European Translational Information and Knowledge Management Services

eTRIKS Deliverable Report

Grant agreement no. 115446

Deliverable 2.3

Requirements document for eTRIKS KM Platform v2.0

Due date of deliverable: Month 9

Actual submission date: Month 12

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:	D2.3
Deliverable title:	Requirements document for eTRIKS KM Platform v2.0
Deliverable version:	1.0
Due date of deliverable:	30/09/2013
Actual submission date:	30/09/2013
Leader:	David Johnson
Editors:	David Johnson
Authors:	David Johnson, Peter Rice, Ibrahim Emam
Reviewers:	Chris Marshall, Leila El Hadjam
Participating beneficiaries:	
Work Package no.:	WP2
Work Package title:	Platform Development
Work Package leader:	ICL
Work Package participants:	
Estimated person-months for deliverable:	
Nature:	PU
Version:	0.7
Draft/Final:	Draft
No of pages (including cover):	
Keywords:	

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 sizes, 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 packagesusing the process described indeliverable 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 the documented set of initial feature requirements for eTRIKS KM Platform v2.0, focussed on added functionality to v1.0 capabilities. Stakeholders were solicited for feature requirements, some of which were directly input into the eTRIKS Product Management (User requirements) Wiki: http://goo.gl/9iBngU and generated from an eTRIKS 'Requirements Hackathon' meeting that was held at Imperial College London.

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 v2.0 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 for eTRIKS to provide an open and collaborative model for scientific knowledge to flourish; and for new approaches for the prevention, diagnosis, and treatment of disease to evolve, ultimately redefining the way biomedical research is translated to better health.

eTRIKS is not going to provide one solution for all, but a commons infrastructure that enables 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 need 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 analysis results of cumulative translational research investigations, where outcomes of basic and clinical research are continually integrated under a common systems-biology context.
2. Study-centric storage for scientific research data providing evidence for 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 create 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 to navigate and visualize translational research knowledge space of biological entities to enhance and support new discoveries and decision-making.

Currently eTRIKS KM Platform v1.0 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 v2.0.

Product features

For the reporting period up until the v2.0 release of the eTRIKS KM Platform (month 18) the following requirements were consolidated and determined of highest priority at this point in time:

- Enhanced data lifecycle management.
- Multiple data type support, including longitudinal studies.
- Semantics and standards adoption.
- Improved automation in ETL.
- Oracle compatibility.
- Project workspaces.
- Other non-functional enhancements (security and performance).

Operating environment

The operating environment in which the eTRIKS KM Platform v2.0 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.

Design and implementation constraints

A key design constraint as set out in the original eTRIKS project proposal and Description of Work is that the eTRIKS KM Platform developed is to use the tranSMART data warehouse software entirely or in part. This decision was based on partner pharmaceutical and IMI project stakeholders already using tranSMART for study evidence storage and analysis. As such, eTRIKS KM Platform v1.0 comprises of 'tranSMART for eTRIKS' – the completed open-source port from Oracle to PostgreSQL, released publicly by the tranSMART Foundation as 'tranSMART 1.1'. By using tranSMART as a core component, the main constricting factors are in management and further development of tranSMART itself, where it is based on the Grails web framework, written in the Groovy and Java programming languages. In its current state, its integration into a wider software platform may require some effort due to tranSMART not yet providing a web API (neither RESTful nor SOAP-based). However, by using tranSMART as a starting point the project has saved significant effort in implementing a study evidence database and ensuring pharmaceutical partner buy-ins by planning to extend a system that is already moderately accepted by the translational research community in IMI.

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.

System features

In this section we provide expanded descriptions of each of the feature requirements listed previously. Note that the bases of theserequirements are taken from the eTRIKS Product Management (User Requirements) Wiki. They are not intended to be a comprehensive list of feature requirements, but rather a set of requirements **beyond**the current eTRIKS KM Platform v1.0 capabilities (i.e. beyond tranSMART for eTRIKS capabilities).

Enhanced data management lifecycle

Description

Current capabilities in eTRIKS for data management are somewhat restrictive. There are yet to be any universal guidelines or technological constraints on the application of data standards, data loading uses legacy and bespoke ETL procedures, and data export is limited to only R scripts and comma-separated value text files. There is an urgent need for an enhanced data management lifecycle beyond that afforded by tranSMART for eTRIKS.

Simple improvements have high benefits so initial workarounds would be a useful delivery.

The key enhancements will be to:

- Adopt, or develop where necessary, industry standard data formats for clinical and –omics data types.
- Allow data curators to publish and manage custom annotation files alongside datasets where bespoke study curation is carried out.
- Develop data loading assisted by standardized formats and semantic annotation.
- Develop data export to additional formats including (but not limited to):
    - Microsoft Excel
    - R inputs
    - S+ inputs
    - SAS inputs
- Enable database migration between eTRIKS KM Platform sites (i.e. between instances of tranSMART for eTRIKS).
- Enable efficient and reliable data loading to a live eTRIKS KM Platform site for very large datasets without interruption to services. Data should be processed in smaller steps so that errors are reported promptly. Ideally, processing should resume from the last successful step.
- Validation of correct data loading should include testing against standards for eTRIKS or for a specific project to enable cross-study query and analysis.

Priority

High.

Projects/Contacts

eTRIKS: Ioannis Pandis
eTRIKS WP4: Serge Eifes

Multiple data type support

Description

For eTRIKS supported projects, each of the features for data management lifecycle as described in the preceding section. While it is already a given that a broad range of clinical and –omics data should be supported based on the requirements of specific supported projects to data (i.e. U-BIOPRED, OncoTrack, RA-Map, ABIRISK, PREDICT-TB have been solicited for data type requirements. The current SmartSheet covers 13 datatypes and 12 projects.

For eTRIKS KM Platform v2.0 the following data types should be supported by a full data management lifecycle:

- RNA sequence
- DNA sequence
- Flow cytometry
- Metabolomics
- Small RNA
- Lipidomics
- Next Generation Sequencing (NGS)
- Variant data (VCF files, SNPs and structural variants)
- Clinical data
- Animal data
- In vitro cell cultures
- Public gene expression data
- Genome Wide Association Studies (GWAS)
- Gene expression arrays
- Methylation

There are further requests for user-uploaded data files to be included in searches. This functionality is included in current developments by Sanofi.

Priority
High.

Projects/Contacts
OncoTrack: Emmanuel van der Stuyft
RAMAP: Chris Marshall
eTRIKS WP4: Serge Eifes
Sanofi

Longitudinal data

Description
Research clinicians need to be able to store and interrogate the data from longitudinal studies while they collect data so that they can investigate time-dependant trends in the data as the trends develop. Prospective studies often require data to be collected on patients at multiple visits. For prompt analyses it is undesirable that investigators should have to wait until all data is collected (a process that may take several years) before investigating their data. To this end the abilities to load, manipulate and investigate parts of a data set as they are collected are required while also being able to add further data to the set. For governance purposes, there is also a need to know what data was in the set at any given time. Without prejudice to possible solutions and for the convenience of describing the required features, it is assumed that some sort of reloading of data sets at agreed time points or when new data becomes available will be involved.

While this document lists all requested data type support, those to be developed within the next development period up to the eTRIKS KM Platform v2.0 release will be prioritised where possible, with longitudinal data support as the highest priority at this time.Requirements to handle multiple samples per patient and animal xenograft models have overlapping needs to extend data models and are therefore included in this topic.

Priority
High.

Projects/Contacts
OncoTrack: Emmanuel van der Stuyft
eTRIKS WP4: Serge Eifes

Semantic and syntactic support

Description
A range of feature request has been made that all fall under the theme of semantic and syntactic support. In particular the following specific requests have been made:

- Automatic checking of the entire dataset being loaded and after loading has ended (i.e. raw data vs.database data) to ensure fidelity and integrity.
- Automatic hypothesis generation, in particular:
    - Allowing a user to select cohorts and compare individual study attributes, where each test is a hypothesis.
    - It has been noted that this process could be performed automatically with significant differences noted as output.
- Versioning/provenance of data and analyses:
    - For studies that may be updated it is imperative for research provenance that we are able to identify exactly what data was in the set at any time in its history and which version of analytics plugins/scripts were used for any analyses.
- Intelligent automated matching of data types to the appropriate analytics tools.
- Searching across studies and projects

Each of these requests can be satisfied wholly or in part through the adoption of standard data formats, standardised terminology, and automated services based on these standardisations. As a target for the eTRIKS KM Platform v2.0 release the following will be developed:

- Standard formats and annotation services based on ISA-tab format and CDISC terminology in the first instance.
- Data loading tools utilising aforementioned standard formats and annotations.
- Study metadata and data visualisation.

Priority
High.

Projects/Contacts
OncoTrack: Emmanuel van der Stuyft
eTRIKS WP4: Serge Eifes
Merck: Fabien Richard

Data Export and Views

Description

Export of data from eTRIKS enables workaround solutions by passing content to external analysis applications and services. As a working principle, export of data in the format originally used for ETL demonstrates that data content has not been lost by loading into eTRIKS. Export of data in raw form (possibly through a link to the original data used for ETL) would be a good way to exploit study provenance and metadata information.

Views of data should be in table form, including values used in querying the content, and selected additional attributes. Data in such a table (e.g. grid view in tranSMART) should be exportable in full or as a selection.

Variation data can be displayed in an external genome browser. These support "tracks" using data exported in a choice of simple data formats with positional information and annotation.

Extended views are required for longitudinal data and to handle samples and animal models.

Priority

High.

Projects/Contacts

OncoTrack: Emmanuel van der Stuyft
RAMAP: Chris Marshall

Analysis and Advanced Workflows

Description

Extensions to the basic analysis functionality in the current eTRIKS release (using tranSMART "advanced workflows") would enable users to make far more effective use of the data in eTRIKS. The development of multiple cohort selections requires modifications to analyse all cohorts, pairs of cohorts, etc. Automated selection of cohorts can be linked to launching of analysis to support automated hypothesis generation. Semantic information about data can be used to suggest statistical tests or data normalisation.

The interface to advanced workflows can be improved, with better management of the parameters in the code (rather than in a database table in the current release). Analysis can be extended to launch user-contributed R scripts or third party applications.

Priority

Medium.

Projects/Contacts

RAMAP: Chris Marshall
Pfizer: Jay Bergeron
Imperial: Ioannis Pandis
eTRIKS WP4: Serge Eifes
OncoTrack: Emmanuel van der Stuyft

External interface requirements

User interface

For eTRIKS KM Platform v2.0, a redesign of the user interface will be undertaken to introduce project workspaces that will allow project teams to share their data, their analysis workflows, their result sets, and new knowledge findings. Building a collaborative environment that spans all components of the infrastructure is essential.

We envision applying the metaphor of a book as the basis of a collaboration environment for the evidence and provenance repository. It is important to note that the study book will encompass both experimental evidence data and research knowledge. At an early stage, each project supported will be provided with a "library" in which they can organise their study books.

A project might constitute a number of related studies that the researcher might want to manage in the same context. The eTRIKS product UI requires a place where the researchers can browse the studies available to them within a particular context of the project. The "library" aims to serve that purpose. It is a place where each study, represented by "study and search functions are provided for fast access to content that matters in a given a particular context.

Acting as the central concept, the "study book" provides a space where it is easy to manage the data, keep track of the study datasets. The study book will offer methods to create dynamic cohorts given a particular dataset, keeping track of all the changes made. This helps to cope with the problem of cohort redefinition in the case of patient panel narrowness. In addition to this it will be possible to create and organise "investigations" to support cross-study work.

An analytics dashboard provides workspaces within which analysis workflows are created that could be shared between users. For example in myExperiment, a scientific workflow sharing social network site developed by the myGrid consortium, workflows can be packaged with data, related scientific publications, and descriptions of the interrelationships between packages. These packages are termed "Research Objects". Sharing workflows in eTRIKS could take a similar approach to myExperiment to enable validation and reproduction of translational research analyses within the eTRIKS product platform.

The collaborative environment will also provide an interface to a common knowledge base where end-users "submit" their research findings as well as "contribute" and "comment" on new or existing knowledge statements, as well as being able to "aggregate" a set of knowledge statements to reflect a complex discovery resulting from different resource and research findings statements. Users could then choose to publish these results or share amongst their community of users.

Priority
Medium.

Projects/Contacts
RAMAP: Chris Marshall
OncoTrack: Emmanuel van der Stuyft
eTRIKS WP4: Serge Eifes

Software interfaces

The eTRIKS KM Platform analytics support will be built on an extensible design that makes it easy to plug in and use third-party tools and services. A framework consisting of a set of web APIs will be provided to enable developers to build services to analyse project data stored in the eTRIKS KM Platform and contribute them to the commons where product developers can install them into the product platform as plugins. Developers will also be able to make use of the eTRIKS KM Platform through programmable APIs to create new plugins (applications/scripts) that are contributed to the commons as new plugins or services. End-users will be able to access plugin/service features via an interactive GUI dashboard that provides a user interface to access the product platform and construct workflows.

The APIs will provide plugin developers targeting scripting languages such as, Python and JavaScript, statistical environments such as R as well as compiled languages such as Java. Users will be able to use these APIs to integrate with existing as well as create new analysis and visualization scripts and integrate them with third party applications. A JavaScript API may enable the integration with a wide range of data visualization tools and libraries. In addition to the APIs, the platform will host a range of common analysis and visualization tools based on R, such as cohort analysis, differential gene expression analysis and data exploratory visualization tools.

Based on the requirements of the IMI project community, eTRIKS will maintain a list of desired analytics applications/tools that could be plugged into the eTRIKS KM Platform. Their development as plugins to eTRIKS will occur partly developed by the eTRIKS project and the open community will be encouraged to utilise the development APIs and submit their own developed plugins to the commons.

Priority
Medium.

Projects/Contacts
RAMAP: Chris Marshall
OncoTrack: Emmanuel van der Stuyft

Other non-functional requirements

Oracle compatibility

Description

Currently tranSMART for eTRIKS (i.e. tranSMART 1.1) is only supported using the open-source PostgreSQL database backend. Previous versions of tranSMART that are used widely in the pharmaceutical industry use Oracle as the database backend, and as such adoption of tranSMART for eTRIKS, and the eTRIKS KM Platform in general, requires Oracle to be supported. Some effort should be put towards enabling tranSMART 1.1 Oracle support to catch up to its respective PostgreSQL version.

This is also a priority for the tranSMART community to have full functionality on both PostgreSQL and Oracle for their next release. Collaboration with the wider tranSMART developer community will reduce the resources required from eTRIKS.

Priority

Medium.

Projects/Contacts

eTRIKS: Peter Rice
tranSMART Core: Peter Rice
Sanofi

Performance requirements

A number of performance improvements have been requested by eTRIKS partners and currently supported projects. In the current release, performance issues have been encountered at different levels using tranSMART for eTRIKS. In particular the following features require significant performance improvements:

- tranSMART SearchApp ETL data loading
- tranSMART SearchApp data retrieval for disease and gene-based queries
- tranSMART Dataset Explorer advanced workflows, specifically:
    - Marker selection
    - Heat map
    - Hierarchical clustering
- tranSMART Dataset Explorer extracting data from database and pivoting.

While these aspects of performance should be improved as they directly impact the usability of tranSMART itself for multiple user groups (i.e. ETL loading impacts curators/data managers, functional performance impacts bioinformaticians/clinicians etc.), scalability and reliability are to be addressed in v3.0 and v4.0 releases respectively as set out in the Description of Work, while the focus of v2.0 will be on feature development. That said, where resourcing allows, these performance requirements that have been noted could be addressed.

Priority

Medium.

Projects/Contacts

eTRIKS Public Server: Ghita Rahal
eTRIKS WP4: Serge Eifes

eTRIKS: Peter Rice

Security requirements

Security features are critical for a stable useable translational research data management solution. The system currently is able to create accounts and challenge for account credentials. Additionally, the system allows privileges to be restricted at a study level. In order to provide a robust level of security for eTRIKS, especially for the proposed multi-use platform, the following enhancements are required.

1. Granular restriction of data access (across all query interfaces) and eTRIKS operations;
2. Methods to facilitate assigning of access privileges including a robust administrative utility and the concepts of groups and roles;
3. Option to use of external access methodologies;
4. Account tracking and reporting.

Priority
High.

Projects/Contacts
Pfizer: Jay Bergeron
tranSMART Core: Peter Rice

Testing and Quality Assurance

Description
The current eTRIKS release required manual testing by the Pfizer testing team. Testing should where possible be automated so that most or all of the use cases are checked routinely during development.

eTRIKS WP2 maintains a server running Bamboo for continuous integration tests on all code changes. Additional tests are needed with testing on multiple platforms, including the latest releases of Postgres and Oracle.

For ETL, tests should validate ETL procedures against the current eTRIKS database schema, and check that appropriate error conditions are detected and reported.

Priority
High.

Projects/Contacts
Pfizer: Jay Bergeron
tranSMART Core: Peter Rice

TranSMART Code Integration

Description
The wider tranSMART developer community has produced a range of extensions which are available under GPL licensing and are free to be integrated in a future tranSMART release. The plan for tranSMART v1.2 is to include the most significant of these code bases. There are several new features implemented in these extensions which potentially address eTRIKS feature requests. Reusing this code, if suitable for eTRIKS requirements, creates a common approach and maintains compatibility between eTRIKS and the core tranSMART code.

Priority
Medium.

Projects/Contacts
eTRIKS: Peter Rice
tranSMART Core: Peter Rice
Sanofi

Glossary

API – Application Programming Interface

CDISC – Clinical Data Interchange Standards Consortium

DNA – Deoxyribonucleic acid

EC2 – Elastic Compute Cloud

ETL – Extract, Transform, Load

eTRIKS – European Translational Information and Knowledge Management Services

GUI – Graphical User Interface

GWAS – Genome Wide Association Study

ICL – Imperial College London

IMI – Innovative Medicines Initiative

ISA – Investigation-Study-Assay

KM – Knowledge Management

NGS – Next Generation Sequencing

REST – Representational State Transfer

RNA – Ribonucleic acid

SearchApp – Search Application

SOAP – Simple Object Application Protocol

UI – User Interface

WP2 – Work Package 2