Lightweight Data Integration Frameworks for Clinical Research

Ronald Shaker, BS[1], Xenia Hertzenberg, BS[1], James F. Brinkley, MD, PhD[1,2,3]
[1]Departments of Biological Structure, [2]Computer Science and Engineering, [3]Medical Education and Biomedical Informatics, University of Washington, Seattle, WA

Abstract

Research data from a single clinical study is often spread across multiple applications and systems. We present a reusable, lightweight, secure framework for automatically integrating and querying study data from heterogeneous sources in order to answer routine, operational questions for researchers.

Introduction

For any given study in clinical research, multiple applications, each with a specialized function, are often used to track information about patients, their specimen samples and lab results. Simple questions about sample availability for patients requiring additional tests or exhibiting a particular set of symptoms can be difficult to answer. A number of bioinformatics integration platforms currently exist that can facilitate such questions (Amalga, i2b2) but are often costly to install, configure and maintain. The goal of our project is to develop an inexpensive, lightweight, and flexible approach to cross-application data integration. Biomediator Data Integration System (BDIS) serves as a reference implementation for our work. It is an open source, web-based application designed to extract, query and re-package heterogeneous data from a collection of remote sources.

Architecture

Data Acquisition: Sources are required to provide an API for downloading their data as XML. Connectivity to each source is established by specifying a network address, HTTP method, authentication credentials and query parameters. Recurring extractions can be scheduled at regular intervals to provide near real-time reporting.

Queries and Semantic Mappings: BDIS uses a mediated[1] approach to integrating heterogeneous data. A common schema, defined as an XSD, unions the elements of interest from each data source. An XQuery is then used to simultaneously filter and map results onto the mediated schema.

Reporting and Security: XSL transformations may be associated with each query and are automatically applied to XML query results, generating user-friendly formats such as CSV and HTML. Original XML results, as well as transformations, are made available to external clients via a RESTful API. Externally, access is controlled using user accounts and single-access tokens over HTTPS. Internally, each piece of sensitive data is assigned to an owner and group. Owners may share their data by adding users to the appropriate group. User access to data is controlled at the ORM level by use of an ActiveRecord plug-in in Ruby and an added security layer above JPA in Java.

Use Case

To illustrate the use of BDIS, we've developed a small set of queries in support of a pediatrics auto-immunity study. The study's researcher stores their patient visit and lab data in an EDC application called REDCap. The locations of patient blood samples taken during visits are stored in a separate freezer inventory system called rLab. In order to answer questions such as "Are any samples available for patients who still need qPCR testing completed?", lab technicians were required to query both systems separately and merge those results by hand. By translating their question into an XQuery across REDCap and rLab, manual reconciliation is avoided. In fact, by providing the end-user with a generic query and transformation that combines all data from both systems into a single spreadsheet, previously difficult questions become easy for end-users to answer without additional programmer support.

Conclusion

BDIS enables researchers to more easily answer cross-application queries about their data. Our framework's inexpensive, minimalist approach to data integration provides re-usable set of building blocks for merging data from collections of heterogeneous sources. *Funded by NIH grants RR0254, AR051545, NS055088.*

References

1. Mork P, Halevy A, Tarczy-Hornoch P. A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases. Proceedings of the AMIA Annual Symposium; 2001 Nov 3-7; Washington D.C..