

## The Cell Centered Database Project: Advanced Informatics for Federation of Distributed Multi-scale Imaging Data

Maryann E. Martone<sup>1,2</sup>, Amarnath Gupta<sup>3</sup>, Yujun Wang<sup>1</sup>, Lihong Sun<sup>1</sup>, Ilya Zaslavsky<sup>3</sup>, Haiyun He<sup>2</sup>, Diana Price<sup>1,2</sup>, Joshua Tran<sup>3</sup> and Mark H. Ellisman<sup>1,2,3</sup>

<sup>1</sup>Department of Neurosciences, <sup>2</sup>Center for Research in Biological Systems and <sup>3</sup>San Diego Supercomputer Center, University of California, San Diego, San Diego CA 92093.

A continuing challenge to structural biologists is the understanding of structures on the scale of 1 nm<sup>3</sup> to 10 μm<sup>3</sup>, a dimensional range that encompasses macromolecular complexes, organelles, and multi-component structures like synapses. Such structures have traditionally been difficult to study because they fall in the resolution gap between technologies, spanning X-ray crystallography, electron microscopy and light microscopy. Structures at this scale represent the heart of information processing in the nervous system and provide a bridge between the molecular information being assembled at one end of the biological continuum and the large scale brain mapping being performed at the other. These structures will have to be solved if the results of the molecular revolution, the protein products of sequenced genomes, are to be situated in their proper subcellular, cellular and tissue contexts.

Technological improvements in both light and electron microscopic imaging have led to major advances in our ability to fill in missing information between the molecular and cellular realms. At the light microscopic level, computational image restoration techniques and the optical sectioning capabilities of the confocal and multiphoton microscopes are increasing the resolution achievable with optical-based imaging techniques. At the electron microscopic level, electron tomography is revealing new information about the three-dimensional ultrastructure of tissues, cells and macromolecular complexes. Using tomography, a 3D reconstruction is obtained from a series of 2D projections with the potential of mapping the location and segregation of functionally important macromolecules at much higher resolution than can be achieved by serial sectioning.

While technological advances in instrumentation and specimen preparation are accelerating the rate at which we investigate biological systems across scales, integrating this mountain of data into unified models of biological systems also requires the simultaneous development of appropriate data management tools. In recent years, as the speed of networks has increased and the cost of computational power has decreased, information technology has moved away from the creation of large scale, centralized resources to a more distributed model. For scientists, the so-called “grid movement” in computer science promises the means to establish virtual communities of collaboration where access to data, computational resources and instruments is provided as if it were local, regardless of where they physically reside.

In this presentation, we will highlight challenges and successes in building neuroinformatics resources for distributed, multi-scale imaging data utilizing grid-based architecture as part of the Cell Centered Database (CCDB), Telescience, and Biomedical Informatics Research Network (BIRN) projects. With support from the Human Brain

Project, the CCDB has been developed and deployed as a web-based relational database for modeling and querying complex microscopic imaging data that utilizes a distributed file collection management system ([www.ncmir.ucsd.edu/CCDB](http://www.ncmir.ucsd.edu/CCDB)). The CCDB is a stand-alone database but has also been used as a prototype for developing methods for promoting interoperability among independently developed resources. For example, the CCDB is serving as the data management system for the Telescience Portal, a web portal-based system that provides researchers with remote access to resources required for electron tomography. It also serves as one of a series of federated databases as part of the BIRN project, a large multi-institution project initiated by the National Institutes of Health to build a biomedical data grid using advanced networks and “Grid” technologies. The BIRN project has developed a mediator system to link databases like the CCDB into virtual data federations so that they can be cross-queried as if a single database. The BIRN mediator also incorporates additional knowledge sources at time of query to provide the necessary “glue” knowledge to query across databases encompassing different anatomical scales and modalities. These knowledge sources include both ontologies and spatial coordinate systems. The mediator is designed to work in conjunction with tools such as the Smart Atlas, a geographical information system (GIS)-based tool for browsing and integrating distributed spatially-referenced, multiscale brain data. The Smart Atlas allows researchers to bring together multi-scale and multi-modal data distributed across multiple sites that has been referenced to a common coordinate system. These projects illustrate how independently generated resources can be linked together in flexible and powerful ways to serve biomedical science.

This work supported by NIH grants from NCRN RR04050, RR08605 and the Human Brain Project DA016602 with support from the National Institute on Drug Abuse, the National Institute of Mental Health and the National Institute of Biomedical Imaging and Bioengineering, and NSF grants supporting the National Partnership for Advanced Computational Infrastructure NSF-ASC 97-5249 and MCB-9728338. The Bioinformatics Research Network is supported by NIH grants RR08605-08S1 (BIRN-CC) and RR043050-S2 (Mouse BIRN).