Initial Lessons for Developing a User-Interface for Querying Federated Heterogeneous Data Sources

Ramkiran Gouripeddi, MBBS, MS\textsuperscript{1,2}, Randy Madsen, BS\textsuperscript{1}, Richard L. Bradshaw, MS\textsuperscript{1}, N. Dustin Schultz, MS\textsuperscript{1}, Ryan Butcher, MS, MBA\textsuperscript{1}, Phillip B. Warner, MS\textsuperscript{1}, Peter Mo, MS\textsuperscript{1}, Bernard A. LaSalle, BA\textsuperscript{1,2}, Julio C. Facelli, PhD\textsuperscript{1,2}

\textsuperscript{1}Department of Biomedical Informatics, \textsuperscript{2}Center for Clinical and Translational Science, University of Utah, Salt Lake City, UT

Abstract: Most query tools for clinical researchers have been designed for querying data that have been homogenized into a common data model. Often a single data source cannot provide all the necessary detail to holistically describe a patient or any other clinical entity of interest to researchers. Likewise, Extracting, Transforming, and Loading several data sources into a common data model may not be possible. Therefore, clinical research of this type requires federation of heterogeneous data sources. Initial results from our user query and behavior analysis is informing our future development of interactive query tools that support the requirements of heterogeneous data.

Introduction: OpenFurther is an informatics platform that supports federation and integration of data from heterogeneous and disparate data sources. It has been deployed at the University of Utah (UU) as the Federated Utah Research and Translational Health e-Repository (FURTHeR)\textsuperscript{1,2} since August 2011 and is available for use by all UU employees and students. OpenFurther links heterogeneous data types, including clinical, public health, biospecimen and patient-generated data; empowering researchers with the ability to assess feasibility of particular clinical research studies, export biomedical datasets for analysis, and create aggregate databases for comparative effectiveness research. With the ability to link unique individuals from these sources, OpenFurther is able to identify cohorts for clinical research. It provides semantic and syntactic interoperability as it federates health information on-the-fly and in real-time and requires neither data extraction nor homogenization by data source partners, facilitating integration by retaining data in their native format and in their originating systems. For our initial development we utilized a modified version of the i2b2 web user interface\textsuperscript{3} which a researcher can use to design a clinical research query. The user interface (UI) consists of complete hierarchies of standard terminologies such as the International Classification of Diseases (ICD) as we currently do not have a priori knowledge of any changes in the data availability and to accommodate source-wise variations in data availability. As of June 2013, FURTHeR has a diverse user base of 324 individuals who have run 2042 queries against its data sources. Users are trained to use the tool through classes provided by the UU’s Health Sciences Library. Additional instructive videos are available within the query tool itself. The main terminology nodes on the ontology navigation panel of i2b2 have information buttons (IB) adjacent to them that provide definitions and the availability of concepts within each data source. In addition, the OpenFurther development team assists researchers with their queries when requested.

Methods: OpenFurther stores logs for all the user queries, which include query attributes such as concepts, data sources and temporal parameters; and query run statistics such as time to completion and results obtained. A query fails to return a result if a user selects a concept that is not supported by a particular data source. We analyzed these logs to understand reasons for failed queries.

Results: We found that 33\% of our users’ queries resulted in failures. Some of these failures (14\%) were due to technical reasons, including errors during query or result translation from data source terminologies to the standard terminologies used within OpenFurther, queries timing out and data source unavailability at the time of execution. However, the remaining majority of query failures (86\%) were due to queries that included a data concept that is unsupported by a selected data source. Such examples include requesting laboratory result concepts from a public health data source that doesn’t have laboratory results or selecting a cause of death concept against a clinical data source that does not support cause of death (Figure 1). Initial usability studies including task analysis\textsuperscript{4,5} and user interactions have shown that many of these failures occur when users search for concepts.
**Discussion:** Knowledge-anchored query paradigms using standard terminology have been studied\(^6\)\(^,\)\(^7\) and deployed in various clinical applications. Most of these applications work against data stored in a common data model. Users of such applications usually have a good understanding of these data and possess accurate mental representations to build queries against those data. In our analysis we found that our users have difficulty understanding heterogeneous data when using the current query tool due to its limited support for presenting context and additional metadata. The context was even more opaque when users searched for their query criteria, as the context that is inherently available when navigating the hierarchy of the ontology was not provided. For example, our UI uses ICD9 for representing diagnoses and ICD10 for representing cause of death. When navigating the terminology hierarchy on the UI, the root nodes of ‘Diagnosis’ at the top of the ICD9 (diagnosis) sub-tree and ‘Cause of Death’ at the top of the ICD10 (cause of death) sub-tree provide context to the underlying concepts. This context provided by the concept hierarchy along with the information provided by the IB is not available within the search results leading to ICD10 concepts being used as diagnosis criteria in queries. Our findings show that tools working with heterogeneous data require sharing more complex data model representations with end users. Results from this study have given us a better understanding of this emergent phenomenon\(^5\) and will inform us in the design and development of improved representations of heterogeneous data on user interfaces (UI). Some considerations for future UI development\(^8\) within OpenFurther include providing the user with sufficient context and metadata (including data availability and quality at each data source) for each concept and query attribute included in a query, and visual feedback at every stage of query building and processing. At the same time our future design efforts must ensure that UI real-estate is not cluttered and workflows consider cognition of the users by not overloading their working memory, forcing them to recall information, and instead rely on recognizing query constructs. Future work will also include capturing the user’s intent in querying in order to understand if the query results met the user’s needs.

**References:**