# A System for Managing Alternate Models in Model-based Mediation

Amarnath Gupta Bertram Ludäscher Maryann E. Martone*

Xufei Qian Edward Ross Joshua Tran Ilya Zaslavsky

San Diego Supercomputer Center
University of California San Diego
{gupta,ludaesch,xqian,eross,trantj,zaslavsk}@sdsc.edu
(*)mmartone@ncmir.ucsd.edu

**Introduction.** In [1,3], we have described the problem of *model-based mediation* (MBM) as an extension of the global-as-view paradigm of information integration. The need for this extension arises in many application domains where the information sources to be integrated not only differ in their export formats, data models, and query capabilities, but have widely different schema with very little overlap in attributes. In scientific applications, the information sources come from different subdisciplines, and despite their poorly overlapping schema, can be integrated because they capture different aspects of the same scientific objects or phenomena, and can be conceptually integrated due to scientific reasons. In the MBM paradigm, a "mediation engineer" consults with domain experts to explicitly model the "glue knowledge" using a set of facts and rules at the mediator. Integrated views are defined in MBM on top of the exported schemas from the information sources together with the glue knowledge source that ties them together. We have successfully applied the MBM technique to develop the KIND mediator or Neuroscience information sources [1,2,3,4]. To accomplish this, sources in the MBM framework export their conceptual models (CMs), consisting of the logical schema, domain constraints, and ***object context*s**, i.e., formulas that relate their conceptual schema with the global domain knowledge maintained at the mediator. Thus model-based mediation has a hybrid approach to information integration – on the one hand at the mediator integrated views are defined over source CMs and the Knowledge Map using a global-as-view approach; on the other hand, object-contexts of the source are defined as local-as-view.

In the demonstration, we present the KIND2 system, a further extension to the MBM paradigm – here the mediator may have *alternate* sources of glue knowledge, often developed by consulting different domain experts. We presume there are no contradictions between recorded or derived facts from different sources, and show how integrated views are defined and queries are evaluated in the presence of alternate knowledge sources in an MBM setting.

## The Demonstration System

**Data and Knowledge Sources.** The KIND2 system mediates across a number of neuroscience data sources provided to us by different partner institutions. The data consists of relational sources containing image and volume-based measurements, XML sources containing protein information and time-series sources containing physiological recordings from neural responses for specific stimulations. The mediator uses two forms of knowledge sources to integrate this information – a graph structured ontology (stored and displayed to the user as a labeled graph) and a spatial atlas of the brain:

- The ontology is constructed from the Unified Medical Language System[1] ontology from the National Library of Medicine and the Gene Ontology from the Gene Ontology Consortium[2]. The two ontologies together store about 2 million concept names (nodes) and 10 million relationships (edges) represented as relational tables in an Oracle8 database. They are accessed through FLORA, an F-logic engine built on top of XSB Prolog [5]. The F-logic engine allows the definition of (often *highly* recursive) views. In our setting, it also pushes certain operations (e.g., SPJ and some hierarchical queries) to the Oracle system below, and performs deductive computations on the results.
- The spatial atlas consists of a number of layers, where a layer is an orthographic cross-section of the brain, on which the observed structured on the section are outlined and labeled. Obviously, since most structures in the brain are three-dimensional, they appear on multiple layers. Our atlas source is created from commercially available brain atlases, by converting line drawings to polylines and polygons in Oracle Spatial Data Cartridge. Using this system, one can perform two-dimensional topological and metric queries on atlas objects. We have developed additional algorithms to simulate some three-dimensional operations as stored PL-SQL procedures on top of the system's native two-dimensional query capabilities.

The demonstration system will show how object contexts are defined from both of these knowledge sources, by illustrating how a user can navigate the knowledge sources themselves "looking for" data sources that are reachable from subgraphs of the first source, or subregions of the second source.

**The KIND2 Mediator.** The primary mediator module in the KIND2 system is built using F-Logic. Data sources register with the mediator by wrapping their native schema into F-Logic. The query capabilities of data sources are modeled by source-specific special predicates. For example, the volume analysis data source for morphometry supports an operation for spine density distribution which, given a user-specified interval along the length of a dendrite, returns an XML document containing frequency histogram of spine density in that interval. The mediator views this operation as a built-in predicate with binding patterns for the input and output parameters.

---

[1] http://www.nlm.nih.gov/research/umls/index.html

[2] http://www.geneontology.org

Integrated views in KIND2 are defined as set of F-logic rules. We illustrate the use of alternate knowledge sources by defining intensional predicates both in terms of logical and spatial operations. For example, the predicate **contains_tc**(object1, object2) can be defined as a transitive closure on the predicate **contains**(object1, object2) maintained by the ontology. Given that the UMLS ontology knows contains(thalamus, 'fasiculi thalami') and contains('fasiculi thalami', 'anterior peduncle') are true, the system infers **contains_tc**(thalamus, 'anterior peduncle'). Alternately, using the spatial atlas (Figure 1), one can use a spatial operator named **inside**: *polygon* → **setof***(polygon)*. The atlas will look for the polygon labeled as thalamus in each slice, and geometrically locate all the other labeled polygons inside it, thus finding the 'anterior peduncle'. Using either method, one may define an integrated view on the data sources. The view may be a selection on all proteins *P* involved in some activity *A* of some neurons *N* and that can be localized in region *R* of the brain. Here, the activity *A* is defined in terms of the time-series recording of neurons, the protein properties are retrieved from a protein database[3], and the protein localization information is available both from whole-brain experiments and neuron-level experiments. The role of the ontology and the atlas is to provide two missing pieces of information to construct the view: (a) the subregional architecture of the brain, and (b) situating specific neurons in specific brain regions. Given a query against this view the mediator needs to rewrite it using the predicate **contains_tc**, the operation **inside** or both. The decision is based on several factors including:

- whether the two knowledge models have equal granularity of information for the query region
- whether there are any data sources that refer to only one model for the query region
- whether other predicates in the query necessitate visiting one source over the other
- the estimated cost of the two operations for the query region
- Often a good solution is to partially execute the query using one source, obtain subregion names, and pass them to the other source to complete the query.

In the demonstration system we will show the system's query evaluation functionality with a *plan-viewer tool*. The plan starts from the user's query and first selects all matching views that support the binding pattern of the query. Once the views are identified, the user of the demo system may choose any view to unfold in order to continue query processing. We will demonstrate the case of the alternate view definitions described earlier. The tool provides a simple graphical interface to demonstrate a trace of the query planning process. Given an arbitrary query against a given view, the demo user may choose to see all generated plans and the plan chosen by the mediator. The demo user may also select a plan from the initial set of plans, and trace when it gets eliminated. In this case, the system will show how chosen plan is selected over other plans or is pruned by a competing plan. This will be shown using a trace of the query processing rules that were applied to prune one plan over another.

[3] For example, consider the web-accessible database for calcium-binding proteins located at http://structbio.vanderbilt.edu/cabp_database/cabp.html

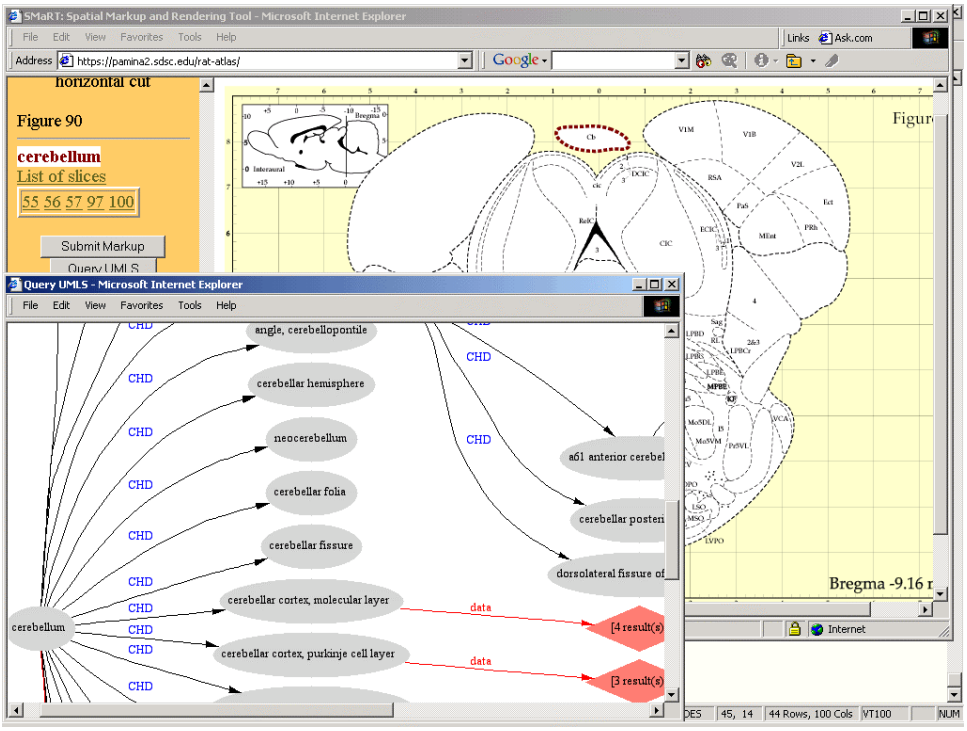


**Figure 1. A query that use the UMLS ontology (in front) and the spatial atlas (behind).**

## References


1. A. Gupta, B. Ludäscher, M. Martone, "Knowledge-based Integration of Neuroscience Data Sources", 12th *Int. Conf. Scientific and Statistical Database Management Systems*, Berlin, 39-52, 2000.
2. B. Ludäscher, A. Gupta, M. Martone, "A Mediator System for Model-based Information Integration", *Int. Conf. VLDB*, Cairo, 639-42, 2000.
3. B. Ludäscher, A. Gupta, M. E. Martone, "Model-Based Mediation with Domain Maps", *17th Intl. Conference on Data Engineering* (ICDE), Heidelberg, Germany, IEEE Computer Society, 81-90, 2001.
4. Xufei Qian, Bertram Ludäscher, Maryann E. Martone, Amarnath Gupta: "Navigating Virtual Information Sources with Know-ME", EDBT 2002: 739-741
5. G. Yang, M. Kifer, "FLORA: Implementing an Efficient DOOD System Using a Tabling Logic Engine", *Computational Logic*, 1078-1093, 2000.