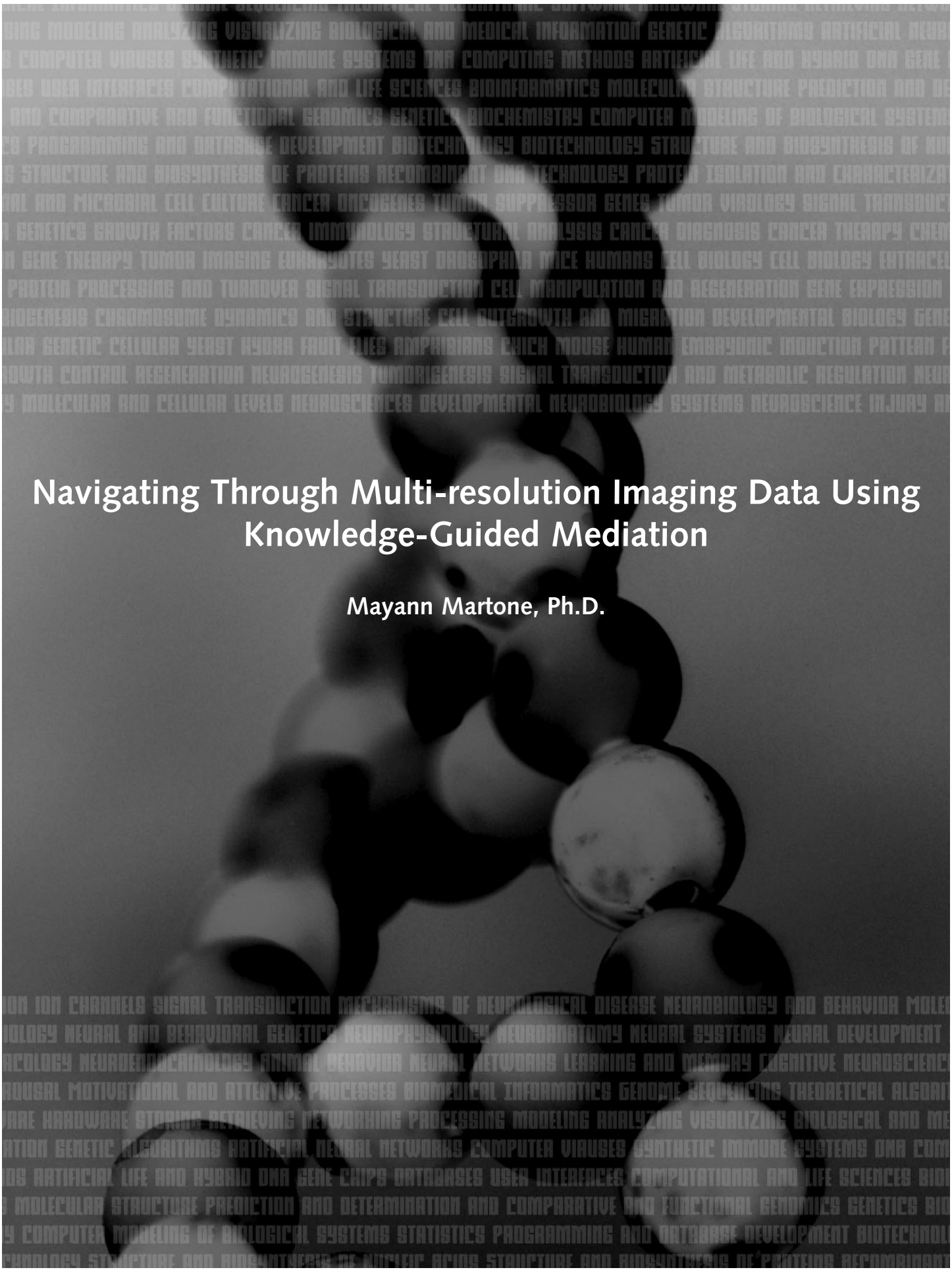


Navigating Through Multi-resolution Imaging Data Using Knowledge-Guided Mediation

Mayann Martone, Ph.D.

Currently there are 2 two figure 5s and no figures 3 or 4

OVERVIEW

The grand goal in neuroscience research is to understand how the interplay of structural, chemical and electrical signals in nervous tissue give 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 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 (1), (2), (3). These databases will complement the vast structural and sequence databases created to catalogue, organize and analyze gene sequences and protein products. This presentation will focus specifically on databases for brain imaging data, from the level of whole brain (4) (5) to those for cellular and molecular structure (2), (6). The creation of these databases is still in the early phase and largely driven by individual efforts. In this presentation, we will specifically discuss our own imaging database, the Cell Centered Database or CCDB (7), and highlight some of the challenges and issues involved in creating such a resource. We will then discuss some of the approaches to linking databases such as the CCDB to information taken at different scales and in different disciplines, to aid the scientist in data integration.

© 2002 Maryann Martone

THE CELL CENTERED DATABASE

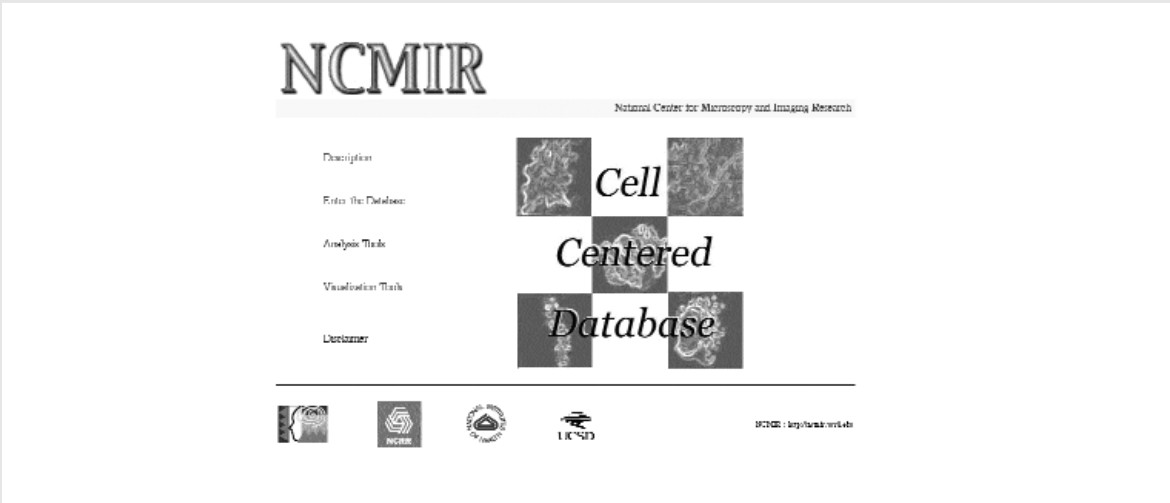
DATA MODELING AND STRUCTURE OF THE CCDB

The CCDB was created to house the types of high resolution 3D light and electron microscopic reconstructions of cells and subcellular structures produced at the National Center for Microscopy and Imaging Research (http://www.ncmir.ucsd.edu). It contains structural and protein distribution information derived from confocal, multiphoton and electron microscopy, including correlated microscopy. Many of the data sets are derived from electron tomography. 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. Many of the data sets in the CCDB come from studies of the nervous system, although the CCDB is not restricted to neuronal information.

The rationale behind creating the CCDB as a publicly accessible data base was several fold. First, we wanted a venue for disseminating the very rich and unique datasets acquired by electron tomography. Because of the superior resolution of tomographic datasets, they often contain much more data than is analyzed by a single researcher. A single electron tomographic study generally relies on a very small sample size because of the labor involved in acquiring and analyzing the specimens. Thus, having a repository where these data sets can be accumulated and reanalyzed will help researchers gain a better picture of variation across structures. Because of the labor intensive nature of the process, tomographers are actively developing algorithms for segmenting and visualizing data. Thus, the CCDB can serve as a resource for those involved in these pursuits. Finally, the CCDB serves a resource for researchers interested in computational modeling by providing high resolution anatomical data in a form suitable for simulations of cellular physiology (8) and molecular dynamics in the cellular microenvironment (9).

The CCDB can be accessed at http://www.ncmir.ucsd.edu/CCDB. It is built on an object-relational framework using Oracle 8i. Although creating and maintaining an Oracle database is considerably more work than other popular database platforms, e.g., mySQL, Oracle

FIGURE 1
Main Web page of the Cell Centered Database. the database is available at www.ncmir.ucsd.edu/CCDB.



© 2002 Maryann Martone

was chosen for several reasons. First, the object-relational features allow us to extend the database in the future to accommodate additional data types. For example, the schema of the CCDB will soon be expanded to include spatial data e.g., positions of vesicles within a synaptic cleft captured during synaptic transmission or the position of spines along a dendritic shaft. Second, by taking advantage of Oracle's Object Relational modeling capabilities, we can create a new set of subclasses of the appropriate existing object class(es) without affecting the rest of the database. Third, the Object Relational modeling capability also allows us to program any analysis or data comparison methods into the database itself. This functionality allows us to move beyond simple retrieval queries to allow us, for example, to create a "compare two protein distributions" function that runs as an integral method from within the database.

The current CCDB has over 80 tables containing a large amount of descriptive data. The CCDB is built around light and electron microscopic 3D reconstructions, including correlated data sets. It models the entire process of 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 that is 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 contours that have not been fitted with a surface; 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, Inc., VT). 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.

By storing the results of any analysis performed on data, we allow the user to query for datasets based on image attributes. For example, because we have the output of analysis programs for neuronal tree structures stored in the CCDB, we can issue queries such as:

Q1: "Find all neurons with dendrites which branch less than 5 μm from the cell soma"

Q2: "Find neurons with curvy dendrites"

A subjective concept such as "curvy" can be translated into a numerical formula by considering the tortuosity measurement generated by Neurolucida, e.g., curvy = tortuosity factor of > 3 in at least 50% of the segments belonging to a dendrite."
Because we have all of the dendritic spines associated with a dendritic shaft stored along with their surface area, volume and length, we can ask

Q3: "Find dendrites having spines with surface area: volume ratios of < 10"

From this type of feature based query, one can begin to construct queries that look at relationships among various data sets. For example, we might ask whether variability in dendritic spine size correlates with any subject characteristic.
In the current version of the CCDB, the image data itself if not stored directly in the database; that is the actual voxel values and spatial arrays are not stored in the database. Instead, the descriptive and analysis data are stored in the database 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

© 2002 Maryann Martone

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 (in our case the CCDB) 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 several databases, file systems, and high-performance storage systems, SRB provides access to the data via a logical SRB identifier. SRB accomplishes this by creating logical collections of physically distributed data objects that are managed by a central Metadata Catalog (MCAT). The MCAT is an Oracle database used by the SRB to keep track of collection attributes and authentication and access control for the data. For each image stored in SRB by the CCDB, the MCAT keeps track of the file format, physical location and image size. Eventually, the SRB will handle additional functions like file format conversions. The CCDB acts as a single client of the SRB/MCAT system, so that separate SRB authentications and accounts do not have to be obtained for each user.

© 2002 Maryann Martone

Evaluating the Validity of Data in the CCDB

A constant concern in creating and maintaining databases of experimental information is the quality of the data stored in the database. At this time, the CCDB accepts both published and unpublished data, and evaluation of the quality and accuracy of morphometric or protein distribution modeling will be up to the user. The data model employed by the CCDB should aid in this process. First, the morphometric data stored with the objects allows 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. Second, the CCDB contains 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. Third, the CCDB notes whether or not the data come from published studies. Fourth, the CCDB contains several evaluation tables to allow users to store estimations of the quality of experimental, imaging, protein labeling and reconstruction results. Finally, because the interpretation of data is often subjective, users are able to supply additional or alternative interpretations of a given data set, indexed under their name. In this way, the CCDB can serve as an interactive forum for data interpretation.

© 2002 Maryann Martone

Spatial and Semantic Registration of Data in the CCDB: The Smart Atlas Tool

One of the goals of our work and informatics in general is to develop systems for integration across scales. In our case, we wish to relate the cellular level data in the CCDB to data acquired at the tissue and molecular levels. As one step in this process, we have created tools to place data contained in the CCDB in both a spatial and semantic context that can be easily related to larger and lower scales. To do this, all data in the CCDB is indexed to a brain atlas to provide the spatial context, and one or more ontology concepts to provide a semantic context. An ontology can be thought of as a network of terms, concepts and the relationships between them. The Spatial Mark Up and Rendering Tool (Smart Atlas) was created to allow users to define polygons on a series of 2D vector images and annotate them with names, relationships and ontology concept IDs. To illustrate the utility of this tool (Fig. 2), we have taken some images from a commercially available atlas of the rat brain (10). The spatial coordinates of the line segments are stored in an Oracle database and the Smart Atlas Tool is used to define and annotate brain regions. The slices currently must be in SVG (Scalar Vector Format) which can be output by Adobe Illustrator by using an SVG plug-in available on the Adobe website. The user can register their data with this coordinate system by drawing a polygon on the appropriate slice representing the location of their data. The resulting geometry file is stored in the coordinate system of the chosen atlas.

Once the user defines a polygon, the Smart Atlas offers the opportunity to link the defined geometry file to available ontologies (Fig. 2). For the CCDB, we are currently indexing the data to the Unified Medical Language System (UMLS), an extensive metathesaurus developed for biomedical sciences by the National Library of Medicine (11). After the user has defined a polygon on the atlas browser, the user is taken to the UMLS Browser where a matching term is selected. By indexing the data to the UMLS, the data inherits all of the concepts associated with the chosen label. The user is therefore relieved from the burden of specifying all potentially relevant concepts in the database.

Users may also draw in polygons specifying the location of their own data. In the example shown in Fig. 3, the user has entered the location of a filled Purkinje neuron. The user is presented with a list of matching concepts by the UMLS browser and the concept ID is entered automatically in the CCDB for that data set. Once this concept is chosen, the data set inherits the parents and children of this node. Thus, the data is indexed to any related term of cerebellum through the UMLS. The advantage of indexing the data in this way is apparent when we consider how to link databases such as the CCDB to other available resources. The Smart Atlas tool is still under development but will eventually be made available to researchers for use with their own spatial templates. The atlas itself is a commercial product and is not available for distribution.

© 2002 Maryann Martone

FIGURE 2

Spatial Mark Up and Rendering Tool developed by Ilya Zaslavsky, Amarnath Gupta, Xufei Qian and Joshua Tran. The atlas slice shown is from Paxinos and Watson rat brain atlas. The user has defined cerebellar lobule iii and attached the concept ID from the UMLS to that polygon.

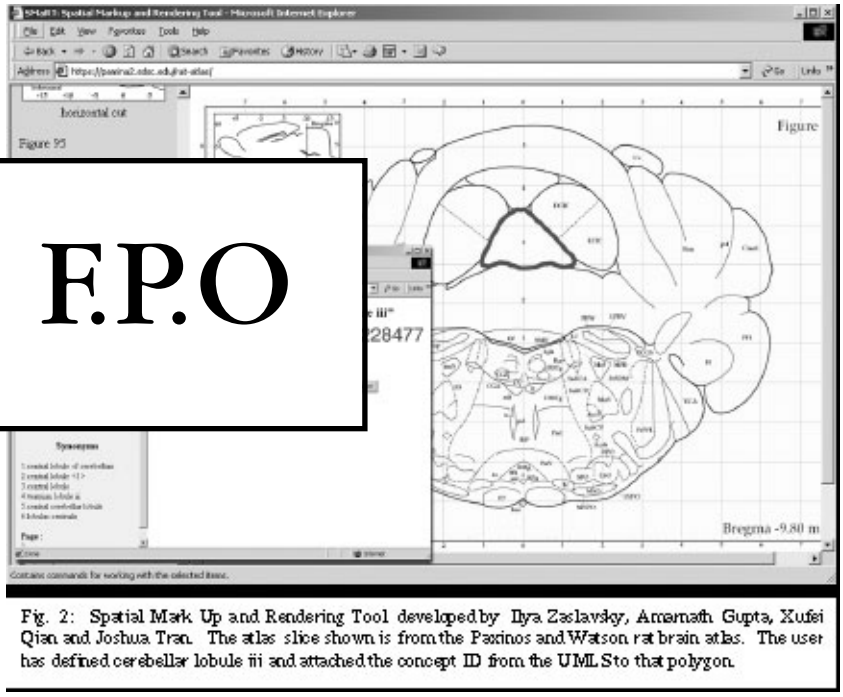


Fig. 2: Spatial Mark Up and Rendering Tool developed by Ilya Zaslavsky, Amarnath Gupta, Xufei Qian and Joshua Tran. The atlas slice shown is from the Paxinos and Watson rat brain atlas. The user has defined cerebellar lobule iii and attached the concept ID from the UMLS to that polygon.

© 2002 Maryann Martone

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 to integrate data from distinct sources. These approaches range from warehousing approaches, where related data is gathered from multiple souces and deposited into a larger database (12), 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 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 cells to brains to behavior and across experimental disciplines, because they possess the requisite knowledge to conceptually relate data at each level. We have developed a novel mediator 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".

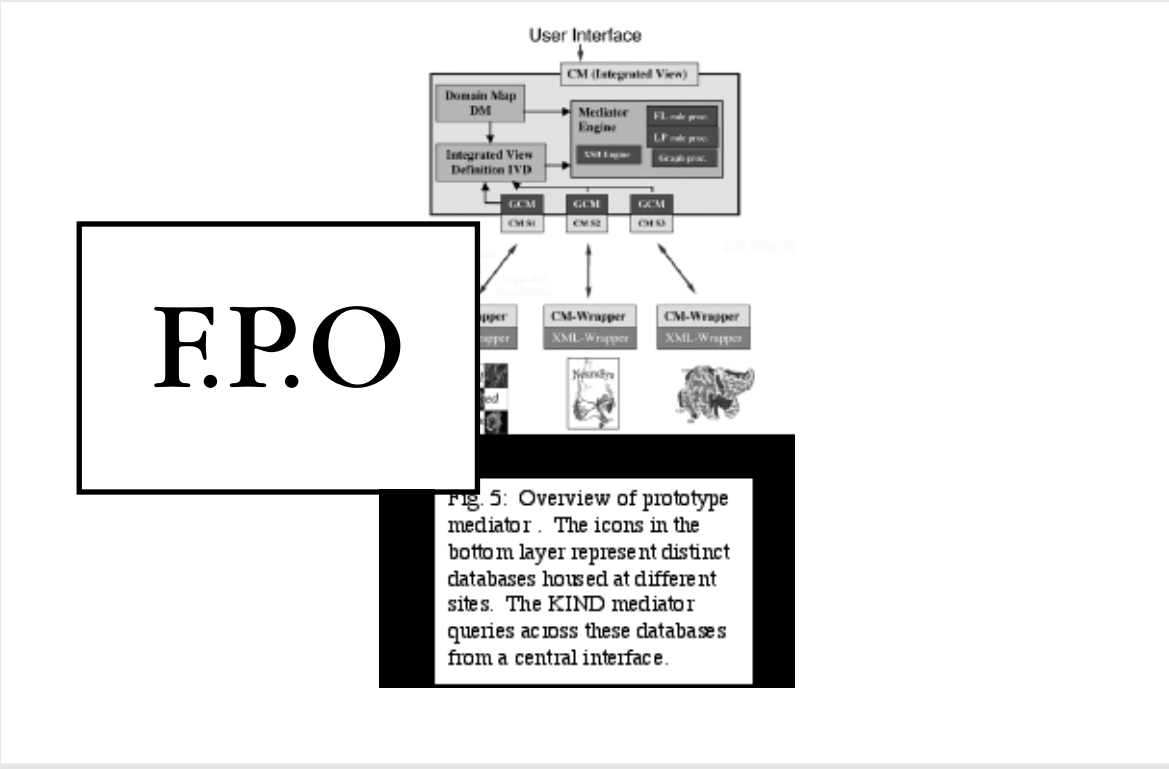
An overview of the prototype mediator, called the KIND (Knowledge Integration of Neuroscience Data) mediator, is shown in Fig. 4 and is described in detail in (13), (14). There are two places in the system where additional knowledge is incorporated. At the source level, the wrapper language exports a 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 an ontology with more formal semantics. The purpose of the domain map is to provide a declarative means for specifiying 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. When the results are returned from a query, the mediator isolates the relevant portion of the domain map, and hangs off the results from the relevant nodes.

An example of how the mediator is used in conjunction with the Smart Atlas is shown in Fig. 5. In this case, the user launches a simple query from the Smart Atlas Tool by clicking on a brain region, in this case, cerebellum. The user wishes to find all data that is associated with this brain region. It turns out that there are no sources hooked to the

© 2002 Maryann Martone

FIGURE 5

Overview of prototype mediator. The icons in the bottom layer represent distinct databases housed at different sites. The KIND mediator queries across these databases from a central interface.



© 2002 Maryann Martone

mediator which have data indexed under the term "cerebellum", but the 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 which they chose. The user then clicks on the data node to retrieve the data.

The KIND mediator is a very flexible system that can be used to address queries that would normally require the individual researcher to go to several different databases to answer. For example, with the KIND mediator, a researcher can pose the query:

Q4: "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 data 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 sequences 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 contained within cerebellum were retrieved. To retrieve results for all rodents, the mediator would go to the taxonomy database to retrieve species under "rodents" and then probes sources.

Knowledge-based mediation can also be used to answer a query by integrating information across sources to derive information that is not present in any single source. For example, the question:

Q5: What is the anatomical overlap between the parallel fiber input into the Purkinje neuron and the distribution of ryanodine receptor?

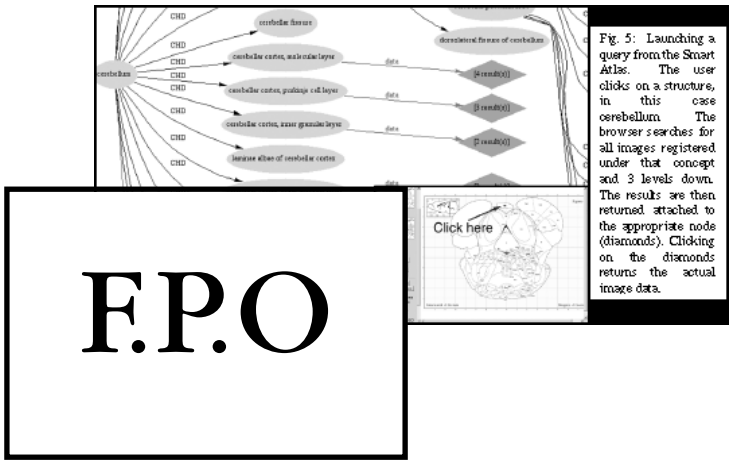
In our current prototype system, two sources registered to the mediator have information about Purkinje neurons, the CCDB and the Yale Senselab Neuron database (2). The CCDB has information on the distribution of the ryanodine receptor, but no information on connectivity. The Senselab database has information 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 KIND mediator is a prototype system, but will form the information integration strategy employed by the newly created Biomedical Informatics Reseach Network (**http://www.birn.ncrr.gov/**).

© 2002 Maryann Martone

FIGURE 5

Launching a query from the Smart Atlas. The user clicks on a structure, in this case cerebellum. The browser searches for all images registered under that concept and 3 levels down. The results are then returned attached to the appropriate node (diamonds). Clicking on the diamonds returns the actual image data.



© 2002 Maryann Martone

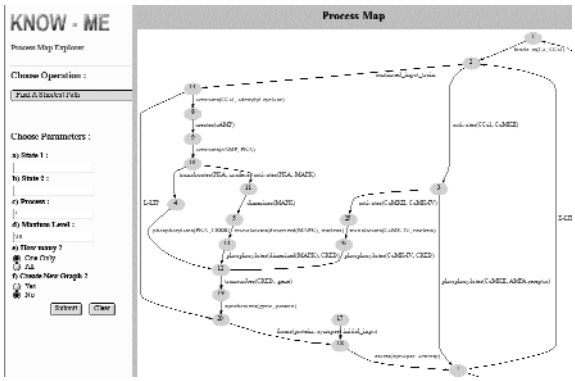
Extending the Domain Map

The power of the KIND system is that it allows the user to take advantage of knowledge sources like ontologies and brain atlases to serve as conceptual bridges between sources. These ontologies can be of local creation or can be generally available, e.g., UMLS and Gene Ontologies. We have been extending the concept of the domain map to include more dynamic processes such as biochemical pathways. 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. A prototype of this Knowledge Map Explorer Tool (Know-Me) is shown in Fig. 6.

© 2002 Maryann Martone

FIGURE XX

Know-Me Tool: A tools for exploring and querying domain maps and process maps. The pathway shown represents a portion of the proposed pathway for LTP. 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 state or transition.



© 2002 Maryann Martone

CONCLUSIONS

Imaging databases like the CCDB are still in their infancy, but it is important for the communities they aim to serve to be involved in their creation and promotion. Their utility will increase as scientists populate these databases and begin to see the possibilities enabled by electronic data representation and access (3). Although biologists can readily point out the difficulties of creating and maintaining these databases (3), (5), the advantages afforded by this new electronic forum for accessing and interacting over data are many. In fact, these 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 they become linked with other web resource 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 across scales and functionalities.

© 2002 Maryann Martone

USEFUL LINKS

The BIRN Project: **http://www.birn.ncrr.gov/**

Federation of Brain Data: **http://www.npaci.edu/DICE/Neuro/**

Semantic Web: **http://www.w3.org/2001/sw/**

The Cell Centered Database: **http://www.ncmir.ucsd.edu/CCDB/**

Montana NeuroSys Project: **http://cns.montana.edu/research/neurosys/**

Computerized Anatomical Reconstruction and Editing Tool: **http://stp.wustl.edu/caret.html**

Yale Senselab Project: **http://senselab.med.yale.edu/senselab/**

The Storage Resource Broker: **http://www.npaci.edu/SRB/**

© 2002 Maryann Martone

REFERENCES

1. Carazo, J.-M., et al., Nucleic Acids Res 27, 280-3. (1999).
2. Miller, P. L., et al., J Am Med Inform Assoc 8, 34-48. (2001).
3. Kotter, R. Philos Trans R Soc Lond B Biol Sci 356, 1111-1120 (2001).
4. Roland, P., et al., Trends Neurosci 24, 562-4. (2001).
5. Toga, A. Nat Rev Neurosci 3, 302-308 (2002).
6. Berman, H. M, et al., Nucleic Acids Res 28, 235-42. (2000).
7. Martone, M. E., et al., J. Struct. Biol. in press (2002).
8. Hines, M. L., Carnevale, N. T. , Neuroscientist 7, 123-35. (2001).
9. Stiles, J. R, et al. in Synapses Cowan, W. M. et al. Eds. (Johns Hopkins University Press, Baltimore, 2001) pp. 681-731.
10. Paxinos, G., Watson, C. The rat brain in stereotaxic coordinates (Academic Press, San Diego, ed. 4th, 1998).
11. Ingenerf, J. et al Int J Med Inf 64, 223-240. (2001).
12. Rachedi, A. et al. Bioinformatics 16, 301-12. (2000).
13. Gupta, A. et al. Proceedings of the 12th International Conference on Scientific and Statistical Database Management (SSDBM'00)
    IEEE Computer Society (2000).
14. Ludaescher, B. et al. Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany. IEEE Computer Society 2001 (2001).

© 2002 Maryann Martone