European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.5 - 1st progress report on Data Curation

Due date ofdeliverable: October 1st 2013

Actualsubmissiondate:April 9th2014

DisseminationLevel		
PU	Public	PU
PP	Restrictedtootherprogrammeparticipants (includingCommission Services)	
RE	Restrictedto a groupspecifiedbythe consortium (includingCommission Services)	
CO	Confidential, only for membersofthe consortium (includingCommission Services)	

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D4.5
Deliverable title:	1st progress report on Data Curation
Deliverable version:	
Due date of deliverable:	October 1st2013
Actual submission date:	April 9th 2014
Leader:	Reinhard Schneider, Manfred Hendlichand Fabien Richard
Editors:	
Authors:	Serge Eifes, Adriano Barbosa, IoannisPandis, Wei Gu, Paul Williams
Reviewers:	GhitaRahal, David Henderson, Fabien Richard, MansoorSaqi
Participating beneficiaries:	
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leaders:	Fabien Richard, Manfred Hendlichand Reinhard Schneider
Work Package participants:	
Estimated person-months for deliverable:	
Nature:	
Version:	
Draft/Final:	Final
No of pages (including cover):	22
Keywords:	Curation, ETL, public data

TABLE OF CONTENT

1.1 Abstract 4
1.2 Aim of the Public server delivery package: 4
1.3 Data sources: 4
1.3.1 National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) 4
1.3.2 The Cancer Genome Atlas (TCGA)........ 5
1.3.3 European Bioinformatics Institute (EBI) Gene Expression Atlas (ATLAS)........ 5
1.4 Curation pipelines 5
1.4.1 Search application 5
1.4.2 Dataset Explorer 9
1.5 tranSMART data upload methodologies 16
1.5.1 Atlas Search pipeline 16
1.5.2 Dataset Explorer pipelines. 17
1.6 Overview of curated studies loaded to tranSMART 18
1.6.1 Search application 18
1.6.2 Dataset Explorer 19
1.7 Discussion and recommendations 20
1.7.1 Lack of source file standardization 20
1.7.2 Lack of a standard study ontology 21

1st progress report on Data Curation

1.1 Abstract

Data curation efforts have been delivered so far in the context of the Public Server Delivery Package (PSDP). Members of the PSDP have developedcuration pipelines for different types of public gene expression repositories, namely EBI Gene Expression Atlas, NCBI Gene Expression Omnibus and The Cancer Genome Atlas. This pipeline as well as the manual curation efforts of the delivery package members has allowed us to make a wide range of gene expression studies available on the eTRIKS Public Server.

In this document, wepresent an overview of the different data sources that have been usedto populate the Public Server. An overview of the curation pipelines as well as the required manual curation methodology will be given.Furthermore, a description of the curated studies is provided.In conclusion, the shortcomings of data management for currently available public data sets as well as recommendations for the future are provided.

1.2 Aim of the Public server delivery package:

The objective of this delivery package is to provide an eTRIKS/tranSMART server making studies of interest available to the public. The curation and homogenization efforts of the eTRIKS curators provide an added value to such data by making studies comparable among themselves and therefore amenable to analysis in an integrated manner.

Public studies have been loaded to the tranSMART Search and Dataset Explorer.

1.3 Data sources:

Three data sources have been used for curating/loading public studies into tranSMART.

1.3.1 National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO)

The GEO[1] is an international public repository for functional genomic data submitted by the research community. As a repository for high-throughput microarray and next generation sequence data, GEO database currently encompasses more than 32000 public studies [1].

Gene expression data can be rendered meaningless unless accompanied by the contextual biological and processing details under which experiments were performed. As a MIAME-compliant infrastructure, GEO supports fully annotated records encompassing biological and other descriptive metadata [2].

GEO represents a repository unifying thousands of valuable public gene expression studies [3], which makes it particularly interesting as a data resource for public studies. These public studies can then be curated and loaded tothe tranSMART Dataset Explorer.

[1] http://www.ncbi.nlm.nih.gov/geo/

1.3.2 The Cancer Genome Atlas (TCGA)

The TCGA is a pan-cancer initiative focused on applying genome analysis technologies for studying the biomolecular basis of cancer. TCGA is a rich resource encompassing different data types including gene expression, single-nucleotide polymorphism, miRNA, DNA methylation among others, along with clinical data [4]. Currently data areavailable for 30 different cancer types on the TCGA Data Portal[2]. The considerable amount of clinical data together with gene expression data makes TCGA an interesting data source for public studies,which can be curated and loaded tothe tranSMART Dataset Explorer.

1.3.3 European Bioinformatics Institute (EBI) Gene Expression Atlas (ATLAS)

The Gene Expression ATLAS[3]is a semantically enriched database that provides information about gene expression in different biological/experimental conditions. The content of this database is derived by curation, re-annotation and differential expression analysis of selected datasets from the EBI ArrayExpress Archive of Functional Genomics [5] thatencompasses approximately 2800 studies (data release 13.07).

ATLAS puts a special emphasison providing clinical important information including disease state, organism part and sample treatment, among others. Data curation efforts in this database are focused on standardizing terminology for describing experimental factors to an ontological level, namely Experimental Factor Ontology [6].

Furthermore, it provides access to high quality gene expression data. Therefore, we decided to focus on this data source for loading public data into tranSMART Search application.

1.4 Curation pipelines

Curation of public data consists of three steps: 1) extracting source data files from public repositories, 2) retrieving data from the source data files, and generatingstandard format filesthat are loaded into tranSMART, 3) completing and/or standardizing annotations of metadata.Once datahave beencurated, they are loaded into the tranSMART data warehouse by using Extract, Transform and Load (ETL) scripts. To facilitate curation, we have developed pipelines for semi-automatic to fully automated retrieval and transformation of GEO, ATLAS, and TCGA data.

1.4.1 Searchapplication

tranSMART Search application is a tool that allows searching across pre-analyzed microarray gene expression data based on one or more search filters (including among others disease, gene, pathway and study identifier) defined by the user. Hereafter we give a detailed description of the Atlas-Search pipeline that has been developed for loading the corresponding data into the tranSMART data warehouse.

1.4.1.1 Atlas-Search pipeline

Atlas-Search pipeline is a fully automated workflow[4]. It retrieves data from different public data repositories (EBI and NCBI) and generates the standard formatfiles for the tranSMART Search application;this means the Gene Expression, study metadata and platform annotation files.The pipeline performsa data consistency and quality checkbyfiltering out problematic studies based on missing standard format files or fold change outliers ($\log_2$ fold change < 15), respectively. Furthermore, the pipeline will only retain studies focusing on primates, mouse and rat. The standard format filescorresponding toa given studythat passes the quality checks are used as input for Search app ETL scripts, and their data is loaded into tranSMART.

[2]https://tcga-data.nci.nih.gov/tcga/

[3] http://www.ebi.ac.uk/gxa/

[4]https://git.etriks.org/serge.eifes/tm_atlas_search_pipeline

ATLAS-Search pipeline

This pipeline is separated into three parts allowing generating the input files required by ETL procedures (see Figure 1).

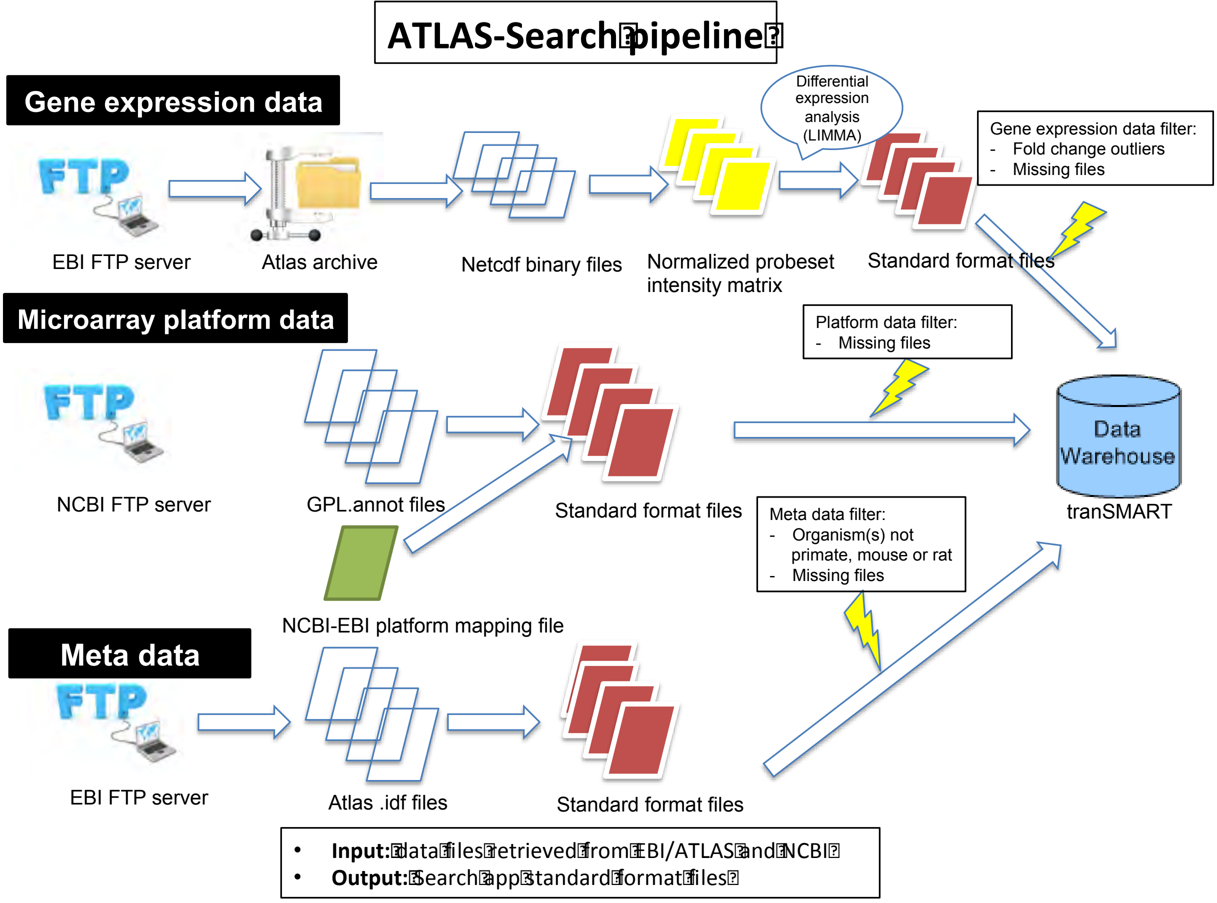


Figure 1: Simplified schema of Atlas-Search pipeline

a. *Gene expression data*

The goal of this part of the pipeline is to generate standard format files for gene expression data. For a given ATLAS data release, a single archive file that containsNetwork Common Data Form (Netcdf) binary filesis downloaded from the EBI ftp server[5]. The ATLASNetcdf file containsgene expression dataas normalized intensity values[6]forall probes or probesets acrossall samples of a study. After decompressing the archive file, normalized intensity valuesare extracted from theNetcdf files, and a matrix (probes/probesets x samples) is constructedfor each study. From this matrix, differential gene expression analysis isrun by performing linear model fittingusing the LIMMA package [7] in R [8]. Differential expression analysis is performed as described in [9].Only genes with a abs(log$_2$ fold change)>1.5 and a false discovery rate < 0.05 are retained as differentially expressed genes (DEG). As a next step the list of DEG isreformatted and stored in a gene expression standard format file (see Table 1).The standard format file provides information on the contrasts that were assessed for differential expression ("ContrastName"), the probe/probeset ("VariableName"), the raw p-value ("RawPValue"), the multiple hypothesis testing adjusted p-value ("AdjustedPValue") and the fold change (FoldChange). Differential expression analysis has been performed as described in Kapushesky et al[9]. It might be noteworthy that for given contrast the

[5] ftp.ebi.ac.uk

[6] http://www.ebi.ac.uk/gxa/help/AtlasFaq#Are_the_expression_values_normalized.3F

experimental factor (e.g. "compound") and the corresponding value (e.g. "transforming growth factor beta"), which was evaluated for differential expression against the overall mean across all factor values, are shown. Columns "Estimate" and "MaxLSMean" are not used as input information by the Kettle script.

Table1: Gene Expression standard format file

ID	ContrastName	VariableName	Raw PValue	AdjustedP Value	Estimate	FoldChange	MaxLS Mean
1	compound => transforming growth factor beta	1007_s_at	0	0.005	NaN	0.2	NaN
2	compound => tumor necrosis factor-alpha	1007_s_at	0	0.005	NaN	0.2	NaN
3	disease_state => osteoarthritis	1007_s_at	0.008	0.032	NaN	0.141	NaN
4	disease_state => rheumatoid arthritis	1007_s_at	0.008	0.032	NaN	0.141	NaN
5	time => 0 hours	1007_s_at	0.096	0.505	NaN	-0.181	NaN
…	…	…	…	…	…	…	…

b. *Microarray platform annotation*

This part of the pipeline retrieves corresponding platform annotation from the NCBI website using appropriate parsers. The NCBI platform identifiers (IDs) are matched to EBI ArrayExpress/ATLAS IDs in an automated manner using a mapping file kindly provided by Robert Petryszak (project coordinator of ATLAS). The output filesgenerated by the corresponding scriptare microarray platform standard format files.

Table 2: Platform annotation standard format file

GPL_ID	PROBE_ID	GENE_SYMBOL	GENE_ID	ORGANISM
A-AFFY-44	1007_s_at	DDR1	780	Homo sapiens
A-AFFY-44	1053_at	RFC2	5982	Homo sapiens
A-AFFY-44	117_at	HSPA6	3310	Homo sapiens
A-AFFY-44	121_at	PAX8	7849	Homo sapiens
…	…	…	…	…

c. *Metadata/Study description*

The metadata file for each study is downloaded from the EBI ftp server[2]. A parser allows reformattingthe file into the metadata standard format. The generatedstandardized metadata files(see Table 3) include among others, study title, description, organism and contact information for data provider. Disease and compound-treatment controlled vocabulary terms are provided byExperimental Factor Ontology (EFO) and Reflect[7][10], respectively.

[7] http://reflect.ws/

Furthermore, to extend the data query capacities for diseases, the pipeline adds parental EFO terms for given disease terms. This should allow the userto retrieve a disease related to a study at different levels of specificity. For instance, a given study related to breast carcinoma can be found back in the same way when querying with various more generic cancer-related terms such as“neoplasm”, “carcinoma”, “cancer” or “breast carcinoma”.

Table 3: Example of a Metadata/Study description file.

Project.Accession	E-GEOD-39
Project.DataSource	Atlas
Project.StudyType	Experiment
Project.Disease	neuroblastoma;neoplasm;nervous system disease
Project.CompoundTreatment	mefloquine
Project.Keywords	.
Project.Organism	Rattusnorvegicus
Project.Title	Transcription profiling of rat NG108 neuroblastoma cell line treated with mefloquine or DMSO
Project.Description	This data series describes expression data for eight paired, control and treated cell cultures obtained on independent occasions. NG108 rat neuronal cell cultures were exposed to either 0.25% DMSO (control) or 4400 ng/ml mefloquine (treated) for two hours. Validation: Modulation of the following transcripts by mefloquine was confirmed by semi-quantitative RT-PCR: U30186, X63594cds_g_at, X63594cds_at, X17163cds_s_at, rc_AI175959 and rc_AA945867 (unpaired, unequal variance, one tailed t-test, p < 0.05).
Project.Design	compound_treatment_design;co-expression_design;transcription profiling by array
Project.Category	compound;dose
Project.ContactAddress	Division of Experimental Therapeutics, Building 503, Walter Reed Army Institute of Research, Robert Grant Ave., Silver Spring, 20910, USA
Project.ContactAffiliation	Walter Reed Army Institute of Research
Project.ContactEmail	Geoffrey.Dow@NA.AMEDD.ARMY.MIL
Project.ContactFirst_Name	Geoffrey
Project.ContactMid_Initials	S.
Project.ContactLast_Name	Dow
Project.Protocol Software	.
Project.Publication Author List	Geoffrey S Dow
Project.PubMed ID	12675948

Project.WebLink	http://www.ebi.ac.uk/gxa/experiment/E-GEOD-39
Project.PlatformType	Affymetrix 3'IVT
Project.Platform	A-AFFY-21
Project.PlatformProvider	Affymetrix
Project.PlatformDescription	RT_U34
Project.ProcessingMethod	Affymetrix_RMA
Project.PlatformOrganism	Rattusnorvegicus

1.4.2 Dataset Explorer

Dataset Explorer enables to compare data between test subjects in two different study groups. The two cohorts/groups are defined based on criteria and points of comparison specified by the user. Dataset Explorer is useful for testing a hypothesis that involves the criteria and points of comparison defined by the user. Hereafter we give a detailed description of the pipelines that have been developed for loading the corresponding data into the tranSMART data warehouse.

1.4.2.1 GEO-Dataset Explorer pipeline

GEO-Dataset Explorer pipeline is a semi-automatic script pipeline for gene expression data[8]. It retrieves data from the NCBI GEO website[9], andgeneratesstandard formatfiles that are used as input for the ETL scripts of the Dataset explorer application[10].

A schematic overview of the complete pipeline can be found in Figure 2. Hereafter, we provide a short description of the four modules that compose the pipeline:

Deleted:

- getData(): refers to a module that accesses NCBI GEO database, and downloads user-selected GEO Dataset archive (GSEXXX_family.xml.tgz files). Such archive consists of gene expression data (GSMXXX-tbl-X.txt), sample description (GSEXXXX_family.xml), and platform information (GPLXX.txt) files. Once this archive file is decompressed, the GEO dataset is handled in the following steps.
- getSamples(): refers to a module that handles the GEO Dataset,extracts sample information and gene expression values contained in such a sample, and copies them in the prepareData() input files. For a givensample, its data (such as expression values) and its descriptionare retrieved from the corresponding GSMXXX-tbl-X.txt file and GSEXXX_family.xml file, respectively.
- getPlatform(): refers to a module retrieves information on the probes/probesetsof a given platform from the corresponding GPLXX.txt file, and copies them in a prepareData() input file.
- prepareData(): refers to the agent script that collects all inputs from the above modules, and populates the tranSMART Dataset Explorer standard formatfiles named: Platform Table, Sample to Subject Mapping, Intensity Table, Raw Clinical Data and Column Mapping File.

[8] https://git.etriks.org/serge.eifes/geo_deapp_pipeline/tree/master

[9] http://www.ncbi.nlm.nih.gov/gds/

[10] https://git.etriks.org/transmart-dse-etl/tree/master/DSE/Kettle/Kettle-ETL

A description and examplesof Dataset Explorer standard format files are available on the website of the tranSMART foundation[11].

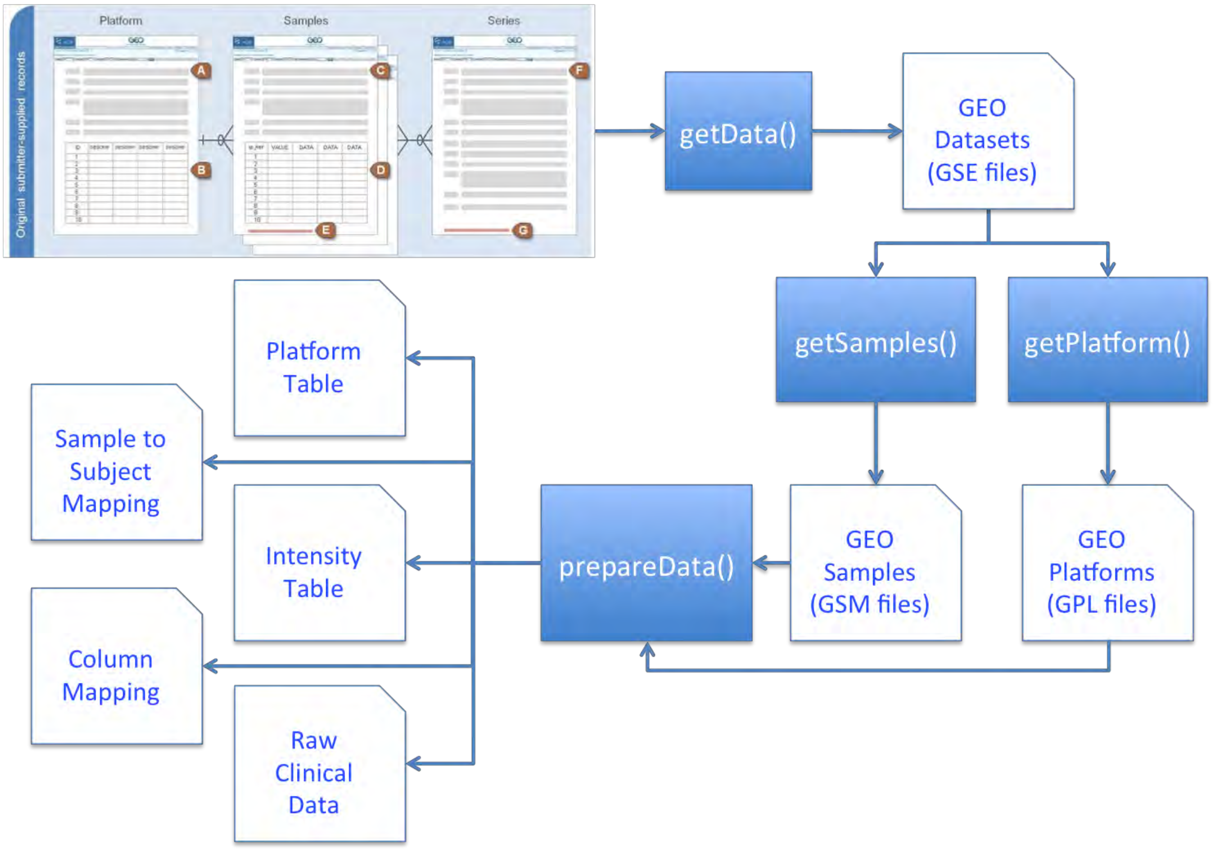


Figure 2: Simplified schema of GEO-Dataset Explorer pipeline. Blue boxes: script modules for data acquisition and processing. White boxes: I/O files. See above text for description.

1.4.2.2 TCGA-Dataset Explorer Pipeline

The TCGA datasets are made publicly available on their website and can be downloaded relatively easily. Transforming data of TCGA source files into the standard formats required by the tranSMART loading procedurewas the main task, and is described in more detail below. Standard formatfiles were loaded into tranSMART using the standard kettle (http://www.pentaho.com/)ETL scripts andthe Framework of Curation and Loading For tranSMART (FCL4tranSMART)[12] application (hereafter referred as the ICE tool) developed by Sanofi. A brief summary of this procedure is provided below.

The initial approach followedto data integration was to manually produce a sample of standard formatfilesusing Excel in order to fully understandthe data structures requiredby tranSMART[13]. Whilst Excel provides substantial flexibility in terms of file data manipulation it becomes slow and unstable when using files of the size of the gene expression files (50-100Mb). It is, however, still a useful tool for viewing the data files and making ad-hoc

[11] https://wiki.transmartfoundation.org/display/TSMTGPL/Simple+Clinical+Data+Example& https://wiki.transmartfoundation.org/display/TSMTGPL/Simple+Gene+Expression+Data+Example

[12] https://github.com/transmart/tranSMART-ETL/tree/master/FCL4tranSMART

[13] https://github.com/transmart/tranSMART-ETL/blob/master/Postgres/GPL%201.0/Standard_Datasets/GSE4382/GSE4382.zip

amendments, and can be used fairly efficiently when the user is proficient with the file formats and some simple data parsing functions.

The second approach used was to construct a Python script [14]in order to completely automate the data transformation process. When the routines have been completely programmed it takes a matter of seconds to automatically download the source files from the TCGA site, and transform the data from the source files into thestandard formatfiles. This is only possible, however, when the exact transformation activity is known.

Unfortunately, as noted, this is not the case with these datasets. There will always, therefore, be a level of manual intervention required to ensure that the correct transformation procedures are carried out, and the study ontology (i.e. the tree/hierarchy of study concepts displayed in the tranSMART user interface)isproduced appropriately. The other major problem in developing stand-alone scripts is that they are not very user friendly either for other users or even the person who has developed the script when she/hehas not used it for a short period of time.

The third approach was to produce a simple prototype application, written in Java[15], to combinein a single placevarious routines that are required for the file transformation. Whilst only a prototype, the benefit of the application is that itprovides the flexibility, in a single place, to carry out a variety of parsing or transformation routines depending on what format the source files arepresented in. As a prototype, the graphical interface has been kept deliberately simple, but provides functions to select the gene expression or clinical source files to be transformed into standard formatfiles. There is also a data entry prompt for the additional information required in the load files. It also provides a process overview narrative and file viewer windows.

The initial objective of this phase of the project was to integrate two to three of the TCGA datasets intotranSMART and develop parsing scripts to help with this process. The end result isthat 5 datasets have been transformed and loaded intotranSMART and both a prototype Python automated script and prototype transformation application have been developed to simplify and, where possible, automate this process.

1.4.2.3 Extract and Transformation Methodology

a. Data Extraction

As noted above, datasets have been made publicly available on the TCGA website, and whilst they are published in a variety of source and compressed formats, the process of downloading the data is not problematic. The lack of standard source data file formats means that each dataset needs to be manually reviewed before any transformation activity is undertaken.

b. Data Transformation

The tranSMART documentation describes in some detail the structure and data contents ofstandard format files[16]that are required by the standard kettle ETLscripts forloading data

[14]https://git.etriks.org/serge.eifes/tcga_deapp_pipeline

[15]https://git.etriks.org/serge.eifes/tcga_deapp_pipeline

[16] https://github.com/transmart/tranSMART-ETL/blob/master/Postgres/GPL%201.0/Standard_Datasets/GSE4382/GSE4382.zip

into tranSMART. Therefore, transformation activity was focused on transforming the TCGA source data files into these standard format files.

The two key source data files ofthe TCGA datasets are the clinical data and the gene expression data files. The clinical data source files contain details on each of the patients studied such as their age, tumour stage classification, and other relevant clinical data. As noted above, there is a substantial degree of heterogeneity in the format and data content of these files. The gene expression source files contain mainly normalized expression for each probe/probeset(rows) and each sample (columns).

The main task withclinical data is to organizeclinical concepts (i.e. the variable names/column headers of clinical data source files) in a hierarchical structurethat formsa study ontology displayed and used for analysis within tranSMART. An example of a study ontology tree as it appears in the Dataset Explorer section of tranSMART is shown in Figure 3.

Deleted:

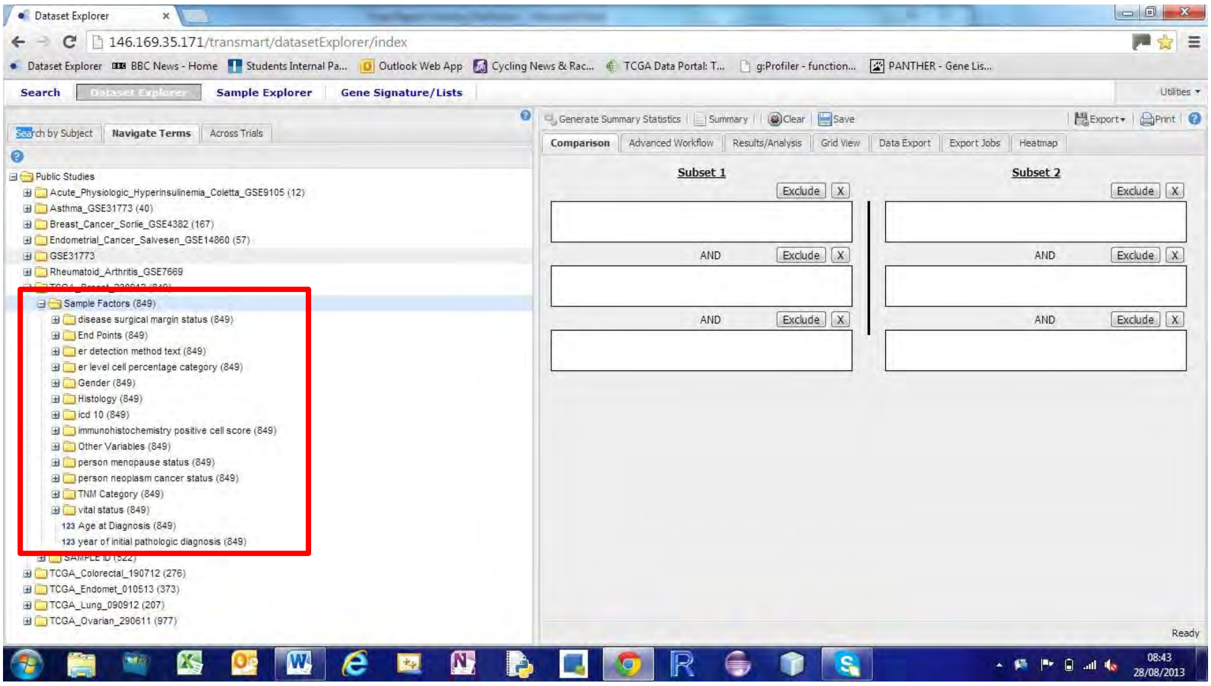


Figure 3: tranSMART Dataset Explorer - Breast Cancer ontology

Three clinicalstandard formatfiles are required for creating a studyontology:

- Raw clinical Data File– a copy of the source clinical data file;
- Column Mapping File – maps the columns in the raw clinical data file to the study ontology to describe the place each column in the raw clinical data file occupies in the hierarchy. It effectively allows the grouping of related data items; and
- Word Mapping File – this is an optional file that allows replacing free text terms by controlled vocabulary terms– for example replacing „NA‟ with „Not Applicable‟.

The main challenge ofsimplifying or automating the creation of clinical standard format files is creating the study ontology. The lack of any real attempt to present the data in a standard format, and the variety of clinical characteristics between cancers make standardization of study ontologiesparticularly challenging.The approach of creating anontology for each of these studies was to mirror some of the existing study ontologies already loaded into tranSMART or referenced in the loading documentation. The key data groupings used in a study ontology were based around time points, histology, TNM categorization, age,or any specific characteristics that was reported as relevant for a cancer type, such as the hormone receptor status in breast cancer.

Transformation of the gene expression file, while still being presented in a variety of formats, was easier to automate. Threestandard format files are required for loading the gene expression data:

- Gene Expression File – this table file containsgene expression valuesof probes or probesets (rows) across samples or patients (columns);
- Subject-Sample Mapping File – it maps the sample IDs used in the gene expression file to the subject IDs used in the clinical data standard format file; and
- Annotation File – it maps the probe IDs provided in the gene expression file to the standard gene symbols and gene IDs.

These standard format files also set the study ID and the platform IDthatareboth used in the tranSMART Dataset Explorer. The microarray platform ID is referenced in the original TCGA paper. The gene ID that is required in the custom annotation file is sourced from the g:Convert tool[17] by reference to the global standard Entrezgene database[18].

c. *Transformation Methodology*

Transforming source filesintostandard format files is not a difficulttask.The most challenging aspect is to build a study ontology thatcategorizes the clinical datawell. New study ontologies were produced by using existing study ontologies as models, and by grouping similar clinical characteristics, such astumour grading and other histological characteristics. Building study ontologies, however, will benefit from data owners‟ feedback.

As noted above, a variety of methodologies were used to transform data. Initially Excel was used to manually curate the data in order to learn about the data provided, data structures and the process of transformation. As described above (see section 1.4.3), spreadsheet applications are limited by their low speed and stability when managing large files.Therefore, as a prototype for demonstration purposes, a Python script[19] was developed to automate the extraction and transformation of the colorectal cancer datasets. The script uses the standard uniform resource locator URL library functionality to download the source clinical data and gene expression files, and the standard CSV parsing functionality to proceed through the file transformations.

When running the Python script,transformingcomma separated values (CSV) files and managinglarge files are relatively quick and easy. Developers, however, have to specify the precise transformation routine to be undertaken. As noted, the fact thatdata from the source files are not presented in a standard format and metadata are not machine readableprevents development ofscripting routinesthat can completely automate this process. The other main disadvantage of such scripting routines is that they are not user friendlyor even for developersto come back to after a short period of time away.For these reasons a specific

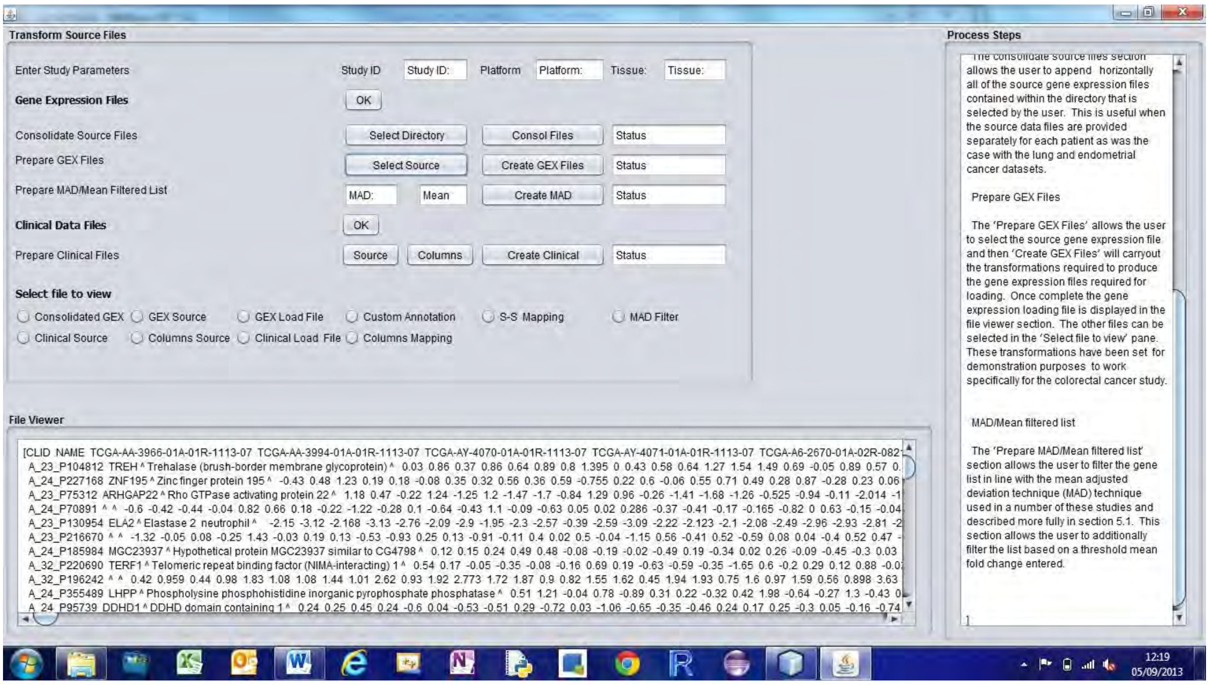
[17] http://biit.cs.ut.ee/gprofiler/gconvert.cgi
[18] http://www.ncbi.nlm.nih.gov/gene
[19]https://git.etriks.org/serge.eifes/tcga_deapp_pipeline

prototype transformation application was developed that allows the user to manage a variety of data transformation routines depending on what transformations may be required. The next section describes the application in details.

1.4.2.4 Prototype Data Transformation Application

A screen shot of the graphical user interface (GUI) of the application is shown on Figure 4. The GUI includes a data entry and push button transformation panel that allows the user to enter the specific data variables that are required by the user, to select the relevant source files, and to initiate the transformation. The bottom section contains a file viewer section so that source or output files can be viewed, and, at the right, a help section provides an overview of the available functions and procedures to be followed.

Deleted:



Transform Source Files

Enter Study Parameters — Study ID: Study ID: — Platform: Platform: — Tissue: Tissue:

Gene Expression Files — OK

Consolidate Source Files — Select Directory — Consol Files — Status

Prepare GEX Files — Select Source — Create GEX Files — Status

Prepare MAD/Mean Filtered List — MAD: — Mean — Create MAD — Status

Clinical Data Files — OK

Prepare Clinical Files — Source — Columns — Create Clinical — Status

Select file to view

○ Consolidated GEX ○ GEX Source ○ GEX Load File ○ Custom Annotation ○ S-S Mapping ○ MAD Filter

○ Clinical Source ○ Columns Source ○ Clinical Load File ○ Columns Mapping

File Viewer

```
[CLID  NAME  TCGA-AA-3966-01A-01R-1113-07  TCGA-AA-3994-01A-01R-1113-07  TCGA-AY-4070-01A-01R-1113-07  TCGA-AY-4071-01A-01R-1113-07  TCGA-A6-2670-01A-02R-082
 A_23_P104812  TREH ^ Trehalase (brush-border membrane glycoprotein) ^  0.03  0.86  0.37  0.86  0.64  0.89  0.8  1.395  0  0.43  0.58  0.64  1.27  1.54  1.49  0.69  -0.05  0.89  0.57  0.
 A_24_P227168  ZNF195 ^ Zinc finger protein 195 ^  -0.43  0.48  1.23  0.19  0.18  -0.08  0.35  0.32  0.56  0.36  0.59  -0.755  0.22  0.6  -0.06  0.55  0.71  0.49  0.28  0.87  -0.28  0.23  0.06
 A_23_P75312  ARHGAP22 ^ Rho GTPase activating protein 22 ^  1.18  0.47  -0.22  1.24  -1.25  1.2  -1.47  -1.7  -0.84  1.29  0.96  -0.26  -1.41  -1.68  -1.26  -0.525  -0.94  -0.11  -2.014  -1
 A_24_P70891  ^  ^  -0.6  -0.42  -0.44  -0.04  0.82  0.66  0.18  -0.22  -1.22  -0.28  0.1  -0.64  -0.43  1.1  -0.09  -0.63  0.05  0.02  0.286  -0.37  -0.41  -0.17  -0.165  -0.82  0  0.63  -0.15  -0.04
 A_23_P130954  ELA2 ^ Elastase 2  neutrophil ^   -2.15  -3.12  -2.168  -3.13  -2.76  -2.09  -2.9  -1.95  -2.3  -2.57  -0.39  -2.59  -3.09  -2.22  -2.123  -2.1  -2.08  -2.49  -2.96  -2.93  -2.81  -2
 A_23_P216670  ^  ^  -1.32  -0.05  0.08  -0.25  1.43  -0.03  0.19  0.13  -0.53  -0.93  0.25  0.13  -0.91  -0.11  0.4  0.02  0.5  -0.04  -1.15  0.56  -0.41  0.52  -0.59  0.08  0.04  -0.4  0.52  0.47  -
 A_24_P185984  MGC23937 ^ Hypothetical protein MGC23937 similar to CG4798 ^  0.12  0.15  0.24  0.49  0.48  -0.08  -0.19  -0.02  -0.49  0.19  -0.34  0.02  0.26  -0.09  -0.45  -0.3  0.03
 A_32_P220690  TERF1 ^ Telomeric repeat binding factor (NIMA-interacting) 1 ^  0.54  0.17  -0.05  -0.35  -0.08  -0.16  0.69  0.19  -0.63  -0.59  -0.35  -1.65  0.6  -0.2  0.29  0.12  0.88  -0.0
 A_32_P196242  ^  ^  0.42  0.959  0.44  0.98  1.83  1.08  1.08  1.44  1.01  2.62  0.93  1.92  2.773  1.72  1.87  0.9  0.82  1.55  1.62  0.45  1.94  1.93  0.75  1.6  0.97  1.59  0.56  0.898  3.63
 A_24_P355489  LHPP ^ Phospholysine phosphohistidine inorganic pyrophosphate phosphatase ^  0.51  1.21  -0.04  0.78  -0.89  0.31  0.22  -0.32  0.42  1.98  -0.64  -0.27  1.3  -0.43  0
 A_24_P95739  DDHD1 ^ DDHD domain containing 1 ^  0.24  0.25  0.45  0.24  -0.6  0.04  -0.53  -0.51  0.29  -0.72  0.03  -1.06  -0.65  -0.35  -0.46  0.24  0.17  0.25  -0.3  0.05  -0.16  -0.74
```

Process Steps

The consolidate source files section allows the user to append horizontally all of the source gene expression files contained within the directory that is selected by the user. This is useful when the source data files are provided separately for each patient as was the case with the lung and endometrial cancer datasets.

Prepare GEX Files

The 'Prepare GEX Files' allows the user to select the source gene expression file and then 'Create GEX Files' will carryout the transformations required to produce the gene expression files required for loading. Once complete the gene expression loading file is displayed in the file viewer section. The other files can be selected in the 'Select file to view' pane. These transformations have been set for demonstration purposes to work specifically for the colorectal cancer study.

MAD/Mean filtered list

The 'Prepare MAD/Mean filtered list' section allows the user to filter the gene list in line with the mean adjusted deviation technique (MAD) technique used in a number of these studies and described more fully in section 5.1. This section allows the user to additionally filter the list based on a threshold mean fold change entered.

Figure 4: Prototype data parsing application visual interface

The application has been developed in Java using NetBeans IDE 7.3.1 with the opensource opencsv library. Its key features are detailed below.

- Help Section – Process Steps

Provides help and instructions on entering the correct parameters in the "Transform Source Files Pane"(Fig. 4).

- Enter Study Parameters

This allows the user to enter the study ID, platform ID and tissue type parameters that are required in the standard format files.

- Gene Expression Files

This section provides a number of file transformation routines that are used when creating the gene expression standard formatfiles.

  - The "consolidate source files" section allows the user togathergene expression data fromindividual gene expression source files of study patientsinto a singlegene expressionsource file of the whole study, as it was the case with the lung and endometrial cancer TCGA datasets.

        - The button “Select directory” (you can set your working directory containing the individual patient GEX files)
        - The button “consolidate files” (used to consolidate individual patient GEX files, into one file, represented by a 2D data matrix; rows = gene probes, columns= individual files)
    - The „Prepare GEX Files“ section allows the user to select the gene expression source file (“select source” button),and, when clicking on the button „Create GEX Files“, to create the three gene expression loading files -gene expression, platform annotation, and subject-sample mapping files- from the source gene expression file selected and input parameters, and tocarry out the transformation of the gene expression source file into the gene expression standard format file. Once completed the gene expression standard format file is displayed in the file viewer section, and the „status“ field updated to „Done“. The other files can be selected in the „Select file to view“ pane. These transformations have been set, for demonstration purposes, to work specifically for the colorectal cancer study.
    - The „Prepare MAD/Mean filtered list“ section allows the user to filter the gene list in line with the meanabsolute deviation[20](MAD). This section allows the user to additionally filter the list based on a threshold mean foldchange entered.

- ClinicalData Files

The „Prepare Clinical Files“ allows the user to prepare the clinical raw data and columns mapping standard format files for loading.After selecting the source clinical data file (the button “Source”) and the columns mapping file (the button “Columns”), the clinical standard format files are produced by clicking on the button “Create Clinical”. Again, as a prototype, the transformations have been set to work with the colorectal cancer dataset only.

- File Viewer

The „Select File to View“ and the „File Viewer“ sections allow the user to select any of the source or standard format files to be displayed in the „File Viewer“ section for review.

Technical Specifications

The application has been developed in Java using the NetBeans IDE to automate the generation of code for the GUI elements. The open source CSV parsing library has been used to manage the CSV file reading and writing objects. When parsing CSV source files, and given the lack of standard format of these source files, it was recommended as safer to use an already developed library that helps to handle the reading and writing rather than developing the basic code separately.The application uses the Java standard ActionListener class to implement action drivers on the various buttons that are used. Data transformation is managed through a conversion of the read data format held in a String Array to a temporary ArrayList structure that can be used to manipulate the data. Before being written the data is then converted into a String Array and written to the output file using the opencsv writer object. Much of the data transformation is then managed using fairly standard loop constructs to construct the standard format files. The key functions that have been implemented are:

viewFile – displays the text file input to either the file viewer section or the „process steps“ section of the GUI, whichever is specified.

createClinical – creates the clinical raw data and column mapping files from the input clinical data and column mapping templates that have been selected.

[20] http://en.wikipedia.org/wiki/Absolute_deviation#Mean_absolute_deviation_.28MAD.29

createGEX – creates the three gene expression loading files – gene expression, custom annotation and subject-sample mapping file from the source gene expression file selected and input parameters.
createMAD – creates a filtered list of genes based on the specified MAD and mean thresholds.
consolidateFiles – consolidates individual gene expression files provided separately for each patient into a single gene expression file based on the input or selected directory.

The potential benefit of an application such as the one we created is that it is very adaptable so that many additional transformation routines can be included as experience is gained. The file viewer and help sections allow review and consultation with specific guidance to be carried out in a single application.

1.5 tranSMART data upload methodologies

In this section we provide an overview on the approaches we used for uploading the standard format files created by the Atlas Search, Geo and TCGA pipelines. Furthermore we provide a short description of the technical problems we encountered.

1.5.1 Atlas Search pipeline

To load a large number of studies, we developed a batch loading script that allowsexecuting the Search application ETL procedure for a panel of studies in an automated way. A major drawback is that the tranSMART data warehouse does not allow performing parallel ETL job execution. Thus the script performs sequential study uploads. As such, this increases considerably the overall runtime of the batch loading script compared to parallel upload of studies to tranSMART.

A first attempt to upload studies remotely from the University of Luxembourg (Esch/Belval - Luxembourg) network to the eTRIKSPublic Server located at CC-IN2P3 in Lyon (France) was considered as a failure due to the very slow data transfer connection between Luxembourg and the Public Server in Lyon as well as the network timeouts encountered, causing frequent failures of the study uploads

Therefore, we decided to perform the study upload procedures in a local network environment setting. Thus,we transferred the whole data (standard format files for approximately 1400 studies) directly to a dedicated storage space located on the eTRIKS Public Server[21]. This allowed us to improve the stability/robustness of the batch loading procedure, even though the runtime for a given study uploadwas/is still not optimal. The execution time of the ETL procedurefor a study is ranging between 10 minutes and a few hours mainly depending on the size of the gene expression standard format file. This limitation is depending on the database design of tranSMART version 1.0. Future developments of tranSMART focusing on materialized views, which are only available since Postgres version 9.3,andNoSQL-based databasesolutions for storage of high dimensional data such as gene expression data might solve this problem.

[21] http://public.transmart.etriks.org/transmart/search

1.5.2 Dataset Explorer pipelines.

1.5.2.1 GEO pipeline

The GEO studies were obtained either using the GEO pipeline or provided by Thomson Reuters.

After curation of the data (for more details see section 1.5.2.3, the studies were loaded to tranSMART using in-house developed batch-loading scripts. These scripts can be run in the background on a linux terminal. At UL, a loading server that hosts all curated data and loading ETLs and loading scripts was setup. Depending on the connection speed between the loading server and the tranSMART database server, the loading of the currently curated GEO data takes between 2 to 10 hours.

For loading of GEO data to the public server at CC-IN2P3, we transferred the software environment as well as the data from our local loading sever at UL to the tranSMART server at CC-IN2P3 to enable a faster connection between the loading environment storing the standard format files and the database server. Using this optimized loading pipeline, we were able to load the curated GEO studies to the public server within a few hours.

1.5.2.2 TCGA pipeline

It was agreed as part of the scope definition of this project that the standard kettle loading scripts would be used in conjunction with the ICEtoolto load the data. The user guide[22] for this application contains a description of the process and steps involved in loading the standard data files.

One of the major drawbacks of data warehouses such as tranSMART is the difficulty and time involved in uploading large datasets over relatively slow upload network connections, especially when working remotely. As an example the colorectal cancer(CRC) dataset took approximately 2.5 hours to load, and the much larger breast cancer dataset sets closer to 4 hours in total. The major part of the upload process is the large gene expression files which can be over 50Mb. Whilst the ICE tool is a useful application in terms of coordinating the upload process – files, connections, load scripts – and it does carry out some useful pre-loading checks to test data integrity, more thorough pre-loading checks would simplify the process. For example, the CRC dataset took approximately 8-9 attempts to load successfully, a total of over 20 hours and often failed at the end of a 2.5 hour load cycle. There was a variety of reasons for these failures such as insufficient storage space being allocated on the virtual machine, the VPN connection failing or inconsistent study names being given across the various files and ICE tool screens, which had not been specified as a criticality previously. Once the FCL4tranSMART application signalled that the loading process had completed successfully, the dataset would be viewed in the tranSMART Dataset Explorer and tested with the summary statistics and limited data export, through the heatmap, functionality.

1.5.2.3 Manual curation procedures for studies loaded to Dataset Explorer

In the context of the tranSMART Public server, data curation can be subdivided into two main steps, namely design of the i2b2 tree and data cleansing. Here below we give an overview on how we have performed manual data curation:

[22] https://github.com/transmart/tranSMART-ETL/tree/master/FCL4tranSMART

a. *Design i2b2 tree structure and create the corresponding column-mapping files.* Frequently it was required to rephrase the leave terms to have appropriate data labels. The structures of the i2b2 tree and the data labels have been harmonized to allow better study comparability. An example of such an i2b2 tree is shownin Table 5.

Table 5: Example of i2b2 data tree structure

Column name (in clinical data file)	Data Label	Category CD
BCRPATIENTBARCODE	SUBJID	Sample_Factors
AgeAtDiagnosis (yrs)	Age	Subjects+Demographics
OverallSurvival(mos)	OverallSurvival(Months)	Subjects+End_Points
breast_tumor_clinical_m_stage	M - Metastasis Stages	Subjects+Medical_History+Cancer_Stage+TNM_Category
breast_tumor_pathologic_grouping_stage	AJCC_Stages	Subjects+Medical_History+Cancer_Stage
…		

b. *Cleanse each column of the clinical data file from all types of errors/inconsistencies/ambiguities.*Here below we show a non-extensive list of such problems that have been encountered:

- Change numeric value from data type string to number (e.g. change “one” to “1”)
- Unify descriptions of numeric values (e.g. change “larger than 3” to “> 3”)
- Replacesemantically identical words by a unique word (e.g. change “NA”, “n/a” and “not available” to “Not Available”)
- Apply the CamelCase rules[23](e.g. change “luminalB" to "Luminal B", “TUMOR FREE" to "Tumor Free")
- Expand non-defined abbreviations to full words (e.g. change "SLN AND NON-SLN BX" to "Sentinel Lymph Node and Non-Sentinel Lymph Node Biopsy")

1.6 Overview of curated studies loaded to tranSMART

In this section we give a survey on the studies that have been curated so far as well as those which have required curation efforts.

1.6.1 Search application

The original ideawas to provide a set of pre-analysed microarray studies that can be browsed in Search app. ATLAS was selected as data source, as it gives access to thousands of studies in a pre-analyzed format.The initial idea was to make all studies for a panel of organisms encompassingprimates, mouse and rat available in the eTRIKS Public Server[24]. It might be important to point out that a considerable amount of these studies have a focus on disease and/or pharmacologicalresearch related aspects.

[23] http://en.wikipedia.org/wiki/CamelCase

[24] http://public.transmart.etriks.org/transmart/search

A first upload of ATLAS data release (Data version 13.05)was intendedto provide access to approximately 1400 studies. Unfortunatelyserver performance issues hindered the use of this data in the eTRIKSPublic Server Search application. Investigation of the performance issues at the database level suggested that the database design in tranSMART release 1.0 is vulnerable to data scalability problems. Future tranSMART releases might help to improve this bottleneck in a similar manner as described above (see section 1.5.1). A further approach to decrease the impact of these performance issues might be to pre-filter the differentially expressed genes based on biological (fold change) and statistical significance (false discovery rate).

We will try to make this data publicly available within the next few months when existing performance problems on the level of tranSMART have been solved.

1.6.2 Dataset Explorer

The initial idea was to provide access to public studies in the Dataset Explorer of the eTRIKSPublic Server[25]. The studies that have been selected are in part related to IMI or other public-private partnership projects covered by eTRIKS. Currently the Public Server Dataset Explorer provides access to 22 curated studies, which can be subdividedaccording to their data sources: GEO (18 studies) and TCGA (4 studies). A complete overview of curated studies by the Public server Delivery Package is shown in Table 4.

Table 4: Overview of curated studies that are currently available in tranSMART

Project/data provider	Data source	Study name(s)	Curation/validation
UBIOPRED	GEO	Woodruff(2005) GSE2125	Curation from scratch
		Woodruff(2007) GSE4302	
		Tsitsiou(2012) GSE31773	
		Choy(2011) GSE23611	
RA-MAP		Julia(2009) GSE12051	
		Yarilina(2008) GSE10500	
		Andreas(2008) GSE10024	
Thompson Reuters		Sorlie(2003) GSE4382	Validation/curation required due to different modification requests/data problems encountered
		Burczynski(2006) GSE3365	
		Gurevich(2009) GSE15245	
		Desmedt(2007) GSE7390	
		Hatzis(2011) GSE25066	
		Arijs(2009) GSE16879	
		Berthier(2012) GSE32583	
		Berthier(2012) GSE32591	
		Mulligan(2007) GSE9782	
		Bienkowska(2009) GSE15258	
		Takeuchi(2009) GSE20690	
-	TCGA	Breast Invasive Carcinoma [BRCA]	
		Colon adenocarcinoma [COAD]	
		Ovarian serous cystoadenocarcinoma [OV]	

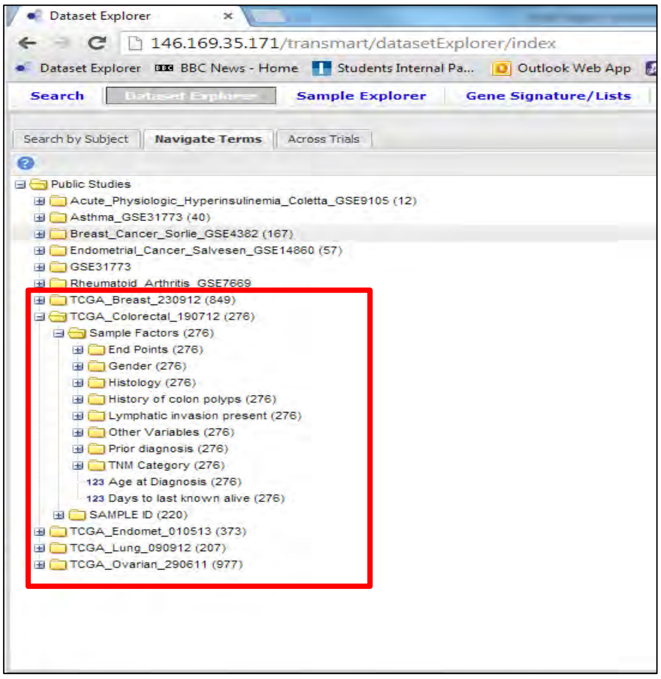
[25] http://public.transmart.etriks.org/transmart/datasetExplorer/index

		Uterine Corpus Endometrial Carcinoma [UCEC]	

On one side, the standard format files for the studies related to the UBIOPRED and RA-MAP projects have beencreated using the GEO-Dataset Explorer pipeline, and their data content has been completely curated and quality checked by the UL team. On the other side, a panel of studies has beenprovided in a curated state, but required additional efforts to make these studies fully compatible with the Dataset Explorer ETL standard format. Furthermore, study ontologies (i.e. the hierarchical structure of the study concepts displayed in tranSMART i2b2 tree) for certain of these studies had to be modified on request of the data provider.

Currently all public studies curated for Dataset Explorer are available on the eTRIKS Public Server. Each of the curated studies has been tested to ensure it is displayed accurately in tranSMART and works with the functionality of the user interface.

The figure below (Figure 5) shows the TCGA datasets displayed in the tranSMART Dataset Explorer.



- Public Studies
  - Acute_Physiologic_Hyperinsulinemia_Coletta_GSE9105 (12)
  - Asthma_GSE31773 (40)
  - Breast_Cancer_Sorlie_GSE4382 (167)
  - Endometrial_Cancer_Salvesen_GSE14860 (57)
  - GSE31773
  - Rheumatoid_Arthritis_GSE7669
  - TCGA_Breast_230912 (849)
  - TCGA_Colorectal_190712 (276)
    - Sample Factors (276)
      - End Points (276)
      - Gender (276)
      - Histology (276)
      - History of colon polyps (276)
      - Lymphatic invasion present (276)
      - Other Variables (276)
      - Prior diagnosis (276)
      - TNM Category (276)
      - 123 Age at Diagnosis (276)
      - 123 Days to last known alive (276)
    - SAMPLE ID (220)
  - TCGA_Endomet_010513 (373)
  - TCGA_Lung_090912 (207)
  - TCGA_Ovarian_290611 (977)

Figure 5: tranSMART Dataset Explorer showing curated TCGA datasets

1.7 Discussion and recommendations

1.7.1 Lack of source file standardization

It is clear that while the use of the loading procedures for tranSMART isstraightforward, the real challenge to automation is the lack of standardization of the provided source files. The transformation process still requires manual and/or ad-hoc processing of the data,as there is currently a lack of data standardization andmachine-readable metadata.The solution seemsto be the use of standard data templates for the publication of such datasets that are embedded with machine-readable metadata. This would substantially simplify the transformation routines and allow the ETL processes to become highly automated. This seems a reasonable and pragmatic compromise to one of the popularalternative suggestions, being explored in the field that is mandated as a pre-publication requirement that the clinical teams pre-load the datasets into a data warehouse. This does not in itself solve the problem of curation and data transformation but simply passes the responsibility to clinical teams who may not have the technical skills or inclination to spend significant amounts of time dealing with a complex data transformation and loading procedure. Therefore, special effort should be put in developing new tools that enable/facilitate the publication/submission of study data in a standardizedand machine-readable format.

1.7.2 Lack of a standard study ontology

Even though our Dataset Explorer-related pipelines allow downloading and extracting relevant clinical and gene expression data in an automated manner for a given GEO study ID, curation at the level of the study ontology is required. The pipelines extract available characteristics e.g. gender, age, mutations status, prior treatment for a given sample in GEO. Based on such a given set of features, the curator manually creates astudy ontology. In this context, special effort has been put in keeping study ontologies homogenous/comparable across all studies, i.e. locating semantically similar variables in identical branches of the i2b2 tree.

References:

[1] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013;41:D991–5.

[2] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 2007;35:D760–5.

[3] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 2011;39:D1005–10.

[4] Robbins DE, Grüneberg A, Deus HF, Tanik MM, Almeida JS. A self-updating road map of The Cancer Genome Atlas. Bioinformatics 2013;29:1333–40.

[5] Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, et al. Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res 2010;38:D690–8.

[6] Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 2010;26:1112–8.

[7] Smyth GK. Limma: linear models for microarray data. In: Gentleman RC, Carey VJ, Dudoit S, Raphael I, Huber W, editors. Bioinforma. Comput. Biol. Solut. Using R Bioconductor, Springer New York; 2005, p. 397–420.

[8] R Core Team. R: A Language and Environment for Statistical Computing 2013.

[9] Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, et al. Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res 2010;38:D690–8.

[10] Pafilis E, O'Donoghue SI, Jensen LJ, Horn H, Kuhn M, Brown NP, et al. Reflect: augmented browsing for the life scientist. Nat Biotechnol 2009;27:508–10.