Available online at www.sciencedirect.com
ScienceDirect
Journal of Structural Biology 161 (2008) 220–231
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

The Cell Centered Database project: An update on building community resources for managing and sharing 3D imaging data

Maryann E. Martone [a,b,*], Joshua Tran [a,b], Willy W. Wong [a,b], Joy Sargis [a,b], Lisa Fong [a,b], Stephen Larson [a,b], Stephan P. Lamont [a,b], Amarnath Gupta [b,c], Mark H. Ellisman [a,b]

[a] Department of Neurosciences, University of California at San Diego, San Diego, CA 92093-0608, USA
[b] Center for Research in Biological Systems, University of California, San Diego, CA 92093-0446, USA
[c] San Diego Supercomputer Center, University of California, San Diego, CA 92093-0515, USA

Received 2 July 2007; received in revised form 4 October 2007; accepted 5 October 2007
Available online 16 October 2007

Abstract

Databases have become integral parts of data management, dissemination, and mining in biology. At the Second Annual Conference on Electron Tomography, held in Amsterdam in 2001, we proposed that electron tomography data should be shared in a manner analogous to structural data at the protein and sequence scales. At that time, we outlined our progress in creating a database to bring together cell level imaging data across scales, The Cell Centered Database (CCDB). The CCDB was formally launched in 2002 as an on-line repository of high-resolution 3D light and electron microscopic reconstructions of cells and subcellular structures. It contains 2D, 3D, and 4D structural and protein distribution information from confocal, multiphoton, and electron microscopy, including correlated light and electron microscopy. Many of the data sets are derived from electron tomography of cells and tissues. In the 5 years since its debut, we have moved the CCDB from a prototype to a stable resource and expanded the scope of the project to include data management and knowledge engineering. Here, we provide an update on the CCDB and how it is used by the scientific community. We also describe our work in developing additional knowledge tools, e.g., ontologies, for annotation and query of electron microscopic data.

Keywords: Electron tomography; Bioinformatics; Ontology; 3D reconstruction

1. Introduction

The Cell Centered Database (CCDB) project was launched in 2002 as an on-line repository of high-resolution 3D light and electron microscopic reconstructions of cells and subcellular structures (Martone et al., 2002, 2003, 2007). The CCDB contains data covering the dimensional range known as the "mesoscale", roughly encompassing the structures that sit between gross morphology and molecular structure, e.g., cellular networks, cellular and subcellular microdomains along with their macromolecular constituents. The study of mesoscale structures, like dendritic spines, continues to present a challenge to experimentalists, because their dimensions fall squarely between the capabilities of current imaging technologies. Investigations of physiology, structural dynamics, coarse molecular distributions, and large scale distributions of dendritic spines are typically accomplished by optical microscopies. Appreciation of the fine structural detail on internal structure, cytoskeletal organization, localization of molecular constituents, location of synaptic contacts, and detailed views of the immediate microdomain such as pre-synaptic boutons and glial processes require 3D electron microscopic imaging. To build a comprehensive understanding of complex tissues in this dimensional range requires the ability to aggregate data obtained by multiple researchers across techniques and spatial scales.

* Corresponding author. Department of Neurosciences and Center for Research in Biological Systems, University of California, San Diego, San Diego, CA 92093-0446, USA. Fax: +1 858 822 3610.
E-mail address: mmartone@ucsd.edu (M.E. Martone).

1047-8477/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2007.10.003

Of current techniques, electron tomography is providing some of the most significant and spectacular information about mesoscale structures, with its ability to situate macromolecules in their 3D cellular contexts (Lucic et al., 2005; Marsh et al., 2004). One of the main motivations in the creation of the CCDB was to provide a forum for the very rich and valuable data sets produced by electron tomography to be made available to the public. The original CCDB was first proposed to the electron tomography community at the 2nd International Conference on Electron Tomography held in Amsterdam in 2001. At that time, the CCDB existed more as a concept than an actual product. By the time the special issue of *Journal of Structural Biology* arising from that conference was published in 2002, however, the first public version of the CCDB was on-line (Martone et al., 2002). The support for the CCDB was provided by a grant through the Human Brain Project (Wong and Koslow, 2001), a program designed to produce computational tools and databases for sharing scientific data with the broader scientific community. Over the past 5 years, we have continued to refine the architecture of the CCDB and have moved it from a prototype to a stable infrastructure. At the same time, we have had to refine our vision of the CCDB in response to community feedback, technological advances in knowledge engineering and our own experiences with sociological, technical and biological aspects of data sharing. In this paper, we present an overview of the current CCDB, our experiences in its creation, and plans for future development.

2. Materials and methods

2.1. Current architecture of the CCDB

The public CCDB is available at http://ccdb.ucsd.edu. The data model of the CCDB is illustrated in Fig. 1, which shows a highly simplified view of the schema. The CCDB was built using a combination of enterprise software components and cyberinfrastructure developed largely in an academic setting. The current CCDB utilizes Oracle 10g as the relational database management system with additional applications written in Java. Data entry forms for the CCDB were built using Gridsphere, an open source project for building secure java-based web portals (www.gridsphere.org). Because Gridsphere components, called portlets, are built to a common specification, the CCDB input forms may be easily incorporated into any Gridsphere-compliant portal.

The CCDB utilized the basic architecture developed by the Biomedical Informatics Research Network (BIRN; Grethe et al., 2005) and Telescience (Peltier et al., 2003) projects for distributed file storage and access. The BIRN project is an example of a so-called "grid" project, predicated on a model of distributed hardware and software. The basic idea behind most cyberinfrastructure projects like BIRN is that it should not matter where a resource is located physically or what hardware it is using. Programmatic access and security should be uniform across all of these resources. This uniformity is provided by software layers, "middleware", that sit between the physical resource and the programs required to access it. The CCDB utilizes both the distributed collections manager called the Storage Resource Broker (SRB; Grethe et al., 2005) and the authentication mechanisms for CCDB files (Peltier et al., 2003).

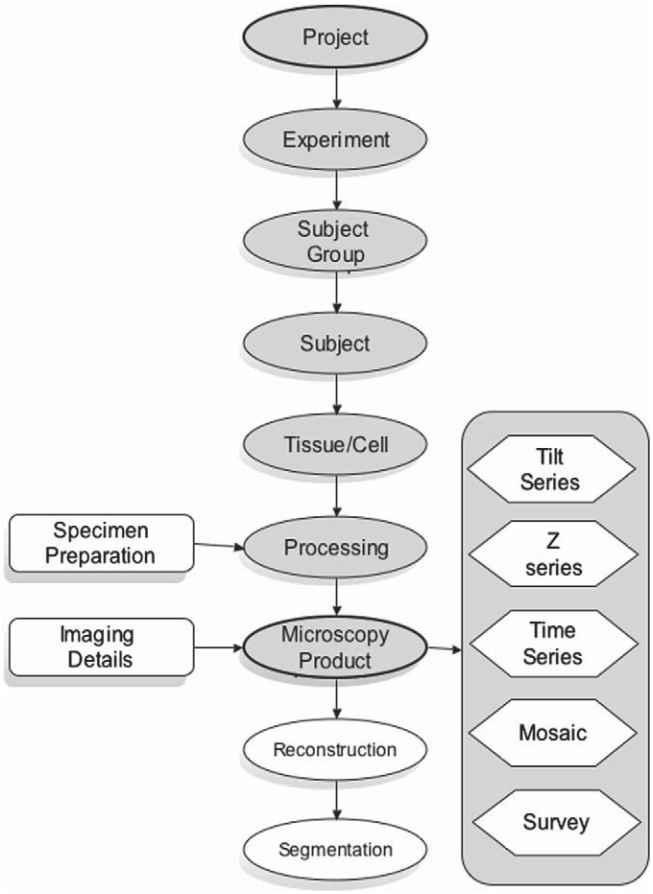

Fig. 1. Simplified view of the CCDB schema, showing the main classes of information contained in CCDB tables. The Microscopy Product provides the unique identifier for the CCDB database. Each oval represents a single table in the CCDB; other shapes are categories of information that are elaborated in multiple tables. The relationship between the core tables of the CCDB (ovals) is one to many, that is, one project can have many experiments and so on. All microscopy products must be registered within their experimental contexts so the first seven tables are required (gray ovals). The tissue and processing tables contain a minimal set of specimen preparation details, while the microscopy product contains a minimal set of imaging details. More detailed specimen preparation and imaging protocols are stored in additional tables. The general classes of microscopy products are illustrated in the gray box to the right.

2.2. Data model

The CCDB was designed around the process of reconstruction from 2D micrographs, capturing key steps in the process from experiment to analysis. The core tables shown in Fig. 1 represent the backbone of the CCDB, each

of which may be further elaborated in additional tables specialized for a particular technique or data product. By designing the CCDB in this fashion, we can easily add new techniques or features without breaking the structure of the database. The full entity-relationship diagram of the CCDB is available on the CCDB web site.

At the center of the CCDB data model is the ''Microscopy Product'' table. The microscopy product refers to a set of related 2D images taken by light (epifluorescence, transmitted light, confocal or multiphoton) or electron microscopy (conventional or high voltage transmission electron microscopy). For each type, the images comprising a single microscopy product are systematically related to each other along a single dimension (Fig. 1). For example, a tilt series contains images that are related to each other through tilt angle; a z series is related through depth and a mosaic contains images that are related through X, Y coordinates. A given set of data may be more than one product, for example, it is possible for a set of images to be both a mosaic and a tilt series. The CCDB also contains a microscopy product type ''Survey section'' which refers to a set of images that were taken in the same session to survey a specimen.

The microscopy product refers to the raw data that comes off the microscope and may be in the form of negatives or digital images, depending on the recording device. The microscopy product identifier serves as the accession number for the CCDB. This value was chosen because a given dataset may only be taken once and therefore each microscopy product represents a unique dataset. The CCDB distinguishes between the original microscopy product and any 2D images that may derive from it. For example, if the microscopy product was a set of negatives, they will have been digitized prior to reconstruction. A microscopy product may give rise to multiple sets of 2D images; however, if a specimen is re-imaged, it is considered a new microscopy product.

3. Results

3.1. Rationale for CCDB design

The CCDB was designed with an eye towards encouraging re-analysis and re-use of tomographic data and for mining the content of these datasets. From its inception, unlike many of the genomics and protein structure databases, the CCDB provided not only the final data product, usually a 3D reconstruction, but the raw data, specimen preparation details, the imaging parameters and any derived data products created as a result of analysis of the reconstruction. Thus, the schema of the CCDB ensures that researchers can trace the provenance of a piece of data and understand the specimen preparation and imaging conditions that led to it, while making raw and derived data available for reanalysis. The ability to access raw or at least minimally processed data is important because as new reconstruction algorithms and techniques are developed, data sets that previously were reconstructed at low-resolution may be improved. For example, CCDB dataset #27 (http://ccdb.ucsd.edu/sand/main?event=displaySum&mpid=27) represents a reconstruction from a tilt series where no fiducial marks were available. The original images are in focus and of high quality, but the subsequent reconstruction was poor. However, if a non-fiducial mark based approach becomes available, this data set may be reprocessed.

The types of imaging data stored in the CCDB are quite heterogeneous, ranging from large scale maps of protein distributions taken by confocal microscopy to 3D reconstruction of individual cells, subcellular structures and organelles reconstructed using electron tomography see (Martone et al., 2003, for more details). Many of the data sets in the current CCDB derive from the nervous system, reflecting our own area of expertise. However, the schema of the CCDB is generic for 3D imaging and may accommodate any cell type. For example, we have recently published a set of tomographic reconstruction of blue green algae (Moisan et al., 2006).

The CCDB stores not only the original images and 3D reconstruction, but also any analysis products derived from these data, including segmented objects and their associated measurements, e.g., surface area, volume, length and diameter. Thus, each data record in the CCDB consists of a set of primary images (referred to in CCDB as Image 2D), and any derived data products, e.g., reconstructions and segmentations (Fig. 2). For each of these data products, the full resolution data files and supporting files, e.g., fiducial mark files, are made available for download.

The CCDB actually splits the storage of information about the microscopic images into two tables: microscopy product and image 2D. The microscopy product is meant to describe the raw data collected from the microscope while the image 2D table is meant to store information about the 2D images that were actually used to create 2D, 3D, or 4D a reconstruction. In the case of digital imaging, the sets of images may be the same, e.g., an optical section series taken from a confocal microscope in which no further processing was performed. In the case of images taken on film, the microscopy product and the image 2D set are not the same; the 2D images are digitized from the negatives. Even in the case of digital images, the original data may be down sampled or cropped before reconstruction.

Because datasets are large and multidimensional, CCDB also stores a set of 2D images and animations that will display easily on the web. Submitters are asked to supply a representative 512×512 images for each of the image data types (Image 2D, Reconstruction and Segmentation) that CCDB stores (Fig. 2). A representative image provides information about the content of the dataset, e.g., for a tilt series might be the zero tilt image, while for a 3D reconstruction, it might be a projection or a single slice through the volume. These display images are meant to give a quick visual guide to the content of a dataset. Because most of

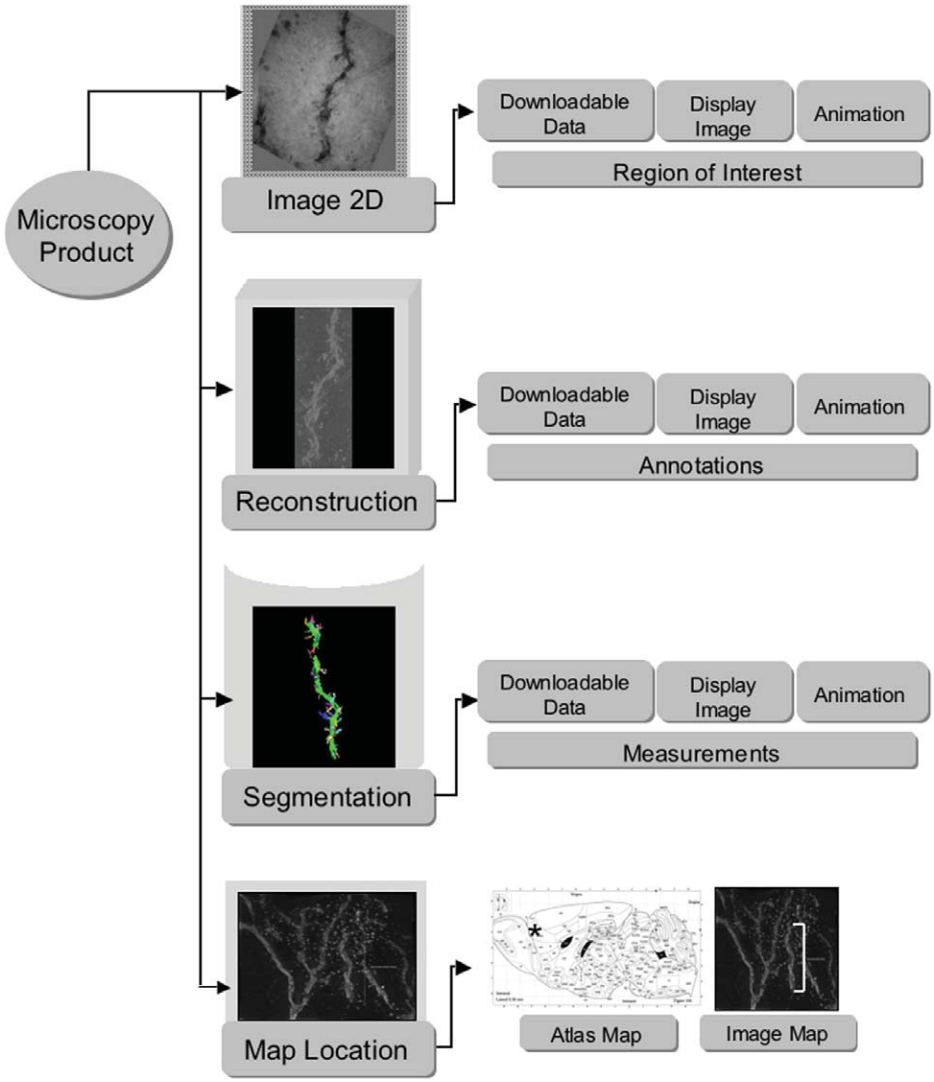

Fig. 2. Overview of types of image data and analysis data stored in an individual record. Each microscopy product may have one or more types of data stored along with it: 2D Image, Reconstruction and Segmentation. Supporting images indicating where the data were taken with respect to an anatomical atlas or an image map providing broader context may also be provided (Map Location). For each of the three main data type, the CCDB makes available the full resolution file for download (downloadable data), a display image and an animation for viewing on the web. Various annotations and measurements are also stored where appropriate.

the data are 3D, however, the CCDB also encourages the submission of animations that allow users to explore the 3D data in more detail on the web. These animations may be aligned tilt series, rotation loops of reconstructions or animations through the slices of a volume. Unlike many journals, the CCDB places no limit on the size of data or supplementary files that can be submitted. For this reason, submitters to the CCDB have found it useful to provide supplementary material for a journal article to CCDB rather than to the journal (e.g., Sosinsky et al., 2005).

The CCDB also provides the means to attach other types of image files to a given record that will aid in the interpretation or display of image content. For interactive browsing of very large 2D images, the CCDB employs the Zoomify package for interactive viewing of large images over the web (http://www.zoomify.com/).

Anatomical information is recorded in essentially three places in the CCDB: (1) in the specimen preparation portion, the general anatomical region from which the section is derived is recorded; (2) at the microscopy product level, the anatomical characterization of the subject of the imaging session is recorded from the level of system down to subcellular structure, e.g., a Purkinje cell from the vermis of the cerebellum; (3) in the segmented object table, each object extracted from the reconstruction is listed along with a description of the object, e.g., dendritic spine segmented from the Purkinje cell. Note that in these three tables, the level of anatomy is specified from coarser to finer anatomical scales, reflecting the progressive subsampling of a biological specimen that typically occurs during preparation, imaging and segmentation. In order to retain this larger anatomical context for the users, the CCDB allows the

storage of maps that show the location of a data set in terms of a standard atlas, e.g., a brain atlas (Fig. 2). It also allows for the storage of an image map that can situate a reconstruction within the larger context of the original image or a light microscopic image of the same sample.

All imaging data deposited in the CCDB must be accompanied by project, experimental, subject and protocol details because proper interpretation of imaging data requires a thorough understanding of the experimental and imaging conditions under which it was acquired. The design of the specimen preparation tables for CCDB has presented quite a challenge because of the flexible nature of specimen preparation protocols. Relational databases derive their power from the highly structured and rigid nature of the table structure; they are not extremely well suited to the type of fluid and changeable nature of experimental protocols. In the original schema of the CCDB, users could store details about specimen processing, but the order in which the steps were performed was lost. Detailed protocols could also be stored as unstructured text. Because the order in which steps are performed is important when comparing protocols across experiments and because unstructured text makes the process of query difficult, the newest version of the CCDB has recently redesigned the specimen preparation tables, implementing a structure whereby key specimen preparation steps such as fixation, staining, embedding, microtomy, and macromolecular localization can be performed in any order and in multiple points in a protocol. The specimen preparation tables are currently in beta testing and will be part of an upcoming release.

3.2. Data input to CCDB

The CCDB accepts data from outside users for dissemination through the CCDB. The CCDB has recently released the first set of input forms for uploading data into the CCDB. The CCDB input forms provide a means for users to submit data for the public CCDB. However, because the data model of the CCDB essentially models the process of 3D reconstruction and analysis from microscopic data sets, the CCDB input forms provide the user with a data management system which can be used to track and organize experimental data during the course of an experiment.

CCDB input forms are available through a secure portal, ''MyCCDB'', that requires a private account, available on request. All data that is entered into the CCDB through the portal is considered private data (Fig. 3). Users may assign group privileges to allow their collaborators to view the data, but this data is not displayed in the public CCDB. While the data are private, users may edit the data and delete records as necessary.

The workflow involved in entering a data set into the CCDB is illustrated in Fig. 3. Users begin the process by registering a project to the CCDB; a project is defined as a group of related experiments or studies that generally lead to a single publication. Once a project has been registered, experiments, subjects, specimens and microscopy products may be registered to the same project in a hierarchical fashion (see Fig. 1). Each of these entities is assigned a unique identifier that can be used to tag the same entities in laboratory records, negatives, grid boxes, etc. Users may add anatomical descriptions, microscope parameters and image set parameters, e.g., tilt range and tilt increment, to a given microscopy product. The image files themselves are uploaded to the SRB. The CCDB automatically creates a link to the stored files upon upload. The CCDB is designed so that all tables do not have to be filled out in the same session. Upon logging in, users may restore a given session and add details and data as they become available.

The process of data entry into the CCDB is currently manual for many types of data. However, we have deployed a set of input forms through a custom portal developed for the National Center for Microscopy and Imaging Research (NCMIR), the P41 technology development center that hosts the CCDB. The Telescience Multiscale Imaging Portal (https://.telescience.ucsd.edu), a Gridsphere-based portal, provides secure access to high voltage electron microscopes, computational resources and tools for electron tomography, all using a single user account (Peltier et al., 2003). The CCDB accepts the authentication certificate issued by Telescience, so that users logged into Telescience are also authorized to enter data into the CCDB. Through the Telescience Portal, we have created a set of services that allow communication between CCDB and resources available at the center, including instruments like our JEOL intermediate voltage electron microscopes, and tools for reconstruction, segmentation and analysis (e.g., Lawrence et al., 2006). Once a user registers a microscopy product, they may acquire a tomographic tilt series through the remote microscopy client. As the images are acquired, they are stored in the SRB while the instrument parameters are input to the CCDB.

3.3. Publishing data to the public CCDB

When data are made public, the CCDB performs the necessary curation to ensure that the forms are filled out correctly and changes the security access in Oracle. As illustrated in Fig. 1, a minimum set of metadata, elaborated in the first seven tables of the CCDB, must be provided for each dataset. If all required fields within these tables are not filled out with valid values, the data set will be returned to the owner for modification. Contributors still own the copyright to their data, but once the data are published in the public CCDB, contributors are no longer able to edit or delete their data. If corrections are necessary, the CCDB curators will add an erratum or addendum to the record, similar to the procedures used by journals.

If the data are from a published article, the CCDB references the original publication in which the data appeared. We are currently working on a mechanism so

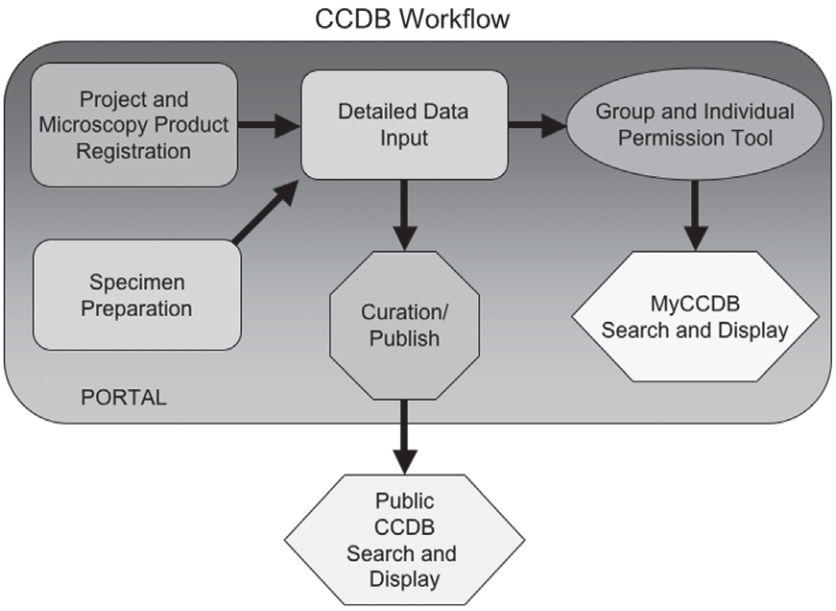

Fig. 3. Schematic overview of data entry and publication process of the CCDB. Data entry occurs within a secure web portal, requiring an account for access. The data entry process is broken up into several steps: Project and microscopy product registration, specimen preparation and detailed data input. Image data files and display images are uploaded during the detailed data input stage. While in the secure portal, users may view their data through the private search and display pages (MyCCDB) and set and manage permissions to allow other users to enter or view data for a given data set. Once the process of data entry is complete, the data set is submitted for curation and publishing to the public site. At that point, the data are available through the public CCDB.

that the availability of data for a given article will be indicated in the "Links" field of Pub Med. However, prior publication in a peer-reviewed journal is not a requirement for depositing data in the CCDB. Many imaging laboratories have high quality data sets that do not make it into a publication but still have value to the community. The CCDB is an ideal place for this type of orphan image data.

3.4. Browsing, searching, displaying, and downloading CCDB data

As of the writing of this manuscript, the CCDB has 185 datasets available to the public, of which 35 come from electron tomography. The CCDB does not require a user name or password to view the public data. Some of the features of the search and display page are shown in the schematic in Fig. 4. In response to a query, data shown on the CCDB display page is dynamically generated. The search results provides a set of thumbnails that represent the three categories of image data: 2D images, reconstructions, and segmented objects, that may be stored for a given microscopy pro duct ID. In this way, users can rapidly scan through the available data and determine what data are available for each product.

The CCDB provides several search and browsing capabilities that allow users to sort the data according to different views (Fig. 4). For example, once a set of microscopy products are returned, the user may choose to view other microscopy products acquired as part of the same project. This project view was implemented for two reasons. First, for data management purposes, users of MyCCDB find it useful to view the structure of a project, e.g., how many experiments were done, how many subjects were in each group, how many microscopy products were generated. Second, for users of the public data, it is often useful to be able to view other datasets taken as part of the same project. These datasets are likely taken under similar conditions and so may be more appropriate for reanalysis as a group than data taken under very different experimental conditions.

Users may select single data sets for download through several mechanisms, or may cache multiple datasets for simultaneous download. Before downloading the data, users must agree to the usage terms, which includes acknowledgment of the contributor of the data to the CCDB and acknowledgement of the CCDB itself. The CCDB does not currently enforce a standard data format but tools for visualizing most of the data formats are listed on the CCDB website under "Tools". In some cases, if no tool is available, the CCDB will convert the data if possible into a format that is more generally readable. Most data can be viewed using the open source program ImageJ by using available plug ins (http://rsb.info.nih.gov/ij/).

Much of the description of file formats and tools used to produce data files is currently in the form of free text. This situation is not optimal because the level of detail provided can vary from record to record. In a future version of the CCDB, we will implement a more structured representation of data properties and software tools. Standards for descriptions of software tools and data are in development in several communities, e.g., the Neuroimaging Informatics Tools and Resources Clearinghouse (http://www.nitrc.org/).

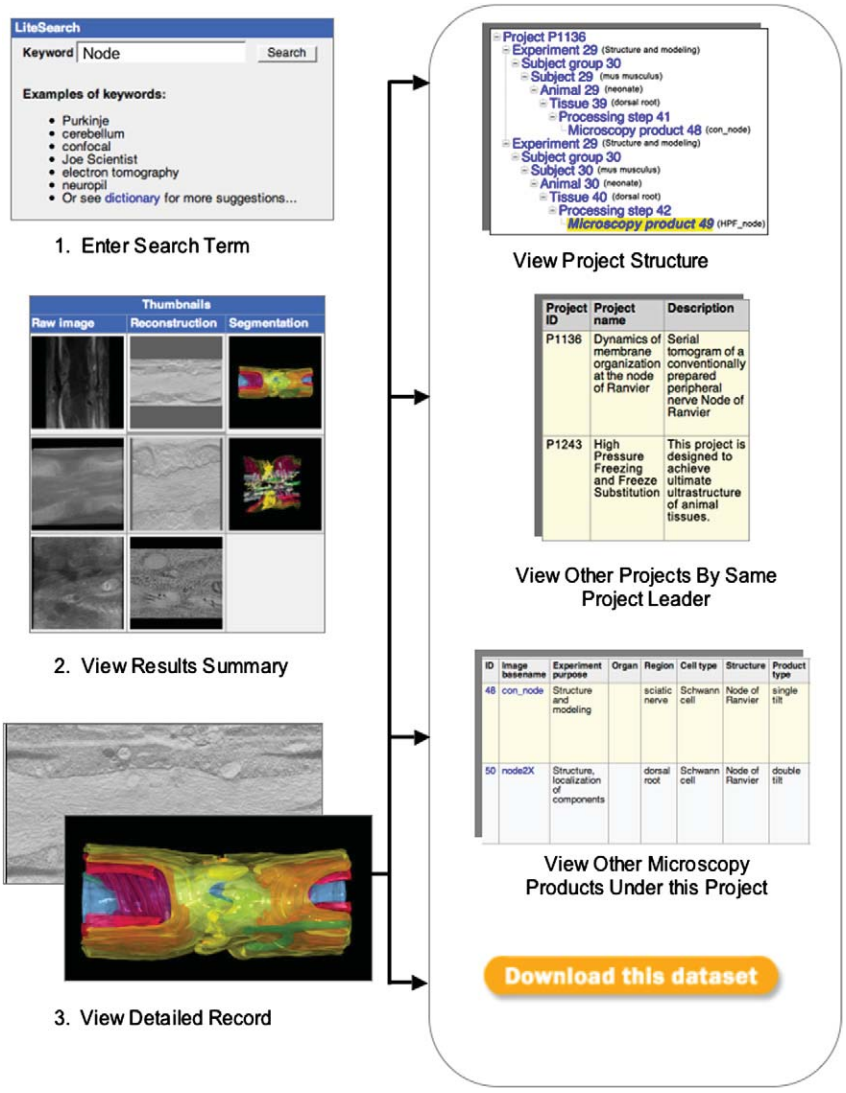

Fig. 4. Summary of the search and display functions of the CCDB. After a search term is entered (1), a summary of the search results is displayed (2) providing a set of thumbnails showing what type of data (Image 2D, Reconstruction, and Segmentation) are stored for each record. Once a data set is selected, a more detailed record is displayed (3), allowing users to browse higher resolution images, animations, image maps and to view all metadata. Within a detailed record, users are given several options for further browsing or downloading (right panel). Users may choose to view the project structure for the project, showing which data sets were acquired from the same subject, for example. Other projects by the same contributor or other microscopy products under the same project may be viewed. Users may also choose to download the data set or add the data to a cache for future download.

The CCDB also allows programmatic access to the CCDB through the creation of a set of web services (see Neerincx and Leunissen, 2005), which allow an application to issue a remote query against the CCDB. The CCDB is also participating in several large scale projects designed to provide software layers that promote cross-query of databases, e.g., the BIRN project (Grethe et al., 2005).

3.5. Ontology for subcellular anatomy

One of the key goals of the CCDB project is to ensure that data acquired through different techniques and at different resolutions can be integrated into a model of the cell. The data model of the CCDB itself does little to achieve this goal, as it mainly elaborates the experimental, imaging and reconstruction details about the data set. In order to provide a more "cell centered" view of data in the CCDB, we have recently released the Subcellular Anatomy Ontology (SAO), a formal ontology that describes the subcellular parts and how they come together to form supracellular domains. An ontology consists of a set of concepts, or entities, within a domain linked by relationships such as "*is a*" and "*has part*", e.g., "*neuron is a cell*" and "*cell has part plasma membrane*". Ontologies are highly valuable in that they provide a formalization of knowledge within a domain in a machine-readable form. Ontologies include a much wider scope of information than taxonomies, which are simply hierarchical representations of the concepts but lack formal descriptions of their properties and the types of relationships they have with one another.

The SAO is available for browsing and download through the CCDB Website (http://ccdb.ucsd.edu/SAO) and has also been recently made available through the BioPortal maintained by the National Center for Biomedical Ontologies (http://www.bioontology.org/ncbo/faces/index.xhtml). The first version of the SAO focuses on the subcellular anatomy of the nervous system, comprising neuronal and glial cells, their subcellular components, structural compartments and macromolecular constituents along with multicellular domains such as neuropil and the Node of Ranvier (Fong et al., 2007; Larson et al., 2007). The SAO was designed with the goal of providing a means to annotate cellular and subcellular data obtained from light and electron microscopy, including assigning macromolecules to their appropriate subcellular domains. The SAO thus provides a bridge between ontologies that describe molecular species and those concerned with more gross anatomical scales. Because it is intended to integrate into ontological efforts at these other scales, particular care was taken to construct the ontology in a way that supports such integration.

A portion of the class structure of the SAO is illustrated in Fig. 5. Details of its construction and content can be found in Fong et al. (2007) and Larson et al. (2007). Briefly, the SAO was organized according to the framework proposed by the Basic Formal Ontology (BFO; Grenon, 2003). The Basic Formal Ontology proposes the fundamental division of biological entities into *continuants* and *occurrents*. *Continuants* are entities that endure through time, e.g., a cell, a mitochondrion *Occurrents* are entities that unfold through time, e.g., mitosis, neurotransmitter release The SAO currently only models continuants. The Basic Formal Ontology additionally divides continuants into multiple upper level classes to help parcellate biological entities into meaningful and useful categories. The main classes specific to the SAO are elaborated under these classes (Fig. 5). They basically summarize cells, parts of cells, molecules and supracellular configurations such as the Node of Ranvier (see below) and synapses.

The SAO contains a rich set of relationships that relate the different classes to each other. The two main relationships are "*regional part of*" and "*component part of*". *Regional part* refers to a parcellation of a structure into multiple domains, e.g., a dendrite *is a regional part of* a neuron, while *component part* denotes a self-contained structure that is contained within another, e.g., mitochondrion *is component part of* dendrite. These relationships are assigned to all classes of the SAO regardless of granu-

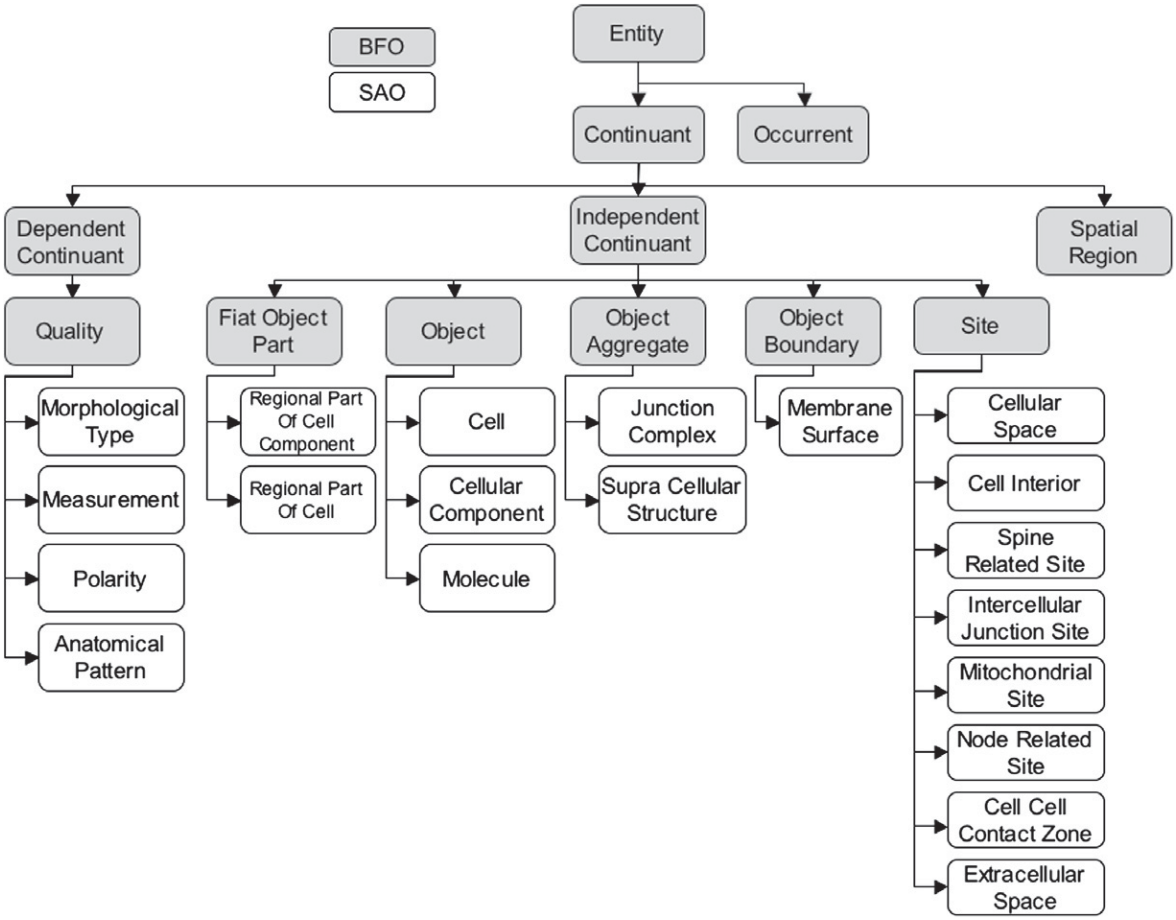

Fig. 5. Basic organization of the SAO. The SAO was constructed using the Basic Formal Ontology (BFO) as a foundation. Classes belonging to the BFO are shown in gray; those belonging to the SAO in white. Only a subset of SAO classes are illustrated in this diagram.

larity, i.e., a dendrite can have regional parts and component parts; mitochondrion can have regional parts and component parts. Molecules may be assigned to either regional parts or component parts.

We are currently undergoing the process of annotating all of the neural data in the CCDB with the SAO (Fig. 6). This annotation is necessary both to provide both a controlled vocabulary for cellular anatomy, but also to take advantage of the features of the ontology for query and analysis (see Section 4).

To aid in the application of the SAO for describing data in the CCDB, we have recently released an alpha version of a segmentation tool, Jinx, based on our prior segmentation program Xvoxtrace (Perkins et al., 1997). Jinx incorporates the SAO so that users create segmented objects as instances of the SAO, rather than supplying their own labels. The implementation goes beyond the use of the SAO as a simple controlled vocabulary by allowing users to define relationships among individual segmented objects. In the example shown in Fig. 6, a manual segmentation of a portion of the axonal and glial components of a Node of Ranvier from mouse peripheral nerve (CCDB dataset #50) is shown Fig. 6. The name of each object is selected from the SAO and named as an instance of an SAO concept, e.g., Schwann_cell_paranodal_termination_0004. The SAO classified the Node of Ranvier as a site, because it is the location on the axon where there are gaps in the myelin sheath. All of the individual objects segmented from the tomographic reconstruction are related to each other through the relationships in the SAO (Fig. 6). The output of this process is to represent the complex content of a tomographic reconstruction as a graph representing all of the segmented objects and their subparts and the relationship among different objects.

4. Discussion

The CCDB has undergone many iterations since its public launch in 2002 and within the last year has finally moved from a prototype to a stable production system. Through the CCDB, users can browse and retrieve 2D, 3D, and 4D datasets from light and electron microscopy. The CCDB welcomes contributions of light and electron microscopic data from the scientific community. While we will continue to develop the functionality of the system and refine the data model, our main focus in the immediate future will involve population of the database, increasing the utility of the CCDB as a data management system and developing the semantic layers necessary to allow more meaningful annotation and query of CCDB data.

4.1. Population of the CCDB

Funding agencies such as the National Institutes of Health have invested heavily in the creation of on-line databases for data dissemination through programs like the Human Brain Project (Koslow and Hirsch, 2004; Ascoli, 2006). Many of these resources are now available, but population of these databases by the communities they are meant to serve remains minimal. Unlike sequence or protein structure data, there is no formal requirement by journal editors or funding agencies for deposition of most data types into a public database, with few exceptions (Van Horn et al., 2004). Also, the amount of effort to submit complex data like 3D images is significantly greater than for sequence and structure, where experimental context is not so important. Finally, because of the rich content of tomographic data, researchers are inclined to hold on to the data for their future re-use. Thus, without a require-

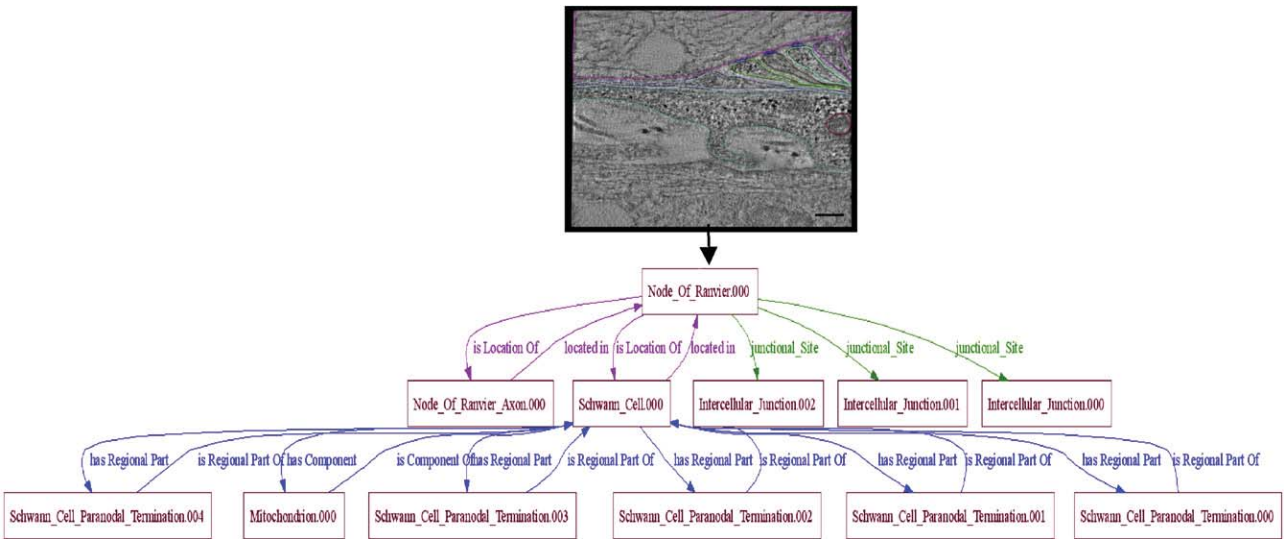

Fig. 6. Segmentation of a tomographic reconstruction of the Node of Ranvier using Jinx to annotate with SAO. Top: Single computed slice through the tomogram with a few of the structures manually outlined. Each color delineates a unique object. Bottom: Instance tree generated from the segmentation in A showing the relationship among the different objects defined. The parent node (Node_Of_Ranvier.000) refers to the entire subject of the reconstruction, and not an individual segmented object. The different types of relationships defined by SAO among the parts of the SAO are indicated by different colors.

ment for data deposition into a public database, researchers currently have little incentive to deposit data.

We have been working to make the CCDB useful for researchers, as an incentive to contribute data. For example, we provide the ability to link data from published manuscripts to original data, high-resolution movies and additional supporting images without size limitations, unlike most journals. Data management through MyCCDB also has appeal to researchers, as it allows them to keep track of their projects, archive their data and easily retrieve it when necessary. As the data in MyCCDB is already structured according to the CCDB schema, the process of releasing data to the public becomes trivial.

The CCDB has seen a steady stream of downloads of datasets over the years, although we do not formally track subsequent usage of the data. As on-line databases and calls for data sharing have proliferated, the question has arisen as to whether complex data truly have any utility beyond the initial purpose for which they were acquired. We think that strong arguments can be made for sharing of information-rich data like images and volumes derived from 3D light and electron microscopy. Indeed, we have several examples where data in the CCDB has been reanalyzed or re-purposed for other studies. For example, Coggan et al. (2005) created a simulation of neurotransmitter release at the ciliary ganglion synapse using data acquired by Shoop et al. (2002) for a morphological study of this synapse (CCDB ID # 3629). At this time, we do not store the results of the simulation. Other major users include computer scientists developing algorithms for image processing and analysis. It is perhaps not surprising that the major requesters of data in the CCDB come from the computational community, as these researchers typically do not have access to the instruments and technical expertise to acquire such data themselves. Yet, as biology moves from a more qualitative to quantitative science, access to data by the computational community will become increasingly important.

The CCDB is not the only public database project for electron tomography data. The Electron Microscopy Structure Database (EMSD) is also accruing electron tomography data, in addition to protein structures derived from electron microscopy (Tagari et al., 2002, 2006). These two database efforts complement each other very well, reflecting their emergence from different biological traditions. The CCDB takes a largely cell biological perspective, integrating data taken across different scales and with different techniques to build models of a cell and the distributions of their molecular constituents. The EMSD takes a structural biology approach, with an emphasis on high-resolution structure. We view the CCDB as one of many distributed resources for cell-level data and have been working on issues of database interoperability through projects like the BIRN (see Section 3). According to this model, contributors should be free to store data where they wish, but through the development of data exchange formats, e.g., XML, for light and electron microscopy (Heymann et al., 2005; Goldberg et al., 2005) users should be able to retrieve, access and use the data regardless of where they live.

4.2. Data management through the CCDB

While interest from the tomography community in contributing data to the CCDB has been minimal to date, interest in utilizing the CCDB as a means of managing data has been much more substantial. Although the data management function was not in the original specification of the CCDB, we have been responsive to community interest and have provided interested groups access to CCDB and its tools. We believe that as data management tools are incorporated into the daily workflow of the modern electron microscopy laboratory, pushing out this data to the broader community will be facilitated. Towards this end, we make the CCDB freely available to any one that would like to use it. For example, we have been working with the tomography groups of Dr. Brad Marsh in Australia, Dr. Abraham Koster in the Netherlands, and Dr. Grant Jensen at the California Institute of Technology to establish the CCDB in these venues. Currently, with its mix of enterprise software and grid-based components, installation of CCDB at a local site requires considerable technical expertise and information technology infrastructure. However, through the MyCCDB portal, we provide a complete web-based data management solution for those who do not have the means or desire to host their own CCDB. As more of the community evaluates and utilizes the CCDB, we hope that the data model and tool base of the CCDB will improve and that a set of standards for sharing cell-level tomography data will emerge.

4.3. Putting the ''cell'' into the Cell Centered Database: ontologies for subcellular anatomy

A major focus of the CCDB project for the past year has been the development of a formal ontology for subcellular anatomy, to describe cells, their parts and how these parts may come together to form supracellular domains like synapses and the Nodes of Ranvier. The first version of SAO was released in May of 2007 (Fong et al., 2007) and largely covers the subcellular anatomy of the nervous system. The SAO was modeled after ontologies for gross anatomy like the Foundational Model of Anatomy (FMA; Rosse and Mejino, 2003). It builds on existing ontologies for cell components (Gene Ontology; Ashburner et al., 2000) and cell types (Cell Type Ontology; Bard et al., 2005), using recommended best practices for ontology construction (Smith et al., 2005).

The SAO was designed to provide the semantic underpinning of the CCDB and related tools to provide a ''cell centered'' view of CCDB data. Each microscopy product, reconstruction and segmented object is annotated as instances of the SAO. At the simplest level, the SAO provides a controlled vocabulary for describing cellular struc-

tures. Currently, most researchers who are segmenting structures from electron tomography employ a terminology short hand that is interpretable to them, but is largely opaque to anyone else. Also, there are many variants of cell terms, e.g., Purkinje neuron vs Purkinje cell, that make it difficult to query. By standardizing the names of structures through enforcement of a controlled vocabulary, the process of query and interpretation of tomographic data are facilitated.

Well-structured ontologies have utility well beyond their use as a controlled vocabulary. Because the SAO contain knowledge about subcellular anatomy in a form that is machine readable, the ontology will provide the means for much more meaningful query of CCDB data. Users will be able to query the CCDB not just through the metadata contained in the CCDB model, but through relationships and concepts contained in the ontology. To give a simple example, the CCDB schema does not contain any knowledge about the cell types represented in CCDB data other than their gross anatomical characterization. If one wanted to retrieve data from CCDB on GABAergic neurons, the user would have to know what neurons use GABA and request data on each of the individual cells. The SAO, however, has the knowledge of what cell types utilize GABA as a neurotransmitter and a mapping between CCDB datasets representing those neurons and SAO entities.

The SAO is also providing the means of drawing relationships among objects contained in the CCDB that are not currently represented in the database. Although the CCDB allows each segmented object from a reconstruction to be entered as an individual object, the CCDB data model can relate each of these objects to a parent reconstruction, not a biologically meaningful entity in the reconstruction. As illustrated in Fig. 6, a segmentation of a structure like the Node of Ranvier may yield of list of parts axon and Schwann cell terminal loop, but there is no way in CCDB to know the relationships among these objects. Through SAO, we can list each segmented object as an instance that belongs to a single Node and relate the individual parts back to their parent cell types, e.g., all paranodal terminations belong to a single instance of Schwann cell. By defining these relationships, we can begin to ask queries such as ''*Find all instances of the Node of Ranvier where the Schwann Cell has a mitochondrion*''.

As with the process of data entry, we are working to make the process of annotation of CCDB data with SAO as easy as possible for the researcher. Just as with MyCCDB, we are trying to incorporate the SAO into the tools that researchers use for their own work. The additional effort to utilize terms from the SAO and define relationships is very minimal compared to the labor-intensive process of segmentation itself. Through continued development and application of SAO, we are hopeful that by the time of the next tomography conference, we will have moved the CCDB from a database for on-line data sharing to a knowledge environment where users can explore and mine cellular information across scales.

Acknowledgments

Supported by NIH Grants NIDA DA016602 (CCDB), NCRR RR04050, and RR08605. The Bioinformatics Research Network is supported by NIH Grants RR08605-08S1 (BIRN-CC) and RR021760 (Mouse BIRN). The authors thank Mr. Daniel Ryan Kloos for help with the figures.

References

Ascoli, G.A., 2006. Mobilizing the base of neuroscience data: the case of neuronal morphologies. Nat. Rev. Neurosci. 7, 318–324.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.. Nat. Genet. 25 (1), 25–29.

Bard, J., Rhee, S.Y., Ashburner, M., 2005. An ontology for cell types. Genome Biol. 6 (2), R21.

Coggan, J.S., Bartol, T.M., Esquenazi, E., Stiles, J.R., Lamont, S., Martone, M.E., Berg, D.K., Ellisman, M.H., Sejnowski, T.J., 2005. Evidence for ectopic neurotransmission at a neuronal synapse. Science 309, 446–451.

Fong, L., Larson, S., Gupta, A., Condit, C., Bug, W., Chen, L., West, R., Lamont, S., Terada, M. and Martone, M.E., in press. An ontology-driven knowledge environment for subcellular neuroanatomy, OWL: Experiences and Directions, Innsbruck, Austria, CEUR Workshop Proceedings, ISSN 1613-0073, http://CEUR-WS.org/Vol-258/, June 6–7, 2007.

Goldberg, I.G., Allan, C., Burel, J.M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P.K., Swedlow, J.R., 2005. The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol. 6, R47.

Grenon, P. (2003). BFO in a nutshell: a bi-categorial axiomatization of BFO and comparison with DOLCE. IFOMIS, ISSN 1611-4019.

Grethe, J.S., Baru, C., Gupta, A., James, M., Ludaescher, B., Martone, M.E., Papadopoulos, P.M., Peltier, S.T., Rajasekar, A., Santini, S., Zaslavsky, I.N., Ellisman, M.H., 2005. Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud. Health Technol. Inform. 112, 100–109.

Heymann, J.B., Chagoyen, M., Belnap, D.M., 2005. Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology. J. Struct. Biol. 151, 196–207.

Koslow, S.H., Hirsch, M.D., 2004. Celebrating a decade of neuroscience databases: looking to the future of high-throughput data analysis, data integration, and discovery neuroscience. Neuroinformatics 2 (3), 267–270.

Larson, S., Fong, L., Gupta, A., Condit, C., Bug, W.J., Martone, M.E., accepted for publication. A formal ontology of subcellular neuroanatomy, Front. Neuroinformatics 2007, in press.

Lawrence, A., Bouwer, J.C., Perkins, G., Ellisman, M.H., 2006. Transform-based backprojection for volume reconstruction of large format electron microscope tilt series. J. Struct. Biol. 154, 144–167.

Lucic, V., Forster, F., Baumeister, W., 2005. Structural studies by electron tomography: from cells to molecules. Annu. Rev. Biochem. 74, 833–865.

Marsh, B.J., Volkmann, N., McIntosh, J.R., Howell, K.E., 2004. Direct continuities between cisternae at different levels of the Golgi complex in glucose-stimulated mouse islet beta cells. Proc. Natl. Acad. Sci. USA 101, 5565–5570.

Martone, M.E., Sargis, J., Tran, J., Wong, W.W., Jiles, H., Mangir, C., 2007. Database resources for cellular electron microscopy. Methods Cell Biol. 79, 799–822.

Martone, M.E., Gupta, A., Wong, M., Qian, X., Sosinsky, G., Ludascher, B., Ellisman, M.H., 2002. A cell-centered database for electron tomographic data. J. Struct. Biol. 138, 145–155.

Martone, M.E., Zhang, S., Gupta, A., Qian, X., He, H., Price, D.L., Wong, M., Santini, S., Ellisman, M.H., 2003. The cell-centered database: a database for multiscale structural and protein localization data from light and electron microscopy. Neuroinformatics 1, 379–395.

Moisan, T., Ellisman, M.H., Buitenhuys, C.W., Sosinsky, G.E., 2006. Differences in chloroplast ultrastructure of phaeocystis Antarctica in high and low light conditions. Mar. Biol. 149 (6), 1281–1290.

Neerincx, P.B., Leunissen, J.A., 2005. Evolution of web services in bioinformatics. Brief Bioinform. 6 (2), 178–188.

Peltier, S.T., Lin, A., Lee, D., Smock, A., Lamont, S., Molina, T., Wong, M., Dai, L., Martone, M.E., Ellisman, M.H., 2003. The telescience portal for tomography applications. J. Parallel Distrib. Comput. 63, 539–550.

Perkins, G.A., Renken, C.W., Song, J.Y., Frey, T.G., Young, S.J., Lamont, S., Martone, M.E., Lindsey, S., Ellisman, M.H., 1997. Electron tomography of large, multicomponent biological structures. J. Struct. Biol. 120 (3), 219–227.

Rosse, C., Mejino Jr., J.L., 2003. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J. Biomed. Inform. 36 (6), 478–500.

Shoop, R.D., Esquenazi, E., Yamada, N., Ellisman, M.H., Berg, D.K., 2002. Ultrastructure of a somatic spine mat for nicotinic signaling in neurons. J. Neurosci. 22, 748–756.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C., 2005. Relations in biomedical ontologies. Genome Biol. 6 (5), R46.

Sosinsky, G.E., Deerinck, T.J., Greco, R., Buitenhuys, C.H., Bartol, T.M., Ellisman, M.H., 2005. Development of a model for microphysiological simulations: small nodes of ranvier from peripheral nerves of mice reconstructed by electron tomography. Neuroinformatics 3, 133–162.

Tagari, M., Newman, R., Chagoyen, M., Carazo, J.M., Henrick, K., 2002. New electron microscopy database and deposition system. Trends Biochem. Sci. 27, 589.

Tagari, M., Tate, J., Swaminathan, G.J., Newman, R., Naim, A., Vranken, W., Kapopoulou, A., Hussain, A., Fillon, J., Henrick, K., Velankar, S., 2006. E-MSD: improving data deposition and structure quality. Nucleic Acids Res. 34, D287–D290.

Van Horn, J.D., Grafton, S.T., Rockmore, D., Gazzaniga, M.S., 2004. Sharing neuroimaging studies of human cognition. Nat. Neurosci. 7, 473–481.

Wong, S.T., Koslow, S.H., 2001. Human brain program research progress in bioinformatics/neuroinformatics. J. Am. Med. Inform. Assoc. 8 (1), 103–104.