European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.2 - Feature Roadmap 1

Due date of deliverable: March 2013

Actual submission data: December 2013

Dissemination Level		
PU	Public	PU
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:	D4.2
Deliverable title:	Feature Roadmap 1
Deliverable version:	1
Due date of deliverable:	March 30th 2013
Actual submission date:	December 1st 2013
Leaders:	Fabien Richard, Manfred Hendlich, and Reinhard Schneider
Editors:	
Authors:	Venkata Satagopam, Fabien Richard, Manfred Hendlich,
Reviewers:	David Henderson, Antigoni Elefsinioti, Peter Rice, Mansoor Saqi, Nathalie Jullian
Participating beneficiaries:	
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Fabien Richard, Manfred Hendlich, and Reinhard Schneider
Work Package participants:	
Estimated person-months for deliverable:	1
Nature:	
Version:	1
Draft/Final:	Final
No of pages (including cover):	
Keywords:	Query, analysis, visualization, export, user space

Table of Contents

TranSMART User Interface Improvement: eTRIKS Analytics and Visualization Roadmap

Version: 1.0 – December 1st 2013

tranSMART (TM) has the potential to become the tool of choice for translational medical research. So far it is serving as a data warehouse with some analytical capabilities. The purpose of this document is to define:

- the eTRIKS high level requirements and approach for improving the TM user interface (UI)
- the TM UI features
- potential technologies to implement those features

This will be the basis of the eTRIKS analytics and visualization roadmap for coming TM versions. The WP4 team will update this roadmap on a regular basis.

1. eTRIKS requirements and approach for TM UI improvement

1.1. Requirements

The TM UI must:

- be friendly, intuitive and logical for end users who are scientists and/or clinicians. The TM UI is currently designed according to clinicians and biologists' needs. The approach to development must however be proactive rather than reactive, and recommendations from bioinformaticians, developers, data managers, and statisticians wil be taken into account when designing UI improvements. It is this community who are likely to know the state-of-the-art in terms of bioinformatics analyses.
- be agile and modular: user must be able to select data for analysis, choose analytical methodologies and visualization features and save, export or perform further processing on the selected data.

1.2. Approach

eTRIKS WP4 proposes developing the TM UI as a workflow. The following are the six main steps for searching/analyzing/visualizing the data:

1. **SEARCH** patients or samples within or across studies in the dataset explorer
2. **SELECT** subsets of the data for analysis

3. **ANALYZE** selected data choosing from a repertoire of available analysis tools with default or user defined analysis parameters,
4. **VISUALIZE** results of data analysis. Define the visualization parameters,
5. **SAVE** the analysis and visualization output, together with the parameters used in individual's user space,
6. **EXPORT** the analysis and visualization parameters, the analysis results, and/or the graphics.

eTRIKS should develop the TM UI as follows:

a. by defining a group of UI features for each step of the workflow,
b. by prioritizing the feature implementation according to 1) users' needs and use cases, 2) interdependency between features of different steps,
c. by selecting the technologies/tools that are required for the feature implementation.

2. TM UI Features

2.1. User session management

TM must provide a password protected session management facility in order to keep track of the search/analysis history and the results over the time. Through this session management, TM can grant various permissions to the users according to their involvement in different IMI projects and/or their rights to access restricted materials.

2.2. Search and data mining

TM requires a semantic layer/knowledge base in order to find/identify "patients or samples within or across studies in dataset explorer and/or historic data by using advanced search features/filters". Here semantic layer means the bio-entities like genes, proteins, chemical, diseases, pathways etc., which should be annotated with their official names, synonyms and with their corresponding hierarchical classification using prioritized ontologies and such well annotated resource in the knowledge base. We need such a knowledge base in eTRIKS, in order to perform efficient and meaningful queries both in TM search applications (or faceted search) as well as in the dataset explorer.

- Integrate improved Search Functionality into eTRIKS core system. Currently independent efforts from EFPIA partners (J&J -> Faceted search, Sanofi -> Combined Search) to improve the Search functionality in tranSMART. Integration of both solutions is planned within the US tranSMART initiative.
- Expand the search functionality to permit simultaneous, multiple factor searches with Boolean operators. Currently, TM can search one gene or one disease at a time; it would

be more useful if the user could submit a list of genes or other bio-entities and retrieve the information related to the given list of entities.

2.3. Selection

Select patients, samples, and/or variables

- While a user selects a patient or sample group across studies or within a study, TM dynamically displays the variables shared in common and the number of patients/samples for each variable. Here, 'variables' are the leaves of the TM trees. If, for example, a user selects disease 'MS', compound 'Rebif', patient 'female', then TM should give the user all the studies that are selected by these criteria as well as all the variables in common between these studies such as age, ethnicity, relapse rate etc.
- Once a user has selected variables, TM calculates and displays the number of patients or samples that have all the selected variables.

2.4. Analysis

TransMART needs to be more than just a data warehouse. There are a number of methodologically straightforward but rather complicated bioinformatics analyses that many biologists and clinicians may want to carry out. transMART should make these classes of analyses accessible to a wider biomedical audience. It must also be agile enough to allow new methods to be included perhaps with a 'plug-in' architecture. We generally deal with three types of data: raw, processed and the derived data from projects of eTRIKS clients as well as public data sources. From the technology point of view, this data derives from different 'omics (transcriptomics, proteomics, metabolomics, lipodomics), NGS, clinical and preclinical experimental systems. We need different tools to handle different types of data at different levels, some are very specific while others are more general. Prioritization of eTRIKS efforts to develop functionality for new data types will be determined after analysis of client project needs.

A typical analysis pipeline will be a user-friendly graphical UI with the following general features.

- A user enters the variables he/she selected in point 2 (drag drop features),
- A user chooses the data level for his/her analysis. TM proposes only data levels that are "compatible" with the technology platforms (e.g. level 1 is proposed only if all the data are from the same platform),
- A user defines the processed data: what primary data and (basic or advanced) transformations are used to obtain processed data. TM proposes normalization methodologies for high dimensional data,
- A user chooses the analysis type (comparison, correlation, pathway analysis) from a library,

- A user chooses the analysis methodology (t-test, ANOVA, for pathway analysis: over-representation approaches, aggregate score approaches, topological approaches, etc.) from a library.

2.4.1. On-the-fly calculation and storage of derived results

TM should provide the necessary analytical methods to derive results, save these in user session and allow retrieval for later use, for example in cohort comparisons. For example, visit and drug injection time points of a longitudinal study are recorded in TM. Users want to assess if there is a correlation between the time period visit - drug injection and the efficacy of the drug measured at the visits. To this, TM has to calculate those time periods on fly.

2.4.2. Cross-study analysis

Enable cross-study analysis between different studies coming 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.

2.4.3. Machine learning methods

TM should be equipped with machine learning methods like SVM, decision trees, random forest etc in order to identify important patterns in the data, especially from the clinical data. [to be provided by IMI project requirements]

2.4.4. NGS

Next generation sequencing (NGS) data is bit special due to of its size, it is not clear what level of NGS data will uploaded into TM, perhaps processed data. In case we plan to provide the possibility to analyze the NGS data within TM, we may need some important tools like SAMtools, Cufflinks etc. There is a possibility to equip these tools within Galaxy as well.

2.4.5. Mass-spec data analysis

Similar to NGS data, what level of proteomics, metabolomics and lipidomics data should be loaded into the TM: spectra level, peptide or protein level. In order to handle the spectra level data, one need to search the spectra against a protein sequence database for example UniProt by using the search tools like Mascot or Phenyx. In order to calculate the peptide or protein level abundances, we may need bioinformatics pipelines like MaxQuant, TPP or commercial software like Progenesis etc within TM.

2.4.6. Improved functional interpretation of Biomarker findings

- tranSMART should facilitate the interpretation of biomarker findings by linking to additional external data bases (e.g. CLINVAR, DRUGBANK, OMIM, PharmGKB [List of databases to be specified])

- tranSMART should provide evidence for variant-disease/phenotype or gene-disease/phenotype associations by integration of text mining results.
- tranSMART should be capable of searching based on different bio-entities (genes/proteins/chemicals/diseases, etc) and their classification. Here classification means for example for chemicals - compounds/drugs, they can have chemical classification or pharmacological classification. TM should allow the user to search either compound name/synonym or any of its parental term in the respective classification (chemical or pharmacological).

2.4.7. Improved Interpretation of Upstream / Downstream effects of variants

In order to perform certain analysis we may also need access to some important biological databases and tools; for example MSigDB for GSEA, or KEGG, Panther, Reactome for pathway enrichment analysis, OpenBEL for causal reasoning. Access to protein-protein interaction (PPI) databases like IntAct, STRING, HPRD, iRefIndex etc. is also required for interaction network analysis. Another simple, yet important problem we face in biology is the database identifiers we use to represent different biological entities: genes, proteins, chemicals etc. We need to have access to resources like bioCompendium (http://biocompendium.embl.de) developed in our group to address this issue. It also comes with pathway and GeneOntology enrichments, protein homology and domain architecture based clustering and certain other bioinformatics analysis functionality.

One possibility is to have external database plug-in or app in TM, that provides the important publicly available databases and proprietary resources like GeneGO, Ingenuity, Ariadne etc with some regular updating mechanism. Another possibility would be to have an app in TM that fetches the data on demand by using SPARQL, an RDF query language in 'Linked data' framework. It can be bit expensive based on the amount of the requested data and the complexity of the queries; this is mainly due to the current limitations of Internet bandwidth.

2.4.8. Literature mining supporting Biomarker Hypothesis

Right now we have over 22million PubMed articles, and roughly 10% of them available as PMC full text articles. In order to take the advantage of this vast store of publicly available literature and some proprietary sources like Elsevier (covers ~25% of PubMed), TM should equip with a text mining plug-in or app. This should enable integration of text mining results on variant-disease/phenotype, gene-disease/phenotype etc associations.

2.5. Visualization

Improved GUI and visual representation of data are the following.

2.5.1. Display of >2 cohorts:

Enable the view multiple (>2) subgroups simultaneously in order to judge the impact of different therapies, genotypes and phenotypes on clinical outcomes.

2.5.2. Improved Patient Centric Views:

- Need improved drill-down capability on data from individual patients and provide better visualization of data from individual patients including longitudinal / time course data.
- Enable search for individual patient data by patient ID.

2.5.3. Improved Presentation of new data types:

- Improved presentation of genetic variant data will be essential. Systems such as cBioPortals OncoPrint visualization could be integrated.
- Improved representation of longitudinal /time course data (e.g. Phenotimer)
- Display of data provenance information

2.5.4. Improved Plots:

- Improve display of the labels, allow re-sizing the diagrams e.g: heat maps.
- Enhance functionalities supporting both low and high dimension time course data, for example enable time scale on line plots.
- Additional plot types e.g. Kaplan-Meier survival curves

2.5.5. Visualization of Network Biomarkers:

The inclusion of methods for identification of network based biomarkers into the available set of transMART analysis tools would necessitate some network visualization functionality. This may be provided by calling external tools (Gephi, Cytoscape, networkX (for network analysis).

2.5.6. Visualization of Network Biomarkers:

- Users should be able to store documents for example MS Word, PDF, images etc for a particular study and should be able to access them as well.
- Improved visualizing of study meta data, definition of a data type in a study, and data normalisation process used for a data entity in order to decide if the same data entities in different studies can be directly compared
- Fully documented data model in application layer is necessary in order to allow the plug-in of new visualization components

2.6. Save

Save the analysis pipelines including the analysis and visualization parameters as well as the analysis results in the user space.

2.7. Export

Export the analysis parameters and results including figures in a report. When exporting data, the provenance of those data (i.e. data owners) must be provided and saved in the export files.

TranSMART should provide an efficient data export tool to export all or the selected analysis results into different formats. The Data Export interface must comply with data privacy restrictions:

- Enabling precise selection of data to export.
- Implement logging functionality for tracking data export (who, which data sets, when)
- Implement functionality to disable export of restricted data sets

3. Technologies/tools

The following technologies/tools would be very useful to have in TM.

- D3.js (http://d3js.org) is a powerful JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.
- Arena3D (http://arena3d.org) provides staggered multi layer concept that allows the analysis of big networks in a three-dimensional space representation. The different layers in the representation correspond to different data types respective concepts like sequences, structures, chemicals, diseases, pathways etc. The data entries for one specific data type, like sequences, can be ordered or clustered on their respective layer by applying a data focused similarity measurement like sequence similarity. Several clustering methods are also supported to cluster several data types in different layers. The different nodes of the layers can be connected according to known or predicted relationships between the nodes. The relationships are typically extracted from available databases and available text mining machineries, and are predicted by various data mining methods or originated from experimentally generated data. Indirect connections that may hide some additional information can also be explored.
- PhenoTimer (http://phenotimer.org) is a visualization tool for time-resolved biological data. It helps make sense out of gene-phenotype relations in a temporal context, by displaying the time-driven phenotypic connections for the given dataset in parallel with networks highlighting the genes or functions corresponding to a particular time point.
- BioJS (http://code.google.com/p/biojs) is an open-source project whose main objective is the visualization of biological data in JavaScript. BioJS provides an easy-to-use consistent framework for bioinformatics application programmers. It follows a community-driven standard specification that includes a collection of components purposely designed to require a very simple configuration and installation. In addition to the programming framework, BioJS provides a centralized repository of components available for reutilization by the bioinformatics community.
- Cytoscape (http://cytoscape.org) is the tool of choice to visualize the data in 2D, especially visualizing complex interaction (network) data and it is also equipped with

data and analysis component through available plug-ins. The current version of Cytoscape 3.0 comes with one of the latest technologies, OSGi framework, which makes the tool much efficient in response point of view.

4. Incremental data loading and backup

- TM should provide the possibility to load data incrementally to e.g. support data management and data mining in longitudinal studies. for example couple of samples related to a study provided at time point one and the study is loaded into TM. But later at time point two few more samples related to the same study arrived, TM should allow to add the new samples to the existing study
- TM should provide a mechanism to rollback at least to the previous version of the study. Lets assume a scenario that something went wrong with the new version of the data or not meeting the specified guidelines and data is already loaded into the system and now the client want to rollback to the previous version. TM should have the mechanism to go back to previous version of the study.