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

Introduction

In clinical research studies, multiple applications (i.e. data sources), each with a specialized function, are often used to track information about patients, their specimen samples and lab results. Simple questions regarding a patient's records can be laborious to answer. A number of bioinformatics integration platforms currently exist that can facilitate the answering of such questions (Amalga, i2b2) but are often costly to install, configure and maintain.

We present a reusable, lightweight, framework and workflow that facilitates the continuous integration and querying of study data from heterogeneous sources in order to answer routine, operational questions. The goal of our project is to develop an inexpensive and flexible approach to cross-application data integration that will provide informaticists with the tools they need to assist researchers. Biomediator Data Integration System (BDIS) serves as the reference implementation for our work.

Fig. 1A.) XML schema for REDCap.

Fig. 1B.) XML schema for rLab.

Fig. 1C.) Combined fields of interest from REDCap and rLab.

Methods

We begin by narrowing requirements for the applications we wish to integrate to a reasonable scope. By keeping expectations simple, data sources not meeting our basic criteria may more easily be altered to provide access to their data.

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 (Fig 1). An XQuery is then used to simultaneously filter and map results onto the mediated schema (Fig 2). XML is used as the output syntax of all data sources. Each source must formally define its output schema in the form of an XSD. Schema definition promotes semantic understanding of the data source, provides validation of input data sets, and can also facilitate the automated generation of user interfaces for filtering and constraining query results.

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.

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.

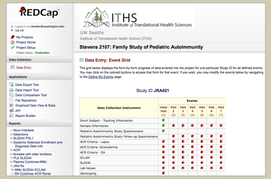

Fig. 2A.) REDCap is a research electronic data capture application used to track patient visit information and lab test results.

Fig. 2B.) rLab is a freezer manage application used to track patient sample types, availability and location.

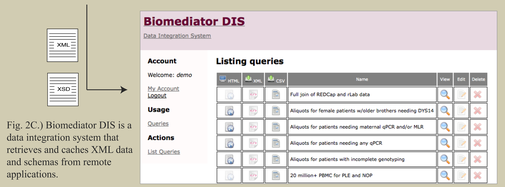

Fig. 2C.) Biomediator DIS is a data integration system that retrieves and caches XML data and schemas from remote applications.

Fig. 2D.) XQuery is used to merge and map multiple data sources onto a mediated schema, and answer researchers' questions about their patient data.

```
let $rl_raw := doc("http://biomdis/query/download?id=1&amp;user_credentials=VGP9o2IvcR7")
let $rc_raw := doc("http://biomdis/query/download?id=2&amp;user_credentials=VGP9o2IvcR7")
let $rows := for $item in $rc_raw/records/item
    return
        for $aliquot in $rl_raw/families/family/subject/sample/aliquot
            where local:is-same-family($item, $aliquot/../../../family-code)
            and local:is-same-visit($item, $aliquot/../sample-date)
            return
                <row>
                    <subject-code>{$aliquot/../../subject-code/string()}</subject-code>
                    <sample-code>{$aliquot/../sample-code/string()}</sample-code>
                    <sample-date>{$aliquot/../sample-date/string()}</sample-date>
                    <sample-type>{$aliquot/../sample-type/string()}</sample-type>
                    <sample-comments>{$aliquot/../comments/string()}</sample-comments>
                    <aliquot-type>{$aliquot/aliquot-type/string()}</aliquot-type>
                    <cell-count>{$aliquot/cell-count/string()}</cell-count>
                    <box>{$aliquot/box/string()}</box>
                    <col>{$aliquot/col/string()}</col>
                    <row>{$aliquot/row/string()}</row>
                    <state>{$aliquot/state/string()}</state>
                    <aliquot-comments>{$aliquot/comments/string()}</aliquot-comments>
                    { for $field in $item/*
                        return
                            $field }
                </row>
return <spreadsheet>{$rows}</spreadsheet>
```

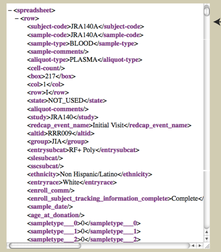

Fig. 2E.) XML result sets may be used to create a new data source that can be consumed by Biomediator and other data integration systems.

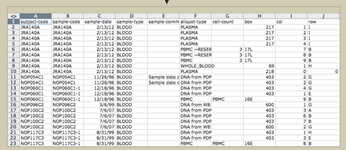

Fig. 2F.) Result sets may be transformed into other formats such as CSV (spreadsheet) and HTML.

Use Case

To illustrate the use of BDIS, we've developed a small set of queries in support of a pediatrics auto-immunity study (Fig. 2C). The study's researcher stores their patient visit and lab data in an EDC application called REDCap (Fig 2A). The locations of patient blood samples taken during visits are stored in a separate freezer inventory system called rLab (Fig 2B). 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 data downloaded from 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 (Fig 2D), we've found that end-users adept with programs like Microsoft Excel are able to answer previously difficult questions with relative ease.

Conclusion

The Biomediator Data Integration System 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.

Acknowledgements

Funded by NIH grants RR0254, AR051545, and 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..