European Translational Information and Knowledge Management Services

eTRIKS Deliverable Report

Grant agreement no. 115446

Deliverable 2.6

Requirements document for eTRIKS KM Platform v3.0

Due date of deliverable: Month 25

Actual submission date: Month 31

Dissemination Level		
PU	Public	X
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:	D2.6
Deliverable title:	Requirements document for eTRIKS KM Platform v3.0
Deliverable version:	0.7
Due date of deliverable:	
Actual submission date:	10/4/2015
Leader:	Ioannis Pandis
Editors:	Ioannis Pandis
Authors:	Ioannis Pandis, Peter Rice
Reviewers:	Fabien Richard, David Henderson, Ghita Rahal, Heike Schürmann, Francisco Bonachela-Capdevila, Jay Bergeron
Participating beneficiaries:	ICL, CNRS
Work Package no.:	WP2
Work Package title:	Platform Development
Work Package leader:	ICL
Work Package participants:	
Estimated person-months for deliverable:	3
Nature:	PU
Version:	1
Draft/Final:	Final
No of pages (including cover):	12
Keywords:	Requirements, Users, Platform development

Table of contents

Introduction

Purpose

As set out in the Description of Work, WP2 aims to develop a scalable, secure and reliable eTRIKS KM platform by extending and enhancing the tranSMART core architecture. Therefore the work package will focus on the development of the eTRIKS core architecture to support petabyte range data sets, four-figure user numbers, secure data, multi-tenancy, and enhanced usability. An initial set of feature requirements has been gathered for the eTRIKS platform in collaboration with other work packages using the process described in deliverable D2.1 'Product Features Decision Making Process', and the plan set out in the D2.2 'eTRIKS Product Roadmap' deliverable document.

Intended audience

The readership of this document is assumed to be familiar with eTRIKS and its overall aims, including being aware of the work completed to date with respect to the tranSMART for eTRIKS software release that currently forms the eTRIKS KM Platform v1.0.

Scope

In this document, we provide a review of the documented set of initial feature requirements for eTRIKS KM Platform v2.0 (D2.3), as well as describe the feature requirements for eTRIKS KM Platform v3.0. Stakeholders were solicited for feature requirements, some of which were directly input into the eTRIKS Product Management (User requirements) Wiki: https://requirements.etriks.org/twiki/bin/view/RequestManagement/WebHome.

The requirements set out here in this document should be treated as a living document, the current version of which represents a snapshot of current requirements that are valid at time of publication. It should be recognized that these might change over the course of the current development scope where testing and minor releases up until the v3.0 will provide feedback from originating feature requestors and users groups.

Overall description

Product perspective

eTRIKS aspires to become the European translational research commons framework to support and enable translational medicine initiatives. It is envisaged that eTRIKS shall provide an open and collaborative model for exchange and analysis of scientific knowledge, supporting development of new approaches for the prevention, diagnosis, and treatment of disease, ultimately redefining the way biomedical research is translated to better healthcare for the patient.

It is not intended that eTRIKS should provide solutions for all problems, but that the commons infrastructure should enable the community to build, expand and share their solutions. From our understanding of the current informatics challenges in translational research and driven by the various IMI projects that request our support, we believe that eTRIKS platform should aim to deliver the following functionalities:

1. A common knowledge base of translational-medicine-related facts and observations resulting from cumulative results of translational research investigations, where outcomes of basic and clinical research are continually integrated under a systems-biology context.
2. Study-centric storage for scientific research data providing ready access to the content of the knowledge base and provenance support for reproducibility of analysis results and reuse of datasets and analysis workflows.
3. Open data and open access[1] services to allow researchers to design different analysis and visualization procedures, to build and reuse analysis workflows and integrate with third party tools and services.
4. A collaborative environment where multiple users share and contribute their data, analyses and interpretations enabling cross-study and cross-domain information sharing and integration.
5. New intuitive methods for the navigation and visualization of translational research knowledge to enhance and support new discoveries and decision-making.

Currently eTRIKS KM Platform v2.0 consists of the study-centric storage described in (2) above, with minimal support for reproducibility of analyses or provenance. The work to be carried out as defined in this requirements document will go towards satisfying all five of the above aims in eTRIKS KM Platform v3.0.

Review of the feature requirements for eTRIKS KM Platform v2.0 (D2.3)

The initial step before gathering requirements for eTRIKS KM Platform 3.0 was to review the requirements set for version 2.0 to determine their status of completion and delivery. This task

[1] Data stored on public server should come from the public repositories and the IMI projects when they are finished. In both cases data must be anonymised. Thus there is not anymore data privacy and security issues, legal terms. eTRIKS will facilitate the data openness. If data access remains restricted, eTRIKS will promote the collaboration by showing publically what variables are. Everyone can know and contact the study owners for access to the study data. We call it "open collaboration".

was carried out initially though a web meeting, which the majority of the stakeholders/requesters attended, followed by a review at the project 2nd annual meeting (Thu 12th Feb 2015) in Barcelona.

*One important point arising from the review meeting was the **lack of formal acceptance criteria for each requested feature**. Such criteria would be used to determine the completion of the requested feature, as well as providing means by which automated functional tests could be applied within the development process.*

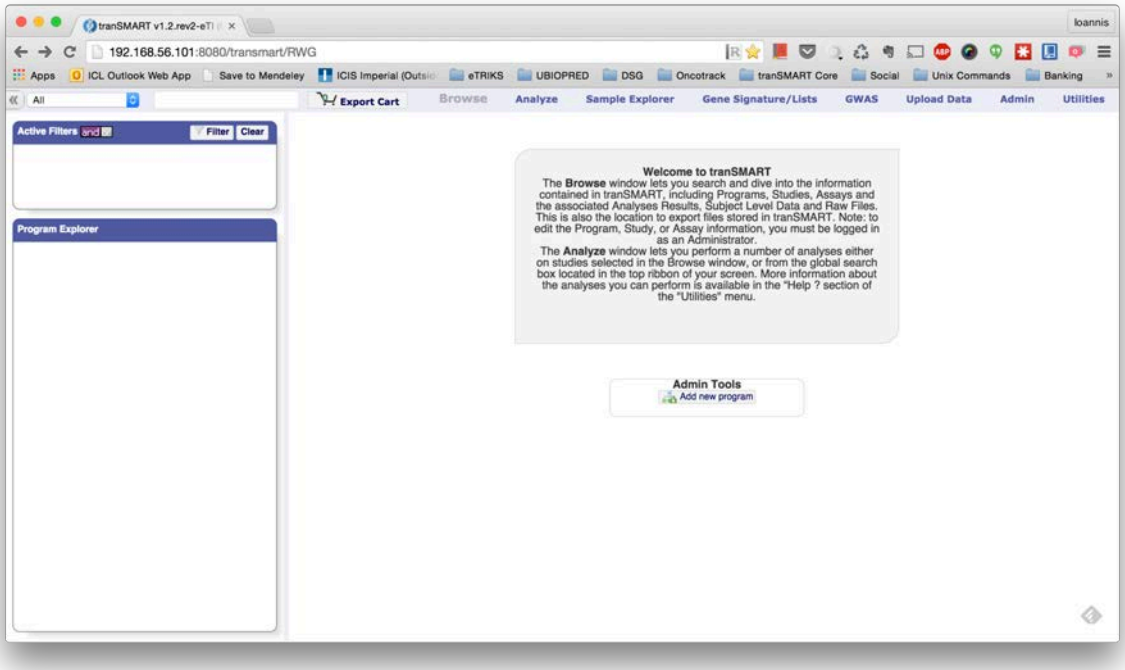
Furthermore, it was recognised that the 18 Consolidated requests, which had been prioritised for development, comprised of 39 sub-features.

Met requirements

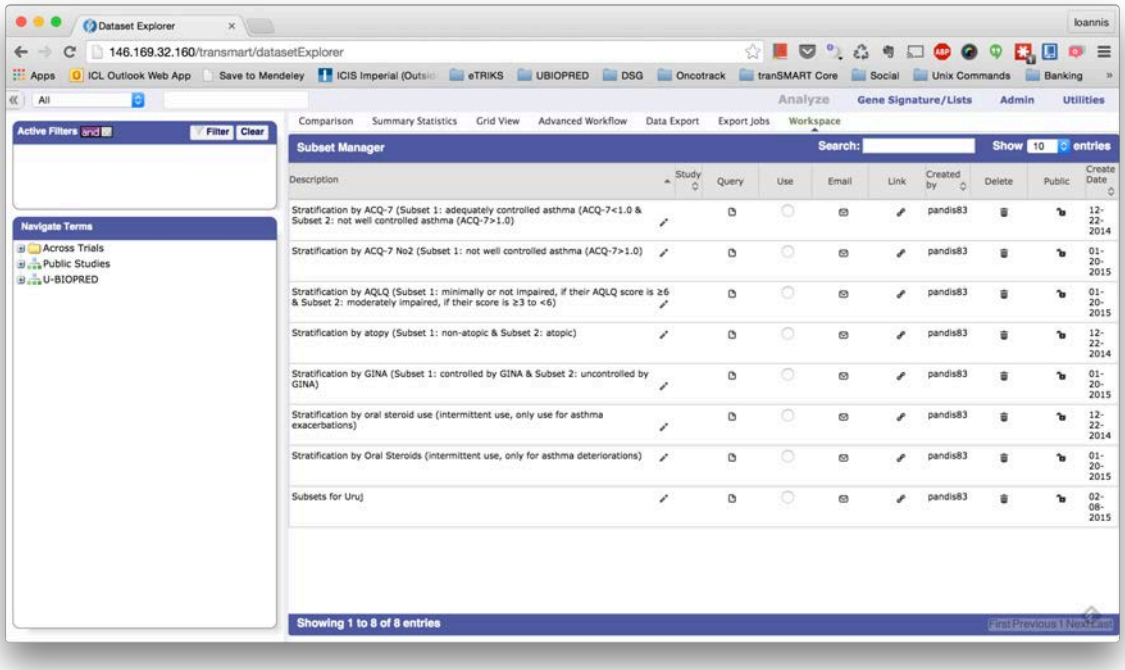
The review process identified 5/18 consolidated requests as being met, namely:

1. Project Workspace - (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq050)

Delivered by the development of Browse panel in tranSMART 1.2 / eTRIKS 2.0 (depicted below), allowing projects to store the metadata, protocols, raw data and other associated documents:



Furthermore, the requirement was further enhanced by the development of the Workspace feature, within the Analyze tab, allowing the storage and management of user queries:

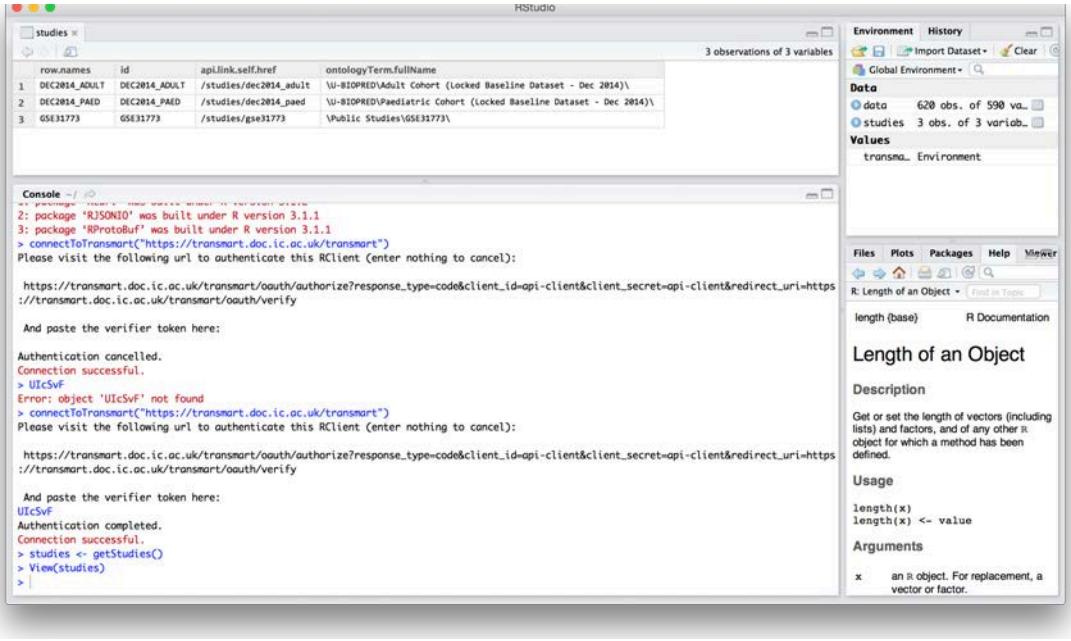


2. Gene Signature/List (Automated Gene mapping and annotation) – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq042)

Delivered by the addition of several content dictionaries allowing automated mapping, covering not only genes (NCBI), but also microRNA (miRbase), proteins (Uniprot) and metabolites.

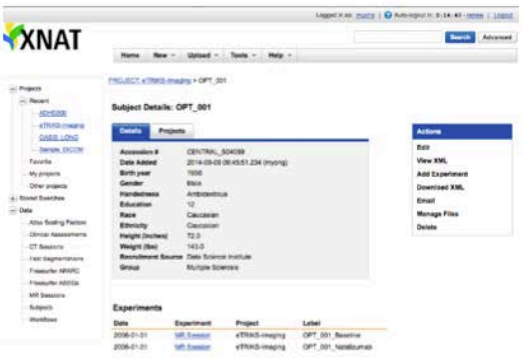
3. Robust R interface – allowing full querying of all data elements hosted in traSMART instances (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq047)

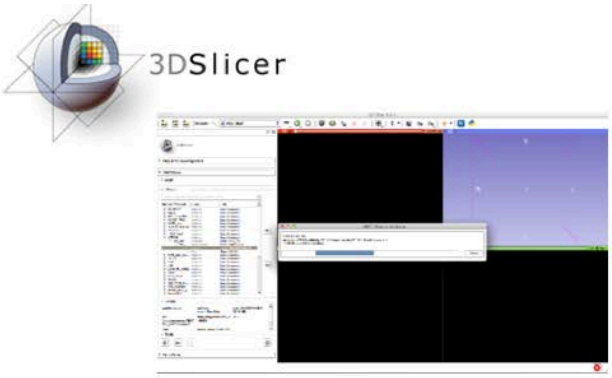
Delivered by the development of the Rclient (https://wiki.transmartfoundation.org/display/TSMTGPL/R+Client).

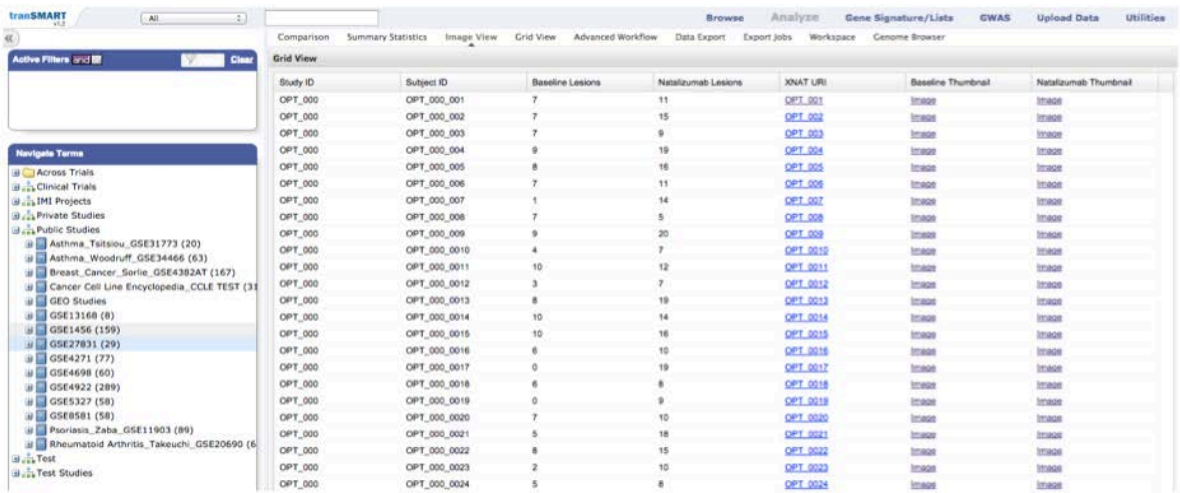


4. Imaging Support – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq048)

Delivered by the development of the XNAT connector:







5. eTRIKS Security – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq029)

Delivered by the core development and LDAP integration.

Partially met requirements

The review process identified 7/18 consolidated requests as being partially met, namely:

1. Support for longitudinal studies - (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq037)
2. Reproducible Research Datasets (versioning / provenance) – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq046)
3. Automated Hypothesis Generation – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq028)
4. Multiple Cohorts – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq036)
5. eTRIKS Export – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq030)
6. Features identified as foundational to support Oncotrack – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq038)
7. Performance (especially ETL) - (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq044)

Details of the individual progress status of each sub-feature components of the above requests is summarised in the following section.

Requirements not met

The review process identified 6/18 consolidated requests as being not yet met, namely:

1. Automated Data Checking post ETL - (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq033)
2. Fail safe data loading / error handling – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq035)
3. Statistical Test Selection – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq034)
4. Easy box plot creation/manipulation – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq043)
5. Date isn't a date – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq045)
6. Consolidated environment for TM/eTRIKS – (https://requirements.etriks.org/twiki/bin/view/RequestManagement/FtRq049)

Details of the individual progress status of each sub-feature component of the above requests:

Title	***Status***
Automated Data Checking post ETL	*Not met*
Multiple Cohorts	*In progress*
To provide the functionality to enable users of TranSMART to compose, execute and review the results for simultaneous comparisons of more that two cohorts.	
To generate and display appropriate multiple cohort statistical results	
To deliver improved query building facilities to streamline the creation of multiple cohort queries.	
eTRIKS Export	
1. Export the results queried by the Data Set Explorer cohort comparison and Grid interface.	*Met*
1.a. Labelling exported attributes in a manor corresponding to distinct field names in the study hierarchy	*Met*
2. Export the tabular results of the Search and Faceted Search	*Not Met*
3. Transfer of a complete study from one eTRIKS instance to another via export and import routines.	*Not Met*
4. Export to file formats associated with key accessory analytical applications.	*Not Met*
eTRIKS Security	
1. Granular restriction of data access (across all query interfaces) and eTRIKS operations	*Met*
2. Methods to facilitate assigning of access privileges including a robust administrative utility and the concepts of groups and roles	*Met*
3. Option to use of external access methodologies	*Met*
4. Account tracking and reporting	*Met*
Support for longitudinal studies	
Load longitudinal data	*Met*
Incremental Data load	*Not Met*
Display	*Partially Met*
Reproducible Research Datasets (versioning / provenance)	*Partially Met*
Fail safe data loading / error handling	*Not Met*
Statistical Test Selection	*Not Met*
Project Workspace	*Met*
Automated Hypothesis Generation	*In progress*

Gene Signature/List (Automated Gene mapping and annotation)	*Met*
Performance (especially ETL)	*Partially Met*
Easy box plot creation/manipulation	*Not Met*
Date isn't a date	*Not Met*
Robust R interface	*Met*
Imaging Support	*Met*
consolidated environment for TM/eTRIKS	*Not Met*
Features identified as foundational to support OncoTrack	
Access from within the clinical hierarchy tree to the concept 'parent test subject'. This information is required as a means to cascade down selections across studies between parent and child test subjects	*Partially Met*
The grid view, as that place where all data relevant to a selection and investigation are available in table view. The grid view will likewise provide critical hooks into further downstream selections	*Partially Met*
Flexibly copy/paste (parts of) the grid view into external applications	*Met*
define and save selection lists, for instantaneous reuse as a cascaded selection clue in other isolated open tranSMART sessions	*Partially Met*
other 1xn relationships exist within the OncoTrack data model, which are crucial and we cannot work our way around	*Not Met*
A user must be able to select a raw or intermediate data file of interest via the grid view and to trigger the opening of a separate web page in OncoTrack DB that contains the file of interest	*Not Met*
include as an integral part of the cascaded querying process a broad range of summarized multidimensional data (directly uploaded from within OncoTrack DB)	*Not Met*
Reuse at different points in time of previously made selections as well as related datasets and analyses *- Query/explore/analyse components, savable into tranSMART for dynamic reuse* *- Parametrizable?* *- Saving static results outcome data alongside tranSMART data upload status*	*Partially Met*
Flexible access to analysis functionality based on a consistent interfacing framework	*Partially Met*
A user must be able to select one (or multiple) raw or intermediate data file(s) of interest via the grid view for interactive viewing/analysis from his PC	*Partially Met*
In order to support further analysis based on native applications, the user must be able to make a selection of both the phenotypical data and the multidimensional data he wants to export and trigger an off-line job to handle that export	*Met*
Integrate an automated reference updating process into tranSMART to ensure summarized multidimensional data remain up to date without requiring the regular re-upload of such data into tranSMART	*Not Met*
Intuitive signal to differentiate dummy-substudies vs effectively fully different studies	

Linkage of interrelated dummy-substudies in such a way that users can deal with cross-dummy substudy selection and analysis transparently	
Handling large complex genetic (structural) variants (translocation, gene fusions)	*Not Met*

Product features

New and existing requirements have been reviewed with each supported project through the Account Manager and coordinated at the eTRIKS Annual Meeting the subsequent requirement-gathering phase it was noted that all requests overlapped with the existing requests. Hence, for the reporting period up until the v3.0 release of the eTRIKS KM Platform, the development team shall work on the existing, aforementioned requests marked as partially met or unmet.

Operating environment

The operating environment in which the eTRIKS KM Platform v3.0 will be deployed may be highly variable and dependent on the supported IMI project's needs. In particular three modes of operation are anticipated:

- Self-hosted by an organisation or project either on servers or in a public cloud-based infrastructure (such as Amazon EC2).
- Self-hosted by an organisation or project on private infrastructure, for example in an IP-restricted intranet.
- Hosting provided by an eTRIKS KM Platform service provider.

As such, the platform will be designed for both stand-alone use and multi-tenancy.

Assumptions and dependencies

As described above, the main dependency for further development of the eTRIKS KM Platform is the reliance on tranSMART as a component of the software platform. This has been planned for from the start of the project.

Several assumptions have been made including the following:

- The eTRIKS KM Platform will be a web-interfaced product that may be configured for use via the Internet or within an intranet infrastructure.
- Hosting the eTRIKS KM Platform software will be possible in multiple modes, as required by the operating environment.
- The software development team in WP2 will determine the implementation strategy, including design decisions and technology dependencies beyond tranSMART.