tranSMART v17.1 - tranSMART Pro Consolidated Requirements

v1.2 27 Apr 2016

This document contains the set of requirements determined to be common among the four sponsoring members of the tranSMART Pro alliance: Roche, Sanofi, Pfizer, and Abbvie. This project will address two of the principal requirements from the tranSMART 17.1 Business Analysis Project: support for longitudinal/EHR/wearables data, and support for large-scale genomic data. Areas that were discussed during that process, but are not addressed in the current project, include general performance and robustness of the platform, support for multiple ontologies in various contexts, and enhancements to security and access control. These areas may be partly addressed by other versions or projects in the future.

The requirements below are divided into Functional and Technical requirements in each area. In addition, there are General requirements that apply to both areas, or provide general guidance to potential implementers.

Longitudinal data support: Functional Requirements

Req No.	Description
F001	The system shall allow a measurement date and time to be optionally associated with a clinical data point (LDD), and support multiple timestamped data points per patient/concept.
F002	Date and time stamps shall allow varying resolutions, for example both date and time, date only, or year-only (the latter to support data privacy regulations).
F003	The system shall allow a named event/encounter, with date and time, to be optionally attached to a clinical data point.
F004	The system shall provide the option to include date and timestamps in query results, data displays, and exports.
F005	The system shall allow patient cohort selection using flexible time and event-based selection criteria. Selection criteria can be absolute (e.g. from *time1* through *time2*, or from *event1* through *event2*) or relative (e.g. within *time1* after *event1*).
F006	The system shall allow Advanced Workflows to operate using time-based data, (e.g. a scatter plot for data between *event1* and *event2*).
F007	The core, R-interface, and RESTful APIs shall be enhanced to allow queries and retrievals that use time-based criteria.
F008	The system shall treat date and timestamps (regardless of resolution) as sortable numeric types in all data displays, exports etc.

Longitudinal data support: Technical Requirements

Req No.	Description
T001	Support for time-based clinical data shall be provided in a manner that is compatible with the data model and usage of i2b2 v1.7. This is intended to preserve interoperability with i2b2 in the future, for example in the event that a tranSMART user also has access to an i2b2 installation. How this compatibility is implemented is left as a matter of technical design.
T002	The system shall allow for data volumes of thousands of studies, and tens of thousands of data points per patient per concept, in order to support for example high-frequency wearables data. It should be possible to load these data with adequate performance, though the system is not intended to be a real-time data collection platform.
T003	The system shall provide an upgrade path from at least the tranSMART 16.2 release, that transfers legacy data (with no timestamps), to the 17.1 data model without the need to reload. Users will need to reload data only if they want to attach timestamps to legacy data. The reload will only be necessary for the affected node, but not for the whole study (incremental data loading). (Specifically: Allow retrieval of time series information that are stored in XML format in the C_METADATAXML column of the table I2B2 in 16.1/16.2.)
T004	In order to support existing external programs, the existing APIs should continue to be supported, but can be deprecated over time.
T005	To the extent possible, the design should anticipate and not foreclose eventual cross-study queries, for example by separating concepts from studies.
T006	To the extent possible, the design should improve the overall stability and robustness of existing functionality that is enhanced or modified.
T007	The system shall preserve the capabilities of existing 16.2 components including specifically SmartR, and shall enable SmartR to include support for time and event-based data.

Scalable Genomics: Functional Requirements

Req No.	Description
F101	The system shall support the import, storage, query, and retrieval of high-volume genomic data including variants and genotypes from microarrays and DNAseq experiments, and expression data from arrays and RNAseq and miRNA experiments.
F102	The system shall allow configurable filters for the import of genomic files, so that low-quality data can be excluded on import
F103	The system shall provide both RESTful and high-capacity APIs (including R API) to support retrieval of genomic data in a flexible and efficient way.
F104	The system shall support the management and query of data from more than one sample per patient, and more than one experiment per sample.
F105	The system shall allow the use of genomic data when constructing patient cohorts.

F106	The system shall allow Summary Statistics to display cohorts stratified by genotype, for one or more variants.
F107	The system shall allow Advanced Workflows to retrieve and operate on combined genomic and patient data with acceptable query and retrieval performance. Workflows may be executed by an external component such as Galaxy or Arvados.
F108	The system shall allow for the storage, maintenance, and annotation of the results of genomic analyses, and provide a means of retrieving those results.
F109	The system shall allow for the creation and maintenance of named lists of genomic entities of interest (genes, probesets, variants, transcripts) so that they can be used in operations such as cohort building and advanced workflows.
F110	The system shall accommodate and separate data (including dictionary/reference data) from different species and different genomic builds, and annotate such data adequately to ensure that users can select, and know which species and build they are working with.

Scalable Genomics: Technical Requirements

Req No.	Description
T101	The system shall support the import and export of GCT files containing expression values for millions of probesets and thousands of samples per file.
T102	The system shall support the import and export of variants, variant annotations, genotypes, and genotype annotations from VCF v4.2 files, including all variant types supported by that version. Annotations may include the standard ones (depth and quality filters), as well as other INFO fields defined by the user.
T103	The system shall support the import and retrieval of CEL and other sample-specific genomics files.
T104	The system shall support the existing Pfizer-developed tranSMART genomics API (to be provided). This may be sufficient for other members of the tranSMART Pro alliance.
T105	The system shall preserve (but not necessarily extend) the existing MongoDB file interface as developed by DEXSTR and currently used by Sanofi.
T106	The system shall interface with the Arvados native APIs to support import of VCF files (complete or partial) stored in Keep, and storage and retrieval (in Keep) of sample-specific files related to a set of patients in tranSMART.
T107	At this time there is no requirement to support the management of functional annotation of variants (except inasmuch as such annotation may be included in imported VCF files), but this capability will likely be needed in the future.
T108	It is expected that a general purpose REST API for genomic data shall be compatible with the GA4GH API definitions, when those are stable and available. It is acknowledged that this may not occur in time for 17.1.

General Requirements

Req No.	Description
G001	The system shall support both Postgres and Oracle with equivalent functionality, but the capacity of the system may be partly determined by the choice of database. Specific capabilities of Oracle for example can be used to provide greater capacity or performance than when using Postgres.
G002	Although it is not an explicit requirement, any data storage approach should support a federated model, and anticipate increased use of such a model as data volumes increase (especially with respect to genomic data).
G003	In general, a functional and robust API in each area is a first priority, over a specific user interface; this allows external apps to utilize the system as a data source for custom visualization and analysis.
G004	The system is not intended to itself provide a high-performance computing platform for primary genomic analyses (e.g. read alignment and variant calling), but can interface with different content and/or workflow management systems that support those analyses (e.g. Galaxy, Arvados Keep/Crunch, ADAM)

Change Log:

v1.1: Additions and modifications from H. Schuermann of Sanofi
v1.2: Updated T102 to indicate that import of VCF files shall include variant and genotype annotations as well as variants and genotypes.