European Translational Information and Knowledge Management Services

eTRIKS Deliverable Report

Grant agreement no. 115446

Deliverable 2.2

eTRIKS Product Roadmap

Due date of deliverable: Month 12

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:	D2.2
Deliverable title:	eTRIKSProduct Roadmap
Deliverable version:	1.0
Due date of deliverable:	30/09/2013
Actual submission date:	30/09/2013
Leader:	Peter Rice
Editors:	Peter Rice
Authors:	Peter Rice, Ibrahim Emam, Ioannis Pandis
Reviewers:	Chris Marshall
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:	0.1
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 this 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 supplementary documentation for the installation and mirroring of the initial release of the eTRIKS KM Platform v1.0. This platform is built upon the first stable Postgres release of tranSMART v1.1 in September 2013.

The information in this document supplements the documentation provided through the tranSMART Foundation website www.transmartfoundation.org including user and administration manuals and installation instructions.

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.

Operating environment

The operating environment in which the eTRIKS KM Platform v1.0 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 can be installed for both stand-alone use and multi-tenancy.

Development

Feature Requests

The process for gathering features/enhancements is described in D2.1 eTRIKS Product features decision making process. Requests were collected from WP6 Account Managers for supported projects as well as projects asking for future support, and from eTRIKS work packages (especially WP4 for curation and data loading request). eTRIKS partner Sanofi also contributed descriptions of their priorities for internal development efforts on the core tranSMART component.

The 49 requests were collated and assessed in a workshop on 15th-16th July 2013 with estimates for benefit, cost and the number of projects and work packages sharing the requirement. These estimates were reviewed by the eTRIKS development team and presented to the Product Management Panel (PMP) in a teleconference on 17th September 2013. Following a review of the benefits and detailed descriptions, a threshold of 5 was agreed for the selection of requests for the next phase of development.

Title	Owner	Benefit	Cost	Projects	Priority
33. Automated Data Checking post ETL	I. Pandis	5	1	5	25
38. Features identified as foundational to support OncoTrack	E. van der Stuyft	5	4	all	18
30. eTRIKS Export	J. Bergeron	3	3	all	14
31. Custom annotation files	I. Pandis	3	3	all	14
36. Multiple Cohorts	C. Marshall	4	4	all	14
29. eTRIKS Security	J. Bergeron	4	5	all	11
37. Support for longitudinal studies	C. Marshall	5	3	6	10
26. SearchApp filter studies by Treatment	S. Eifes	4	2	4	8
46. Reproducible Research Datasets	C. Marshall	4	5	8	6
35. Fail safe data loading / error handling	I. Pandis	5	4	5	6
34. Statistical Test Selection	I. Pandis	3	3	6	6
50. Project Workspace	C. Marshall	5	5	6	6
28. Automated Hypothesis Generation	J. Bergeron	3	3	5	5
42. Gene Signature/List (Automated Gene mapping and annotation)	F. Richard	3	3	5	5
44. Performance (especially ETL)	S. Eifes	5	4	4	5

Benefit: 1: low 5: high
Costs: 1: 1-3 days; 2: 1-2 weeks; 3: 1+ month; 4: 3+ months; 5: 1+ years
Projects: Count of IMI projects, plus points for overall eTRIKS impact
Priority: Ranking score with threshold of 5 for first round of development

The priority requests were converted to development topics, some covering parts from more than one original request, some combining multiple requests, and presented to the Resource team meeting on 26th September 2013 by Peter Rice (Development Manager) and Ibrahim Emam (Architect) on behalf of the PMP, where the development proposals were agreed.

System functional requirements

Security features
Initial efforts to address a subset of the full request to cover the project priorities for the next development phase. The remainder of the feature requests will be implemented at a later stage.

Longitudinal study
Combining requests for longitudinal studies and part of the OncoTrack request covering support for multiple samples. Both requests require work on the data model to load and query information.

Analysis functionality
Requests for hypothesis generation and statistical test selection were combined with the analysis implications of other requirements and requests from the tranSMART community. Development on this topic is to start later in the initial phase when analysis inputs are available.

Datatype support
Datatype requirements are continuously collected from WP6 project account managers and other stakeholders. The initial requirement was defined by the OncoTrack request.

ETL Procedures
Validation of ETL inputs, a breakdown of ETL procedures into smaller functions that can be restarted, and validation after ETL has completed. Validation will use the controlled vocabularies from the "Semantics" development topic.

Semantics
Creating a framework for semantic markup and validation, to include ETL validation and the definition of common data types for cross-study query and analysis. Largely defined by the OncoTrack request and extended to requirements from other supported projects.

Workspace for users/projects
Initial work to provide a prototype area for users and projects to save, export and share results. The original request covers a large development effort, which will be continues in succeeding phases.

Extended search/views
Extension to the viewing, selection and reporting of data, defined by OncoTrack and extended to other projects and new data types.

Testing/QA
The ongoing effort to provide continuous integration and unit testing combined. The release cycle will include a 4-week testing period involving developers and end users.

Data export
Export of raw data files from processed data or from links to the original data, required for OncoTrack and other projects. Maintenance of provenance information covering database versions and dates.

Multiple cohorts
Selection of more than 2 cohorts, manually or by automated procedures. Reporting and analysis of all selected cohorts

Performance and scalability
Performance issues identified for large data volumes may be addressed by the release of Postgres 9.3 which promises improved support. Initially test to what extent this resolves the issues, then improve performance further in critical areas.

Multi-omics
Comparable metadata and summary data for multiple data types. Requested by OncoTrack and other projects.

Oracle compatibility
The initial eTRIKS release supports Postgres. The tranSMART Foundation is working to provide full functionality when running on Oracle, which continues to be used internally by EFPIA members. ETRIKS will contri8bute to this effort and aim to maintain Oracle compatibility in new developments.

Study metadata
Gene/Signature lists and related collections for the selection of gene expression and other data. Requested by OncoTrack and other projects.

Development priorities

The total effort estimated for all the prioritised and agreed requests exceeds the available resources for WP2 development.

Selection for the initial phase of development took into consideration the urgency of the requirements to support ongoing projects and the complexity of the task.

Security
The initial request provides a comprehensive description of security requirements. In the first phase priority is given to the needs of ongoing individual projects. It is anticipated that other stakeholders may also contribute in this area. The remaining part of the request will be reviewed in the next stage.

Workspace for users/projects
In the first phase development will provide a prototype workspace to guide a further round of requirements gathering. A full solution will require development separate from the core tranSMART component, making use of new interfaces currently under development within the tranSMART community.

ETL and datatype support

Initial support will be for datatypes for which there is an immediate need from supported projects. These will be reviewed and priorities set for further developments in later stages.

Critical factors

Testing and performance/scalability will be an ongoing effort across all development topics.

Unaddressed areas

The roadmap covers development driven by the features requested for eTRIKS. There will be benefits from developments by the tranSMART community which will be reviewed at each stage of development. Where possible eTRIKS platform releases will be at the same time as releases of the core tranSMART component.

Implementation

The Resource team meeting approved development for the first 3 months, with an outline of planned development to complete the above topics and deliver a new release for eTRIKS.

Future Perspective

Towards the end of this phase of development eTRIKS will carry out a further round of feature requests, evaluation and approval. Postponed developments on requests for security, workspaces, ETL and datatype support will be included.

Glossary

ETL – Extract, Transform, Load
eTRIKS – European Translational Information and Knowledge Management Services
GUI – Graphical User Interface
ICL – Imperial College London
IMI – Innovative Medicines Initiative
KM – Knowledge Management
NGS – Next Generation Sequencing
SearchApp – Search Application
UI – User Interface
WP2 – eTRIKS Work Package 2 Development
WP4 – eTRIKS Work Package 4Curation and Analysis