A Practical Approach for Microscopy Imaging Data Management (MIDM) in Neuroscience

Shenglan Zhang[1], Xufei Qian[2], Amarnath Gupta[2], Maryann E. Martone[1,2]

[1]National Center for Microscopy and Imaging Research, Center for Research in Biological Structural and Dept. of Neuroscience, University of California, San Diego, La Jolla, CA 92093-0608 and [2]San Diego Super Computer Center, University of California, San Diego, CA 92093-0505, USA
{szhang, maryann}@ncmir.ucsd.edu, {xqian,gupta}@sdsc.edu

Abstract

Current data management approaches can easily handle the relatively simple requirements for molecular biology research but not the more varied and sophisticated microscopy imaging data in neuroscience research. We developed a project-oriented experimental imaging data management system through integration of the object-relational Oracle DBMS and a distributed file management system, the storage resource broker (SRB). The data model we developed on Oracle9i supports semantic and analytical queries and image content mining. The MIDM provides comprehensive descriptive, structural, spatial and administrative information on microscopy image datasets. The current MIDM is web accessible at http://ncmir.ucsd.edu/CCDB. This paper describes the MIDM architecture and data mode in MIDM.

1. Introduction

Over 300 databases are now available in the field of molecular biology [1]. Current data management approaches include structured flat files or XML-based methods, relational Database Management Systems, object-relational DBMSs and object-oriented DBMSs [2]. Database resources for scientists engaged in research at the cellular and tissue levels using microscopy imaging are scarce. Although some on-line biology image databases have tried to facilitate image exchange and management, e.g., the QBIC [3], BioImage [4] and PSLID systems [5], most systems do not extract and model the content of the image produced by scientific instruments. Maintaining and managing all of the rich image and image metadata acquired by light and electron microscopy techniques can not be accomplished by any of the current management systems.

The design and implementation of image management systems for neuroscience data faces several scientific and technological challenges including: maintaining large image sizes (typically 3-10GB) and a variety of image types on different storage location; performing semantic queries with obscure scientific nomenclature and heterogeneity; performing analytical queries on tree structured neuron object obtained by different scientific instruments, different preparation methods and multiple microscopy image processing steps; performing spatial queries on multi-resolution and multi-scale microscopy images. This paper describes a microscopy imaging data management system (MIDM) and an object-relational data model which address data grid, data federation, image content retrieval and data lineage issues for managing 2D and 3D microscopic imaging data.

2. Architecture of MIDM

The MIDM was designed to store 2D and 3D light and electron microscopy images, reconstructed image, image analysis and image descriptors, image related experimental data. The microscopy data resources contain heterogeneous multimedia information. The potential multimedia information in the MIDM includes: (i) Still images: Individual micrographs or derived 2D data that are encoded in standard formats (e.g. JPEG). These images may form an orderly sequence related to one another through one or more parameter, e.g., tilt angle, time; (ii) Mixed multimedia data: Compressed image files and parameter files that are bundled together as one 3D volume; (iii) Animations: A sequence of images (e.g. MPEG) that were taken at different tilt angles to illustrate a reconstructed 3D cell structure; (iv) Graphics: Drawings or illustrations that are encoded using some descriptive standards (e.g. PICT); (v) Spread sheets: Formatted cell structure analysis files from a stored data set (e.g. ASCII). The above different types of image and analysis files may be stored on distributed archival resources. Storage resource broker (SRB) as a middleware is able to manage our MIDM different formats of heterogeneous data files distributed on different types of storage devices over the network. SRB provides access to data stored on distributed archival

resources such as HPSS, UniTree and ADSM, file systems, such as the Unix File System, NT File System, and Mac OSX File System and databases such as Oracle, DB2, and Sybase, etc [6]. MIDM takes advantage of the SRB to store variety of image files on distributed archival resources and transparently and securely retrieve multimedia files or related files across network that supporting the web access of MIDM.

The description of each image object is essential to provide identification of the images and information about their content and make interdisciplinary usage, image retrieval and image federation possible. Every image object in MIDM is accompanied by descriptive, structural, and administrative metadata. The descriptive metadata includes the experimental description, file description, image annotation, image evaluation, and image object measurement and analysis. Experimental description includes experimental subject properties, sample preparation, instrument parameters, and data processing methods. Structural metadata is the coordinates of a given image object within a whole brain coordinate system based on a standard brain atlas. The structural metadata is important for the federation of images in MIDM with internal correlated microscopy images or images external to the MIDM. The administrative metadata contains the access control information used to implement authentication and authorization for data access in the object level. In MIDM, all the above image descriptive metadata is managed using the Oracle 9.0.1.4.0 platform. The logical SRB addresses of the image files are stored in the Oracle database and link image files in SRB to metadata in Oracle. Any image data query performed on MIDM will first retrieve the logical address of the images from Oracle, and the SRB sends the image to the client (Figure 1).

The Oracle database in MIDM is designed as a structured representation of the accomplished projects. The data model is mainly to serve the semantic and analytical queries and information management. However, scientists need a record keeping staging area to store the large variety of experimental descriptions and experiment records that can't be mapped to the data model in MIDM. In addition, the data model adds complexity for designing a user friendly interface for the scientist to input data during the experiment because of constraints. We are planning to implement a semi-structured (XML) data representation for record keeping of experimental data. The semi-structured data model allows researchers to document their data in an organized, regularly formatted, and network accessible manner. Using two databases (XML database and Oracle database) allows us to separate the semi-structured experiment raw information from the final structured project information. The data that are stored in the XML database can be parsed into the Oracle database when the project is completed (Figure 1). The additional semi-structured XML database provides a necessary "buffer" to reduce the transactions and traffic to the Oracle database. The validation can also be processed before a complete metadata set is loaded into the Oracle database.

3. Data Model in MIDM

The current Oracle database, Cell Centered Database (CCDB) comprises 65 tables, as illustrated in a simplified overview diagram in Figure 2. We model the entire process of 3D microscopy reconstruction, from specimen preparation to segmentation and analysis. Most of the data types and aggregates for CCDB are self-defined by the neuroscientists. The data model for microscopy imaging data composes the data lineage, which includes the pipeline of processing steps (tissue processing, microscopy product, reconstruction and segmentation; Figure 2). With this design, post facto analytical queries on comparison of neuronal structure can be performed by progressive series of queries through the processing metadata.

The vision of developing a data model in MIDM is to integrate different categories of cell level imaging data and derived image data products to provide a resource for cell biologists and to provide the means for further database federation. Additional description of the purpose of the CCDB and the types of data it was designed around can be found in Martone et al. [7]. The federation of our Oracle database with other cell level databases (e.g., a neurotransmission senselab database developed at Yale University; http://www.med.yale.edu/senselab) and multi-scale data such as the protein data bank developed at Rutgers (http://www.rcsb.org/pdb), creates a system that will allow the scientist to discover information that is not present in a single information source. For example, after federation of multiple biological databases, a neuroscientist can retrieve information about the structure of a protein that is expressed in certain compartments in particular nerve cells from a brain region. In this query, the CCDB will serve as one source in as part of the BIRN data mediation framework [8].

For more than 30 years, the base functionality for information retrieval remained a query by a set of words; those items in the collection that contain those words are returned. The fundamental technology for searching large collections finally changed, so that information retrieval in the next century will be far more semantic than syntactic, searching concepts rather than words [9]. The current approach to concept search is to create a semantic translation database (e.g. ontology database) to supply the

necessary expert knowledge to bridge concepts across databases. Additional information on this system can be found in Ludaescher et al., [10]. Our CCDB contains three entities to allow the CCDB to link to an ontology database. The ontology entity (ontology_id, ontology_object_relationship, ontology name) stores information on an ontology object that defines an object in CCDB database. A CCDB look up entity (ccdb_id, table_type, table_name, primary key name) was designed to link the ccdb_id with a primary key in the particular table in the CCDB database. The relationship between a ccdb_id and an ontology_id is stored in the CCDB entity (ontology_id, ccdb_id, SRB_id, srb_path, atlas_id) to link CCDB database with ontology database. The CCDB, ontology, and CCDB lookup entities allow scientists to search subject domains in unfamiliar areas; an intermediary such as an ontology database can often translate the term in one subject into standard terms within another.

The CCDB was designed not only to serve as an image repository for simple text-based retrieval, but also provides image content retrieval. Although content-based image retrieval has been widely researched to retrieve desired images on the basis of features (such as color, texture, shape) that can be automatically extracted from the images themselves, it still lacks a universally accepted methodology for evaluation measures. Reliance on predefined requests, with little end-user involvement or interaction, has been criticized [11]. MIDM provides an approach to query image attributes by designing queries around quantitative data derived from particular datasets. The current version of the schema in CCDB implemented tree structured object types (e.g. dendrite) that allow user to perform the queries on neuronal structure, such as dendritic branching pattern. The tree structured object type in CCDB is mapped to the computed output file from the program, Neurolucida (Microbrightfield, VA). Neurolucida is used to address specific research questions that require quantitative information on neuronal processes in 3D by performing segmentation of neuronal branching structures and anatomical mapping. As the neuron is traced, a battery of measurements (e.g. tree_number, branching_order, coordinate etc.) is made automatically and parsed into the Oracle database for different tree structured objects that represent different neuronal compartments including spine, cell body, axon, dendrite etc. In the future, user defined functions will allow statistical queries on the top of these defined tree structured objects. MIDM uses this approach to allow the user to retrieve images on the basis of measurement or statistical analysis of objects that are segmented from an image. These queries may be quite specific, for example, the user may query for the primary dendrite's left child dendrite's average length and standard deviation among a group of purkinje cells, or the user may query for a significant difference in dendritic branching pattern for Purkinje neuron cell between two different species in the CCDB database.

4. Conclusion

MIDM addresses the data federation, data grid and data lineage issues for microscopy imaging management by integrating Oracle, XML and SRB servers. The data model in MIDM is proving to support scientific analytical and semantic queries and image content mining.

5. References

[1] A.D. Baxevanis, "The Molecular Biology Database Collection: 2002 update", *Nucleic Acids Research,* 30, 2002, pp. 1-12.

[2] F. Achard, G. Vaysseix, and E. Barillort, "XML, bioinformatics and data integration", *Bioinformatics*, 17, 2001, pp. 115-125.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, H. Qian, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the QBIC system", *Computer*, 28, 1995, pp. 23-32.

[4] J.M. Carazo, and E.H.K. Stelzer, "The BioImage database project: Organizing multidimensional biological images in an object-relational database", *Journal of Structural Biology*, 126, 1999, pp. 97-102.

[5] K. Wang, J. Lin, J. Gajnak, and F. Murphy, "Image content-based retrieval and automated interpretation of fluorescence microscope images via the protein subcellular location image database", *IEEE*, 2002, pp. 325-328.

[6] A. Rajasekar, M. Wan, and R. Moore, "MySRB & SRB – Components of a Data Grid", In *The 11th International Symposium on High Performance Distributed Computing (HPDC-11*. Edinburgh, Scotland, 2002.

[7] M.E. Martone, A. Gupta, M. Wong, X. Qian, G. Sosinsky, B. Ludaescher, and M.H. Ellisman, "A cell centered database for electron tomographic data", *Journal of Structural Biology***,** 138, 2002, pp. 145-155.

[8] A. Gupta, B. Ludaescher, and M.E. Martone, "Registering Scientific Information Sources for Semantic Mediation", Proc. 21st International Conference on Conceptual Modeling, (ER), Tampere, Finland, October, 2002.

[9] B.R. Schatz, "Information Retrieval in Digital libraries: Bring Search to the Net", *Science*, 275, 1997, pp. 327-334.

[10] B. Ludaescher, A. Gupta, and E.M. Martone, "Model-Based Mediation with Domain Maps", 17th Intl. Conference on Data Engineering (ICDE), Heidelberg, Germany, IEEE Computer Society, 2001.

[11] J.P. Eakins, and M.E. Graham, "Content Based Image Retrieval: A report to the JISC Technology Applications Program", Inst. for Image Data Research, Univ. of Northumbria at Newcastle, 1999.

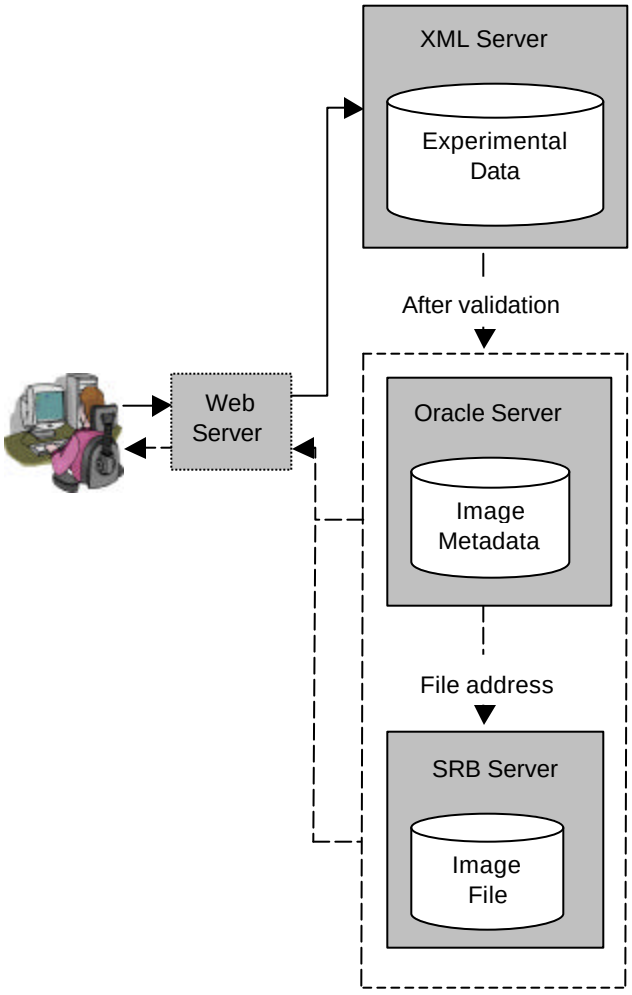

Figure 1. Architecture of MIDM

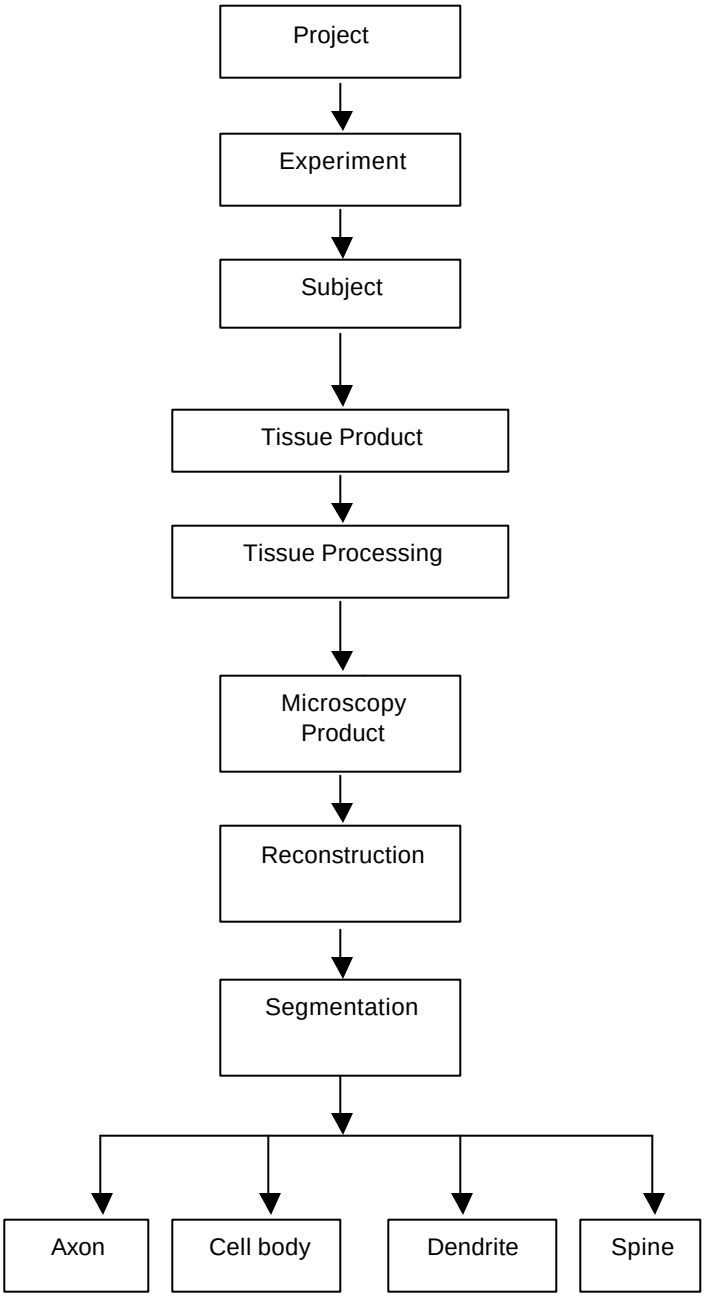

Figure 2. Overview of data model implemented on Oracle9i. The detail schema is at http://pamina2.sdsc.edu/CCDB/IMG/ER071602.jpg