

DATA SHARING FOR COMPUTATIONAL NEUROSCIENCE CENTRAL SERVICES

Fritz Sommer, Bruno Olshausen, Jarrod Millman
Redwood Center for Theoretical Neuroscience, UC Berkeley

COLLABORATIVE RESEARCH IN COMPUTATIONAL NEUROSCIENCE WORKSHOP

June 6-7, 2007

University of Maryland
University College

Contents

General Scope.....	3
1) Guiding priorities and goals.....	3
2) Proposed organizational structure.....	5
3) Relation to and interaction with existing data sharing efforts.....	5
Specific Description of Centralized Services.....	7
1) Basic functions for data-sharing.....	7
2) Services for data users on repository website and support.....	7
3) Services for data contributors.....	8
4) Governance of central services.....	9
5) Time line for establishing central services.....	9
6) Service details.....	10
6) Partnerships.....	12

General Scope

1) Guiding priorities and goals

This proposal is for providing the central services to enable and facilitate the planned NSF initiative: Data sharing for Computational Neuroscience, targeted to foster teaching and research in this area. There are three guiding principles for our proposal:

- Market driven: The NSF initiative has the goal to enable demand-driven scientific foci in the area of computational neuroscience. Thus, our ambition is to create an interactive market place for resources of particular significance for the field rather than a large repository for everything.
- Service: Our goal is to lessen the burden on contributors to make their data/resources available and to optimize the ability of the user community to identify and use the data/resources.
- Community input: This proposal contains initial suggestions and many questions. We are here to listen to your priorities as potential contributors and users. We recognize that there are different ideas about how to do this and together we may come up with the approaches that work best for the field.

The main purpose of the central services is to enable and facilitate sharing of neuroscience resources, *such as raw data and stimuli from physiological experiments, data analysis tools and computational models*. An important aim of the central facility is to provide searchable and transparent access to the shared resources in a manner that scales up to large data sets. While the majority of use will be by students and modelers looking for specific types of neurophysiological data, the central facility may also become a resource for experimental labs looking for certain types of sensory stimuli, or specific protocols to employ in experiments. Also, because of the considerable effort that will be put into organizing datasets, the central facility will be able to help experimental labs develop effective data infrastructures for their own data. Perhaps most importantly, the central facility will provide a crucial service to the computational neuroscience community by helping to identify new opportunities for collaborations between experimental and theoretical labs. Therefore, our proposal will include activities to elicit direct community feedback, for example, the organization of data analysis challenges and the creation of benchmark environments for comparing new analysis methods and models for neuroscience data. Further, the growth of the

repository content will be both usage- and demand-driven. In addition, it will also make legacy data available by rescuing important datasets when labs retire.

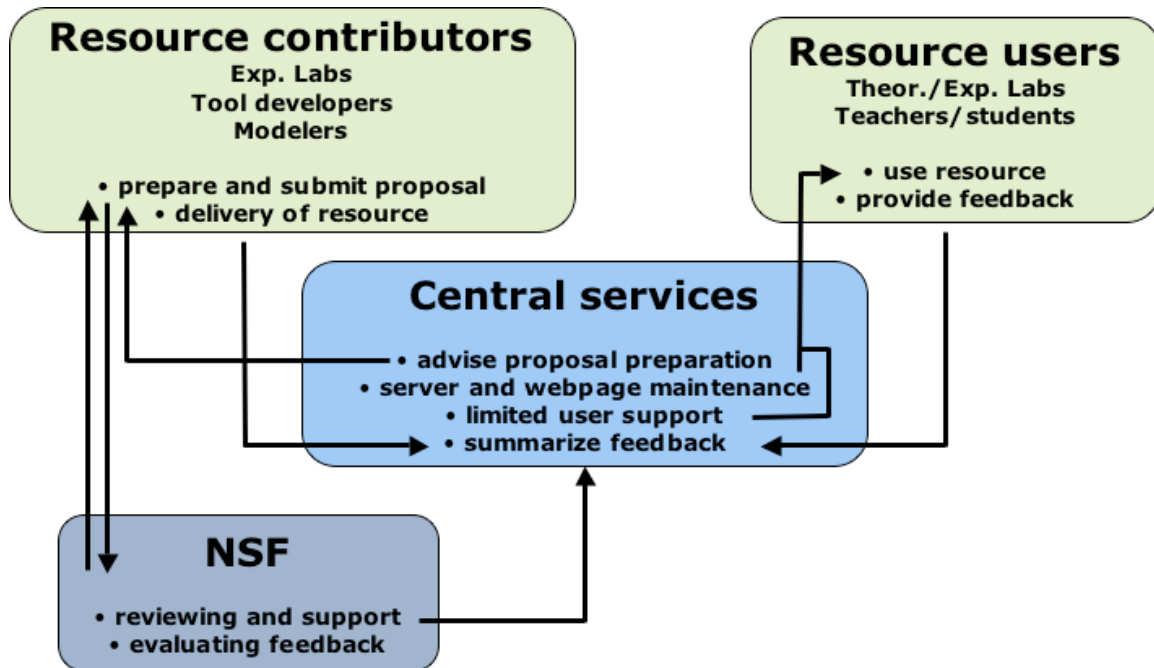
Activities in the starting phase of the program:

- Establish basic functionality for data sharing: web site, privacy/security issues, data hosting, data contributor/user instructions, search and browse options using online data access protocol.
- Organize data analysis challenges around specific empirical/computational questions that could be addressed by specific datasets. (How to use community input to define challenge task?)
- Establish benchmark environments for evaluation of analysis algorithms and models such as spike sorting. (How to identify data sets with benchmark quality? Voting mechanism? How to encourage the contribution of data sets with some ground truth?)

Activities in the advanced phase of the program:

- Improved online data browsing and visualization
- Phase in support for novel data analysis and modeling tools such as visualization and multivariate analysis and data mining tools etc. (Next funding rounds of NSF initiative)
- Links to freely available data analysis tools with FAQ section for each linked resource

2) Proposed organizational structure



3) Relation to and interaction with existing data sharing efforts

There have been some pioneering efforts for sharing neurophysiological data on the web. The planned data-sharing initiative will seek to build on these efforts and learn from their experience.

- Neurodatabase (D. Gardner, Cornell, <http://neurodatabase.org/>) 10 data sets of spike trains from labs at Cornell (e.g., from J Victor's lab).
- Audimotor Spike train database (CD Woody, UCLA) is a database of spike trains from about 5000 neurons along the auditorimotor pathway from the cochlear nuclei to the neocortex, cerebellum, and elsewhere.
- Initiatives by individual labs:
 - Jack Gallant Lab: Neural Prediction Challenge, <http://neuralprediction.berkeley.edu/>
 - EPFL data analysis challenge,

<http://icwww.epfl.ch/~gerstner//QuantNeuronMod2007/challenge.html>

- Dario Ringach Lab (currently offline)
- Henry Markram, Neocortical Microcircuit database:
<http://microcircuit.epfl.ch/>
(currently no electrophysiology)

b) Recently, a number of broader neuroinformatics initiatives have been established. These initiatives involve the sharing of many different resources/data types and are usually more of a coordinating nature. It is important that the planned program will be tied in into these broader initiatives.

- Neuroscience Database Gateway (NDG) is a portal of the Society for Neuroscience to make neuroscience resources broadly available (<http://ndg.sfn.org/>). Their section with experimental data that includes some of the physiology databases listed above.
- Neuroscience Information Framework (NIF) funded by NIH Blueprint (Dan Gardner, Cornell; Paul Sternberg, Caltech; Giorgio Ascoli, George Mason; Maryann Martone, UCSD; Gordon Shepherd, Yale. <http://neurogateway.org/>). This effort aims to catalog electronic and non-electronic neuroscience research resources, and make them searchable by content and usage. A further goal is to enable more uniform access to those resources that are online.
- International Neuroscience Coordinating Facility (INCF) funded by European member countries and the European Commission (Jan Bjaalie, Raphael Ritz). INCF coordinates and fosters international activities in neuroinformatics. Contributes to the development and maintenance of database and computational infrastructure and support mechanisms for neuroscience applications.

Specific Description of Centralized Services

To enable and facilitate the sharing of resources in computational neuroscience, services will be provided in the following three areas: 1) Establishment of basic functionality for sharing resources, 2) services for all potential users of resources in the neuroscience community and 3) services for contributors of resources.

1) Basic functions for data-sharing

What about lab security? How to exchange data?

- a) Signup and authentication procedures for access to resources. Two security levels will be provided for browsing data versus for full access to meta information about experiments (How important are such security mechanisms?)
- b) Provide a secure, web-based collaborative infrastructure for (i) creating documentation, (ii) contributing data, and (iii) accessing, searching, and managing all this content. (How important is search functionality? What search functions?)
- c) Provide fast data web server for download and data-shipping service in collaboration with Google. In addition, maintenance of independent data mirror (Should all data be centrally hosted?)
- d) As far as possible, unify the management of experimental metadata in an accessible, secure, and searchable database. (Experiences with BrainML?)
- e) Simplify the sharing of neuroscience research data by providing a simple, reliable, secure, and unified way to transfer and share data (e.g., actual experimental data).

2) Services for data users on repository website and support

How can neuroscientists find in the repository what they are looking for? What will they see in typical data package and what helps them to get oriented/get started? How can data integrity and service availability be ensured? What mechanisms available to ask user groups? How can technical problems be solved?

- a) Administering of rights of access (if required)
- b) Build web portal to shared resources with functions that help to get users oriented quickly

- c) Flexible online data access protocol (DAP) permitting configurable browsing of data subsets and online visualization (How important is ability to browse the data before download?)
- d) Other interactive features on website
 - Frequently asked questions sections for each particular resource
 - Organization of data analysis challenges
 - Establishing of benchmark environments for testing analysis methods and models.
 - Classified sections: "Match making" - for direct collaborations, "Wanted" - calls for data/analysis technique etc.
 - Discussion forum (somehow moderated?)
 - Usage statistics and feedback mechanisms (Make publicly available or only to contributor of resource?)
 - Links to freely available data analysis tools with FAQ section for each linked resource, e.g., Cronux, <http://www.chronux.org/>; Spike train analysis toolkit <http://neuroanalysis.org/>
- e) Some limited user support (How to limit that?)

3) Services for data contributors

How to prepare proposals? How can labs maintain control over who uses their data? How to avoid/reduce burden of technical user support?

- a) A signup procedure will be designed for browsing and administer access control to meta information about experiments (How important are these security mechanisms?)
- b) Advice with working out data-sharing agreements will be provided in collaboration with legal expert (Creative Commons or NSF?).
- c) Technical advice will be provided to individual labs for preparing grant proposals (Clearance letters?)
- d) Help will be provided to align data format and metadata used in individual lab. Provide general data reader/data writer and tools that make sanity checking and quality assessment easy
- e) Assemble standard guidelines for data delivery/self-testing

- f) On data delivery, check completeness of individual meta information necessary to understand data sets
- g) Host data on data server
- h) Contributing labs will be provided with information about usage tracking/IP addresses, etc.
- i) Maintain FAQ for each data set and provide limited technical support for users (To relieve contributors of the burden of answering low-level questions)

Capabilities for large data sets: Hard-drive shipping and online data access protocol with subsetting operations to allow "online peaking" into data sets

4) Governance of central services

How will decisions be made? How will disputes be resolved?

It is planned to establish a governance board for the central services, including active data contributors, noninvolved experimentalists, contributing and noninvolved theoretical neuroscientists and some other major representatives. (How to elect board?)

5) Time line for establishing central services

Available in first 9-12 month:

- Basic data sharing functionality/website with FAQ for users and data contributors
- Basic services for proposal preparation and resource delivery.
- First data analysis challenge

Available in later phase:

- Full functionality of web site
- Flexible browsing and online visualization of raw data
- Links to freely available data analysis tools with FAQ section for each linked resource
- Support of analysis/visualization tools (next funding rounds of NSF initiative) other types of resources, such as data analysis tools and models

6) Service details

Data format for representing raw and metadata

Neuroscience experiments involve the generation and manipulation of large quantities of both raw and processed data. Without knowing the precise acquisition parameters and conditions, the raw data is meaningless. Moreover, it is important to track the exact history of the analytic processing so that the analyses can be reproduced. Such metadata are typically scattered and too often lost. The ability to reanalyze one's own data depends critically on access to the totality of this disparate, unorganized information; the problem compounds when sharing this data. To preserve and organize this metadata, we propose one of two options:

- 1) XML wrapping (BrainML). Meta data is provided as text in XML format. The binary raw data files are wrapped in a XML data container structure.
- 2) Structured data format and format independent access protocol (e.g., HDF5 and DAP). Create standard profiles for data based on and extending definitions of BrainML.

Option 1 is the simplest solution since we can just use the BrainML metaformat. BrainML, <http://brainml.org/>, is a XML schema to facilitate information exchange between user application software and neuroscience data repositories. It allows for common shared library routines to handle most of the data processing, but also supports use of structures specialized to the needs of particular neuroscience communities. BrainML also enables the sharing of data structures.

Option 2 uses the hierarchical data format HDF5, a general purpose library and file format for storing scientific data developed by the National Center for Supercomputer Applications at UIUC. HDF5 is extensively used in many scientific fields but has yet to be widely adopted for neuroscience applications. The adaptation could be based on data structures defined in BrainML and will be relatively straight forward. One side-benefit of HDF5 is that it supports common data access protocols, such as OPeNDAP and PYDAP allowing flexible partial access of the data for online browsing and online visualization. Further, HDF5 library and format emphasize storage and I/O efficiency. For instance, the HDF5 format can accommodate data in a variety of ways, such as compressed or chunked. And the library is tuned and adapted to read and write data efficiently on parallel computing systems.

(HDF5 can store two primary objects: datasets and groups. A dataset is essentially a multidimensional array of data elements, and a group is a structure for organizing objects in an HDF5 file. Using these two basic objects, one can create and store almost any kind of scientific data structure, such as images, arrays of vectors, and structured and unstructured grids. You can also mix and match them in HDF5 files according to your needs.)

Sanity checking of items to be published

Sanity checking and quality control will be a crucial step before resources will be made public. For experimental data the central services will develop a general data reader and writer that can be configured to translate between an open data format (such as HDF5) and the data formats in which the data were acquired. Labs can cycle their data through the data reader and writer and use their own analysis/visualization tools to ensure that the data translation process was flawless.

Serving vs pointing to resources

The central services will make resources available for data hosting. However, whether all data will be centrally hosted should be discussed at the workshop.

Large data sets

The central services will prepare the infrastructure that even very large data sets can be shared. For data sets that are too big to be transferred over the Internet there will be a service for mailing hard disks (in collaboration with Google).

Updating correction/version control

There will be a version control for all available resources (CVS).

Development/maintenance of repository website

The centralized services will develop and maintain a shared web-based system for managing documentation, code, and data content based on an existing application called Plone. Plone is a powerful, widely used, open source Content Management System written in Python. All content will be available through accessible sites as defined by the US Rehabilitation Act Section 508, W3C Web Accessibility Initiative as well

as similar guidelines. In addition, we will use standard technologies such as XHTML and CSS.

Traditionally to create new web pages, a user tells a web designer or programmer what they want on the website. The programmer then creates the web page or site and the user reviews the work. This process is tedious and results in web sites that rapidly become outdated. Content management systems were designed to overcome the shortcomings of these original methods by allowing users to directly update the site without needing the assistance of a programmer.

A content management system allows users to build and update web pages with no knowledge of Internet protocols or technologies. Users interact directly with a browser interface enabling them to create web content and pages with point-and-click ease. Anyone who can use a standard word processor can create and edit web content. There are multiple types of users: administrators, publishers, and authors. The administrator adds, modifies, and deletes user accounts as well as specifying which web pages each user can modify. Any user with privileges to the website can author new content. However, the administrator of the website can protect certain content so that new content requires approval before publication.

By using Plone, we will gain access to Plone's many existing plug-ins including modules specifically designed to facilitate creating user documentation such as:

- Frequently Asked Questions (FAQs): Short questions with 2-4 sentence answers, often organized into categories.
- How-Tos: Brief one-page descriptions of how to accomplish a specific task, which are direct and to the point.
- Tutorials: Multi-page articles (more detailed than How-Tos) covering conceptual material in addition to the technical steps required for accomplishing a particular task.

Facilitating integration across data sets -> provide stimuli data

Meta analyses over several experimental data sets are often not possible because the experimental conditions are dissimilar. It is planned to make also the stimuli available that were used in the experiments. This information is not only crucial for comparing computational models with the experimental data. It can also be used in designs of new experiments that could lead to the availability of data sets amenable to meta analyses.