Lightweight Distributed XML-Based Integration of Translational Data

Joshua D. Franklin, MS [1], Landon T. Detwiler, MS[1], and James F. Brinkley, MD, PhD[1]
[1]Structural Informatics Group, Department of Biological Structure, University of Washington, Seattle, WA

Abstract
A distributed XQuery engine sends sub queries to separate XML data sources, and then combines the results into a single XML composite result. The system is lightweight in that it is very simple to add a new data source. An illustrative example is given for integrating data from an electronic data capture (EDC) system and a separate specimen management system.

Introduction
As one attempt to fill the gap between common ad hoc methods for data integration and larger-scale but more resource intensive methods, we are developing a lightweight distributed query based approach based on our DXBrain data integration system for brain mapping[1]. In this system the user writes a distributed XQuery, which is then sent to various XML sources by a distributed XQuery engine we have developed. XML snippets from each source are then combined by the distributed XQuery engine into a single XML result which can optionally be transformed into an HTML table or exported as a CSV file.

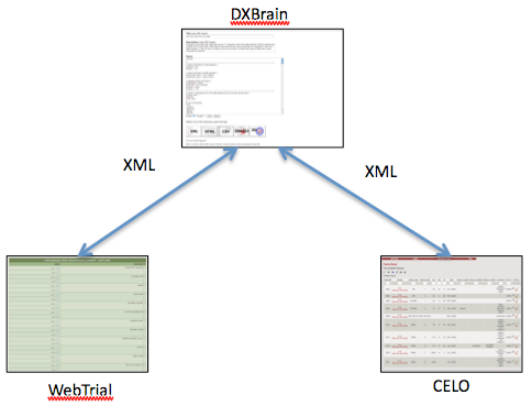
Example Application
In an example study, case report form (CRF) data from individual subjects are stored in our WebTrial electronic data capture (EDC) system, whereas the freezer locations of blood sample aliquots acquired from these subjects are stored in a separate specimen tracking system called CELO. On occasion the investigator would like to retrieve aliquots for further study that correspond to specific clinical conditions captured in the CRF's. An example query would be to find the freezer locations of all unused aliquots taken from pediatric lupus (PLE) patients who have experienced Raynaud's phenomenon. Our distributed XQuery first queries the WebTrial database using its CDSIC Operational Data Model (ODM) web service interface to find the ID's of all PLE subjects with Raynaud's (highlighted in the XML result below), and for each of these subjects queries the CELO database for unused aliquots.

Discussion
Although the above query is very simple, in our brain mapping work we have shown that any number of sources can be added. Unlike heavier weight data integration systems it is very easy to add a new source, (including RDF/OWL ontologies queryable via SparQL since the returned RDF is also XML), by simply including a subquery to the new source in the distributed XQuery. The tradeoff is that the user must know the XQuery language and the schemata of the sources. Although we are developing methods for addressing this tradeoff our experience has shown that it is not a great burden for a knowledgeable user with a small number of sources. For EDC data, adoption of the CDISC ODM standard allows us to leverage XQuery code across separate systems. Thus, this approach may be a viable method for filling the gap between small-scale ad hoc data integration methods, and larger-scale but heavyweight approaches.

Supported by NIH grants HL087706 and UL1RR025014.

References

1. Detwiler LT, Suciu D, Franklin JD, Moore EB, Poliakov AV, Lee ES, Corina D, Ojemann GA, Brinkley JF. Distributed XQuery-based integration and visualization of multimodality data: Application to brain mapping. Frontiers in Neuroinformatics 2009;3(2).




```
<sample subject_id="PLE 20" sample_id="422">
  <Aliquot>
        <AliquotID>6354</AliquotID>
        <aliquot_type>WB</aliquot_type>
        <box>02</box>
        <row>A</row>
        <col>09</col>
        <state>NOT_USED</state>
  </Aliquot>
…
```