

## Chapter 19

# Federation of Brain Data through Knowledge-guided Mediation

Maryann E. Martone<sup>1</sup>, Amarnath Gupta<sup>2</sup>, Bertram Ludäscher<sup>2</sup>, Ilya Zaslavsky<sup>2</sup> and Mark H. Ellisman<sup>1,2</sup>

<sup>1</sup>*National Center for Microscopy and Imaging Research and Department of Neurosciences, University of California, San Diego, San Diego CA,*

<sup>2</sup>*San Diego Supercomputer Center, University of California, San Diego, San Diego CA*  
Correspondence to: [mmartone@ucsd.edu](mailto:mmartone@ucsd.edu)

**Abstract:** In this chapter, we describe a data integration framework for multi-resolution data and some of the tools we have created for achieving integration of distributed data sources. The chapter is divided into two parts. In the first part, we describe a recently developed database for high resolution 3D data on cell structure and protein localization, the Cell Centered Database or CCDB. In the second part, we describe how we can link data sources such as the CCDB to other databases using a mediation-based approach. By incorporating additional knowledge into the system, we are able to link together sources that are related conceptually, even if they do not contain similar elements in their schemata. For each of the tools and resources described, we indicate its current state of development and availability.

**Key words:** database, electron tomography, protein localization, confocal microscopy, electron microscopy

## 1. INTRODUCTION

The grand goal in neuroscience research is to understand how the interplay of structural, chemical and electrical signals in nervous tissue gives rise to behavior. Experimental advances of the past decades have given the individual neuroscientist an increasingly powerful arsenal for obtaining data, from the level of molecules to entire nervous systems. Scientists have begun the arduous process of adapting and assembling neuroscience data at all scales of resolution and across disciplines into computerized databases and other easily accessed sources (Wong and Koslow, 2001). These databases will complement the vast structural and sequence databases created to catalogue, organize and analyze gene sequences and protein products. The utility of such databases has been amply demonstrated, as an entire new field, bioinformatics, has been inaugurated to organize and mine the data (Persidis,

1999, 2000). The question of whether equal rewards will accrue from similar efforts for histological data, whole brain imaging, physiological data, etc., remains to be answered. Nevertheless, at the very least, neuroinformatics efforts will provide organizing frameworks and computational tools to manage neuroscience data, from the lab notebook to published data (Gardner et al., 2001; Miller et al., 2001). In the best case, it will provide new opportunities to integrate large amounts of biological data into unified theories of function and aid in the discovery process (Chicurel, 2000; Fox and Lancaster, 2002).

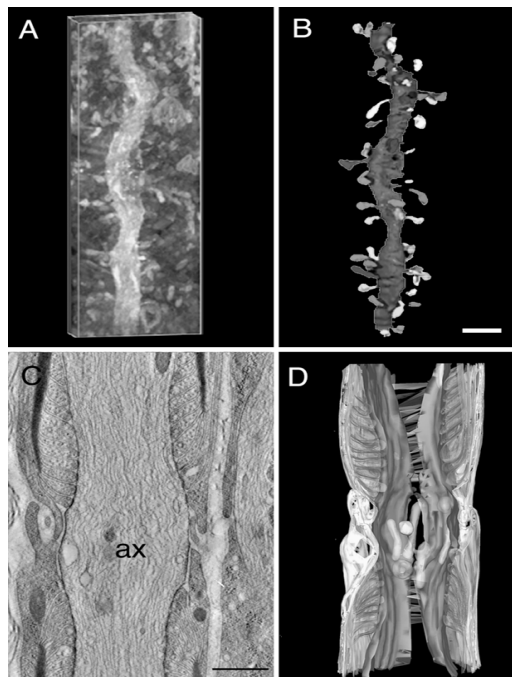
In this chapter, we describe a data integration framework for multi-resolution and multi-disciplinary data and some of the tools we have created for achieving integration of distributed data sources. The chapter is divided into two parts. In the first part, we describe a recently developed database for high resolution 3D data on cell structure and protein localization, the Cell Centered Database or CCDB (Martone et al., in press). In the second part, we describe how we can link data sources such as the CCDB to other databases using a mediation-based approach. In a mediated system, the user does not interact directly with an individual data source but poses a query to a mediator, which then breaks down the query and retrieves data from the individual sources. By incorporating additional knowledge into the system, we are able to link together sources conceptually so that information can be linked and assembled from disparate sources to achieve integration of biological data across scales and disciplines. For each of the tools and resources described, we indicate its current state of development and availability.

## **2. THE CELL CENTERED DATABASE**

### **2.1 Data Modeling and Structure of the CCDB**

The CCDB was created for high resolution 3D light and electron microscopic reconstructions of cells and subcellular structures. It contains 3D structural and protein distribution information from confocal, multiphoton and electron microscopy, including correlated microscopy. Many of the data sets are derived from electron tomography, a powerful 3D reconstruction technique for electron microscopic specimens. Electron tomography is similar in concept to medical imaging techniques like CAT scans and MRI in that it derives a 3D volume from a series of 2D projections through a structure. In this case, the structures are contained in sections prepared for electron microscopy, which are tilted through a limited angular range. Projections through the section are captured at regular tilt increments, aligned to a common origin and then back projected into a 3D volume (McEwen and Frank, 2001). When combined with high voltage electron microscopy,

detailed reconstructions can be made of relatively large expanses of tissue at EM resolution. For example, the CCDB contains reconstructions of extended structures such as spiny dendrites from selectively stained neurons contained within 2-5  $\mu\text{m}$  thick sections (Fig. 1). Single reconstructions can contain up to a 20  $\mu\text{m}$  stretch of dendrite with upwards of 60 spines. Electron tomography is also providing new insights in the 5–40 nm range. Structural features in this range have traditionally been difficult to study because they are smaller than the thickness of a thin section produced for electron microscopy (60-100 nm thick) and so tend to be obscured within the depths of the section. In electron tomography, computed sections through a volume can be generated with an effective resolution of 5-10 nm, thereby uncovering structural detail within this range. Tomography has led to a complete revision of our understanding of even well-studied structures such as mitochondria (Perkins et al., 1997; Mannella et al., 1997) and the neuromuscular junction (Harlow et al., 2001).



*Figure 1.* Examples of tomographic data sets in the CCDB. A and B show a selectively stained spiny dendrite from a Purkinje cell. A is a projection of the volume reconstruction (dendrite appears as white against dark background). B is the segmented dendrite. C and D show a tomographic reconstruction of the node of Ranvier. C is a single computed slice through the volume. D is a surface reconstruction of the various components comprising the node. Scale bar in B = 1  $\mu\text{m}$ ; in C = 0.5  $\mu\text{m}$

The CCDB is built on an object-relational framework using Oracle 8i. The current CCDB has over 80 tables containing a large amount of descriptive data. It models the entire process of 3D reconstruction, from specimen preparation to segmentation and analysis. A volume reconstruction is stored along with all of the raw images and the processing details required to reconstruct the volume from the raw data. Each object segmented from the 3D volume is stored as a separate object indexed to the parent reconstruction. Four types of segmented objects are currently modeled in the CCDB: 1) surface objects: polygonal surface meshes representing 3D objects in the reconstruction, extracted using either isosurfacing methods or manual contouring; 2) contour objects: a series of 2D contours defining an

object; 3) volume objects: subvolumes containing an object of interest; and 4) tree objects: skeletons of branching objects like dendrites and axons, derived from NeuroLucida (MicroBrightfield, Colchester, VT, USA). Each object is stored along with any measurements like surface area, volume, length, number and labeling intensity. Whenever possible, parsers are written for the output of analysis programs so that results can be uploaded directly into the CCDB. For example, measurement summaries for tree objects are uploaded directly from the output of NeuroExplorer, an analysis program for NeuroLucida-derived data.

The data contained in the CCDB will be derived from both published and unpublished data. A constant concern in creating and maintaining databases of experimental information is the quality of data retrieved from a query. In the CCDB, the evaluation of the quality and accuracy of morphometric or protein distribution modeling is up to the user. Thus, the CCDB contains several features to aid in this process including: 1) the inclusion of morphometric data stored with the objects to allow the user to compare the statistics of a given data set to other stored and published data to determine whether they fall within expected ranges; 2) access to the raw data along with all the imaging and processing steps to allow the accuracy and quality of the final reconstruction to be assessed by an experienced user; 3) field to indicate whether the data come from published studies; 4) evaluation tables for users to store estimations of the quality of experimental, imaging, protein labeling and reconstruction results; and 5) the ability for users to supply additional or alternative interpretations of a given data set, which will be retrieved along with the data. In this way, the CCDB can serve as an interactive forum for data interpretation.

## 2.2 Accessing the CCDB

The CCDB can be accessed at <http://www.ncmir.ucsd.edu/CCDB>. The home page is shown in Fig. 2. The CCDB is still under development but is available on-line for limited use. We welcome feedback from interested users, although some of the functions described in this chapter may not be fully implemented by the time of publication. Information on the current functionality and updates to the site are available from the “Status/Update” link off of the main page. The main page also contains links to a description of the CCDB and to visualization and analysis tools available for data contained in the CCDB. The CCDB itself is accessed by the “Enter the database” link which takes the user to the log in page. The log in page allows users to create new accounts, log in to the database or view the CCDB schema by clicking on the appropriate links. Some sample query results are also available. The CCDB is freely accessible, although users must register for an account. Accounts are created by clicking on “Create Account.”

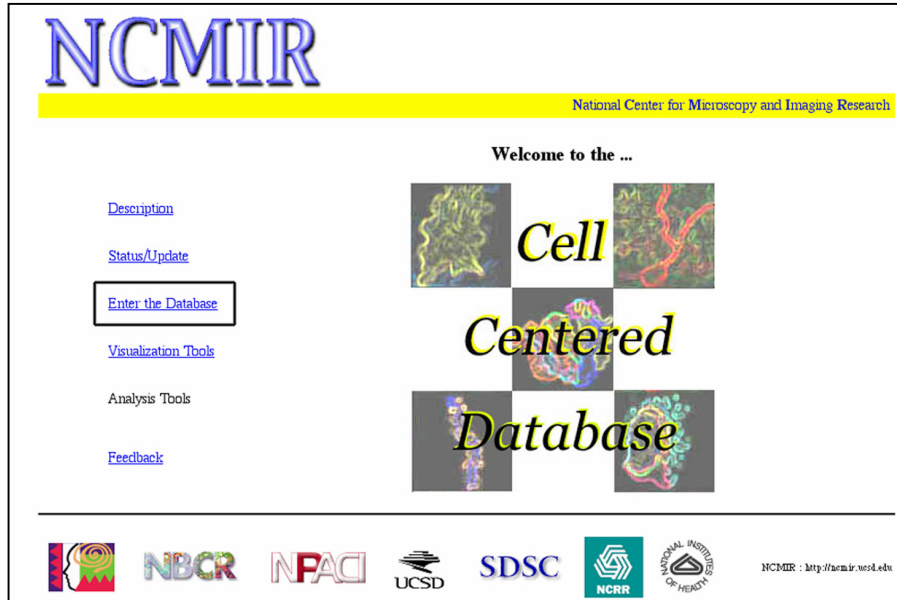


Figure 2. Home page of the CCDB. Clicking on “Enter the database” takes the user to the log in page.

Upon logging in, the user is presented with a simple query interface shown in Fig. 3. Users may query on a limited number of descriptive fields such as cell type or structure, by typing in values or selecting from the pull down menus. More advanced query interfaces are currently being developed for the CCDB which will take advantage of the rich data model of the CCDB to retrieve data based on morphometric features contained in the data. An example of such a query form is shown in Fig. 3 (right panel) for neuronal branching data contained in NeuroLucida files. Users can retrieve neuron data based on branching patterns, tortuosity of dendrites and other metrics. This form can be viewed, along with sample queries, by clicking on the “Advanced Query Form” from the “Enter the database” page, although at the time of writing, this form is not linked to the CCDB. From these types of feature based queries, one can begin to look at relationships among various data sets. We might ask, for example, whether variability in neuronal branching patterns correlates with any subject characteristic.

In response to a query, the CCDB informs the user of the total number of data sets matching the criteria and presents a short set of descriptors. For each data set returned, the CCDB presents thumbnails of the types of image data available for that particular result (Fig. 4). The user may view a 2D image by clicking on the thumbnail or may download the data. In the current version of the CCDB, the image data itself is not stored directly in the

database; that is the actual voxel values and spatial arrays are not stored in the

The figure shows two web-based query forms. The left form, titled "Simple Retrieval/Query", includes instructions and fields for ProjectID, Project Keyword, Person's Name, Reconstruction Technique (Optical section), Cell Type (All), Structure (All), Organ (All), and Species (All). The right form, titled "Search Neuronal Branching Data", features a list of attributes with search criteria: Diameter (10 micrometer), Tortuosity (1.2, Percentage 50%), Branch Order (3, Percentage 50%), Branch Distance (10 micrometer), Number of Primary Dendrite (1), Spines per Segment (3, Percentage 50%), and Dendrite Order (15). It also includes a "Search Type" section with radio buttons for "AND" and "OR", and "Submit" and "Clear" buttons.

Figure 3. Simple query form for the CCDB (left). An example of the type of advanced query forms being developed for attribute queries is shown on the right.

database.

Instead, the descriptive and analysis data are stored along with pointers to the image file. To store the images for the CCDB, we have opted to use the Storage Resource Broker (SRB), a data management system for storing and

accessing distributed data (<http://www.npaci.edu/DICE/SRB/>). The SRB is sophisticated client-server middleware that provides a uniform interface for connecting to data resources over a network. Unlike conventional access methods, e.g., file servers, ftp, or http, SRB is grid-based software providing transparent access to data, relieving the user from dealing with aspects such as physical location of imaging data, concrete storage devices, and device-dependent access protocols. Thus, regardless of where the data lives, whether in a single location or distributed across

The "Query Result" page displays "4 records were found". Record 1 details are as follows:  
**Project:** P1170  
**Description:** NeuroLucida tracing of filled Purkinje neurons  
**Species:** mouse  
**System:** CNS  
**Organ:** brain  
**Region:** cerebellum  
**Cell type:** Purkinje  
**Instrument:** BioRad Radiance  
 A "View Detail" button is present below the record information. At the bottom, four image types are listed: Volume (with a "Download!" link), Segmented File (with a "Download!" link), Animation (with a "click here to obtain the image, click on the image to view the animation" instruction), and Atlas Map.

Figure 4. Return of query results from the CCDB. All of the types of image files associated with a given data set are listed at the bottom. "Atlas Map" shows the location of a data set with respect to a brain atlas (see below).

several databases, file systems, and high-performance storage systems, SRB provides access to the data via an SRB identifier. When a user requests a file from the CCDB, the file is retrieved via its SRB ID. Eventually, the SRB will handle additional functions like file format conversions so that the user can request the data in most common formats. The CCDB acts as a single client of the SRB, so that separate SRB authentications and accounts do not have to be obtained for each user.

Data in the CCDB are stored in a variety of formats currently used at NCMIR and may be viewed with tools available for download. A list of these tools is available under the “Visualization Tools” link on the CCDB main page (visible in Fig. 2), which provides the platform dependencies of each tool along with its utility for the different types of data available (volumes, contours, tree structures and surfaces). A general purpose java-based viewer for CCDB data, Jviewer, is being developed, which can be used to view images in the CCDB directly from the web interface. Also available at the bottom of the Visualization Tools web page is a link to file format converters available through NCMIR.

### **2.3 Spatial and Semantic Indexing of Data in the CCDB**

One of the goals of our work, and informatics in general, is to develop systems for integration of data across scales and disciplines. In our case, we wish to relate the cellular level data in the CCDB to data acquired at the tissue and molecular levels. In order to achieve this goal, we must develop shared frameworks in which to place individual data sets so that they can be related to data taken in other experiments. One obvious framework is spatial, using standard brain atlases to provide registration to a common coordinate system. A second is semantic, taking advantage of efforts in various communities to develop shared ontologies. An ontology can be thought of as a network of terms, concepts and the relationships between them.

To aid in the registration of data in databases like the CCDB, we are developing a web-based tool for situating brain data in both a spatial and semantic context. The Spatial Mark Up and Rendering Atlas Tool (“Smart Atlas”), based on a geographical mapping tool, is a web-based tool that allows users to define polygons on a series of 2D vector images and annotate them with names, relationships and ontology concept IDs. Using this tool, we are creating a “Smart Atlas” and spatial database for neuroscience data using a brain atlas as the interface. For the purposes of demonstrating the utility of this tool, we are using a commercially available brain atlas, Paxinos and Watson (1998), as the interface. Because the atlas is a commercial product, the web page is password protected. The Smart Atlas tool, itself, though will be made available to other scientists for use with their own atlases.

Therefore, we describe its utility in some detail below. A demonstration page is being set up and will be available from the log in page of the CCDB.

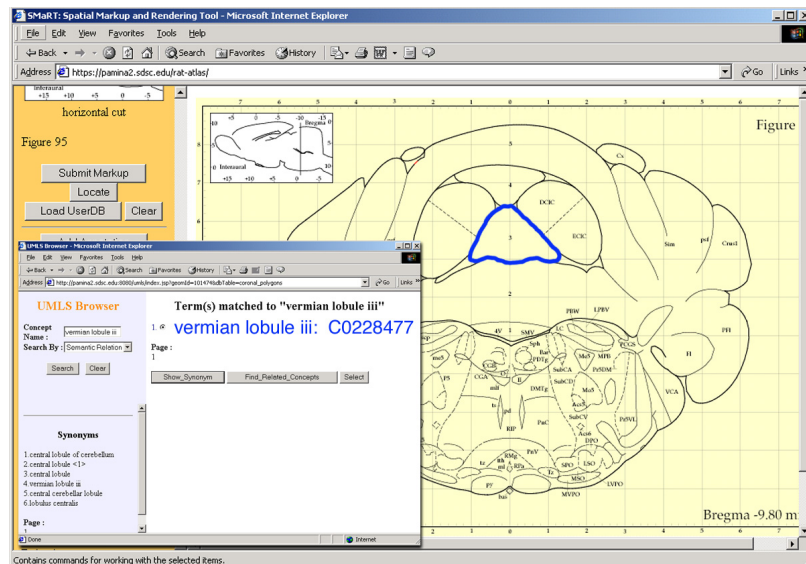


Figure 5. Spatial Mark Up and Rendering Tool for spatial and semantic registration of brain data. The atlas slice shown is from the Paxinos and Watson rat brain atlas. The user has defined cerebellar lobule iii and attached the concept ID from the UMLS.

The Smart Atlas interface is illustrated in Fig. 5. The Adobe Illustrator files supplied with the atlas were converted into SVG format using a plug-in available at <http://www.adobe.com/support/downloads/>. The spatial coordinates of the line segments in each image were stored in an Oracle database, which serves as the back end of the Smart Atlas. The user then “marks up” the atlas using the Smart Atlas tool. First, the user selects those line segments comprising a brain region to define a polygon (Fig. 5). The user then annotates the polygon with the name of the brain region. In addition, the user links the brain region to one or more ontologies through a browsing tool accessed through the Smart Atlas. In the case of the CCDB, we are using the Unified Medical Language System (UMLS), developed for biomedical sciences by the National Library of Medicine (Ingenerf et al., 2001). Strictly speaking, the UMLS is a metathesaurus, but for many purposes it can be used like an ontology. The user browses the UMLS and selects the relevant concept. The polygon is now tagged with both the name of the region and the UMLS concept.



Once the atlas has been marked up, it can be queried from the same browser. Clicking on any point in the atlas will return the stereotaxic coordinates. Clicking on a brain region will return the name of that region, along with any synonyms, and highlight all planes containing that structure (Fig. 6).

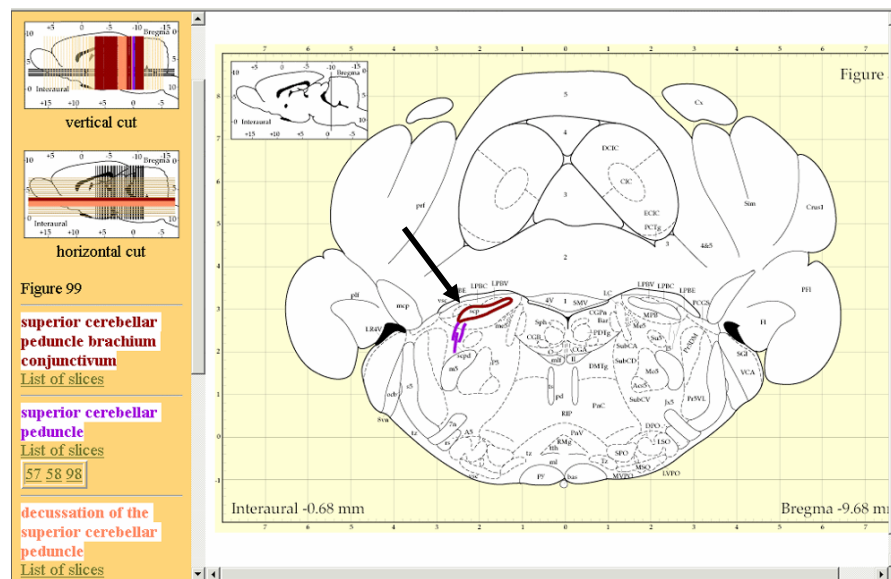
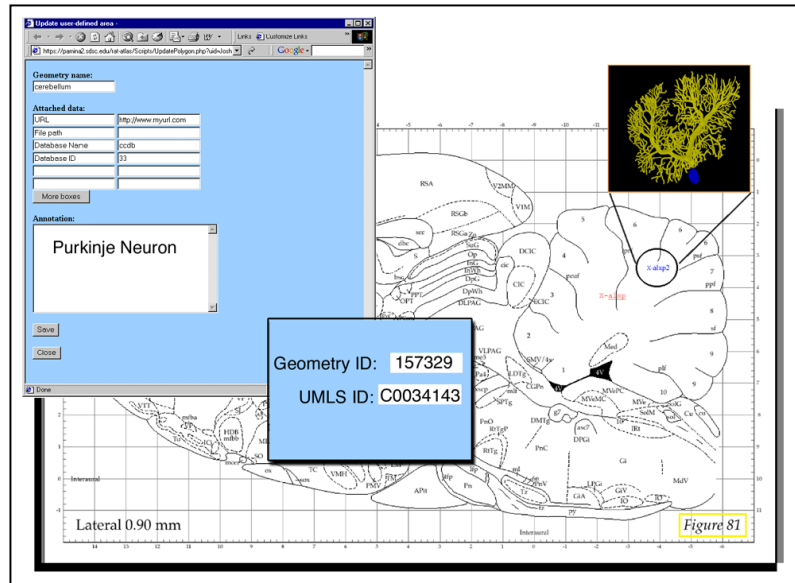


Figure 6. Querying the marked up brain atlas. The user has highlighted the superior cerebellar peduncle (arrow). The list of slices containing the structure are highlighted in the image maps in the upper left corner.

The Smart Atlas can now be used to register a researcher's data to a specific spatial location. In this case, the user draws an arbitrary polygon representing the approximate data location on one of the atlas planes (Fig. 7). The user is then presented with a form which can be used to add annotations or to link to an ontology concept. Although the UMLS is used in the examples shown here, the user will eventually be able to use multiple ontologies, including those of their own creation, for semantically indexing data. Tools are also being developed to define new terms and relationships in existing ontologies.

The utility of indexing data in a database using the Smart Atlas tool is several fold. By using an ontology concept ID for indexing CCDB data, the user is relieved from having to specify multiple terms to characterize the data. Once the concept is chosen, the data set inherits the parents and children of this node and is thereby placed in a larger semantic context. The full utility of this system will be illustrated in the next section, which describes our approach to linking brain databases into a federated, cross-queriable system.



*Figure 7.* Registration of a data set with the Smart Atlas. The user draws a polygon representing the location of a data set, in this case a filled Purkinje neuron. The user then specifies the database containing this data, enters an annotation and selects a concept from the UMLS or some other ontology. The concept ID is stored back in the data base.

Spatial indexing of data to a coordinate system will allow the user to pose spatial queries, e.g., find data associated with this region; find all structures within 1 mm of this area. It also allows the user to take advantage of efforts to develop spatial transformations and alternative representations of brain topology, e.g., flat maps of cortex and cerebellar cortex (Dickson et al., 2001; Van Essen et al., 2001). The locations of filled or immunolabeled neurons from the CCDB can be mapped onto these flat maps to give additional information on the distribution of neuronal classes or compare features such as dendritic arbor size as a function of location. Thus, by indexing to a common coordinate system, we can facilitate interoperability of informatics tools.

### 3. FEDERATION OF BRAIN DATA

No matter how sophisticated and complete the CCDB is in modeling and querying the data, it only covers a very small portion of the biological spectrum. The oft stated goal of informatics research is to create computer-based approaches to allow scientists to integrate and relate data obtained at different scales, experimental systems and sub-disciplines. To tackle such an ambitious goal, various approaches have been taken. These approaches range from data warehousing, where data is gathered from multiple sources and

deposited into a larger database (Rachedi et al., 2000), to the creation of “virtual federations” where individual data sources remain separate but are “wrapped” in a standard language such as XML (extensible markup language) that allows their content to be understood by a federation engine known as a mediator. The mediator is responsible for selecting, restructuring, and merging information from autonomous sources and for providing an integrated view of the information.

Although mediator-based systems have been around for some time, their use to integrate biological databases presents a challenge. When one is linking

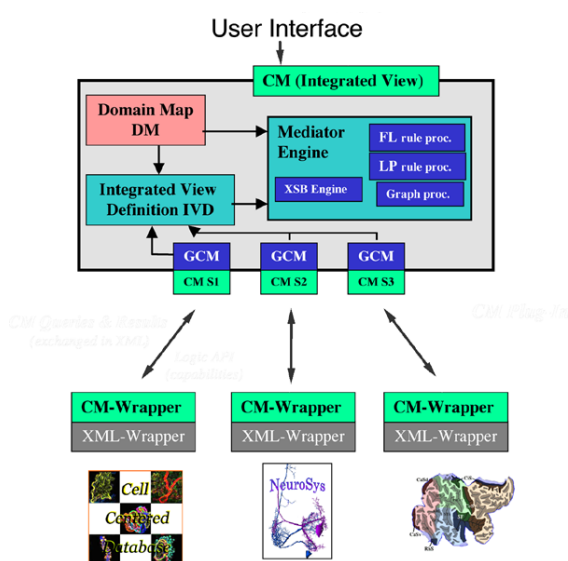


Figure 8. Overview of the prototype KIND mediator. The lower icons represent 3 different sources linked to the mediator. The box in the upper panel represents the mediator layer.

databases containing similar information, e.g., gene sequence databases, the linkage between data sources are straightforward and can be expressed as relatively simple structural statements which identify common elements in their respective schemata. The task becomes more difficult, however, when attempting to relate data sources which cannot be joined on a purely structural level but which nevertheless contain relevant data. For example, a database at Montana State University on sensory neurons encoding wind direction in the cricket may not share any attributes in common with a database of anatomical structures in the monkey visual cortex.

Despite the lack of common semantic links, most neuroscientists can easily relate these two data sources at the conceptual level, by recognizing that they both are sensory systems and may share certain aspects of population coding of sensory stimuli. In fact, neuroscientists usually can navigate with relative ease from the level of individual molecules to brains and across experimental disciplines, because they possess the requisite knowledge to conceptually relate data at each level. We have developed a novel mediator-based integration paradigm which exploits such expert knowledge to begin to address the problem of data integration of heterogeneous neuroscience data, a system which we call “knowledge-guided or model-based mediation”.

An overview of the prototype mediator, called the KIND (Knowledge Integration of Neuroscience Data) mediator, is shown in Fig. 8 and is described in more detail in Gupta et al., (2000) and Ludaescher et al., (2001). Additional references and demonstrations are provided at <http://www.npaci.edu/DICE/Neuro/>. There are two places in the system where additional knowledge is incorporated. At the source level, the wrapper language exports a conceptual model (CM) of each data source containing information about relationships, classes and values using an object-oriented language, F-logic, as the deductive engine. At the mediator level, conceptual knowledge is encoded in the form of a semantic network of terms and relationships which we call a “domain map”. Domain maps can be thought of as ontologies with more formal logic-based semantics. The purpose of the domain map is to provide a declarative means for specifying additional knowledge that is not present in the source but is required to bridge two information sources. When a standard SQL query is launched, the mediator breaks it down into its component parts, accesses the appropriate data sources using knowledge sources contained in the wrapper or the domain map, and reassembles the results of the query into an integrated view.

In the following examples, we illustrate the utility of the KIND mediator for answering queries which require bridging of multiple data sources, ranging from retrieval of multi-resolution data to true data integration where a new piece of information is derived by combining data from two sources. The current system consists of the KIND mediator with the following sources registered: 1) a prototype of the CCDB; 2) the EF-Hand Calcium binding protein database, a database of sequence data available from Vanderbilt University ([http://structbio.vanderbilt.edu/cabp\\_database/](http://structbio.vanderbilt.edu/cabp_database/)); 3) the NeuronDB database from the SenseLab project (Shepherd et al., 1998; Miller et al., 2001); 4) TAXON, a database created from taxonomy information available from the National Center for Biotechnology Informatics (NCBI).

The queries below are written in natural language for the sake of clarity, but in all cases were executed in standard SQL. Similar demo queries can be found at: <http://www.npaci.edu/DICE/Neuro/> (go to “Demonstrations: Knowledge based mediation” and click on “Protein Localization Query”) and must be executed with an XML-capable browser such as Internet Explorer.

### **3.1 Examples of data federation using the KIND mediator**

#### **3.1.1 Retrieval and navigation through multi-scale data**

An example of how the mediator is used in conjunction with the Smart Atlas is shown in Fig. 9. In this case, the user launches a simple query from the Smart Atlas Tool by clicking on the cerebellum in order to find all data associated with this brain region. It turns out that there are no sources hooked to the mediator which have data indexed under the term “cerebellum”, but the

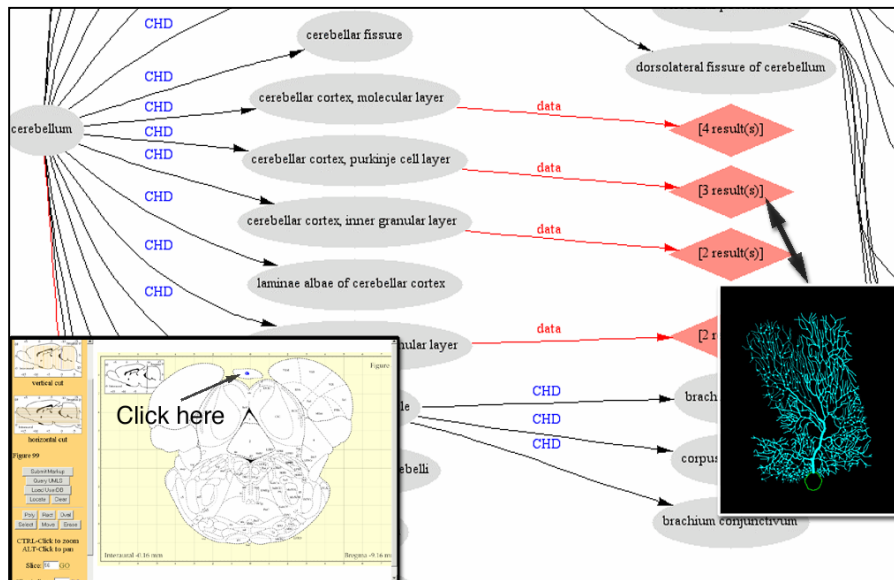


Figure 9. Results of query launched from Smart Atlas. The ovals represent concepts from the UMLS and the diamonds where data is found. Clicking on the diamond returns the data.

CCDB has data on Purkinje neurons and other structures found within the cerebellum. The mediator locates the terms under “cerebellum” in the domain map, queries the sources and returns the relevant portion of the domain map to the user with the available data placed in context. In this way, the user can immediately see how the available data relates to the concept they chose. The user then clicks on the data node to retrieve the data.

### 3.1.2 Making use of ancillary data sources to increase efficiency of data retrieval

“What is the cerebellar distribution of proteins with 90% homology to human NCS-1 in the rat? In all rodents?” To answer this query with single sources, the researcher would first go to the calcium binding protein database and retrieve the sequences with 90% homology to human NCS-1. Each retrieved protein would then have to be used to query the available histological data sources. Using the KIND mediator, however, the researcher specifies the protein, degree of homology and anatomical region in the mediator and launches the query. First, the mediator retrieves the list of proteins with homology to NCS-1 from the Calcium Binding Protein database, probes the available sources and then returns the results. Because the user specified “cerebellum”, all structures defined as belonging to cerebellum are retrieved. To retrieve results for all rodents, the mediator goes to the taxonomy database to retrieve species under “rodents” and re-probes the sources.

### 3.1.3 Integrating information from multiple sources

*“What is the overlap between the parallel fiber input into the Purkinje neuron and the distribution of ryanodine receptor?”* In the third example, we show how knowledge-based mediation can answer a query by integrating information across sources to derive information that is not present in any single source. Two registered sources have information about Purkinje neurons, the CCDB and the Senselab Neuron database. The CCDB has information on the distribution of the ryanodine receptor, but no information on connectivity. The Senselab database has information on connectivity, but no information on the ryanodine receptor. To address this query, an integrated view definition must be created which structurally and conceptually links the two data sources at the level of the domain map. First, the mediator retrieves data from the SenseLab database on the distribution of the parallel fibers. It then locates the relevant portion of the domain map and queries the CCDB via the domain map to find the intersection. The full demonstration of this query can be viewed at: <http://www.npaci.edu/DICE/Neuro/>, under “Demonstrations: Knowledge-based mediation” by clicking on “Protein Localization Query (with domain map)” and viewing query #3.

## 3.2 Extending the Domain Map

We have been extending the concept of the domain map to include more dynamic processes such as signal transduction. In a process map, the nodes represent states and the edges represent transitions. For example, we have taken some of the pathways involved in the production of long term potentiation and represented them in a graphical form. As ontologies and process maps are developed, they become not only a useful adjunct to database indexing and data management, but an important source of knowledge in their own right. To take advantage of these graphical bodies of knowledge, we have developed a tool for browsing, querying and tying data to these process maps. The Knowledge Map Explorer Tool (Know-Me) is shown in Fig. 10 and can be demonstrated at <http://pamina2.sdsc.edu:18080/knowme/pm.jsp>. Users can select nodes to elaborate processes, execute queries like “compute the shortest pathway between two nodes” and query for experimental evidence supporting a given relationship.

## 4. CONCLUSIONS

The KIND mediator is a prototype system, but will form the information integration strategy employed by the newly created the Biomedical

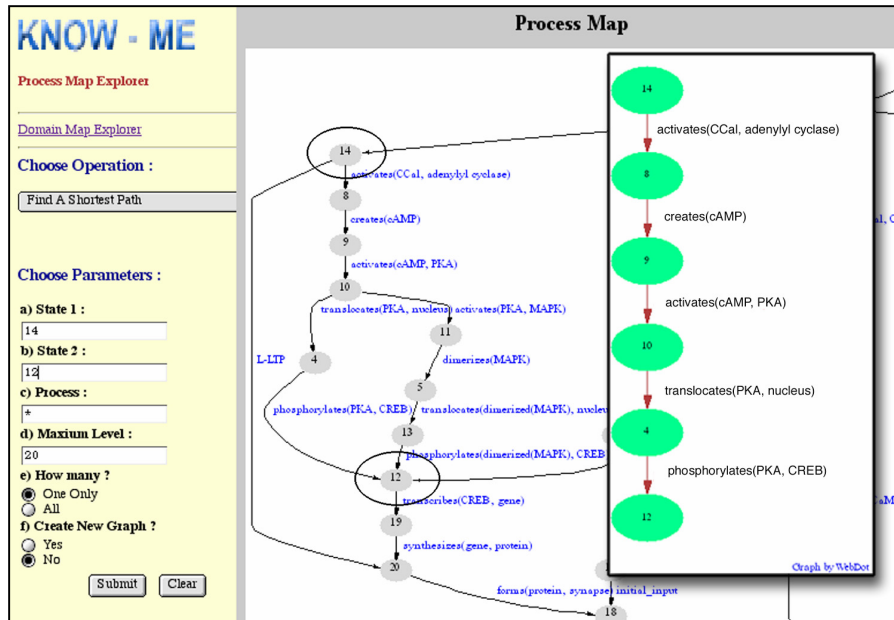


Fig. 10: The Know Me Tool: Part of the process map for long term potentiation is shown in the main window. Each step in the process is identified by a number while the edges are labeled with the process name. In the example shown, the “compute shortest pathway” function was used to determine the shortest pathway between step 14 and 12.

Informatics Research Network (<http://www.birn.ncrr.gov/>) or BIRN. The BIRN project will ultimately link together technology development resources supported by the National Center for Research Resources of the NIH into a virtual data grid, where data can be shared and accessed through a mediator-based system over Internet2. Initiatives like the BIRN will serve as test beds for creating the infrastructure necessary for scientists to adapt the scientific process to the new electronic forum enabled by network connectivity and high performance computing. The challenges involved in adapting the scientific enterprise for this global forum are many. They range from sociological issues involved in data sharing to issues of curation and validation to technical issues of interoperability and moving large data sets (Kotter, 2001; Toga, 2002). However, we should not let these challenges blind us to the advantages afforded by immediate access to data. New methods for visualizing, indexing and exploring data will likely help to accelerate the discovery process by identifying inconsistencies, controversies and knowledge gaps much more rapidly than is possible through the literature. As more data sources become linked through technologies such as database federation, we will be able to navigate through many levels of biological



complexity and come closer to the goal of understanding biological systems in their entirety.

## ACKNOWLEDGEMENTS

This work supported by NIH grants from NCRR RR04050, RR RR08605 and the Human Brain Project DC03192, an NSF grants supporting the National Partnership for Advanced Computational Infrastructure NSF-ASC 97-5249 and MCB-9728338. The authors would like to acknowledge the contributions of Ms. Xufei Qian, Shenglan Zhang, Mona Wong and Mr. Joshua Tran for their work on the tools presented here.

## REFERENCES

- Chicurel, M. (2000) Databasing the brain. *Nature*, 406, 822-825.
- Dickson, J., Drury, H. and Van Essen, D. C. (2001) 'the surface management system' (sums) database: A surface-based database to aid cortical surface reconstruction, visualization and analysis. *Philos Trans R Soc Lond B Biol Sci*, 356, 1277-1292.
- Fox, P. T. and Lancaster, J. L. (2002) Opinion: Mapping context and content: The brainmap model. *Nat Rev Neurosci*, 3, 319-321.
- Gardner, D., Abato, M., Knuth, K. H., DeBellis, R. and Erde, S. M. (2001) Dynamic publication model for neurophysiology databases. *Philos Trans R Soc Lond B Biol Sci*, 356, 1229-1247.
- Gupta, A., Ludaescher, B. and Martone, M. E. (2000) Knowledge-based integration of neuroscience data sources. *Proceedings of the 12th International Conference on Scientific and Statistical Database Management (SSDBM'00) IEEE Computer Society*.
- Harlow, M. L., Ress, D., Stoschek, A., Marshall, R. M. and McMahan, U. J. (2001) The architecture of active zone material at the frog's neuromuscular junction. *Nature*, 409, 479-484.
- Ingenerf, J., Reiner, J. and Seik, B. (2001) Standardized terminological services enabling semantic interoperability between distributed and heterogeneous systems. *Int J Med Inf*, 64, 223-240.
- Ludaescher, B., Gupta, A. and Martone, M. E. (2001) Model-based mediation with domain maps. *Proceedings of the 17th International Conference on Data Engineering, IEEE Computer Society 2001*, Heidelberg, Germany, April 2-6, 2001.
- Mannella, C. A., Marko, M. and Buttle, K. (1997) Reconsidering mitochondrial structure: New views of an old organelle. *Trends Biochem Sci*, 22, 37-38.
- Martone, M. E., Gupta, A., Wong, M., Qian, X., Sosinsky, G., Ludäscher, B. and Ellisman, M. H. (in press) A cell-centered database for electron tomographic data. *J. Struct. Biol.*,
- McEwen, B. F. and Frank, J. (2001) Electron tomographic and other approaches for imaging molecular machines. *Curr Opin Neurobiol*, 11, 594-600.
- Miller, P. L., Nadkarni, P., Singer, M., Marengo, L., Hines, M. and Shepherd, G. (2001) Integration of multidisciplinary sensory data: A pilot model of the human brain project approach. *J Am Med Inform Assoc*, 8, 34-48.
- Paxinos, G. and Watson, C. (1998) *The rat brain in stereotaxic coordinates*, Academic Press, San Diego.
- Perkins, G., Renken, C., Martone, M. E., Young, S. J., Ellisman, M. and Frey, T. (1997) Electron tomography of neuronal mitochondria: Three-dimensional structure and organization of cristae and membrane contacts. *J Struct Biol*, 119, 260-272.
- Persidis, A. (1999) Bioinformatics. *Nat Biotechnol*, 17, 828-830.
- Persidis, A. (2000) Data mining in biotechnology. *Nat Biotechnol*, 18, 237-238.



- Rachedi, A., Rebhan, M. and Xue, H. (2000) Gabaagent: A system for integrating data on gaba receptors. *Bioinformatics*, 16, 301-312.
- Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., Nadkarni, P. M. and Miller, P. L. (1998) The human brain project: Neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci*, 21, 460-468.
- Toga, A. (2002) Neuroimage databases: The good, the bad and the ugly. *Nat Rev Neurosci*, 3, 302-308.
- Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D. and Anderson, C. H. (2001) An integrated software suite for surface-based analyses of cerebral cortex. *J Am Med Inform Assoc*, 8, 443-459.
- Wong, S. T. and Koslow, S. H. (2001) Human brain program research progress in biomedical imaging/neuroscience, 2001. *J Am Med Inform Assoc*, 8, 510-511.

