European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.9 – 3rd progress report on Data Curation

Due date of deliverable: September 2015

Actual submission data: October 2015

Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D4.9
Deliverable title:	3rd progress report on Data Curation
Deliverable version:	
Due date of deliverable:	October 2015
Actual submission date:	
Leader:	Reinhard Schneider, Manfred Hendlich
Editors:	
Authors:	Adriano Barbosa; Serge Eifes; Wei Gu; David Henderson; Nathalie Jullian; Ioannis Pandis; Venkata Satagopam; Emmanuel Van der Stuyft; Francisco Bonachela-Capdevila
Reviewers:	Chris Marshall, Gino Marchetti, Philippe Rocca Serra
Participating beneficiaries:	
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Manfred Hendlich and Reinhard Schneider
Work Package participants:	
Estimated person-months for deliverable:	
Nature:	
Version:	
Draft/Final:	Final
No of pages (including cover):	11
Keywords:	

3rd Progress Report on Data Curation

1 Abstract

Data curation has been provided so far for different Innovative Medicines Initiative (IMI) projects as well as for the European Translational Information and Knowledge Management services (eTRIKS) Public server.

A set of IMI projects have so far collaborated with the eTRIKS consortium on the level of data curation, namely UBIOPRED, OncoTrack, RA-MAP, ABIRISK, APPROACH and AETIONOMY.

In this document, we update an overview of the different projects that eTRIKS has provided support for curation since October 2014. A short project description along with information on what data types have been curated as well as the methods and algorithms applied to the data and specific problems that had to be solved during curation.

2 Introduction

The Innovative Medicines Initiative (IMI) is Europe's largest public-private initiative. It is focused on developing better and safer medicines for patients. Data intensive translational research as it is the case for the IMI projects requires a knowledge management (KM) environment that allows providing sustainable access to the data in an integrated manner.

eTRIKS as a project is specifically focusing on building a sustainable KM platform and able to provide support on the level of data management throughout the life cycle of a given translational research projects. In this context, the value of data curation allows guaranteeing the sustainability of the data and facilitates the analysis and integration of highly complex clinical and multi-omics data.

In this report we describe the curation efforts provided by the eTRIKS consortium during the third year of the project (time frame October 2014 – October 2015). A global overview on the support we have provided to the different collaborating IMI projects will be given. Therefore, we focus on the following information for each of the projects:

- generic project description,
- information on the data types that have been curated,
- methods and algorithms applied to the data and
- specific problems that have been faced during curation.

Besides the curation support for IMI projects, we give here an overview on the data that has been curated for the eTRIKS Public server and details on the curation training that has been provided to the different projects.

3 Curation training and documentation

To provide a better support to the curators of the different IMI projects, eTRIKS has provided an IMI curator training. During the time span of this report, two trainings have been offered. A detailed overview on the curator training and documentations can be found in Table 1

Table 1: Overview on curator training/documentation

Title	Presenter	Organization	Type training	Data type	Date
OncoTrack tranSMART user training	Emmanuel Van der Stuyft	JnJ	Webinar	1) Low dimensional data: patient clinical data, Xenograft and cell culture experimental data, *in silico* model data 2) High dimensional data: VCF data with DNA variant info, methylation data, RNA-Seq data, micro RNA data	6/3/15
Galaxy Training	Wei Gu	UL	Webinar	Selection from above-referred data	11/5/15

4 Overview on eTRIKS supported projects

4.1 APPROACH

Wei Gu (UL) and Andreas Tielmann (Merck)

4.1.1 Project description

IMI APPROACH aims to implement a comprehensive and high quality biomarker assessment to characterise osteoarthritis (OA) patient subsets and support future regulatory qualification and endpoint validation.

The project will provide a framework to identify the "right patient" to treat for a given drug by linking OA patient subsets to potential DMOAD targets based on phenotypic biomarkers, highlight specific disease drivers and progression criteria.

Finally, APPROACH wants to build a stronger collaboration within and among academic and industrial groups to enable future OA therapeutic development.

4.1.2 What data types have been curated?

Until the time of this deliverable, eTRIKS has been working on the curation of two public available datasets:

- FNIH Osteoarthritis Biomarkers Consortium Project
- Cohort Hip and Cohort Knee (CHECK cohort)

For the FNIH cohort, there are 600 subjects, each with more than 350 variables collected. The first round of curation is finished with a full-dataset and a reduced-dataset (a subset filtered based on the full-dataset) both loaded to the APPROACH-tranSMART working server hosted at the University of Luxembourg.

For the CHECK cohort, there are 631 subjects. So far we have finished the curation of a subset of 16 variables. This subset has been also loaded to the APPROACH-tranSMART working server hosted at the University of Luxembourg.

4.2 AETIONOMY

Contributors: Adriano Barbosa (UL) and Wei Gu (UL)

4.2.1 Project description

AETIONOMY[1] is novel in terms of both, its scientific approach and its scale. There is a lot of published literature on the potential causes of Alzheimer's and Parkinson's disease and a significant number of major collaborations already funded and working well. The majority of these are looking at individual hypotheses or approaches to the problem e.g. genetic association studies, imaging studies, non-motor Parkinson's disease or familial Alzheimer's disease. Rather than start another similar approach, AETIONOMY will identify all of the available datasets either from published literature, publically available datasets or datasets from our collaborators.

A common framework will be developed which will allow the integration of data relevant for modeling and mining. Once this data has been curated (re-annotated and quality controlled) and put into the common framework, novel data mining and visualization approaches will be used to identify the pathophysiological changes occurring in the disease process at a molecular level. The knowledge extracted from the datasets will be used to cluster individual patients into separate mechanism based sub-groups leading to a new taxonomy of Alzheimer's disease and Parkinson's disease.

eTRIKS support will happen mainly regarding the activities of AETIONOMY's WP2, such as to acquire, to curate and to build the data cube infrastructure which will integrate the available data. So far, the eTRIKS curation workflow has been adapted to cope with the curation needs of AETIONOMY.

4.2.2 What data types have been curated?

AETIONOMY will use tranSMART as one of the main components of the AETIONOMY Knowledge Base[2] (AKB). For that purpose, eTRIKS support is needed so set-up the AETIONOMY tranSMART server[3] as well as to load the selected studies to the system.

So far, AETIONOMY has selected public studies that are relevant for its purposes and used some studies previously loaded at the eTRIKS public server have been re-uploaded to the AETIONOMY tranSMART server (PD studies displayed on Table 2).

[1] http://www.aetionomy.eu
[2] http://aetionomy.scai.fhg.de/
[3] https://aetionomy.uni.lu/transmart

4.2.3 What methods/algorithms and/or pipelines have been developed/used?

The methods applied to AETIONOMY are the same described for projects mentioned on previous reports.

4.2.4 What problems have been encountered?

No major problems once that the PD Studies from eTRIKS loaded at the AETIONOMY server were already ready for upload.

4.3 RA-MAP

Contributors: Denny Verbeeck (JnJ) and Francisco Bonachela-Capdevila (JnJ)

4.3.1 Project description

RA-MAP[4] is a public-private collaborative project into early Rheumatoid Arthritis (RA).

RA-MAP seeks[4]:

- To identify the key predictors of clinical response and remission in RA patients, and;
- To identify those individuals at high risk of developing RA.

By understanding the human immune system in RA through the study of biological samples from RA patients we plan to develop an 'immunological toolkit' measuring the immune status of healthy individuals and patients.

The goal of RA-MAP[4] is to identify predictors of remission in RA.

There is a major need to identify the characteristics of those individuals most likely to achieve clinical remission so that both new and existing therapies can be targeted to the right patient populations.

4.3.2 What data types have been curated?

For the moment, mostly clinical data and gene expression data. In the future, small RNA and metabolomics data will be included.

Clinical data comes from the TACERA study and includes 273 patients. Since the project is still ongoing, the study in tranSMART is updated every three months. Currently, gene expression data available in tranSMART comes from the "Pilot" experiment, which includes a subset of 12 TACERA patients and 8 controls. More microarray data for TACERA patients are expected by the end of 2015.

The RA-MAP curated public studies have been loaded into the RA-MAP tranSMART instance. Studies with less than 10 patients have been filtered out.

[4] https://research.ncl.ac.uk/mrgnewcastle/translationalprojects/ramap/

4.3.3 What methods/algorithms and/or pipelines have been developed/used?

- The clinical data curation is at an early stage. Checks are performed to ensure that the provided data fall within valid types and valid ranges. It is also checked possible inconsistencies that might lead to data reformatting when necessary. Any reformatting is agreed with the data provider at King's College.
- Derived data columns are obtained based on the feedback of tranSMART users to increase cohort selection flexibility.
- As for gene expression, both raw and normalized data are uploaded to tranSMART. Raw data are extracted with GenomeStudio from the idat files provided by Tepnel Pharma Services and normalized data are obtained using neqc method in Limma/Bioconductor over the raw data.
- Currently, an R script is being developed to check that data uploaded to tranSMART are consistent with and equivalent to the original gene expression data. In this way, the data owner or data user can be sure that the data has not been altered while being uploaded.

4.3.4 What problems have been encountered?

Data curation is a slow process since it involves several actors. It needs to be ensured that any data changes produced during the curation process are agreed and approved by the data donor. It also needs to be ensured that any curation script is shared within the RA-MAP community.

4.4 ABIRISK

Contributors: Wei Gu (UL), Nathalie Jullian (CNRS) and Serge Eifes (UL)

No major update since the 2nd progress report release.

4.5 OncoTrack

Contributors: Adriano Barbosa (UL); Serge Eifes (UL); Wei Gu (UL); David Henderson (Bayer); Gino Marchetti (CNRS); Nathalie Jullian (CNRS); Ioannis Pandis (ICL); Anthony Rowe (JNJ); Venkata Satagopam (UL); Emmanuel Van der Stuyft (JNJ)

No major update since the 2nd progress report release.

4.6 U-BIOPRED

Contributors: Ioannis Pandis (ICL), Kai Sun(ICL) and Florian Guitton (ICL)

No major update since the 2nd progress report release.

5 eTRIKS Public server

Contributors: Serge Eifes (UL), Wei Gu (UL), Adriano Barbosa (UL), Venkata Satagopam (UL) and Nathalie Jullian (CNRS)

5.1 Introduction

As described in deliverable D4.5 (1st Progress report on Data Curation, section 1.2 "Aim of the Public server delivery package"), the main objective of the eTRIKS

Public server is to provide a public eTRIKS/tranSMART server[5] giving access to highly curated and standardized public studies. This server should make public studies that are of interest for the different IMI projects accessible to the public. An added value for this public data is the application of eTRIKS data curation and quality standards facilitating the integrated analysis of these studies in the eTRIKS tranSMART software.

In this section, we give an overview on the different data resources and studies that have been curated and uploaded to the Public server for the reported period. We provide details on which data types have been curated, the methodologies and tools that we have used during data curation and upload as well as the problems and limitations that have been encountered.

5.2 Advertisement for Public study curation requests

To reach out to the other IMI projects and make them aware of the capabilities of the eTRIKS project in general and the Public server more specifically, we decided to provide an advertisement[6] for Public study curation requests to the IMI consortium, which was announced in their newsletter.

This advertisement gives detailed information on the eTRIKS data curation and loading service for public studies that we offer to other IMI projects. Following curation by our curation team, the studies are made available at the Public server. In conjunction with the advertisement and to allow to properly collecting the incoming curation requests we developed an online curation request form. The corresponding URL has been made available in the advertisement[7].

5.2.1 Description

The Gene Expression Omnibus (GEO)[8] is a public functional genomic repository for data submitted by the research community. GEO contains high-throughput microarray and next generation sequence data, and currently encompasses more than 32000 public studies (Barrett et al., 2013).

For gene expression data, it is critical that contextual biological and processing details under which experiments were performed are also made available. A lack of such "experimental metadata" can render the data itself meaningless. GEO is a MIAME-compliant infrastructure and supports fully annotated records encompassing biological as well as descriptive metadata (Barrett et al., 2007).

GEO is a database that allows a unified access to thousands of valuable public gene expression studies (Barrett et al., 2011). This makes it particularly interesting as data resource for study curation in the context of tranSMART Dataset Explorer.

Since the last curation report, the eTRIKS Public server team have made available a new dataset containing 16 Parkinson's Disease-related and 8 Asthma-related GEO studies. Curation has been performed according to the standards defined in

[5] https://public.etriks.org
[6] https://portal.etriks.org/Portal/pdf/IMI_Public_server_advertisement_v1.pdf
[7] http://tinyurl.com/qfpxtgd
[8] http://www.ncbi.nlm.nih.gov/geo/

deliverable "D4.5 1st Progress report on Data Curation". A full overview of these studies can be found in **Table 2**.

Table 2: New GEO studies loaded to the eTRIKS/tranSMART Server. "Study ID" indicates the name of the GEO Series for the selected. "Domain" indicates the disease domain of the study: Asthma or PD. "Gene Expression Platform" shows the IDs of the platforms used for gene expression study. "Samples" corresponds to the number of samples deposited on GEO for for each study. "Variables" shows the number of clinical variables collected for each sample of one study. "Data points" the total number of clinical values collected for each study.

Study ID	Domain	Gene Expression Platform	Samples	Clinical Variables	Data points
GSE19301	Asthma	GPL96	685	37	25345
GSE27876	Asthma	GPL6480	18	5	90
GSE31773	Asthma	GPL570	40	8	320
GSE43696	Asthma	GPL6480	108	4	432
GSE45251	Asthma	GPL4133	16	2	32
GSE46171	Asthma	GPL6480, GPL16981	91	5	455
GSE46238	Asthma	GPL14550	12	4	48
GSE63142	Asthma	GPL6480	155	2	310
GSE20141	PD	GPL570	18	2	36
GSE20146	PD	GPL570	20	3	60
GSE20153	PD	GPL570	16	4	64
GSE20163	PD	GPL96	17	4	68
GSE20164	PD	GPL96	11	5	55
GSE20168	PD	GPL96	29	4	116
GSE20291	PD	GPL96	35	35	1225
GSE20292	PD	GPL96	29	4	116
GSE20295	PD	GPL96	93	4	372
GSE20314	PD	GPL96	8	4	32
GSE20333	PD	GPL201	12	2	24
GSE23676	PD	GPL5188	27	3	81
GSE35642	PD	GPL96	18	3	54
GSE54282	PD	GPL17047	33	4	132
GSE6613	PD	GPL17047	105	1	105
GSE7621	PD	GPL570	25	1	25
TOTAL	**2**	**9**	**1621**	**150**	**29597**

5.2.2 What data types have been curated

We have curated the available clinical data in conjunction with the gene expression data.

5.2.3 What methods/algorithms/pipelines have been used or developed?

The GEO-Dataset Explorer curation pipeline[9] (for more details see deliverable "D4.5 1st Progress report on Data Curation", section "1.4.2.1 GEO Dataset Explorer pipeline") was used to generate the corresponding standard format files. The curated files have then been uploaded to tranSMART using the Kettle Pentaho[10] ETL scripts[11].

5.2.4 Problems encountered

The metadata uploaded by the study data owners to GEO are currently not following any standards and come in various, inconsistent patterns. For a number of studies, metadata are not provided for many key fields or standard categories. These data are frequently found back in other fields. Hence, the lack of standardization requires a lot of manual curation to resolve such data problems.

[9] https://git.etriks.org/serge.eifes/geo_deapp_pipeline/tree/master
[10] http://community.pentaho.com/projects/data-integration/
[11] https://git.etriks.org/transmart-dse-etl/tree/master/DSE/Kettle/Kettle-ETL

5.3 Discussion on encountered problems

The major achievement of the present report was the inclusion of 24 new GEO studies to the public server. The studies are relatively rich in terms of clinical variables of interest (150 in total), which challenged the curation efforts to consolidate the inclusion of these datasets under the eTRIKS availability.

The difficulties regarding the curation of these datasets are similar to those mentioned at the previous report, in summary: lack of standardization of the terminology used in various studies; lack of standards regarding the format in which the datasets are represented in the source repository (i.e. NCBI GEO).

However, the difficulties encountered in the reported period are better approached due to the expertise accumulated so far in the project.

5.4 Bibliography:

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., … Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Research*, *39*(Database issue), D1005–10. doi:10.1093/nar/gkq1184

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., … Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Research*, *35*(Database issue), D760–5. doi:10.1093/nar/gkl887

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991–5. doi:10.1093/nar/gks1193