doi:10.1017/S143192761100225X
Microsc. Microanal. 17 (Suppl 2), 2011

Petabyte Data Management and Automated Data Workflow in Neuroscience: Delivering Data from the Instruments to the Researcher's Fingertips

J.C. Bouwer, V. Astakov, W. Wong, T. Molina, V. Rowley, S. Lamont, H. Hakozaki, Ohkyung Kwon, A.W. Kulungowski, M. Terada, S. T. Peltier, M. Martone, and M.H. Ellisman

University of California, San Diego, National Center for Microscopy and Imaging Research, 9500 Gilman Dr., BSB 1000, La Jolla, CA 92093-0608

Abstract:
The advent of new large area detectors and high-speed data acquisition technologies in light and electron microscopy are now producing data sets on the order of multiple terabytes. With the development of detectors such as the 8k x 8k lens-coupled CCD camera system developed at the National Center for Microscopy and Imaging Research (NCMIR), it is clear that it is no longer possible to store, view, and visualize these large data on a single desktop machine. With instrument automation, it is now possible to collect million by million pixel images with a data size of just over 2 terabytes. Additionally, with automated montaged tomograms up to 24k x 24k pixels per image, 120 tilts per rotation axis, and up to six rotation axis, we can produce just under 1 terabyte of raw data. The final computed volumes are approaching 10 terabytes and could exceed 100 terabytes for serial section tomography. With data this size, large memory, multi-node computation clusters must be used for processing, presentation, and visualization of this data. The results can then be displayed over the Internet, using dynamic binning to match the native screen resolution. With these large data sets, transfer of a single completed data set from the instrument where the data is acquired to networked storage units can bring the network and disk access to a standstill. With multiple users on various instruments and processing workstations all trying to access the network disks simultaneously, the technical challenges only multiply. It vital that these data be stored in databases and shared with the community as a whole, whether to use reuse the data for an alternate analysis or for educational purposes. We report on a new petabyte data management, workflow, and database system that allows us to overcome these challenges. This system is built upon the mature code base of the Cell Centered Database (CCDB), an integrated rule-oriented data system (IRODS) for distributed storage, scientific workflow engine - "Kepler", a set of advance hardware, newly developed code to manage the data and metadata from each microscope, and a NCMIR developed Web Image Browser (WIB). This system also utilizes service oriented architecture (SOA) to provide integration with external projects such as the Whole Brain Catalog (WBC), the International Neuroinformatics Coordinating Center (INCF), and Neuroscience Information Framework (NIF), which are benefiting from NCMIR data.

A key objective for NCMIR has been to develop our current information technology infrastructure for managing distributed microscopy imaging data to serve the requirements of the mesoscale technological research and development and associated collaborative activities. Through the Cell Centered Database (CCDB), Telescience, and the Neuroscience Information Framework (NIF), NCMIR has developed a robust infrastructure for storing, sharing, searching, and disseminating microscopy image information via the CCDB and a secure web portal. CCDB has made available to the scientific community light and electron microscopic data sets contributed by NCMIR, NCMIR collaborators, and outside contributors. Over 50 data sets were released to the public in 2010, including live cell imaging, electron tomography, serial tomography, and correlated LM and EM.

Data access rates are up to 15,000 views per month. At least 8 papers have been published re-using CCDB data from groups outside of NCMIR.

We have completed the beta version of the data workflow system to meet most of our data management requirements. This workflow utilizes IRODS as the data grid middleware system, which provides a uniform interface to heterogeneous data, and is available through the new OpenCCDB portal, which provides authentication and authorization. Every piece of data acquired on every microscope is now registered and stored in the CCDB. The following provides a list of the current workflow features:

1) Data acquisition – When researchers want to use the microscope, they submit the basic information about the project, experiment, and subject information to the CCDB through the instrument user-interface that we created. Then, the CCDB generates the unique ID for tracking this data. All of the microscope and detector metadata is extracted using each instrument's custom APIs. The data and metadata file are automatically moved to the temporary network drives in real time during acquisition time using the rsync protocol. Image viewing and annotation services through the web image browser (WIB) are automatically created.

2) Data migration – network performance and network bandwidth are monitored by the management software, which provides a way to monitor data flow rates from the microscope machines to the network storage.

3) Validation and archiving – Users log into the portal to find their data easily. From the portal, users can launch the WIB to visualize their massive data over the web. Upon approval, the data and metadata are moved to the archival storage in IRODS and a program creates the meaningful folder structure for storing this data.

4) Improved data access – Data is processed by users by mounting the IRODs managed network drives and results are synced to the IRODs distributed storage system and registered automatically to the CCDB.

5) Application Integration – Various software packages in the processing pipeline are accessed through a unified protocol for application integration with a Kepler workflow engine..

6) External project integration – Integration with other projects such as CRBS, WBC, NIF, INCF, and CAMERA can be accessed through WPS/REST web services.

7) High performance computing access – We are integrating our current workflow with the UCSD Triton Resource, a 256 node 4 petabyte storage resource, so that portal users can launch their computation-intensive program easily. The system architecture is flexible enough to be integrated with other clusters like INCF cluster the one at Karolinsky Institute of Technology (Sweden).

In this poster we will present the development strategy and operation of this new data management and workflow system in detail. This system is currently in active use at the NCMIR.

https://doi.org/10.1017/S143192761100225X Published online by Cambridge University Press