Research Data Services for Biomedical Science
Program status and upcoming opportunities

David E. Hudak, Prasad Calyam, Kevin Wohlever, Ashok Krishnamurthy
Ohio Supercomputer Center, 1224 Kinnear Road, Columbus OH 43212
{dhudak, pcalyam, kevin, ashok}@osc.edu

1 Introduction

High-performance computing (HPC) centers traditionally maintained adequate storage to meet the needs of researchers actively using their systems. Recently, the rise of computational methods in the biomedical sciences coupled with the dramatic increase in data generation capacity of laboratory instruments has created a research community looking (or desperate?) for large-scale storage and data-intensive computing solutions. In addition to storage and computation, these groups require higher-level features such as integration with established laboratory experiment workflow, data sharing within research groups and data export to outside groups. Also, NSF and NIH requirements for research data management will lead to retention of data far past its active use at an HPC center. The long-term management of research data is a major challenge that will impact researchers, universities, state and federal agencies.

The Ohio Supercomputer Center (OSC) is actively supporting data-intensive biomedical research groups located at The Ohio State University, the Comprehensive Cancer Center (CCC) at The Ohio State University Medical Center and the Research Institute at Nationwide Children's Hospital (RINCH). These organizations contain a number of core facilities, common laboratories providing analysis to a collection of research and clinical groups. Currently, OSC is supporting the following core facilities: Campus Microscopy and Imaging Facility (CMIF), Small Animal Imaging Shared Resource (SAISR) and the Plant Microbe Genomics Facility (PMGF). CMIF offers services that range from routine microscopy to leading edge, live animal, multiphoton microscopy. The SAISR supports a variety of in vivo small animal imaging modalities as a non-invasive tool for studying morphological, biochemical, or genetic perturbations in small animal models of cancer. The SAISR generates multidimensional data from various microimaging modalities including microCT, microMRI, nuclear (microPET microSPECT), ultrasound, and optical. PMGF offers DNA sequencing services and operates high-end gene sequencers. Despite the variety of instruments and data types employed by this biomedical research community, common workflow, data storage, processing and distribution requirements exist across these cores. OSC supports these core facilities through its Remote Instrumentation Services (RIS) program.

RIS is a successful program meeting the needs of active groups at these core facilities. We see two critical issues for the long-term viability of the program that must be addressed. First, as these groups begin to complete projects, the data remains at OSC. Policies and mechanisms for long-term research data retention and retrieval are needed. Second,
investigators want assistance in meeting research data management requirements from funding groups.

2 Biomedical Research Data Services

2.1 Remote Instrumentation Services (RIS) Program

Biomedical core facilities are generating ever-increasing amounts of data. They need large-scale storage, high performance computing (HPC) processing to reduce analysis time, visualization and data distribution services. These services largely lie outside their capabilities, funding and resources. OSC supports these facilities through a program called Remote Instrumentation Services (RIS). RIS provides production services for large-scale data import, HPC data processing and remote visualization. We have developed the Data Import, Storage and Collaboration (DISC) service to automatically transfer data from individual core facilities to OSC.

2.2 OSC Data Import, Storage and Collaboration (DISC) Service

The cornerstone of RIS is the Data Import, Storage and Collaboration (DISC) Service. (http://www.osc.edu/disc) DISC allows core facilities to schedule and track data transfers and alerts OSC support staff in the event of transfer failures or errors. DISC allows for easier data sharing and works with Linux, Mac, or Windows systems. Given the various user workflows with research data, DISC has been developed to support multiple workflows and either synchronous (scheduled on data sources) or asynchronous (user pushes data manually) control. A number of data sharing models are supported, including:

- One to One - Individual OSC user’s data can be imported, stored and accessed from their OSC home directory
- Many-to-1 - An entire project team’s data can be imported, stored and accessed from their OSC project directory
- Many-to-Many - An entire project team’s data or data from multiple machines with common users (i.e., Research Instrument Lab Groups) can be imported, stored and accessed from a set of OSC project directories.
3 Upcoming Challenges

As the number of RIS clients increases, data automatically imported by DISC continues to accumulate at OSC. In order to manage this data for longer periods of time, we must investigate technologies, policies and funding options. We are conducting investigations with iRODS and CNRI’s Digital Object Repository (http://dorepository.org/) for managing raw and processed data as well as providing additional search and retrieval options. Long-term provenance of published material is sponsored at the institutional level (i.e., universities pay for their own libraries) with some sharing at higher levels (e.g., OhioLink http://www.ohiolink.edu/). We believe that the research data management policies of funding agencies will spur discipline-specific standards for presentation and preservation of raw research data. Supporting such standards will require commitments from institutions and funding agencies.