



European Translational Information and Knowledge Management Services

eTRIKS Deliverable Report

Grant agreement no. 115446

Deliverable 2.4

Requirements document for eTRIKS Installer and Mirroring

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:	D2.4
Deliverable title:	Requirements document for eTRIKSInstaller and Mirroring
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, Keith Green, Florian Guitton
Reviewers:	Chris Marshall and 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:	0.1
Draft/Final:	Draft
No of pages (including cover):	
Keywords:	

Table of contents

Introduction	4
Purpose	4
Intended audience	4
Scope	4
Overall description	5
Product perspective	5
Operating environment	5
Demonstration server	6
Public Server	6
Installation	6
Standalone installation.....	6
Prerequisites.....	6
Mirroring.....	8
Virtual Machines	11
Glossary.....	11

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 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 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.

Demonstration server

Public Server

A collection of public study datasets is provided by eTRIKS and available at <http://demo.transmart.etriks.org>

Installation

TranSMART 1.1 can be installed on any Linux system. Installation requires additional packages and PostgreSQL database configuration and therefore should be performed by a system administrator.

Operating systems

The platform has been tested on the following systems:

System	Version tested	Comments
Ubuntu	12.04, 13.04, 14.04	
openSUSE	12.3	
CentOS	6.2	

Prerequisites

The eTRIKS KM Platform v1.0 requires the installation of tranSMART 1.1. This product in turn depends on several additional packages which are required or recommended for full functionality.

The installation instructions below note additional issues that may be encountered on some operating system variants.

The platform has been tested on the following versions.

Package	Version tested	Comments
Java	1.6	
Grails	2.2.4	
Apache	2.2, 2.3	
Tomcat	7.0.42	Also tested with 7.0.35-1
Postgres	9.1, 9.2	Minor change to installer script for 9.1
R	2.15.2	Also tested with 2.14.1 and 3.1.0
Rserve	1.7-3	Launched as an R server
Other R packages	Biobase - 2.20.1 BiocGenerics - 0.6.0 BiocInstaller - 1.8.3 bitops - 1.0-6	

	Cairo - 1.5-2 caTools - 1.16 cluster - 1.14.4 codetools - 0.2-8 colorspace - 1.2-4 data.table - 1.8.10 dichromat - 2.0-0 digest - 0.6.3 doParallel - 1.0.6 dynamicTreeCut - 1.60-1 flashClust - 1.01-2 foreach - 1.4.1 Formula - 1.1-1 gdata - 2.13.2 ggplot2 - 0.9.3.1 gplots - 2.10.1 gtable - 0.1.2 gtools - 3.1.1 Hmisc - 3.12-2 impute - 1.34.0 iterators - 1.0.6 KernSmooth - 2.23-10 labeling - 0.2 lattice - 0.20-24 limma - 3.14.4 matrixStats - 0.8.12 multicore - 0.1-7 multtest - 2.16.0 munsell - 0.4.2 plyr - 1.8 proto - 0.3-10 RColorBrewer - 1.0-5 reshape2 - 1.2.2 reshape - 0.8.4 R.methodsS3 - 1.5.2 rpart - 4.1-3 Rserve - 1.7-3 scales - 0.2.3 stringr - 0.6.2 survival - 2.37-4 WGCNA - 1.34 base - 2.15.2 compiler - 2.15.2 datasets - 2.15.2 graphics - 2.15.2 grDevices - 2.15.2 grid - 2.15.2 MASS - 7.3-23 methods - 2.15.2 parallel - 2.15.2	
--	--	--

	splines - 2.15.2 stats4 - 2.15.2 stats - 2.15.2 tcltk - 2.15.2 tools - 2.15.2 utils - 2.15.2	
--	---	--

Optional packages

The following packages are optional and can provide additional functionality:

Package	Version tested	Comments
Kettle	4.2.1	Used by ETL procedures, alternatives are available
GenePattern	3.2.3	License required

Developer packages

Developers working on the source code will also require:

Package	Version tested	Comments
git	1.8	Just to fetch the code

As well as a Grails compatible IDE or simply the grails package.

Installing tranSMART

TranSMART is downloaded and installed by an automated script. This script checks for essential components, downloads the software and demonstration datasets and performs the installation.

It is considered safer to install the tranSMART manually by simply downloading the compiled archives on the eTRIKS website :<http://go.transmart.etriks.org/download/>.

This website also provides virtual appliances, preconfigured and ready to go (see below)

ETL scripts

Scripts and stored procedures for data loading are included in the tranSMART distribution. Additional scripts are made available through the eTRIKS Wiki (<http://requirements.etriks.org/twiki/bin/view/Curation/WebHome>) for projects and through the tranSMART Wiki (<https://wiki.transmartfoundation.org/display/TSMTGPL/Data+ETL>) for public use.

Sanofi ICE Tool

Data loading

Security administration

For the security model to work correctly, certain database values needs to be present and these are detailed below:

Secure Access Levels are Present

By default, tranSMART is expecting three access levels to be present to allow users to OWN, EXPORT or VIEW a given study. There is no UI aspect for viewing, modifying and deleting these access levels and that could be part of a future request along with any underlying controller logic to make these modifications. The current version of tranSMART expects these values to be in the database and if not present the security model will not work properly. Future version of tranSMART could allow other access level along with the necessary controller and UI logic to make the modifications to the access levels in the Admin interface. To check if the values are present perform the following query:
SELECT * FROM SEARCHAPP.SEARCH_SEC_ACCESS_LEVEL;+

SEARCH_SEC_ACCESS_LEVEL_ID	ACCESS_LEVEL_NAME	ACCESS_LEVEL_VALUE
263801	OWN	255
263802	EXPORT	8
263803	VIEW	1

Figure 1 - Secure Access Levels

If you do not see results as shown in figure 1, please run the following SQL to insert the necessary values (hat tip, Paul A!):

```
INSERT INTO SEARCHAPP.SEARCH_SEC_ACCESS_LEVEL  
(SEARCH_SEC_ACCESS_LEVEL_ID, ACCESS_LEVEL_NAME,  
ACCESS_LEVEL_VALUE) VALUES (263801, 'OWN', 255);
```

```
INSERT INTO SEARCHAPP.SEARCH_SEC_ACCESS_LEVEL  
(SEARCH_SEC_ACCESS_LEVEL_ID, ACCESS_LEVEL_NAME,  
ACCESS_LEVEL_VALUE) VALUES (263802, 'EXPORT', 8);
```

```
INSERT INTO SEARCHAPP.SEARCH_SEC_ACCESS_LEVEL  
(SEARCH_SEC_ACCESS_LEVEL_ID, ACCESS_LEVEL_NAME,  
ACCESS_LEVEL_VALUE) VALUES (263803, 'VIEW', 1);
```

Load Study

The study must be loaded into tranSMART as Charlotte E notes [here](#) using Kettle with the “SECURITY REQUIRED” set to “Y”.

Secure Objects

The purpose of secure objects is to add a layer of security to the i2b2 data model that will allow tranSMART to use authentication rules to control access for a given study. The underlying i2b2 model by default allows access to all studies for all users and the secure object methodology is an attempt to overlay this layer of security on the i2b2 model without modifying any of the underlying internal i2b2 code. Obviously, other options are available and it will be interesting to see how this evolves especially with the deployment of the core API.

The first step is to create a secure object in the Admin interface by choosing **Add Study:**

Create SecureObject

Bio Data Id:

Data Type:

Bio Data Unique Id:

Display Name:


 Create

Figure 2 - Adding a Secure Object

All of these values should be present in the BIOMART.BIO_EXPERIMENT as Charlotte R mentions [here](#). These are summarized in the following table:

Create Secure Object Field	BIOMART.BIO_EXPERIMENT
Bio Data Id	BIO_EXPERIMENT_ID
Data Type	BIO_EXPERIMENT_TYPE
Bio Data Unique Id	"EXP:" + ACCESSION
Display Name	BIO_EXPERIMENT_TYPE + ":" + ACCESSION

Table 1 - Mapping SecureObject Fields

You will need to copy the fields from the BIOMART.BIO_EXPERIMENT table and enter them in the corresponding fields in the Adding a secure object section shown in figure 2. After the creation process is complete, you should have the values populated into SEARCHAPP.SEARCH_SECURE_OBJECT table and you can also verify this in the **Study List** view in the Admin interface.

Secure Object Paths

Secure objects are merely containers or wrappers around a given study in the BIOMART.BIO_EXPERIMENT table. The next step is to map the secure objects to a given path in the Dataset Explorer tree. This is completed by creating a secure object path using the Add SecureObjectPath function in the Admin interface.

Create SecureObjectPath

Concept Path:

Secure Object:

 Create

Figure 3 - Creating a Secure Object Path

Select the Secure Object that was just created in the previous step and then enter the concept path for the top node in the i2b2 tree for the given study including all forward slashes. If you are unsure of the concept path, you can retrieve it with the following query:

```
SELECT C_FULLNAME FROM I2B2METADATA.I2B2 WHERE C_HLEVEL = 1 AND SOURCESYSTEM_CD = <ACCESSION>
```

Please, note that the <ACCESSION> field is the

BIOMART.BIO_EXPERIMENT.ACCESSION field that was discussed in the creating a secure object section.

Once, this is complete you have now linked the secure object to a concept path in the Dataset Explorer tree representing the top level node of a study.

Post-installation checks

There are no scripts or procedure to thoughtfully test a transmart installation. Browsing to the address will be a good enough gage.

You should be able to test the security of the study with a user account that has the `ROLE_DATASET_EXPLORER_ADMIN` role set and a user without this role. The DSE admin user should be able to see this study while the study should be disabled (but visible) to the other user. The reason for the ability to see the study is to allow a user to right click and obtain more details about the study itself without allowing them to access any of the study data.

The final step is to provide the user or group with the appropriate level of access (i.e. OWN, EXPORT or VIEW) to the study in question. This can be done through either the Access Control by Group or Access Control by Study sections in the Admin interface

Mirroring

Virtual Machines

TranSMART 1.1 can be installed as a virtual machine with no further installation requirements.

Glossary

EC2 – Elastic Compute Cloud

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 – Work Package 2