# Table of Contents

- Executive Summary
- Introduction
- Process
- Interviews
- Requirements
- Use Cases
  - Summary of use cases:
    - i2b2 Integration
    - Technology Evaluations
- Development Proposals
  - Summary of Development Proposals
    - Harvard Proposal
    - The Hyve Proposal
    - ConvergeHEALTH by Deloitte Proposal
- Summary and Conclusions
Executive Summary

The tranSMART Foundation has embarked on the development of an infrastructure to manage the specification, development, release and support of the tranSMART platform. In 2015, the Foundation published a roadmap for the platform, which calls for two annual releases of the platform at ~6 month intervals. The first release under this new paradigm is the 16.1 release (first half of 2016), which is the first version of the platform to be produced by the Foundation by following its new code governance and quality standards process. As a part of the roadmap, the Foundation has also defined the set of community-developed features that are candidates for the 16.2 (second half of 2016) release of the platform. The Foundation has not played a direct role in the development of these features (with some minor exceptions), but has worked with the community to ensure that these features are developed to the new quality standards of the Foundation so that they can be incorporated into the release.

The Foundation has now entered into a collaborative process for specifying the features and requirements for the 17.1 version of the platform, and has conducted this project to gather requirements and use cases among the Foundation membership, to prioritize these requirements, and to engage with developers who could develop software under the direction of the Foundation and its membership. This report summarizes the conduct and results of this project.

Process

The Foundation assembled a team to conduct a 6-week requirements gathering project. The team had meetings with 11 out of 22 of its members of the Foundation. Each group was asked to evaluate and prioritize the features that are candidates for 17.1 development. Candidate requirements were categorized into several key areas based on previous input from the community. Members were asked to prioritize those requirements for their organizations, and to develop use cases to further refine these requirements. Table 1 shows the consensus priority of each category, and Table 2 shows the detailed breakdown by organization.

The team solicited implementation proposals from interested developers. Proposals were submitted by Harvard, the Hyve, and ConvergeHEALTH by Deloitte. The developers were also asked to provide proposals that can address the project in stages, in the event that the full development budget is not available in 2016.

Conclusion

The highest priority functional requirements were in order, improvements to the robustness and performance of the platform, support for longitudinal data, support for cross-study analysis, and support for high-volume genetic variant data. In addition, a separate proposal to fully re-integrate i2b2 1.7 and tranSMART 17.1, and to maintain that integration in the future, was put forward by the i2b2 team at Harvard. Although this integration was not specifically identified as a priority during this effort, it has a number of advantages, and in particular includes both longitudinal data and cross-study support in a way that is interoperable with i2b2.
Introduction

The tranSMART Foundation Board of Directors recommended that the Foundation management team embark upon a project to define the requirements for the tranSMART 17.1 project, to prioritize these features by interviewing members of the Foundation, and to gather use cases that exemplify the functionality required by these members. The Foundation formed a team that included the VP of Engineering, John O’Hara, the Director of Community Relations, Keith Nangle, and an experienced member of the tranSMART community, Ken Kubota. The team developed a process for determining requirements, prioritizing these requirements and capturing use cases, and embarked on a 6-week program to complete these tasks. This report outlines the key findings of that project, and presents the results in summary form. All of the source documents behind these findings are available on the Foundation wiki here.

Process

The process began with key areas of concern that have been identified in previous discussions with stakeholders in various venues including online community meetings, Foundation annual meetings, and 3C working groups. These areas were:

- Improvements to the stability, robustness, and performance of the platform and its APIs
- Improvements to ETL processes and performance
- Support for longitudinal and event-based clinical data
- Support for cross-study analyses
- Support for ontologies in ETL and data query
- Support for high-volume genetic variation data (e.g. whole-exome or whole-genome sequencing)

In addition to the above, a proposal to integrate tranSMART with i2b2 v1.7 is on the table, and if implemented will by default address some of the above functionality.

From this starting point, the process was to then:

- Conduct interviews with key stakeholders to define specific requirements in each area
- Solicit use cases and prioritization of these requirements from stakeholders
- Perform initial evaluations of candidate technologies and infrastructure that is currently in use by stakeholders (e.g., Arvados)
- Provide requirements and constraints to potential implementers, and work with them to formulate technical proposals and associated cost estimates

The development of 17.1 differs from the processes used in previous releases, in that the Foundation will contract directly with organizations able to produce a system that meets predefined requirements. This is in contrast to earlier releases, in which functionality is contributed by community members to match their particular requirements, and integrated into a release by the Foundation after the fact. The 17.1 approach relies on sufficient funding being secured prior to the start of implementation. Because of schedule constraints, the technical proposals were developed concurrently with requirements.
gathering and prioritization process, so the requirements and design must be further refined as part of implementation.

**Interviews**

Interviews were held with the following key stakeholders:

**Academic**
- University of Michigan (Gold member)

**Pharma**
- Sanofi (Gold member)
- Pfizer (Gold member)
- Roche (Gold member)
- Takeda (Gold member)
- AbbVie (Silver member)
- Boehringer Ingelheim (membership in process)

**Other**
- IO Informatics (Silver member)
- The Hyve (Silver member)
- Rancho Biosciences (Silver member)
- Thomson Reuters (Silver member)
- Perkin Elmer (Silver member)

Additional stakeholders were approached and either declined to participate or were unable to meet within the available schedule. Some of these will have an opportunity to participate during the implementation process as requirements and designs are refined.

The interviews were an important part of the process as they provided an opportunity to hear directly from stakeholders about their views on issues not specifically related to requirements and priorities. Some of these are:

- The strength of tranSMART is in making data available in an intuitive way to the scientists who are best placed to derive insights from it. It is also used by more expert users to perform specific analyses, often by exporting to external tools. The design of 17.1 should keep both of these categories of user in mind.
- Performance limitations (ETL, query and export) are a barrier to wider strategic adoption, even with respect to clinical (low-dimensional) data. The system must be able to efficiently handle thousands of studies and tens of thousands of subjects without the need for local optimizations.
- It is important that development timelines are realistic; it is better to defer some functionality to a later release than to jeopardize the on-time production of a high-quality 17.1 release.
- It is important that the design supports a clear distinction between samples and subjects, so that data can be accommodated (and distinguished) from multiple samples per subject.
- Besides performance issues, ETL needs to be less brittle, with better error handling and ideally, drag-and-drop loading of standard file types. Support for ETL tools other than Kettle (KNIME, Talend, Pipeline Pilot) would be useful for organizations that already use them.
● The ability to combine data easily across studies is very important: if it’s not easy for scientists to do without significant ETL efforts, it won’t happen, and insights will be missed. There are dangers inherent in aggregating and analyzing data from different study designs, but the platform should not stand in the way. This capability requires robust support for ontologies in order to make the process reliable.

● With respect to genetic variation data: tranSMART should not attempt to provide primary processing (read processing, alignment, variant calling) of sequencing data. Instead, it should integrate with platforms that perform those functions, and focus on integrating genotype data with clinical data to enable scientific insights.

● Genetic analysis workflows often proceed from large to small sets of variants, and the platform should support both points of view at the appropriate times. For example, a large-scale GWAS of a categorical phenotype will result in a small number of significant variants to be further investigated in detail along with a richer set of clinical data.

Requirements

The requirements resulting from this project, and their overall prioritization (1-4, with 1 being Low and 4 being High) are categorized and tabulated in Table 2 below. Table 3 contains the detailed breakdown of priorities by stakeholder.

Some non-functional requirements emerged during discussions, notably:

● 17.1 must provide an upgrade path from earlier version(s), so that data can be migrated without having to be reloaded. In some cases this will require explicit design of backward compatibility with existing data (for example, when extending the data model to support longitudinal data)

● Integration with existing IT infrastructure, for example Arvados (genomics) and Active Directory Services (security).

● Native support for a federated data model would leverage the reality of data distribution in large organizations, as well as provide a means of sharing data without the need for additional ETL activity by each user.

● Not all analyses can be performed on just two cohorts; 17.1 should allow for the creation and analysis of multiple cohorts in one workflow.
**TABLE 1: SUMMARY OF REQUIREMENTS AND PRIORITIES**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
<th>Priority</th>
</tr>
</thead>
</table>
| Platform Robustness and Performance      | • Make the platform more supportable and upgradable  
• Make queries more reliable and have better performance on scaling  
• Robust and well-documented APIs  
• Tune queries for Oracle backend  
• Support for federated data model  
• Improved ETL performance  
• Security and access control          | 3.77     |
| Longitudinal Data Support                | • Ability to align clinic or data capture visits across trials  
• Support unscheduled data (e.g., EHRs)  
• Support for I2b2-style ‘encounters’ table  
• Support for relative / elapsed time queries and comparison operators (e.g., patients with AE within 24 hours of dosing)  
• Support for data and event selections in advanced workflows and visualizations  
• Backward compatibility with existing study data, such that existing studies do not have to be reloaded | 3.23     |
| Cross Study Support                      | • Ability to merge data from multiple trials for analysis without losing data origins  
• Allow concepts to be independent of studies so that the same concept can be utilized and compared across studies | 3.10     |
| Support for High Volume Variant Data     | • Allow loading, querying, exporting, and cohort-building based on genetic variation data, whether derived from sequencing or other platforms (e.g., array-based assays)  
• Support for whole-genome volumes of variants (10s millions per patient) from 100s of thousands of samples  
• Support all variant types and annotations supported by VCF 4.2 spec (including structural variants)  
• Support for ARVADOS, and GAIMH API | 3.00     |
| Upgrade path / I2b2 Integration           | • Backward compatibility from previous versions of tranSMART  
• Ensure data support from previous versions of I2b2  
• I2b2 Integration must not mean that data have to be loaded twice | 3.00     |
| Continuation of SmartIR or other plugin visualization/analytic tool (e.g., Spotfire) | • Harmonize workflows: Need to maintain SmartIR workflows and tranSMART workflows  
• Ability to create a cohort and data set, then be able to apply multiple workflows against it | 2.85     |
| Support use of standard and internal proprietary ontologies | • Embedded support for ontologies to help make data curation (ETL) and cross study data queries and normalization easier | 2.85     |
| Better support for flexible ETL          | • tranSMART should recommend preferred ETL tool  
• Improved error handling | 2.23     |
TABLE 2: REQUIREMENTS FOR 17.1 PRIORITIZED BY COMMUNITY MEMBERS.

| Topic                                      | Average | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|--------------------------------------------|---------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|
| Platform robustness and performance       | 3.8      | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Longitudinal data support                  | 3.2      | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Support for high volume variant data      | 3.1      | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Continuation of benefit or shortfall for type of plugin visual analytic interface such as Tableau and others. | 3.0      | 4 | 4 | 4 | 4 | 4 | 1 | 4 | 4 | 3  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Upgrade path                               | 2.9      | 3 | 3 | 4 | 4 | 4 | 4 | 2 | 4 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Cross-Model Support                        | 2.9      | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Support of standard and internal proprietary terminologies | 2.9      | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 1 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Fine-grained access control                | 2.8      | 4 | 4 | 4 | 1 | 4 | 1 | 4 | 4 | 1  | 4  | 1  | 4  | 1  | 5  | 1  | 2  | 4  |
| Support for GA4GH API                      | 2.5      | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| FASTQ, SAM/BNF file store                  | 2.4      | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Support for ARVADOS (Mongo DB or similar platform e.g. ROC55) | 2.3      | 1 | 4 | 1 | 4 | 4 | 1 | 4 | 4 | 1  | 2  | 1  | 1  | 1  | 1  | 1  | 1  | 4  |
| Better support for flexible ETL            | 2.3      | 1 | 4 | 1 | 3 | 4 | 1 | 3 | 4 | 4  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |
| Enable meta analysis, e.g. clinical cohort normalisation | 2.2      | 3 | 4 | 3 | 4 | 4 | 4 | 3 | 4 | 4  | 2  | 2  | 1  | 1  | 1  | 1  | 1  | 4  |
| Variants/Functional Annotation             | 2.3      | 1 | 1 | 1 | 1 | 3 | 4 | 1 | 3 | 4  | 4  | 1  | 3  | 4  | 4  | 4  | 4  | 4  |
| More prelabeled/feat to triage and genomic data | 2.1      | 1 | 3 | 1 | 4 | 1 | 1 | 3 | 1 | 4  | 1  | 1  | 1  | 1  | 4  | 1  | 4  | 4  |
| GIMV analysis results                      | 1.9      | 1 | 1 | 1 | 1 | 4 | 1 | 2 | 1 | 2  | 1  | 4  | 4  | 4  | 4  | 4  | 4  | 4  |
| Support for somatic and germline variants  | 1.6      | 2 | 2 | 1 | 3 | 4 | 1 | 2 | 1 | 1  | 1  | 4  | 1  | 4  | 1  | 1  | 1  | 4  |
| Integration with LabSAFE                  | 1.9      | 1 | 4 | 1 | 1 | 1 | 4 | 1 | 3 | 1  | 1  | 1  | 1  | 1  | 4  | 1  | 1  | 4  |
| Ability to link files to a specific patient| 1.8      | 4 | 1 | 1 | 1 | 3 | 4 | 1 | 1 | 2  | 1  | 1  | 2  | 1  | 1  | 1  | 1  | 1  |
| ACDH status                               | 1.8      | 3 | 3 | 1 | 4 | 1 | 1 | 1 | 1 | 4  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |
| Make expression data easy to normalize(e.g. in a cross-study analysis) | 1.8      | 2 | 2 | 1 | 3 | 1 | 1 | 3 | 1 | 1  | 1  | 2  | 4  | 2  | 4  | 2  | 4  |
| Interoperability with bioinformatics/LIMS systems | 1.8      | 4 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 5  | 1  | 5  | 1  | 5  | 1  | 5  | 1  | 5  |
| Support for analytics                      | 1.8      | 4 | 1 | 1 | 1 | 1 | 3 | 1 | 3 | 1  | 3  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |
| Histology of data, access to updated data without data read. | 1.7      | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 1 | 1  | 1  | 2  | 4  | 2  | 4  | 2  | 4  |
| Imputed data                               | 1.6      | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4  | 1  | 3  | 2  | 1  | 1  | 1  | 1  | 1  |
| Integrated genome database, be able to call a single gene within one graph (NEO4J) lightning | 1.6      | 1 | 3 | 1 | 3 | 1 | 3 | 2 | 1 | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |
| Ability to further process belonging to dimensional file and store into RC/RIS (e.g Charon) | 1.6      | 1 | 2 | 1 | 3 | 1 | 1 | 4 | 1 | 1  | 1  | 2  | 1  | 2  | 1  | 2  | 1  | 2  |
| Ability to return mutation calling pipelines to update the data stored in text START | 1.4      | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1  | 1  | 1  | 1  | 2  | 1  | 2  | 1  | 2  |
| Access control integrated with informed consent | 1.6      | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 1  |

Note that the requirements are rated from 1-4 in each category with 4 being the highest. Community members are organized by Gold, Silver and Blue (membership pending) status.

*Sanofi rates support for Mongo DB, Arvados is not as high priority.

Comment [1]: do the sanofi priorities reflect the updated document from Christophe?
Use Cases

The Use Case Document is a business document which provides a story of how a system, and its actors, will be utilized to achieve a specific goal. An effective Use Case should provide a detailed step-by-step description of how the system will be used by its actors to achieve the planned outcome. The purpose of the Use Case is to tie the business needs of the system to the design parameters of the system to ensure that the completed system achieves the goals established by the business requirements. The level of detail in Use Cases may vary greatly depending on the size and complexity of the system being designed.

The template shown in the example below was used to document use cases in several topic areas.

Figure 1: Use Case Example

<table>
<thead>
<tr>
<th>Name of Use Case:</th>
<th>Query tranSMART for cancer specific genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created By:</td>
<td>Keith Nangle</td>
</tr>
<tr>
<td>Date Created:</td>
<td>17 Mar 2016</td>
</tr>
<tr>
<td>Last Updated By:</td>
<td>Keith Nangle</td>
</tr>
<tr>
<td>Last Revision Date:</td>
<td>17 Mar 2016</td>
</tr>
</tbody>
</table>

Description: An analyst queries the data in tranSMART to retrieve cancer specific genotypes from a study or a group of studies, e.g. the genetic variant is present in cancer samples but not in normal samples (from the same patient). In addition, when outcome data from drug treatment is available, analyst should be able to optionally retrieve the genotypes along with associated outcome data linked to the patients.

Actors: Bioinformatician

Preconditions:
1. Cancer variant (mutation) data has been loaded into tranSMART
2. Clinical data has been loaded into tranSMART

Postconditions:
1. Allow user to export mutations in various formats, or to pass the data to an analysis workflow or visualization tool e.g. Spotfire.

Flow:
1. A query interface will allow user to search studies by keywords such as disease, drug name, or study owner.
2. User identifies cohort(s) of interest using a combination of clinical attributes and variant attributes, including:
   - Gene name(s)
   - Variant ID(s) (rsid, probeset ID)
   - Specific mutation, e.g. V600E of BRAF gene
3. User chooses whether to export the corresponding data, in VCF or tab-delimited format, or to pass to an analysis workflow

Comment [2]: we should have these fields completed.
**Alternative Flows:**

1. A query interface, e.g. “advanced search” will allow users to search studies together with other attributes such as gene name(s).
2. User chooses whether to export the corresponding data, in VCF or tab-delimited format, or to pass to an analysis workflow.

**Exceptions:** If the query returns no data, the system should provide meaningful error message(s) that help to narrow down the problem.

**Requirements:** Query performance should be within the norm of database size with proper indexes and the hardware it runs on.

### Summary of use cases:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal Data</td>
<td>BMI rate of change for genomic variants</td>
<td>Determine the number of carriers of the Gln354-&gt;Glu variant for a particular gene and how they differ from non-carriers, for example in rate of change of BMI over time.</td>
</tr>
<tr>
<td>Cairo User study (TraIT)</td>
<td>Validate MSI as a prognostic biomarker for disease progression in metastatic colorectal cancer</td>
<td>Combine experimental and clinical data to perform analysis and make the data publicly available.</td>
</tr>
<tr>
<td>Cairo2 User study (TraIT)</td>
<td>Validate MSI as a prognostic biomarker for disease progression in metastatic colorectal cancer</td>
<td>Combine experimental and clinical data to perform analyses (survival- and frequency plots), and to make data publicly available.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Import VCF Files</td>
<td>User has one or more VCF files to load into tranSMART, for a single study.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Create Variant Set</td>
<td>Create a set of variants of interest to a study.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Create patient cohort based on variant data</td>
<td>Use genotype data to define patient cohorts</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Summary Statistics for variant data</td>
<td>Show summary statistics for a set of variants and a patient cohort.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Export genotype data</td>
<td>Export genotypes for a patient cohort and variant set.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Perform genotype association analysis</td>
<td>Perform genotype association for a variant set and pair of cohorts.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Perform phenotype association analysis (PheWAS)</td>
<td>Perform phenotype association for a variant set, set of phenotypes, and cohort</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Query tranSMART for cancer specific genotypes</td>
<td>The genetic variant is present in cancer samples but not in normal samples (from the same patient)</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Query tranSMART for gene fusion events</td>
<td>An analyst queries the data in tranSMART to retrieve gene fusion data e.g. from a cancer study or a group of studies.</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Query tranSMART for copy number variation</td>
<td>The CNV may be specific to disease samples over normal samples.</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Apply gene expression signature to predict novel indication for drug</td>
<td>User applies a predictive signature to a set of gene expression profiles from subjects with a variety of disorders in order to identify potential new indication for drug</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Derive Gene Expression Signature to predict drug efficacy</td>
<td>User creates a predictive signature for drug efficacy based on gene expression data</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Microarrays: Samples with Specific aberration</td>
<td>Identify which samples have an ERBB2 amplification</td>
</tr>
<tr>
<td>Differential gene expression</td>
<td>Find out if the STK6 gene is differentially expressed between adenomas and carcinomas in Carvalho et al. 2012</td>
<td>Query aCGH and expression array data from GEO studies to determine differential expression of selected gene between adenomas and carcinomas</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Calculate survival times in colorectal cancer, test biomarker validity in CRC liver metastases and primary tumor.</td>
<td>Combine clinical and immunohistochemistry data, create Kaplan Meier plots, and make data available for others</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Biomarker discovery and validation for colorectal cancer (utilize OpenClinica)</td>
<td>Marker discovery and validation for Colorectal cancer (CRC) in stool samples</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Discovery and validation of biomarkers in blood, tissue and stool samples.</td>
<td>Use a beta-binomial test to normalize, check and determine the cut-off value for over-expression, the fold change and the p-value for protein expression.</td>
</tr>
<tr>
<td>Security</td>
<td>Partitioned Access to trial data use case</td>
<td>Administrators are able to grant access within and across clinical trial data sets.</td>
</tr>
</tbody>
</table>
i2b2 Integration

tranSMART was originally derived from i2b2, but over time has accumulated changes that result in the two systems no longer being compatible. It is a goal of both the Foundation and the i2b2 community to re-integrate these two platforms in a way that utilizes the strengths of both, and that would ensure their ongoing compatibility. Doing so would automatically address some of the requirements above, especially support for longitudinal data and ontologies.

For organizations that already manage an i2b2 installation, the ideal would be for tranSMART to use clinical and other data already in i2b2 as a source of these data, without having to reload them into a tranSMART instance. For organizations that do not have the need for i2b2 itself, there would still be an advantage to having i2b2-compatible data and functionality. It would also make it possible for the organization to adopt i2b2 later, or to access a remote i2b2 installation, for example one being used by a collaborator, in a federated manner.

During interviews, the desire was expressed by organizations who do not currently use i2b2, that tranSMART not require the separate installation and management of an i2b2 instance, if there is not already a reason to install it. From a technical perspective, there are different ways to achieve these somewhat contradictory goals, and the technical proposals address how they would make this possible.

Technology Evaluations

TranSMART conducted technology reviews with many of the potential partners for genomics workflow processing. Several viable technology options have emerged within the last year that manage genome processing (read alignment, variant calling, storage and query). Workflow Systems (such as Arvados and Adam) have particular relevance because they can also store and query the resulting processed data.

The 17.1 Release will focus on the storage and retrieval backend and not on the work flow processing. Releases post 17.1 may have some capability in this space, in particular support for advanced analytics, but 17.1 will focus on storage and retrieval, with a focus on pre-existing standalone or cloud based systems.

Our review confirmed that a unifying theme was that these systems planned to implement the Global Alliance for Genomic Health, GA4GH, APIs that support these functions. In addition to backend systems, cloud providers, and in particular Google, plan to implement these APIs into their cloud offerings.

As part of our evaluations TranSMART evaluated the following offerings and activities:

- Adam / Spark
- Broad Institute GATK and Google Cloud
- Google Genomics platform
Additionally our member meetings confirmed that there was a strong preference for systems that support open API’s or had committed to do so. As a result release 17.1 will focus on systems that support the GA4GH API’s.

**Development Proposals**

The team communicated with a number of potential development partners to solicit proposals. Detailed discussions were held with Harvard, the Hyve, ConvergeHEALTH by Deloitte, Collabora and Igalia. Of these, three groups submitted proposals to the Foundation for consideration: Harvard, the Hyve and ConvergeHEALTH. These proposals are available for review, and are summarized below:

<table>
<thead>
<tr>
<th>Project</th>
<th>Deloitte</th>
<th>Harvard</th>
<th>Hyve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systems Integration and Platform Evolution</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Longitudinal Data</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Scalable Genomics</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>I2b2 integration</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Summary of Development Proposals**

<table>
<thead>
<tr>
<th>Service Provider</th>
<th>Hours</th>
<th>Budget</th>
<th>Blended Rate</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard</td>
<td>3,264</td>
<td>$610K</td>
<td>$167/hr</td>
<td>Systems integration, Scalable Genomics, Longitudinal Data</td>
</tr>
<tr>
<td>Hyve</td>
<td>6,048</td>
<td>$1,067K</td>
<td>$176/hr</td>
<td>Systems integration, Scalable Genomics, Longitudinal Data</td>
</tr>
<tr>
<td>Deloitte</td>
<td>17,000</td>
<td>$2,300K</td>
<td>$135/hr</td>
<td>Systems integration, Clinical data integration, Scalable genomics, Longitudinal data, Cross-study analysis, Performance tuning, regression testing and documentation</td>
</tr>
</tbody>
</table>
Harvard Proposal

The process for re-integrating i2b2 and tranSMART was evaluated at a hackathon in Dec 2014. The goals of the hackathon were to:

- Remove calls from browser to i2b2 through tranSMART proxy.
- Add a communication layer between tranSMART and i2b2.
- Identify differences in database schemas between the two platforms.

The hackathon showed that the reintegration of i2b2 with tranSMART was feasible and tractable, and resulted in an early prototype for this re-integration. The Harvard proposal is to use the results of the hackathon to develop a complete i2b2 integration, with additional work to enable effective use of both platforms.

Installation Scripts
Streamline the installation process and offer automated deployment scripts on AWS and Dockers.

Consolidated Schemas
Ensure that the i2b2 schema remains entirely separate from tranSMART schema, and that both can be updated independently.

Develop i2b2 to tranSMART API
Using results of the hackathon to enable the tranSMART API to allow tranSMART to utilize all the functionality currently available in i2b2 1.7+

tranSMART UI Updates
Ensure a user can generate the same kinds of queries in tranSMART that i2b2 is already enabled to handle.

Incorporate Harvard’s tranSMART Enhancements
- Variant Explorer
- Sample Explorer linked to Dataset Explorer (Analyze tab) – Concept Overlap diagram (D3.js)
- Family Relations
- PheWas
- NLP Data Integration
- System Administrator page

What else do we get?
- Modifiers to support complex data storage and querying.
- More use of temporal facts to create time based queries and proper visit handling.

The Hyve Proposal

The Hyve proposal offers a software development project to the tranSMART Foundation for creating a small, stable (i.e. reliable) and improved tranSMART Core (database + APIs) that can serve as a strong basis for further tranSMART developments:
Improvements to the current data model

- Longitudinal & EHR Data Support: the new clinical data model should be fully compatible with longitudinal and EHR data, leveraging i2b2 (this implies an application-wide change from patient to sample as the granular source dimension).
- Cross Study Support: The current tranSMART data model has a hierarchical structure per study, which is a good principle. However, tranSMART has limited support for doing cross trial analyses. The data model (and ETL blueprint) needs to be updated to allow for proper cross-study concept mapping by design.
- Whole Genome Sequencing Data Support: The data model should have a scalable data layer for supporting the calls and metadata from whole genome sequencing data.
- ETL blueprint: The data model design should provide a clear, unambiguous description of the complete data model, including implied dependencies between tables. This description should provide a complete specification for ETL tools.

Changes to the current APIs

- Core API: Adapt Core API to incorporate support for:
  - longitudinal queries
  - queries based on cross study ontology terms
  - whole genome sequencing data
- RESTful API: Improve performance of the RESTful API implementation, to be measured via the R client.

Non-Functional Requirements for the Core

- Reliability: the result of this project will provide a solid basis for further adoption and growth of the tranSMART community. It will allow for a diverse range of GUI and analysis applications which can be built on top of the APIs, as well as a stable data model than can be used for integration with neighbouring tools in an enterprise IT environment. This will be made explicit by the following:
  - Built-in Security: the data model should contain a clear platform wide study authorization model. The APIs should support integration of enterprise security policies (e.g. the authentication should support several authentication mechanisms such as Kerberos/LDAP).
  - Scalability: the data layer should support storing hundreds of thousands of whole genome sequences.
  - Performance: performance benchmarks for the RESTful API should be defined for the result of this project.
  - Automated Testing: The Core API should have unit and integration tests with a minimal test coverage of 70%. The RESTful API should have automated functional tests for all API calls.
  - Documentation: Comprehensive documentation of the new core to enable easy use of the data core and the APIs
Interoperability: the data model should be aligned to i2b21.7, such that i2b2 and tranSMART share the same basic clinical data model. The data model should allow referencing terms from external (e.g. company wide) ontology servers.

Data Migration: it must be possible to migrate data from an existing tranSMART instance to the new data model (but it should also be suitable for a fresh start).

ConvergeHEALTH by Deloitte Proposal

The proposal from Deloitte addresses the five key areas identified by the requirements process:

- Clinical Data Enhancements
- Support for high volume variant data
- Longitudinal data support
- Cross-study analysis support
- Platform robustness and performance

The figures below show the target architecture and a summary of the features included in each of the five areas.
Clinical Data Enhancement – Overview

Business Problem

- Users need the ability to perform temporal and event-based queries
- Users need the ability to store and query samples and time points and unscheduled visits
- Modifiers are not present in the user interface (e.g., user wants to see which patients had a severe (modifier)/fever (facet)
- Foundation members need backward compatibility from previous versions of trendSMART

Solution Approach

- It’s a shared understanding that the current core API enabled 322 integration meets community member’s need
- Proposed approach is based on extending the core API to make it functionally compatible with 322 1.7 query, and to enable trendSMART to leverage 322 features such as modifiers and event based longitudinal queries for cohort selection
- Data migration scripts will be created to allow in-place data upgrade from the previous trendSMART 16.2 release

Key Activities

- Update to 322 1.7 schema
- Design and fix sample, time points, and categorical concepts in clinical data model
- Core API and REST API extension
- ETL pipeline enhancement
- Produce data migration scripts
- Unit and integration testing
- Documentation of feature and APIs

Support for High Volume Variant Data – Overview

Business Problem

- VariantSMART application does not provide the ability to handle large volumes of genomic variants
- Users need to use variants to define patient cohorts similar to clinical data
- Users should be able to create new variant sets, view them in concept trees, and share access to studies to which the users have access
- Users should be able to query and integrate genomic datasets managed in commercial or open source platforms such as Avadara/ADAM
- Users should be able to use API such as G440H to exchange data with other systems

Solution Approach

- We propose an approach to manage and support variants based cohort selection in the following three operations:
  1. Small, less than 1,200 variants – Load variant set as observation fact in the clinical data store (262)
  2. Less than 9,000 variants – Use SOLR with jaxbb indexing, similar to Harvest’s variant explorer functionality
  3. Full genome sequencing / FASTQ / BAM / VCF – Leverage GMCHAP API and integrate with Avadara/ADAM

Key Activities

- Develop REST API to enable genomic data query and cohort selection
- Develop a genomic data core to enable federated data storage, query, and analytic model for NGS data types via GMCHAP
- Enable integration with genomic analytic platform such as AmiQOSS via G440H API
- Update ontology and cohort selection API for variant set
- Enhance gene signature to support variant signature
- Build SOLR variant index schema and implement variant explorer
- Update existing advanced workflows to support variant sets
Longitudinal Data Support – Overview

Business Problem
- TranSMART application does not provide the ability to perform longitudinal trend analysis because there is no capability to represent data time consistently across concepts
- No summary options for calculating a rate of change across the time course
- Time-based (vs. named categorical attributes) are not supported

Solution Approach
- Temporal query will be implemented as additional filters in the existing tranSMART cohort selection – this is similar to the 252 UI
- Time series and trend analysis will be implemented via the Smart plugin
- To enable existing studies for longitudinal data analysis, the concept of time point needs to be converted to the correct concept type
  - An algorithm needs to be developed to group and convert relative time points to events, and link facts and events together
- Clinical and onco samples will be mapped to the correct time point

Key Activities
- The following 4 use cases will be implemented:
  1. Clinical event based temporal queries similar to existing 252 1.7
  2. Time series and trend analysis using clinical event related attributes for both clinical trials and observational data
  3. Update all advanced workflows to work with longitudinal support-based database changes
  4. Migration / data fixing scripts for studies in existing tranSMART deployment

Cross-Study Analysis Support – Overview

Business Problem
- Users need the ability to merge data from multiple studies for analysis
- Data nodes in each study are coded with study specific concept codes, even though they may represent the same concepts in other studies, this prevents data from being analyzed across multiple studies using the 252 query
- A modifier based cross study explorer was implemented, but its functionality is limited

Solution Approach
- Study specific node concept approach is a workaround for implementing study based query and access control
- Since tranSMART has implemented its own API for querying 252, it becomes possible to use the same concept codes across studies and still maintain the granular access control at the study level
- Cohort selection, summary statistics, and advanced workflow pages need to be updated to add a study condition filter automatically every time a node is added into the query selection
- To enable existing studies for cross-study analysis, all nodes need to be remapped and updated using a standard vocabulary; database scripts will be developed to accomplish this

Key Activities
- Design cross study ontology - adopt standards for cross study concept codes
- Update UI to enable both single study tree and cross study tree in the same explorer
- Update cohort selection, summary statistics, and other UI components to support cross study nodes and "hidden" study filter
- Update application security and access control for single or cross study data access
- Update JS to use sample concept across studies
- Create data migration scripts to update existing study specific concepts with cross study concept codes
Summary and Conclusions

The Foundation has undertaken a project to identify and prioritize the community requirements for the 17.1 release project, and has worked with its members to develop key use cases to exemplify these requirements and new features, and to guide development. The Foundation has also evaluated new technologies that could potentially be incorporated into the platform, to assess their suitability for the 17.1 project. The Foundation also worked with development service providers from the community to obtain an estimate of scope and resources required for the project. A critical remaining task, is to define the resources available for the development project, to finalize the scope of the project based upon these resources, and to organize and manage the development plan. During the course of the project, a number of groups have offered to contribute ‘in kind’ development resources, which will be evaluated and incorporated into the project on a case by case basis.