LabKey Server: An Open Source Platform for Large-Scale, Translational Research

Peter Hussey1, Elizabeth K. Nelson1, Britt Piehler1, Matthew Bellew1, Mark Igra1, Josh Eckels1, Adam Rauch1

LabKey Software, Seattle, WA USA (Contact: peter@labkey.com)

Abstract

Translational research networks engaged in large-scale, collaborative projects face difficult data management challenges. To speed progress towards disease therapies, researchers need software systems that provide robust support for not just clinical and specimen data, but also the complex information produced by modern molecular and cellular techniques. These data sets must be gathered from often geographically distributed labs, screened for quality and consistency, linked with curated public data sources, and shared across the network. Statistical and visualization tools must be able to access the combined data as a basis for forming and testing translational hypotheses. Upon publication, the software system should offer public access to the data, metadata, and tools such that independent reviewers can reproduce the research results– a crucial requirement for the advancement of translational science..

LabKey Server stands out as an open source system that meets these challenges. LabKey Server offers a robust and flexible platform for diverse data integration, secure sharing, direct access from statistical and visualization tools, and assay data annotation for establishing provenance of results. These capabilities make LabKey Server a valuable tool that helps project teams discover insightful answers to translational research questions.

Integration of Diverse Data Types

Lab and clinical data systems differ substantially:

Lab-based data (*e.g.* gene expression, MS2 spectra)
- Evolve rapidly with instrument generations
- Require flexible storage (spreadsheets)
- Require large data storage (binary 'omics files)

Clinical data (*e.g.* demographic, blood tests, biopsies)
- Usually smaller in size and simpler in structure
- Or image data (not usefully aggregated)
- Designed for compliance (21CFR Part 11, HIPAA)

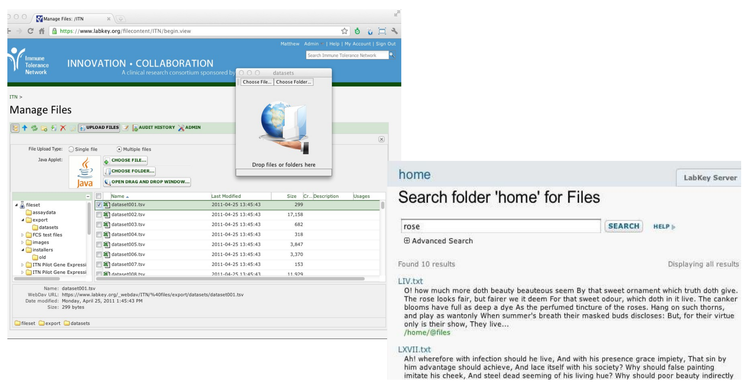

Figure 1: File management and search services
The data produced by lab assays starts out as files, usually of a type specific to the technology of the instrument. LabKey Server supports the standard WebDAV protocol for transferring and managing files over HTTP. The built-in LabKey File Manager provides browser-based UI for uploading files to the server and optionally associating arbitrary metadata properties with the files. At the server, file directories are by default indexed by a built-in Lucene search service, allowing users to find files by content using familiar internet search syntax. The File Manager also supports importing of uploaded files into assay data sets, thereby linking files to data sets that can be tagged with metadata, queried, joined and aggregated by the Query Service.

LabKey Server's two-part solution consists of a file management system that provides upload and search capability (Figure 1), and an ontology-based tabular data store that supports integration and aggregation of results using SQL queries (Figure 2). These two services combine to provide the capacity of a file share, the flexibility of a spreadsheet, the robustness of a SQL database, and the accessibility of a web server.

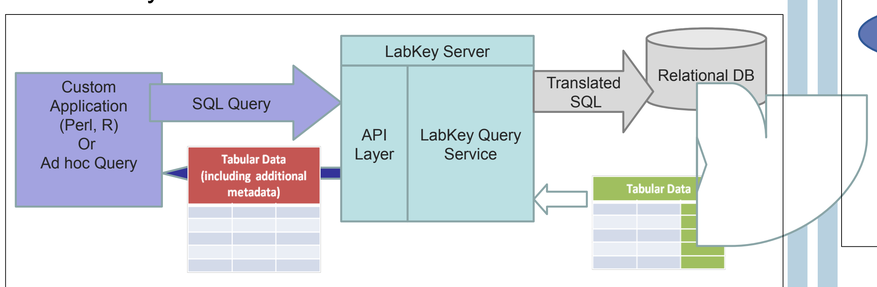

Figure 2: LabKey Server Query Service: *(Clockwise, starting from the left) The user creates a SQL SELECT Query, either implicitly by setting filters and sorts in a Data Grid or custom application, or explicitly by typing ad hoc SQL into the Query Editor. The server verifies the user's permission to read the selected data. The query is then converted into the native SQL of the underlying relational database, including generation of join syntax and expansion of columns whose values are actually stored as rows in the property store. Sending this query to the relational database returns a tabular result to the Query Service. The Query Service then wraps additional metadata around each column and the entire set, for example the base concept domain from an ontology or a formatted link to the original data file. The user then sees the result of the query, and the metadata can be shown by the UI, e.g. as hover text or as a link.*

Secure Sharing Among Collaborators

Distributed teams need to share project data both flexibly and securely. LabKey's role-based security model (Figure 3) supports this requirement. With LabKey, individual labs can have private data folders, then "promote" data to be visible to all.

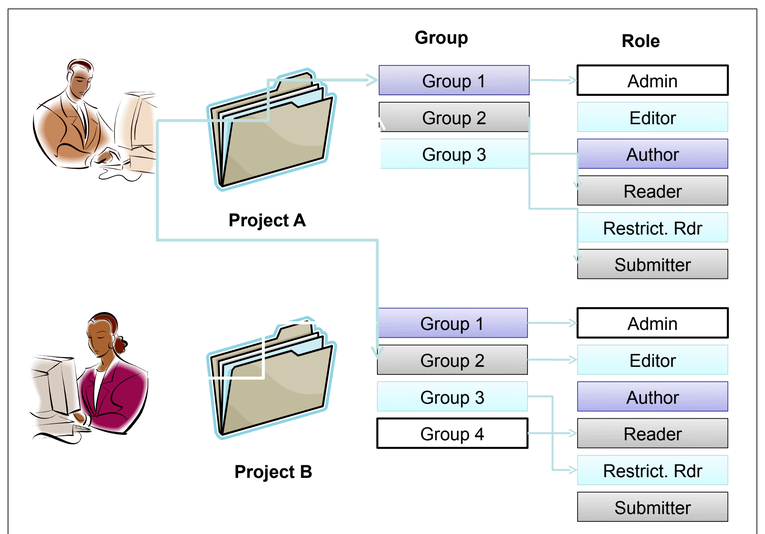

Figure 3: Role-based security model
Users on the left have been authenticated by one of multiple supported mechanisms. When a user accesses any folder in the project, the User ID is checked for membership in one or more groups. Then the User ID and all the groups that contain the User ID are checked for role assignments in the current folder. The combination of the role assignments determine the user's permissions in the folder, from no access to full administrator privileges. When the same users access a different project or folder, their group assignments and roles may be completely different.

Framework for Reproducible Lab Assays

A typical laboratory assay involves the following steps:

- Sample preparation
- Instrument run
- Processing of the raw instrument data to produce biologically meaningful results

The end result data is not complete and reproducible without recording the context elements from the steps above:

- Identity and properties of the original sample
- Sample preparation protocol and experimental conditions
- Identity and settings of the instrument
- Post-processing steps taken, including the versions of tools used and their parameters.

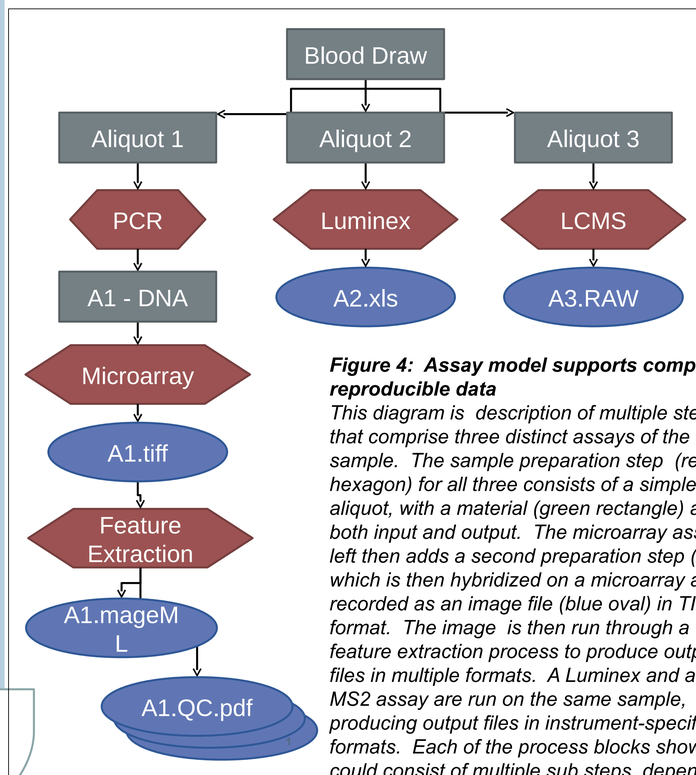

Figure 4: Assay model supports complete, reproducible data
This diagram is description of multiple steps that comprise three distinct assays of the same sample. The sample preparation step (red hexagon) for all three consists of a simple aliquot, with a material (green rectangle) as both input and output. The microarray assay at left then adds a second preparation step (PCR) which is then hybridized on a microarray and recorded as an image file (blue oval) in TIFF format. The image is then run through a feature extraction process to produce output files in multiple formats. A Luminex and an MS2 assay are run on the same sample, producing output files in instrument-specific formats. Each of the process blocks shown could consist of multiple sub steps, depending on the protocol design.

The purpose of LabKey Server's Assay Service (Figure 4) is to provide a mechanism for collecting and storing all of those elements of context. Since the steps, inputs and output vary by assay type, recording the relevant context requires a type-specific template, called an "assay design" in LabKey. The goal of the Assay Service is that the end result data are complete, traceable to their origins, and (ultimately) reproducible.

Access From Scripting and Statistical Tools

LabKey Server's Assay Service matches the assay samples to participants and dates in a clinical research study. Within this context of a study, lab assays can be statistically analyzed across cohorts, ad-hoc groups, or demographic factors.

Taking full advantage of the integrated lab and clinical data requires easy and fast access to the data from familiar tools. The LabKey Server API (Figure 5) is designed to enable tools like R and languages like JavaScript to select data sets that have been loaded into the LabKey Server database. The LabKey Server API works equally well for accessing external relational databases that are maintained by other applications within the research organization.

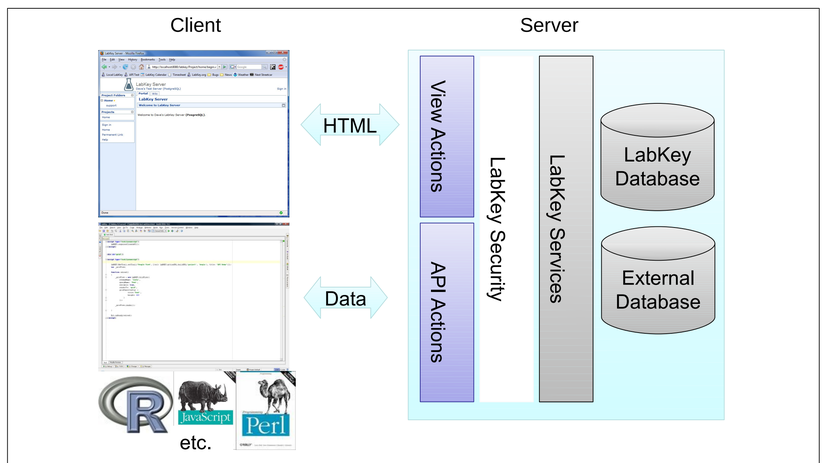

Figure 5: The LabKey API
The core services that make up LabKey Server– query, security, assays, studies– are accessed through a modern browser and/or through the LabKey Server API. The LabKey Server API is a simple, HTTP-based protocol for invoking server actions and for sending and receiving tabular data sets . Protocol libraries implement the client API in the native syntax of popular scripting languages, including R, Perl, Python and Javascript. The Javascript library incorporates the popular ExtJS framework for building dynamic, responsive web pages that request data asynchronously from the server as users interact with the page.

The LabKey Server API makes it easy for researchers to run statistical models in R to assess whether there is enough data to support a hypothesized result from a series of assay runs to a target level of confidence. R and JavaScript can also use the LabKey Server API to generate visualizations of the data (Figure 6) that make it easy to spot anomalies and trends, possibly leading to new insights into disease processes.

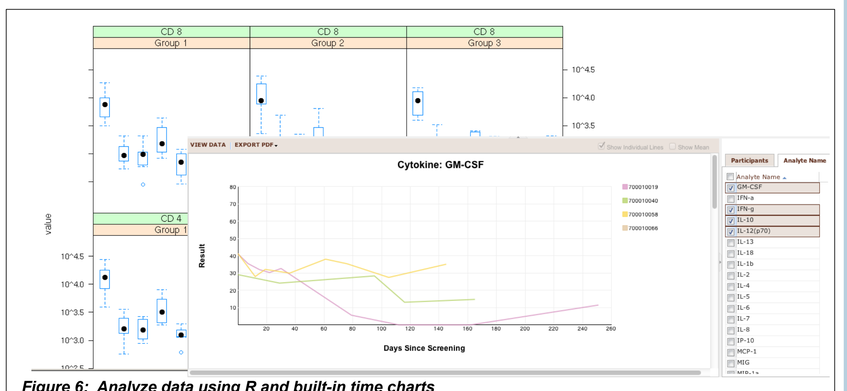

Figure 6: Analyze data using R and built-in time charts
The upper picture demonstrates the tight integration of R with tabular data in LabKey. R analyses and charges can be performed at the server in the context of a tabular data set in the Data Grid, from an R console on a desktop, or on an R Server accessible to the LabKey Server. The lower picture Illustrates a time chart of Cytokine levels in several patients. Similar charts can also be drawn for the mean levels of entire cohorts, or the mean levels in ad-hoc groups of Time charts are written in JavaScript and use the LabKey API to access Query result sets.

Conclusions

LabKey Server's data storage and query capabilities, assay framework, secure accessibility across the internet from tools and scripts as well as browsers, all combine to make the platform an important asset to translational research projects. Its usefulness extends beyond the research phase to publication, where the same capabilities support reproducibility of results by independent reviewers.

Since its launch in 2005, this open source system has been adopted and customized by organizations across the globe, including the Immune Tolerance Network and consortia within the Global HIV Enterprise. As of early 2013, over 70 active installations of LabKey Server leverage more than $15 million of past investment in the platform. Source code, compiled binaries, documentation, and tutorials are professionally maintained by LabKey Software (http://www.labkey.com) and freely available under the Apache 2.0 license at http://www.labkey.org.