European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

Deliverable 2.7

Requirements document for eTRIKS KM Platform v4.0

Due date of deliverable: Month 48

Actual submission date: Month 50

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:	D2.7
Deliverable title:	Requirements document for eTRIKS KM Platform v4.0
Deliverable version:	1
Due date of deliverable:	September 2016
Actual submission date:	November 2016
Leader:	Denny Verbeeck
Editors:	Denny Verbeeck
Authors:	Denny Verbeeck, Peter Rice
Reviewers:	Chris Marshall, Jay Bergeron, Francisco Bonachela Capdevila
Participating beneficiaries:	ICL, CNRS
Work Package no.:	WP2
Work Package title:	Platform Development
Work Package leader:	ICL, Pfizer
Work Package participants:	
Estimated person-months for deliverable:	2
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 feature requirements for eTRIKS KM Platform v3.0 (D2.6), as well as describe the feature requirements for eTRIKS KM Platform v4.0. Document analysis was performed on existing requirements documentation in order to identify unmet requirements. 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 v4.0 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 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.

Requirements analysis

The following section will go list the requirements that were met between the release of v2.0 and v3.0. After that an overview of the unmet requirements will be presented. These form the basis of the feature requirements for v4.0.

The Twiki platform for requirements gathering has been discontinued due to lack of support by the developers behind the platform. We have archived existing documentation from the wiki at https://biosciconsulting1.teamwork.com/#/projects/74582/files?catid=333306. The requirements that originate from this archive are annotated in this document with (FtRqXXX), where XXX is the number of the requirement. Interested readers are referred to the archive, where PDF files following this naming scheme will be found.

Met requirements

This is a list of requirements met for the release of eTRIKS v3.0. They originate from existing documents and feature requests received over the period between the two releases. In the following table, the left column denotes the added functionality, implemented feature or fixed bug, and the right column lists the requirements that were met by this work.

Browse tab Browse tab functionality has been restored, leaving researchers the opportunity to add any and all relevant metadata to studies loaded into TranSMART. This metadata includes a set of predefined tags, as well as user-defined tags and even full files.	1. Reproducible Research Datasets (FtRq046) 2. Search and data mining (FtRq002) 3. Improved patient centric views (FtRq012) 4. Improve consistency and synchronization of data trees in 'Browse' (Program Explorer panel) and in 'Analyze' (Navigate Terms panel): a) When clicking on the "Open in Analyze view" button from Browse, display a filter on the study ID in the Active Filter panel and restrict the data trees in both tabs to that study. Also highlight and open that study in the data trees of both tabs. b) There should be an easier way to navigate from the Analyze tab back to the Browse tab. c) Data trees and right panels should be kept as they were in a tab when switching to the other tab (unless a new filter is applied, in which case both data trees are refreshed). 5. Ability to relate analysis results in Browse tab to subject level data in Analyze tab 6. Visualization of study metadata improvement is needed (FtRq016) 7. Implement security rules and user permissions in Browse tab, which are consistent with the current security rules and permissions in Analyze tab.
SmartR With the addition of SmartR, an intuitive and interactive way of visualizing the data is now realized.	8. Easy box plot creation/manipulation (FtRq043) 9. Improved plots in transmart (FtRq014) 10. D3.js in TM (FtRq021)

Public server The eTRIKS public server is up and running, serving over 70 public datasets, using the latest eTRIKS version of TranSMART. Available at https://public.etriks.org/.	11. Consolidated environment for TM/eTRIKS (FtRq049) 12. Automated scripts to upgrade database schema between versions. They were tested on the public server.
Incremental data loading New routines have been developed that can add new data to an existing study.	13. Add ability to load data incrementally for a given study and to delete or overwrite some previously loaded data (for certain subjects only, or certain variables only) without having to reload the study entirely. (FtRq056) 14. Allow loading of 'serial' high and low dimensional data (time course, dose response, different sampling conditions, etc.). 15. Improve subcategorization of high dimensional data (tissue, timepoints, etc.) in the high dimensional data node selection screen in Advanced Workflows
Additional data types A range of new high dimensional data types have been implemented, together with loading and export routines. Currently TranSMART supports the following high dimensional data types: aCGH, Metabolomics, miRNA, mRNA, Proteomics, RBM, RNASeq, RNASeqCog and VCF (mutations).	16. Presentation of new data types is needed in transmart (FtRq013) 17. Add ability to load and analyse RBM subject-level data as high dimensional data. Manage protein/peptide identifiers and ensure unit display. 18. Add ability to load and analyse microarray miRNA subject-level data, as high dimensional data. 19. Add ability to load and analyse qPCR mRNA and miRNA subject-level data 20. Enable metabolomic subject-level data loading and analysis, as high dimensional data. 21. Improve SNP subject-level data handling: a) Develop more automated procedures for processing and loading large sets of SNP data and thus accelerate loading time. b) Make sure the infrastructure is well-dimensioned to support large SNP data volumes. 22. Add ability to load and analyse RNA sequencing subject-level data (gene-level expression quantification) 23. Optimize the management of annotation files (relating probes to genes) for omic data: same information should be used for gene expression data and gene expression analyses, and gene lists. 24. Add dictionaries for miRNA, proteins, metabolites

	25. Enable subject-level RNA-seq data loading and analysis (transcript-level) 26. Allow loading and analysis of mass spectrometry subject-level data.
LDAP Plugin It is now possible to configure the login mechanism to use LDAP as an alternative to the built-in TranSMART user list	27. Set up user authentication through the company's Active Directory
R API An R package to allow interfacing to the TranSMART data warehouse was developed 'connectToTransmart'. This allows access from R scripts to the data, where the complete R bio-informatics ecosystem can be leveraged.	28. Allow analysis of RBM data using existing analytics for high dimensional data 29. Allow analysis of microarray miRNA data using existing analytics for high dimensional data 30. Allow analysis of qPCR miRNA and mRNA data using existing analytics for high dimensional data 31. Allow analysis of metabolomic subject-level data using existing analytics for high dimensional data 32. Allow analysis of RNA sequencing data using existing analytics for high dimensional data
Improved advanced workflows Several improvements were made to the advanced workflows to meet user requirements.	33. For the Boxplot analytics, make individual box plots for each variable when dragging multiple nodes in field 'Dependent Variable', and present output in table format 34. Ability to run jobs in the background (marker selection could take a while) 35. Ability to view list of genes included in a pathway or gene list when using high dimensional data in advanced analyses 36. Ability to combine gene lists when using high dimensional data in advanced analyses
Improved UI in Comparison and Summary statistics tabs These two tabs have been redesigned and rewritten to produce SVG plots and	37. 'And'/'or' filters are too small – some people couldn't read them. 38. Allow removal of nodes one by one in boxes of cohort selection and variable selection in advanced workflows
Password management Users can now change their own password	39. Non-LDAP users can change their password.
Data loading A new data loading system transmart-batch is available for testing. This avoids the use of Pentaho Kettle scripts. The ICE tool has been contributed by Sanofi to	40. User friendly design of study tree and data loading 41. Fail safe data loading/error handling (FtRq035)

manage data loading ETL and to ensure Browse study metadata is loaded. ICE tool extensions are planned for eTRIKS v4.	
Admin panel improvements New admin report to check status of associated services – solr for metadata queries, R for analysis workflows, and GWAVA for the Pfizer GWAS functionality.	42. Better status checking capabilities for admins
Data export Data export is now available in two formats. The user can export the contents of the Grid View tab to an excel file. The user can export all clinical and high dimensional data as an archive.	43. Export feature (FtRq020)
Statistical test selection In summary statistics, an appropriate statistical test is automatically selected based on the type of data being compared.	44. Statistical Test Selection (FtRq034)
Date conversions Exact dates can be converted to 'time since' or other features that are independent of an exact date. Storing exact dates for visits or other events can cause issues with respect to privacy.	45. Date isn't a date (FtRq045)
Load user-defined gene lists It is now possible to load user-defined gene signatures through the web interface. This should be used when a user wishes to repeat advanced workflows with the same gene sets.	46. Keep last entries in High Dimensional Data screen of Advanced Workflows

Partially met requirements

Based on the work listed above, we can identify the following requirements as partially met. A requirement is defined as partially met if either one or more sub requirements are not met, a feature is not yet fully tested and ready for production, or a feature is not yet complete.

1. eTRIKS export	
1.1 Export the results queried by the Data Set Explorer cohort comparison and Grid interface.	**Met**
1.2 Labelling exported attributes in a manor corresponding to distinct field names in the study hierarchy	**Met**
1.3 Export the tabular results of the Search and Faceted Search	**Not met**
1.4 Transfer of a complete study from one eTRIKS instance to another via export and import routines. (FtRq055)	**Not met**. Will be covered by eHS in the future.
1.5 Export to file formats associated with key accessory analytical applications.	**Met**, export to R is possible. Recent developments by TheHyve have facilitated exporting to Python as well. From these platforms it should be straightforward to export to any desired file format.
2. Improve the Correlation Analysis analytics	
2.1 Combine Correlation Analysis and Scatter Plot Linear Regression into one workflow	**Met.** The Correlation Analysis advanced workflow also shows a linear regression between the selected variables
2.2 Allow Correlation Analytics to run with high dimensional data	**Not met**
2.3 Allow correlation of one variable against many	**Met.** An arbitrary amount of variables can be included in the correlation analysis
2.4 Improve the table output when many variables	
3. Improve Grid View	
3.1 Enable categorical variables in a single column	**Met**
3.2 Enable column deletion, row or column selection	**Met**
3.3 Enable export of selection	**Partially met.** This is now available in the Data Export tab
3.4 Automatically include variables used in Advanced Workflows	**Not met**
4. **Machine Learning Methods**	

4.1 Include machine learning methods (FtRq008)	**Partially met.** A Shiny app exists that performs SNF on transmart data.
5. User session management	
5.1 Keep track of the search/analysis history and the results over the time (FtRq001)	**Partially met.** Audit trails are being recorded, however they cannot yet be viewed by the user or used as a history.
6. Cross study analysis	
6.1 Enable cross-study analysis between different studies coming various IMI projects. It is very important to use some standard terminology when loading the data obtaining from above mentioned projects, so that it will allow the user to perform cross study analysis based on for example disease, age, sex, demographic location, certain population etc. It should also equip with necessary statistical methods to normalize the user selected cross study data and perform necessary analysis (FtRq007)	**Partially met.** Cross study analysis pipelines are being developed. eHS can also aid in standardizing clinical variables in multiple studies.
7. Improve the Line Graph analytics	
7.1 Enable Line Graph to use high dimensional data	**Not met.**
7.2 Better handle x axis variable (time, numerical or categorical), i.e. have a scaled x axis when time or numerical variable is used. Should work even for time course data without date/time	**Met.** This has become available with the addition of SmartR
7.3 Add option to plot individual data in addition to group means or medians	**Not met.** This is already available in HiDome for high-dimensional data, and will be made available for all numerical data.
8. Audit logging	
8.1 Keep a log of all user-generated events. This includes a log of which nodes have been opened, which analyses have been performed, which summary statistics have been run etc.	**Partially Met.** Audit logging is turned on by a configuration setting. We are testing this on the test server to see what information is available for a user logged in or for a guest user. There are some gaps in the coverage – for example advanced workflows. We can look to extend the coverage and look for additional log information for 4.0.
9. GWAS Plugin	
9.1 Improved GWAS functionality with code fixes from Pfizer for Oracle and postgres – to be further updated in eTRIKS v4.	**Partially met.** GWAS functionality on Postgres is greatly improved. The feature to upload files through the web interface was removed, so that GWAS analysis will be loaded by a database administrator using the same privileges as loading study data.

Unmet requirements

The following requirements are not yet met. These requirements are listed with their priority and development efforts for eTRIKS v4 will be assigned accordingly.

10. Save feature. Save the analysis pipelines including the analysis and visualization parameters as well as the analysis results in the user space. (FtRq018)
11. Allow better analysis of 'serial' high and low dimensional data using existing analytics:
a. Ability to analyze a complete 'serial' data matrix (e.g., gene expression at all time points using Heatmap),
b. Ability to analyze individual columns (e.g., plot gene expression of a gene at one time point against gene expression at another time point using Scatter Plot),
c. Ability to use the 'serial' dimension as a variable (e.g., plot gene expression of a gene in function of time using Line Graph).
12. Display sample ID related to patient ID in Grid View.
13. Build cohort by filtering high-dimensional data

New requirements

eTRIKS is in close contact with all of its supported projects. For eTRIKS Knowledge Management Platform v4 no additional requirements have been brought forward. WP2 will focus on the outstanding partially met and unmet requirements, with the emphasis on reliability and scalability of the platform. Given the remaining time in the project, we believe this is the path with the most realistic goals to take.