Project Proposal

EN605\_452 - Summer 2010

Fred Criscuolo

fcriscu1@jhu.edu

**Summary: an ontology-specific data source based on core data from the National Cancer Institute (NCI) supplemented by an expandable list of federated bioinformatics data sources. Access to the proposed data source will be via SOAP-based Web services. Retrieved data will be persisted in structured XML format. A client application that will demonstrate a subset of the data source's capabilities will also be provided**.

Primary Project Objects:

* Use NCI published cancer data to establish a core relational database
* Use federated data sources and services to supplement these data
* Persist mined data in structured XML format and support queries via XQuery
* Provide client access to data via SOAP-based Web services
* Demonstrate that an XML persistence model improves database and application flexibility
* Database: Oracle 11*g*R2
* Primary coding language: Java (JDK 1.6)
* Application server: Apache Tomcat & Axis or Sun Glassfish v3
* Database design tool: Oracle JDeveloper or Oracle SQL Developer
* Single developer

The National Cancer Institute publishes the Cancer Gene Index (https://cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index), through the distribution of two (2) structured XML files. These files contain data for approximately 7000 human genes that have been associated with cancer. These data are acquired by mining and validating Medline entries. Individual gene entries have cross-references to several major bioinformatics data sources including HUGO, EntrezGene, RefSeq, UniProt, and OMIM. Most of the data within the XML elements are based on controlled vocabularies supported by NCI's Enterprise Vocabulary Services (http://evs.nci.nih.gov).

The proposed project application would utilize the data within the XML files to establish the core of a relational database[[1]](#footnote-1). The application would then provide access to data and services provided by other bioinformatics resources identified by the noted cross references. To a limited extent, these cross reference would in turn be used to access additional data and services as a proof of concept. Data obtained from these federated sources will be cached within the application's database for a limited period of time. In addition to obtaining these data in response to queries, the data will be available for subsequent service requests. For example, if a query response includes sequence data, that data can be used for a supported sequence-based service (e.g. BLAST). If these data are no longer cached within the local database, they will be retrieved without user involvement.

The intended end user is a bioinfromatician who wants to utilize an oncology-focused portal to obtain access to a variety of bioinformatics services and data. It is anticipated that this user will develop client applications that will deliver these federated data and services to scientists through end user applications.

An important implementation characteristic of this project will be to persist much of the mined data in structured (i.e. schema-based) XML format. Both Oracle 11*gR2*  and Microsoft SQL Server 2008 provide support for persisting and querying data in native XML. After reviewing the capabilities of both products, it would appear that Oracle provides more support for XML data and represents the better choice. This implementation decision was made in order to address a valid observation noted in the Lacroix and Critchlow text: that the long development times and inherent inflexibility of large databases and supporting applications are incompatible with the more rapid pace of scientific discovery. The result, as the authors note, is critical data "hidden" within spreadsheets on a scientist's computer. An objective of this project will be to utilize XML as a means of increasing database and application flexibility. This will be demonstrated by comparing the level of effort required to add additional federated data sources as well as changing the data content provided from these sources to a traditional relational model.

The provenance of all data retrieved from federated sources will be persisted and made available to clients.

As noted, client access to this application's services will be via SOAP-based Web services. It is assumed that users will register and provide an id within service requests. This will facilitate access to cached data and allow a user to annotate data. A sample client application will be provided to demonstrate a subset of published services.

Planned Deliverables:

1. Database ERD

2. XSD documents for all schemas

3. All SQL code (i.e. create scripts, stored procedures)

4.All Java source code (server and client)

5. WSDL files

6. XQuery scripts

7. Design document including screen shots of client operation.

8. Possible remote access to application

I will be the sole developer for this project.

1. The percentage of the NCI data supported may be limited based on available computer resources. [↑](#footnote-ref-1)