European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.14 – Document describing the refinements and optimizations of the roadmap implementation of eTRIKS curation and analytics support

Due date of deliverable: September 2017

Actual submission data: August 2017

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:	D4.14
Deliverable title:	Document describing the refinements and optimizations of the roadmap implementation of eTRIKS curation and analytics support
Deliverable version:	1
Due date of deliverable:	September 30th 2017
Actual submission date:	August 2017
Leader:	Reinhard Schneider, Manfred Hendlich
Editors:	
Authors:	Wei Gu; Venkata Satagopam; Sascha Herzinger; Adriano Barbosa; Mansoor Saqi; Irina Balaur; Samiul Hasan; Bertrand De Meulder; Alexander Mazein; Charles Auffray; Manfred Hendlich; Reinhard Schneider
Reviewers:	Jay M. Bergeron; David Henderson
Participating beneficiaries:	CNRS, UL
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Manfred Hendlich and Reinhard Schneider
Work Package participants:	Wei Gu; Venkata Satagopam; Sascha Herzinger; Adriano Barbosa; Mansoor Saqi; Irina Balaur; Samiul Hasan; Bertrand De Meulder; Alexander Mazein; Charles Auffray; Manfred Hendlich; Reinhard Schneider
Estimated person-months for deliverable:	2
Nature:	Report (PU)
Version:	1.0
Draft/Final:	Final
No of pages (including cover):	14
Keywords:	Curation, Analytics, Roadmap, Engagement

1 Abstract

Data curation and analytics support has been provided to different Innovative Medicines Initiative (IMI) projects via direct curation, curation training, implementation of specific analytics tools, analytics tools training using the project specific tranSMART (TM) servers or via eTRIKS Public Server (https://public.etriks.org).

A set of IMI projects have been supported by eTRIKS on the level of direct data curation, namely UBIOPRED, OncoTrack, RA-MAP, ABIRISK, APPROACH and AETIONOMY. While general analytics support has been provided to all supported projects, implementation of additional analytical functions has been provided at the request of ABIRISK, APPROACH and OncoTrack.

In this document, we provide a short project description of eTRIKS and summarise the overall refinements and optimizations of the roadmap implementation of curation and analytics support that eTRIKS has provided to the engaged IMI projects during the entire eTRIKS project period.

2 Introduction

The Innovative Medicines Initiative (IMI) is Europe's largest public-private partnership in the life sciences. The IMI is focused on developing better and safer medicines for patients. Data intensive translational research, as needed by IMI projects, requires a knowledge management (KM) environment that provides sustainable access to the data in an integrated manner.

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

In this document, we describe the evolution of eTRIKS curation and analytics support efforts during the five years of the project, together with the refinements and optimizations preformed in terms of process and tools developed to provide better and more efficient support. A global overview on the support model we have provided to the different collaborating IMI projects is given. The report focuses on the following information for each of the projects:

- general problems and requests in data and knowledge management for supported IMI projects in terms of curation and analytics,
- initial approach and challenges,
- refinements and optimizations of the approaches.

3 Refinements and optimisations of the roadmap implementation of curation support

3.1 General problems and requests

The key driver for translational research is to "translate" data and findings from fundamental research into medical practice to improve health outcomes. In practice, this involves complex multidisciplinary research teams working collaboratively to integrate data and knowledge arising from ever-increasing types of assays from pre-clinical and clinical research studies. This involves a significant number of strategic considerations with regard to managing data. This includes cost effectiveness, composition of the right skills, making unbiased data-driven decisions, assuring data quality, having access to prior informative data and knowledge and ideally for maximizing value of the data being generated for the research organization. Curation activities that enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time are needed.

3.1.1 Problems

Data curation refers to a set of planned activities and processes focused on delivering quality data to support evidence based science. The cost for achieving these desirable goals is, however, very often underestimated. Lack of sufficient resources (time) and expertise make efficient curation very challenging. Typical data curation problems in translational research projects are:

- Lack of a proper Data Management Plan (DMP) that documents the data collection, storage, sharing and quality control processes and, most importantly, allocation of resources to implement the DMP.
- Variation in the data quality and non-standardized data coding of data coming from multiple sites.
- Lack of metadata that is essential for curators to understand the pre-curated data (mostly for retrospective data)
- Underestimation of required resources and expertise.
- Legal issues that delay or even prevent external experts from providing curation support.

3.1.2 Requests and needs

Typical needs from supported IMI projects in terms of data curation are:

- Data extraction and preparation to be machine useable (mostly for retrospective data)
- Data pre-processing from raw measurements (mostly OMICS data)
- Data cleansing and quality control
- Data normalization/standardization
- Data transformation (to different formats, different shapes: e.g. denormalized or normalized, that would ease the analysis)

- Data integration (clinical + OMICS data) and loading to eTRIKS platforms (tranSMART)

3.2 Initial approach and challenges

3.2.1 Initial curation support model and process

Initially, eTRIKS allocated personnel previously trained and experienced in preparing data for the eTRIKS platform to curate data for client projects (i.e. the "full support" model). Under the full support model, a client project was not charged for these curation services.

However, the full support model soon faced some challenges. Firstly, negotiation of material transfer or data processing agreements between eTRIKS and client consortia proved to be complicated and inordinately time consuming. Secondly, the number of curation personnel in work package 4, would have limited eTRIKS to engaging only five to eight client projects which would fall far short of eTRIKS' engagement goals. To tackle this, an alternative "light weight" support model was implemented, in which eTRIKS provided data curation guidelines and training to enable scientists working within the client projects to curate their own project data. This light weight model also faced challenges, in that general data curation guidelines/training could not always address project-specific data curation needs to support specialized analysis. Furthermore, many projects did not have personnel with the requisite skills to perform curation.

3.2.2 Initial curation support tools

The initial curation support tools were comprised mostly of custom-purpose scripts written in R/Python/Shell. Data loading ETLs (Extract Transform Load routines) were configured using third-party tools such as Pentaho Kettle. The challenge was to reuse the scripts for other datasets and overcome the steep learning curve of Pentaho.

3.3 Refinements and optimizations of the approaches

To provide better and more sustainable curation support to other IMI projects, we have refined and optimised the initial approach at different levels.

3.3.1 Curation support model

The curation support model of "full support" and "light support" was kept but adopted to a more sustainable and practical manner. A full support model was provided for several consortia (U-BIOPRED, OncoTrack, RA-MAP, ABIRISK, APPROACH and AETIONOMY), for which the legal issues had been resolved or circumvented (e.g. an eTRIKS partner that was also a partner in the client project could process client data without the need for an additional legal agreement). For other projects that chose light support, eTRIKS has developed an IMI curator training course. During the time span of the eTRIKS consortium, several training events have been offered to train colleagues from supported projects so that they can start curation themselves. Questions and problems encountered during the curation and data loading have been discussed and solved by eTRIKS supporting team via the on-line eTRIKS Support Portal.

3.3.2 Curation support process

For the full support model, the support process consists of the following steps:

1) Trigger the work to check and establish legal agreement (when needed) for the curation support.
2) In parallel to 1), eTRIKS supports and is involved in gathering as much metadata as possible for each dataset from data controller (CRO etc.), as they provide essential information for the curation resource planning.
3) In parallel to 1) but after 2), both sides estimate and plan the resources needed to perform the curation. A basic rule is to estimate person-hours needed based on the number of variables in the dataset, rather than the number of subjects. The estimation also depends on the source type, data quality, availability of a data dictionary and requirement of the curation.

In the meanwhile, the following steps are performed based on either dummy data that have the same variables, structure and format as the actual data (before legal issues are resolved) or actual data (after legal issues solved):

4) Transform the source data to a machine-readable format
5) Design a template for the data dictionary and populate the data dictionary using the same template for each dataset that will be curated. This template should be variable-based and should include:
    a. Data source (file name, DB host/port)
    b. Column name/variable name in each data source
    c. Data type of each variable
    d. Data range (for numeric/time variables)
    e. Allowed values or regular expression (for categorical variables)
    f. Validation rule (for values derived from other variables like Body-Mass-Index that should reflect the body size and weight)
    g. Primary key (or part of a primary key) in the dataset, meaning unique value for each row and no NULL value allowed

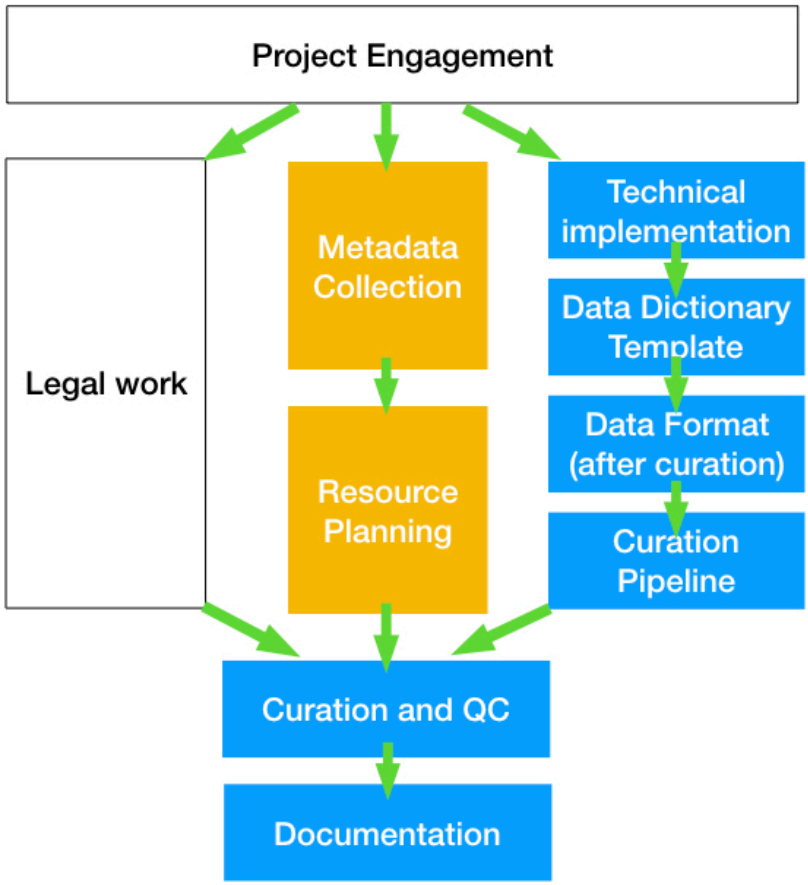
    An example of data dictionary is shown as below:

Data Source	Column/ Variable Name	Short Description	Data Type	Value and Mapping	Ranges and Rules	Primary Key
DM.txt	SUBJ_ID	ID of Study Subjects	String			Y
DM.txt	SITE	Clinic site	Integer	1;2		N
DM.txt	AGE	Age at baseline	Numeric		20-120	N
DM.txt	AGECAT	Age at baseline, 3 categories	Integer	1=Age 50-59; 2=Age 60-69; 3=Age 70-79		N
DM.txt	ETHNICITY	Ethnicity	Integer	0=No; 1=Yes; 8=Do not know/Refused		N

DM.txt	RACE	Racial background	Integer	1=White or Caucasian; 2=Black or African American; 3=Asian; 4=American Indian or Alaskan Native; 5=Hawaiian or Other Pacific Islander; 6=More than one race; 7=Other; 8=Do not know		N
DM.txt	SEX	Sex/Gender	Integer	0=Female; 1=Male		N
LB.txt	KNEE	Knee	String	L=Left; R=Right		N
LB.txt	Reading	Image Assessment moment	String	RV1; RV2		N

6) Design templates for the resulting (curated) dataset before any curation tasks should be performed
7) Develop pipelines that use source data (machine readable format) and data dictionary to automate, as best as possible, the curation (including quality control, reshape, compute derived data, map to ontology).
8) Perform the curation of each dataset using the pipeline(s) developed in step 7
9) Document the curation steps, which should include
    a. Placing the pipeline source code into a version control system
    b. Avoiding Manual steps as much as possible favouring the development of automated curation routines coded into the pipeline to promote data quality and traceability.
    c. Placing the curated dataset into a version control system
    d. Documenting the rationale which necessitated any alteration of data values during the curation process, as well as the approval process associated with such alterations including the date, reason and personnel having modified, and approved modifications to, data values.

These steps are also shown as a diagram as below:



Projects supported by the lightweight model sent their own researchers / data managers to eTRIKS curation training sessions. Problems and issues that subsequently arose were solved via the eTRIKS supporting portal.

3.3.3 Curation support tools

To facilitate the curation support, eTRIKS has developed the following tools:

GEO data retrieval pipelines for public server: this tool takes a GEO_ID as input to retrieve sample information and expression data from the GEO server and formats the retrieved data into tranSMART ETL standard format for further curation and corrections before loading to tranSMART DB.

Data dictionary template (for IMI-APPROACH): this is an excel template developed together with IMI-APPROACH to collect and format metadata for each variable collected in IMI-APPROACH. The populated metadata using this template can be read by curation scripts to generate tranSMART ETL standard format files.

OncoTrackDB exporter and automatic loading (together with OncoTrack colleagues): this is a plugin function for OncoTrackDB co-developed between eTRIKS and OncoTrack to automatically export new data stored in the OncoTrackDB into tranSMART ETL standard format files.

eTRIKS Harmonisation Service (eHS: together with WP2 and WP3): This tool expedites data loading using a new repository for harmonized data and meta-data

for translational research projects. The repository and associated interfaces will facilitate the configuration of study designs, preparation of study data for system import and transformation of study data into semantically consistent representations.

eTRIKS tranSMART Master Tree (together with WP3): eTRIKS Master Tree is based on the non-redundant coverage of SDTM (Study Data Tabulation Model) to represent clinical data. The Master Tree follows a basic and easy-to-understand logic, which was built on the premises of tranSMART rules for data loading. This means that multiple data collected for one patient for the same domain should be distinguished based on *Data Labels*. Moreover, multiple results for the same test, should also be distinguished based of the *Visit Names*. Respecting these two basic rules, any laboratory tests results, and even results for any other domain, can be easily represented in tranSMART via the eTRIKS Standard Master Tree (Barbosa-Silva and Bratfalean et. al., manuscript in preparation).

4 Refinements and optimisations of the roadmap implementation of analytical support

4.1 General problems and requests

eTRIKS has developed its knowledge management platform around the tranSMART platform. This has become the repository of choice for many eTRIKS supported projects. Initially, this platform served as a data warehouse with some simple analytical capabilities. However, there are much higher expectations of this platform in terms of visual-analytics so that it can be used as a data exploration, hypothesis generation, and more importantly a collaboration platform.

4.1.1 Problems

In the earlier versions of TM, the user interface (UI) is not intuitive. This needs to be improved, including functionality for the browsing, querying and extraction of curated data sets.

In these earlier versions, the analytical functions are very basic. Results of the workflows are static pictures with very limited contextual information and lack links to external knowledge bases.

Users can't interactively alter the filters for the data without running the complete workflow again. This makes hypothesis testing and data exploration a time-consuming task.

4.1.2 Requests and needs

Intuitive for end users: the intended user community comprises non-IT scientists and clinicians. The TM functionality needs to be designed according to clinicians and biologists' requirements. Browsing the data and searching for particular types of information is a common requirement and should be facilitated.

Allow dynamic and interactive visual-analytics: The user should be able to modify the filter and scope on the results of the initial analysis to get an in-depth look at the data. Re-running the analysis on-the-fly without the need to repeat the complete workflow should be possible.

Support new data types: genetic variants, time-series data and overlay of different data types in the same visual-analytics should be supported by the platform.

Agile and modular: a user must be able to select data for analysis, choose analytical methodologies and visualization features, and save, export or perform further processing on the selected data. The algorithms and associated parameters used for any analysis must be clearly recorded. A plug-in architecture will allow new methods to be added to the platform.

4.2 Initial approach and challenges

Initially eTRIKS focused on fixing issues in existing analytics functions and reacting to user requests in a "first come first serve" model. Many of the requests were implemented as new "advance workflows" that suffered from the same static analysis approach. Developers and support team created new workflows based on initial understanding of the user's need and implemented on a by-request base. Due to the framework used in TM, implementation of new workflows is very time consuming. External tools like Galaxy have been connected to TM so that the workflow implementation can be more focused on the analysis itself rather than struggling to fit it to the framework TM is using. However, despite the many efforts in this initial approach, user needs were not completely satisfied.

4.3 Refinements and optimizations of the approaches

4.3.1 Analytical support model

To have a sustainable analytical support model, the approach to development must be proactive rather than reactive. Recommendations from bioinformaticians, developers, data managers, and statisticians were therefore taken into account when including new functionality and UI improvements, as is this community is familiar with the state-of-the-art in bioinformatics analysis.

In the refined and optimized analytical support model, planning is carried out according to collective user requests and new developments in tranSMART. Developers use more systematic tools (JIRA and eTRIKS service portal) to receive and track issues and requests. Development and support teams work closely with users to adjust development to reflect their requirements. Documents, training, webinars, videos are provided to users as early as possible.

To allow the exploration of new technologies without user's explicit requests, a new concept, "eTRIKS-labs" was introduced. Developers that have insight in a particular area of bio-visual-analytics can develop new proof-of-concept (PoC) tools and demo

them to the user community. Once positive feedback is received, a production level version can be implemented based on the PoC version.

4.3.2 Analytical support process

In the new analytical support model, the support process consists of the following steps:

- Systematic collection of user requests.
- Then developers (IT and bioinformatics background) and support team (bioinformatics and biology background) harmonize their understanding of the initial requests.
- Together they evaluate, merge and rank requests.
- The development plan (roadmap) is made based on the evaluation and ranking.
- Implementation in tandem with user tests and feedback.
- Training and documentation are provided once a new function is stable.

4.3.3 Analytical tools developed

Under the refined and optimized analytical support model, the following tools have been developed:

SmartR: SmartR provides a highly dynamic and interactive way of visualizing and analysing data within tranSMART. Using recent web technologies, SmartR generates visual analytics within the web browser rather than making use of static images. This provides the user with the possibility to explore and interact with the data while background scripts ensure that heavier computations are carried out by the tranSMART server, maintaining a responsive user interface.

eAE (eTRIKS Analytical Environment): eAE aims to provide a scalable analytical environment that allows efficient analysis of (big) translational research data.

Disease Knowledge Base: The Disease Networks module has been developed as a multi-scale framework (using the graph database approach) to facilitate management (integration, exploration, visualisation, interpretation) of diverse types of biological and biomedical data. Disease Networks employ the popular graph database Neo4j (http://neo4j.com/), which provides a persistence mechanism that is robust and has powerful functionality (the Cypher query language) that allows the user to query networks, to find connections between particular data entries using graph traversal techniques.

Disease maps: a collection of interconnected signalling, metabolic and gene regulatory pathways relevant to a particular disease. Disease mechanisms are depicted on the level of molecular processes and represented in standard computer-readable formats. The approach has grown into a large-scale community effort with multiple projects involved (http://disease-maps.org/projects) and continues scaling up in partnerships with such friendly projects as Virtual Metabolic Human

(https://vmh.uni.lu/), GARUDA (http://www.garuda-alliance.org) and the Physiome Project (http://physiomeproject.org/).

HiDome: The core objective of this delivery is to allow researchers to create patient cohorts by setting constraints on components in high dimensional data sets, e.g. select all patients with expression value of a particular gene greater than a user-selected threshold. Additionally, the values for these components can be viewed like regular numerical concepts in the Summary Statistics and Grid View tabs. To this end, the tranSMART web application will be extended with GUI features that will allow researchers to specify constraints over these components. By allowing researchers to directly compare clinical data with high dimensional data, the integration of HiDome in tranSMART will expedite translational research.

Interactive implementation of the Weighted Gene Co-expression Network Analysis (WGCNA): The eTRIKS WGCNA lab was developed using existing R packages wrapped in a R shiny framework. This made possible the use of the WGCNA R package through a webpage so that users can perform this complex analysis without having to type a single line of code. This implementation makes it easier for biologists and clinicians to use this powerful algorithm despite its relative complexity of use.

Interactive implementation of the Similarity Network Fusion (SNF) algorithm: SNF is a novel computational method for genomic data integration that was developed by Wang et al., in the lab of Anna Goldenberg. SNF constructs patient similarity networks for each of the data types and in a second step iteratively integrates them until it converges to a final fused network. In order to make the approach more accessible, a Shiny web app for SNF was developed, where a user has the ability to integrate various data types, adjust the parameters, view the results and download network files and group assignments.