Data Integration in the Biomedical Informatics Research Network (BIRN)*

Vadim Astakhov, Amarnath Gupta, Simone Santini, and Jeffrey S. Grethe

University of California San Diego, La Jolla, CA 92093, USA
{astakhov, gupta, ssantini}@ncmir.ucsd.edu

Abstract. A goal of the Biomedical Informatics Research Network (BIRN) project sponsored by NCRR/NIH is to develop a multi-institution information management system for Neurosciences, where each participating institution produces a database of their experimental or computationally derived data, and a mediator module performs semantic integration over the databases to enable neuroscientists to perform analyses that could not be executed from any single institution's data. This demonstration paper briefly describes the current capabilities of Metropolis-II, the information integration system for BIRN.

1 Introduction

The goal of the data integration system for the Biomedical Informatics Research Network (BIRN) (`www.nbirn.net`) is to develop a general-purpose information integration framework which diverse groups of neuroscientists can use for a variety of application problems that arise from different scientific research needs. This framework is designed to support a number of neuroscience research test beds. In the setting of the *mouse BIRN* test bed, a large number of very different information integration applications may need to be designed over a slowly increasing set of very heterogeneous data sources. The data to be integrated range from 3D volumetric data of nerve components, to image feature data of protein distribution in the brain, to genomic data that characterize the anatomical anomalies of different genetically engineered mouse strains and so forth, and there are a number of integrated schemas over different combinations of these sources designed for different study groups. In contrast, the integration requirement of the *human morphometry BIRN* and *human functional imaging BIRN* test beds have a single virtual schema collectively developed by the participating research groups, and an increasing number of research universities are contributing their data to this schema. The data provided by these test beds are mostly deidentified patient records for patients with neurodegenerative diseases, containing, for instance, demographic data, psychological evaluations and medical imaging analyses.

* This work is supported by NIH BIRN-CC Award No. 8P41 RR08605-08S1, NIH Human Brain Project Award No. 5RO1DC03192.

B. Ludäscher and L. Raschid (Eds.): DILS 2005, LNBI 3615, pp. 317–320, 2005.

Given this application context, the data integration framework of BIRN consists of a *global-as-view* mediator called Metropolis-II, a number of specialized tools for schema registration, view definition and query building, a number of domain-specific clients, and a set of tree and graph structured ontologies that supply intermediate information such that integrated views can be defined over the sources. Using the external ontologies to integrate information is our way of implementing *semantic integration* [1]. The overall architecture of the system is given in Figure 1.

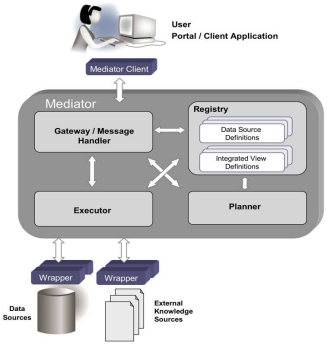

Fig. 1. The general architecture of Metropolis-II

2 Integration Framework

Data Sources. Metropolis-II makes the assumption that a data source is relational that may have a binding pattern for every exported relation. Every schema element $relations, attributes$ has a descriptor for keyword search, and a so-called *semantic-type* that can be used to map the element to an ontology [2]. Further, a data source may export a set of functions that are internally treated as relations with the binding pattern $(\bar{b}, f)$ where $\bar{b}$ represents a set of bound arguments and the single f is the free output variable of the function. In Mouse BIRN, for example, specialized functions are used to compare the distributions of proteins in a set of user-specified regions in the brain. Using this model also enables us to treat computational sources such as the R statistical package as a "data source" that contributes only functions and no relations. Integrated views are written using standard data sources as well as these functions. We have also designed source-specific wrappers for sources such as Oracle, Oracle Spatial, and Postgres, where a generic query can be translated into the appropriate flavor of SQL, and functions supported by the specific systems.

Ontological Sources. We use the word ontology here to mean a term-graph whose nodes represent terms from a domain-specific vocabulary, and edges represent relations that also come from an interpreted vocabulary [3]. The nodes and edges are *typed* according to a simple, commonly agreed upon set of type produced by test bed scientists. The most common interpretation is given by rules like the transitivity of *is-a* or *part-of* relations, which can be used, for example, to implement inheritance. However, there are also domain specific rules for relationships such as *volumetric-subpart: brain-region* $\rightarrow$ *brain-region* and *measured-by: psych-parameter* $\rightarrow$ *cognitive-test* that need special rules of inference. For example, if a brain-region *participates-in* a *brain-function* (like "working memory"),

and the brain-function is *measured-by* a *cognitive-test*, then the cognitive-test *functionally-tests* the brain-region. Currently, ontologies are represented as a set of relations reflecting the set of nodes and their properties, a set of edges, and a set of edge properties. Also other operations, including graph functions such as path and descendant finding, and inference functions like finding transitive edges are implemented using an API of functions, as described in the previous paragraph.

View Definition and Query Languages. The query language for Metropolis-II is the union of conjunctive queries, which may contain function terms, as well as the standard aggregate functions. The syntax of the language, expressed in XML, is essentially that of Datalog with aggregate functions [4]; essentially, a query has the form $q(X, F(Y)) : -r_1(X, Z), r_2(Z, Y)$ where $F(Y)$ is the aggregate function operated on sets of Ys and X is a (in reality, a set of) group-by variable. The query planner and execution engine in Metropolis-II translates this expression to

$$q'(X, Y) : -r_1(X, Z), r_2(Z, Y)$$

$$q(X, W) : -F(gb(q'(X, Y))$$

where the the group-by function gb followed by the aggregate function F is pushed together to the data source whenever possible, and are otherwise evaluated at the mediator. The language also admits nested queries, where inner queries are assigned to intermediate relation variables, that are used by the main query. The view definition language for the system, on the other hand, does not allow aggregates and nested queries at the present time. The language allows only safe negations, where all variables in negated predicates are bound.

Mapping Relations. In the current GAV setting of the mediator, the burden of creating proper integrated views over data sources is on the integration engineer who works with the domain scientists to capture the requirements of the application at hand. This often leads to the pragmatic problem that the relationships between the attributes exported by different sources and those between the data values are, quite often, not obvious. To accommodate for this, the recent version of the system [5] has created additional mapping relations. Currently there are three kinds of mapping relations. The *ontology-map* relation that maps data-values from a source to an ontology term of a known ontology (like the Unified Medical Language System from the National Library of Medicine). The *joinable* relation pairs attributes from different relations if their data type and semantic types match. The *value-map* relation maps a *mediator-supported data value* or a *mediator-supported attribute-value pair* to the equivalent value (resp. attribute-value pair) supported by the source. For example, the mediator may export a demographic attribute called gender with possible values {male, female}, while one source may refer to it as attribute sex with possible values {0, 1}, while another may call it kcr_s57 with the domain {m, f}. The Metropolis-II planner uses a look-up function to make a substitution before dispatching the query plan to the execution engine.

Authentication and Authorization. Access control is a very important aspect of a practical information integration system. For BIRN, this is accomplished in two stages – defining authenticable users, and the implementation of authorization that enables a user to perform only the tasks she is permitted to. The authentication function is handled outside the mediator by a community authorization service. The authorization is handled through an additional access control database that is implemented inside the mediator.

3 The Demonstration

The demonstration will present to the user the information integration system together with the different clients for tasks performed by the submitter of a newly joining source, and integration engineer. These tasks include schema registration, integrated view design, ontology browsing and query design. A number of different query clients are designed for different user groups, and walk through the different stages of query execution in the system. This will include the XML-encoded query language and the view-definition language of the mediator, the plan generated by the system, the communication between the mediator and the different wrappers. As part of this walkthrough, we would also demonstrate how we have used the statistical package R as a computation resource accessed through the mediator. In this process, we will also illustrate the different kinds of data sources and different classes of queries the system can handle.

Acknowledgments. David Little, Maryann Martone, Robin Park, Xufei Qian, Edward Ross, Joshua Tran, Yujun Wang, Wai-Ho Wong, Aylin Yilmaz, Ilya Zaslavsky are the BIRN R&D team. Bertram Ludäscher contributed to the basic research and first version of the system.

References

1. Ludäscher, B., Gupta, A., Martone, M.E.: Model-based mediation with domain maps. In: Proc. 17th Int. Conf. on Data Engineering (ICDE), Washington, DC, USA, IEEE Comp. Soc. (2001) 81–90
2. Gupta, A., Ludäscher, B., Martone, M.E.: Registering scientific information sources for semantic mediation. In: Proc. 21st Int. Conf. on Conceptual Modeling (ER), London, UK, Springer-Verlag (2002) 182–198
3. Gupta, A., Ludäscher, B., Martone, M.E.: Knowledge-based integration of neuroscience data sources. In: Proc. 12th Int. Conf. on Scientific and Statistical Database Management (SSDBM'00), Washington, DC, USA, IEEE Comp. Soc. (2000) 39–52
4. Zaniolo, C., Arni, N., Ong, K.: Negation and aggregates in recursive rules: the ldl++ approach. In: Proc. 3rd Int. Conf. on Deductive and Object-Oriented Databases (DOOD), Springer-Verlag (1993) 204–221
5. Santini, S.: Metropolis MkIII: Prolegomena to design. Technical Report 04-04, Dept. of Neurosciences, Univ. of California San Diego (2004)