Framework Curation & Loading For tranSMART – FC&L4tranSMART

Version 1.1.0

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

This application is provided under license GPL3 by Sanofi.

Compliance

This application has been tested with tranSMART version 1.0 RC2 and version 1.0 GA.

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 in the property file of this application, by adding the following lines at the end of the file called *FCL4tranSMART.ini*, placed at the root of the application folder:

```
-Dhttp.proxyPort=<proxy port>
-Dhttp.proxyHost=<proxy host>
-Dhttp.proxyUser=<user domain>\<user name>
-Dhttp.proxyPassword=<user password>
```

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.

Disk availability

It is possible to see disk free space, for the workspace and for the different tablespaces of the selected database, in the menu item ‘Free space’ of the ‘?’ menu.

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

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”.

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.

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.

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.