European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

Deliverable D1.6

A high value, stable repository of curated and annotated translational studies

Due date of deliverable: September 2017

Actual submission date: September 2017 for the first draft version

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)	

Project	
Project acronym:	eTRIKS

Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D1.6
Deliverable title:	A high value, stable repository of curated and annotated translational studies
Deliverable version:	1.0
Due date of deliverable:	Sep 30th 2017 (Month 60)
Actual submission date:	7 September 2017
Leader:	Ghita Rahal
Authors:	Ghita Rahal, Wei Gu, Kavita Rege, Venkata Satagopam, Gino Marchetti, Justin Bussery, Benjamin Guillon
Reviewers:	Chris Marshall, Jay Bergeron, Francisco Bonachela-Capdevila
Participating beneficiaries:	CNRS, UL
Work Package no.:	1
Work Package title:	WP1: Platform Service Delivery
Work Package leaders:	Ghita Rahal
Work Package participants:	Ghita Rahal, Pengfei Lui, Gino Marhcetti, Benjamin Guillon, Andreas Tielmann, David Henderson, Denny

	Verbeeck, Florian Guitton, Wei Gu, Chris Marshall, Peter Rice
Estimated person-months for deliverable:	12
Nature:	Report
Version:	1.0
Draft/Final:	FINAL
No of pages (including cover):	12
Keywords:	eTRIKS, Data Curation, GEO studies

DELIVERABLE INFORMATION

I. Introduction

The eTRIKS consortium hosts a publicly-accessible instance of the eTRIKS platform (the "Public Server": https://public.etriks.org) that provides, to the research community at large, a collection of well curated public studies together with applications capable of providing visualization and analysis tools to exploit this important source of data. The Public Server also provides a data-rich resource to meaningfully demonstrate the features of the

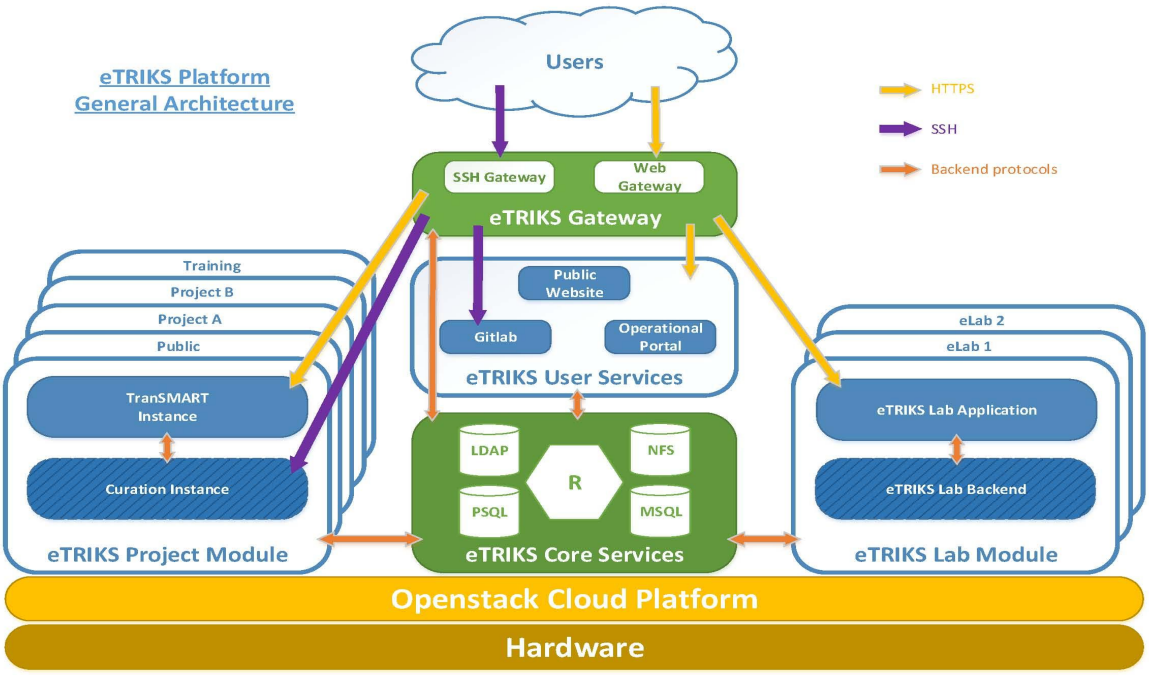
eTRIKS/tranSMART platform for potential clients as well as for current clients wishing to better utilize the platform..

This document describes the creation of this study collection which, at the time of this writing, contains more than 180 public domain studies encompassing almost 15000 patients.

II. Executive Summary

The goal of this deliverable is to provide an overview of the catalog of public studies that have been cleaned, curated and loaded into a database to be exploited by researchers through the tranSMART-based eTRIKS platform and other scientific applications. This deliverable provides (in Section III) a brief description of the platform that has been discussed in detail elsewhere[1]. Section IV describes the data processing methods to transform raw experimental/study data to a format that is analysis-ready and readily available from the eTRIKS data warehouse. Section V finally details the repository itself including the study content available at the time of this writing.

III. Platform



[1]Bussery J., Denis L-A., Guillon B., Liu P., Marchetti G., Rahal G. *eTRIKS platform: conception and operation of a highly scalable cloud-based platform for translational research and applications development*. (to be submitted)

Fig 1: Schema of the eTRIKS platform deployed at CC-IN2P3. A Cloud technology (Openstack) layer is deployed over the base hardware. On the left is shown the sketch of a generic module as deployed for each project. On the right are the more specific modules built up to satisfy the needs of every new in-development application. In the middle are all the core services that are shared between the modules. On the top are the possible ways for the users to access to the services of the platform.

The platform design and operation at CC-IN2P3 has been described elsewhere [Deliverables D1.1 and D1.2 and reference in the note below]. Figure 1 represents the eTRIKS platform's architecture, identifying each of the components and representing the component interactions that comprise the platform in its entirety.
The data curation-related system that is used to provide the repository of annotated translational studies is composed of:

- A physical cluster dedicated to postgreSQL databases, with a separate storage stack with systems redundancy at both the hardware and application levels.
- a Virtual Machine (VM) accessing the data on the storage disk and on the PostgreSQL database. Only a small set of users, the "curators", have the required SSH access privileges to create, modify and delete data and files. It provides an environment where curators may manipulate the raw data loaded onto the disk storage and import the curated sets to the database.

System security is implemented at both user and data levels and these security measures ensure the identification of users (Authentication), providing them with the rights pertinent to their role (Authorization), and protect the data from unauthorized access. Network communications employ SSL encrypted protocols.

IV. Method

a. Data Source

Four data sources have been used for curating/loading to the eTRIKS public server as described in "III. Platform". These data sources provide datasets that are publicly accessible.

National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO)
The GEO 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, the GEO database currently has some of the the richest public datasets (http://www.ncbi.nlm.nih.gov/geo/) available. 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. GEO represents a repository unifying thousands of valuable public gene expression studies, which makes it particularly interesting as data resource for study curation in the context of tranSMART Dataset

Explorer.

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 along with clinical data (https://tcga-data.nci.nih.gov/tcga/). Currently data is available for 30 different cancer types on the TCGA Data Portal. The considerable amount of clinical data altogether with gene expression data makes TCGA an interesting data source for curating studies for tranSMART Dataset Explorer.

Cancer Cell Line Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (CCLE) (https://portals.broadinstitute.org/ccle/home) project is an effort to conduct a detailed genetic characterization of a large panel of human cancer cell lines. The CCLE provides public access analysis and visualization of DNA copy number, mRNA expression, mutation data and more, for 1000 cancer cell lines.

GlaxoSmithKline (GSK) Cancer Cell Line Genomic Profiling Data

The GlaxoSmithKline (GSK) Cancer Cell Line Genomic Profiling Data (Greshock et al., 2010) provides access to the genomic profiling data for more than 300 human cancer cell lines. This data initially generated by GSK has been made publicly available and encompasses a large variety of cancer cell types.

b. **Data Curation and Loading**

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 generating standard format files that are loaded into tranSMART (including data cleansing and construction of data tree)
3) completing and/or standardizing annotations of metadata.

Once data have been curated, 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, TCGA, CCLE and GSK data.

The difficulties regarding the curation of these datasets are similar to those mentioned in the previous report (see deliverables D4.5, D4.7, D4.9 and D4.12). Some of the data and metadata

provided by the source, for example GEO, are currently not following consistent standards. Metadata are incomplete, and not provided for some studies in the intended fields. These data are frequently found back in other fields. This lack of standardization requires a lot of manual curation to cover such data problems. However, the effort in tackling the difficulties encountered during the process are better approached due to the expertise accumulated in the project. The data and metadata correction effort extended by eTRIKS curators in making the data ready for use in the eTRIKS Knowledge Platform means that the study data sets are more robust, accurate and suitable for cross-study analysis than the original data sets.

Details of the curation pipelines have been reported in D4.5, D4.7, D4.9 and D4.12.

V. eTRIKS data studies repository

The major achievement of the present report was the curation and loading of 180 studies from the above-mentioned data sources to the eTRIKS public server. The studies represent a wide variety of disease areas, data types and clinical variables., Curating such varied datasets required substantial effort by WP4 personnel who, if and as possible, harmonized the data such that metadata and data values would be as consistent as possible across studies.

A summary of the curated studies in the eTRIKS public server ins shown in Table 1.

Table 1. List of curated studies in the eTRIKS public server

Study Area	Datasets curated and loaded
Acute Myocardial Infarction(AMI)	GEO Studies: GSE62646 GEO Studies: GSE11947 GEO Studies: GSE29532 GEO Studies: GSE49925 GEO Studies: GSE34198 GEO Studies: GSE66360 GEO Studies: GSE48060 GEO Studies: GSE97320 GEO Studies: GSE28454
Asthma	GEO Studies: GSE45111 GEO Studies: GSE59339 GEO Studies: GSE41861 GEO Studies: GSE41862 GEO Studies: GSE41863 GEO Studies: GSE23611 GEO Studies: GSE67472 GEO Studies: GSE63383 GEO Studies: GSE35643 GEO Studies: GSE56553

	GEO Studies: GSE46171 GEO Studies: GSE52074 GEO Studies: GSE43696 GEO Studies: GSE31773 GEO Studies: GSE22324 GEO Studies: GSE65163 GEO Studies: GSE65204 GEO Studies: GSE65205 GEO Studies: GSE45847 GEO Studies: GSE61225 GEO Studies: GSE40240 GEO Studies: GSE31773 GEO Studies: GSE44037 GEO Studies: GSE2125 GEO Studies: GSE4302
Cancer	Cancer Cell Line Encyclopedia GEO Studies: GSE7390 GEO Studies: GSE25066 GEO Studies: GSE4382 GEO Studies: GSE9782 TCGA_BREAST TCGA_COLORECTAL TCGA_OVARIAN TCGA_ENDOMET
Cell Lines	GSK Cell Lines
Chronic Obstructive Pulmonary Disease	GEO Studies: GSE8581
Inflammatory Bowel Disease	GEO Studies: GSE20881 GEO Studies: GSE16879 GEO Studies: GSE47908 GEO Studies: GSE3365 GEO Studies: GSE6731 GEO Studies: GSE51785 GEO Studies: GSE9686 GEO Studies: GSE10616 GEO Studies: GSE36807 GEO Studies: GSE9452 GEO Studies: GSE38713 GEO Studies: GSE10714 GEO Studies: GSE48957 GEO Studies: GSE48958 GEO Studies: GSE48959
Lupus Nephritis	GEO Studies: GSE32583 GEO Studies: GSE32591
Macrophages Related Studies	GEO Studies: GSE11886 GEO Studies: GSE23622

	GEO Studies: GSE10220 GEO Studies: GSE24780 GEO Studies: GSE7138 GEO Studies: GSE15907 GEO Studies: GSE17761 GEO Studies: GSE19236 GEO Studies: GSE3982 GEO Studies: GSE18275 GEO Studies: GSE5099 GEO Studies: GSE14769 GEO Studies: GSE6054 GEO Studies: GSE1432 GEO Studies: GSE9820 GEO Studies: GSE8512 GEO Studies: GSE16385 GEO Studies: GSE21548
Multiple Sclerosis	GEO Studies: GSE15245
Parkinson's Disease	GEO Studies: GSE35642 GEO Studies: GSE20333 GEO Studies: GSE7621 GEO Studies: GSE20146 GEO Studies: GSE54282 GEO Studies: GSE6613 GEO Studies: GSE23676 GEO Studies: GSE20168 GEO Studies: GSE20291 GEO Studies: GSE20292 GEO Studies: GSE20295 GEO Studies: GSE20141 GEO Studies: GSE20153 GEO Studies: GSE20163 GEO Studies: GSE20164 GEO Studies: GSE20314
Psoriasis	GEO Studies: GSE32620 GEO Studies: GSE13355 GEO Studies: GSE26952 GEO Studies: GSE53552 GEO Studies: GSE9120 GEO Studies: GSE38039 GEO Studies: GSE14905 GEO Studies: GSE47598 GEO Studies: GSE7216 GEO Studies: GSE36700 GEO Studies: GSE30999 GEO Studies: GSE27887 GEO Studies: GSE30355 GEO Studies: GSE30768

	GEO Studies: GSE31652 GEO Studies: GSE32407 GEO Studies: GSE42305 GEO Studies: GSE52471 GEO Studies: GSE27628 GEO Studies: GSE36287 GEO Studies: GSE50790 GEO Studies: GSE6710 GEO Studies: GSE6932 GEO Studies: GSE41905
Rheumatoid Arthritis	GEO Studies: GSE10024 GEO Studies: GSE36757 GEO Studies: GSE15602 GEO Studies: GSE15615 GEO Studies: GSE11827 GEO Studies: GSE15258 GEO Studies: GSE21959 GEO Studies: GSE12051 GEO Studies: GSE15316 GEO Studies: GSE9329 GEO Studies: GSE30662 GEO Studies: GSE27390 GEO Studies: GSE3592 GEO Studies: GSE13026 GEO Studies: GSE21537 GEO Studies: GSE11575 GEO Studies: GSE25160 GEO Studies: GSE12653 GEO Studies: GSE20690 GEO Studies: GSE15573 GEO Studies: GSE33377 GEO Studies: GSE19821 GEO Studies: GSE10500
Systemic Lupus Erythematosus	GEO Studies: GSE46923 GEO Studies: GSE23076 GEO Studies: GSE45923 GEO Studies: GSE63829 GEO Studies: GSE78193 GEO Studies: GSE11909 GEO Studies: GSE36941 GEO Studies: GSE27895 GEO Studies: GSE51997 GEO Studies: GSE61635 GEO Studies: GSE37356 GEO Studies: GSE59217 GEO Studies: GSE41825 GEO Studies: GSE36700 GEO Studies: GSE39088

	GEO Studies: GSE63829 GEO Studies: GSE72754 GEO Studies: GSE72754 GEO Studies: GSE23076 GEO Studies: GSE82218 GEO Studies: GSE13887 GEO Studies: GSE50635 GEO Studies: GSE55447 GEO Studies: GSE13887 GEO Studies: GSE20864 GEO Studies: GSE30153 GEO Studies: GSE57869
Epilepsy	GEO Studies: GSE12196 GEO Studies: GSE13428 GEO Studies: GSE14763 GEO Studies: GSE1643 GEO Studies: GSE16969 GEO Studies: GSE26246 GEO Studies: GSE27015 GEO Studies: GSE27268 GEO Studies: GSE47516 GEO Studies: GSE47752 GEO Studies: GSE4917 GEO Studies: GSE50628 GEO Studies: GSE5320 GEO Studies: GSE57585 GEO Studies: GSE63808 GEO Studies: GSE6614 GEO Studies: GSE66762 GEO Studies: GSE73878 GEO Studies: GSE7486 GEO Studies: GSE77578 GEO Studies: GSE81024

VI. Conclusion

The eTRIKS project has provisioned a stable repository of curated and annotated public domain translational studies on a scalable, flexible and secure Openstack cloud-based IT platform. The detailed curation and, where possible, standards-based harmonization of these data greatly increases their value and utility of these relative to the originally published versions. Moreover, the eTRIKS public server platform, with its advanced analysis and visualisation tools, provides investigators at large immediate access to robust search, analysis and export features that can be used directly on these datasets. These features save the time consuming and resource intensive data processing costs that would otherwise be necessary to make use of these

datasets. eTRIKS had an initial goal to provide the research community with a large number of datasets (goal of ~120). This deliverable not only satisfies, but exceeds that goal.