optional):
This step allows choosing a word mapping file, by indicating the path or by choosing it with the 'browse' button.
When he button 'Add file' is clicked, the format of the file is checked, and then the file is copied in the workspace with the extension '.word_mapping'

-Set subject identifiers:

This step allows choosing the column corresponding to the subject identifier. If a column mapping file is already existing, the identifiers are retrieved in this file and given as parameters by default

-Set other identifiers (optional):
This optional step allows choosing the column corresponding to site, visit and sub-visit identifiers.
A sub-visit name column can not be set if visit name column is not.
If a column mapping file is already existing, the identifiers are retrieved in this file and given as parameters by default

-Set terms:
This step allows replacing terms of raw files by other terms.
A column of a raw file has to be chosen on the dropdown list. A field is then displayed for each term of this column. A new term can be indicated in this field
It is possible to map terms for a property with terms from another property, by selecting a property in the 'Copy mapping from column' list and click on the 'OK' button
By clicking on the 'OK' button, all the new terms of all columns are saved in a word mapping file

-Set study tree:
This step allows defining the study ontology tree for clinical data, from the study root.
A node can be added by selecting the parent on the tree, filling the field 'New node' and clicking on 'Add node'.
A node or a property can be removed by selecting on the tree and clicking on the button 'Remove a node'.
A property can be added by selecting the parent on the tree, then the property on the dropdown list, and by clicking on the 'Add property' button
An operation on a numerical property can be added by selecting a free text parent on the tree, then properties and operations in the lists, and by clicking on the 'Add operation' button. An operation can only be set for a numerical property (column containing only numbers and dot characters), considering the word mapping file.
When the button 'OK' is clicked, the column mapping file is updated

-Set labels ontology:
This step allows choosing labels for the study parameters, and optionally mapping these labels with a controlled vocabulary code.
These controlled vocabulary codes can be find with bioontology portal:
http://bioportal.bioontology.org/
Only labels which have been put in the data tree are presented.
-Set units:
This step allows setting units for properties.
The value column list has to be set to the column for which a unit will apply, and the value column has to be set to the column containing the units.
The 'add a line' button allows setting several units for several properties. If there are more than two lines, the 'Remove line' button present for each line except the first allows removing a line.
If a line is empty (except for the first), it has to be removed to update the column mapping file.

-Set units

This step allows setting units for properties.
The value column list has to be set to the column for which a unit will apply, and the value column has to be set to the column containing the units.
The 'add a line' button allows to set several units for several properties. If there is more than two lines, the 'Remove line' button present for each line except the first allows removing a line.
If a line is empty (except for the first), it has to be removed to update the column mapping file.

-Load data:
This step allows loading clinical data from raw files and mapping files, using a Kettle job.
The place of the study in the dataset explorer tree has to be indicated in the 'Study description' data type. The tree is displayed, with the study to load in orange, to check that the study tree is well defined.
If security is required for this study, please check the ‘Security required’ line.
A database connection is needed for this step.

-Monitoring:
This step allows accessing error logs for clinical data loading.
If an error has occurred while the kettle job was running, it is indicated, but details are given in an error file saved in the workspace
If an error has occurred while the stored procedure was running, this error is detailed.
A database connection is needed for this step.

-Quality control:
This step allows controlling clinical data loading quality.
Values for each property are got in the raw files and in the database, and displayed. It is also indicated if theses values are the same in the two cases.
A database connection is needed for this step.

Gene expression data steps
-Select raw file:
This step allows choosing a raw file, by indicating the path or by choosing it with the 'browse' button.
When the button 'Add file' is clicked, the format of the file is checked, and then the file is copied in the workspace.

-Select Subject To Sample Mapping File:
This step allows choosing the subject to sample mapping file, by indicating the path or by choosing it with the 'browse' button.
When the button 'Add file' is clicked, the format of the file is checked, and then the file is copied in the workspace with the extension '.subject_mapping'
.Warning: the subject to sample mapping file has to contain a header line, the first line is not considered as data, and so is not read.
The columns of the subject to sample are the following: study identifier, site identifier, subject identifier, sample identifier, platform, tissue type, attribute 1, attribute 2, category code

-Set subject identifiers:
This step allows defining subject identifiers for samples.
A subject identifier has to be defined for each sample.

The button 'Apply' allows setting all selected fields to the value in the field names 'Value'. All fields can be selected or deselected at the same time with buttons.
The button 'OK' allows creating or updating the subject to sample mapping file.

-Set platforms:
This step allows defining platforms for samples.
The button 'Apply' allows setting all selected fields to the value in the field names 'Value'. All fields can be selected or deselected at the same time with buttons.
The button 'OK' allows updating the subject to sample mapping file.

-Set tissue type:
This step allows defining tissue type for samples.
The button 'Apply' allows setting all selected fields to the value in the field names 'Value'. All fields can be selected or deselected at the same time with buttons.
The button 'OK' allows updating the subject to sample mapping file.

-Set site identifiers (optional):
This step allows defining an optional site identifier for samples.
The button 'Apply' allows setting all selected fields to the value in the field names 'Value'. All fields can be selected or deselected at the same time with buttons.
The button 'OK' allows updating the subject to sample mapping file.

-Set attribute 1 (optional):
This step allows defining an optional attribute for samples.
The button 'Apply' allows setting all selected fields to the value in the field names 'Value'. All fields can be selected or deselected at the same time with buttons.
The button 'OK' allows updating the subject to sample mapping file.

-Set attribute 2 (optional):
This step allows defining a second optional attribute for samples.
The button 'Apply' allows setting all selected fields to the value in the field names 'Value'. All fields can be selected or deselected at the same time with buttons.
The button 'OK' allows updating the subject to sample mapping file.

-Set study tree:
This step allows defining the study ontology tree for gene expression data, from the study root.
A node can be added by selecting the parent on the tree, filling the field 'New node' and clicking on 'Add node'.
A node or a label can be removed by selecting on the tree and clicking on the button 'Remove a node'.
A property can be added by selecting the parent on the tree, then the label on the dropdown list, and by clicking on the 'Add property' button
When the button 'OK' is clicked, the subject to sample mapping file is updated

-Check annotation:
This step allows checking that the Affymetrix platform used in the study has annotation already loaded in the database. The identifier of the platform (e.g. 'GPL6244') has to be indicated.
If annotation for this platform is not loaded yet, an annotation file can be chosen to be loaded. A title has to be provided, and optionally a date and a release number.

The file with platform annotation has to contain the following headers, in the right order:

GPL_ID PROBE_ID GENE_SYMBOL GENE_ID ORGANISM

The gene identifier has to be a unique numeric identifier, and can be used to associate a pathway with a gene.
A database connection is needed for this step.
Warning: This step can be very long.

-Load data:
This step allows loading gene expression data from raw files and mapping files, using a Kettle job.
The place of the study in the dataset explorer tree has to be indicated in the 'Study description' data type. The tree is displayed, with the study to load in orange, to check that the study tree is well defined.
If security is required for this study, please check the ‘Security required’ line.
A database connection is needed for this step.

-Monitoring:
This step allows accessing error logs for gene expression data loading.
If an error has occurred while the kettle job was running, it is indicated, but details are given in a error file saved in the workspace
If an error has occurred while the stored procedure was running, this error is detailed.
A database connection is needed for this step.

-Quality control:
This step allows controlling gene expression data loading quality.
The intensity values for a probe of the microarray are displayed for all samples, coming from the raw data file and from the database.
A probe name has to be chosen. This probe names has to be present in the raw gene expression data file.
Values for each sample are get in the raw files and in the database, and displayed. It is also indicated if theses values are the same in the two cases (with an allowed round of 0.001).
A database connection is needed for this step.