Data Sharing II: Standards and Practices for Interoperability

Daniel Gardner, Michael Abato, Kevin H. Knuth, and Adrian Robert

Laboratory of Neuroinformatics, Dept. of Physiology & Biophysics, Weill Cornell Medical College, NY, NY

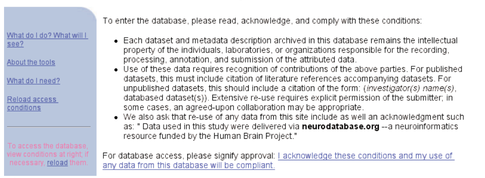
Interoperability Will Aid Data Sharing

The Human Brain Project's Program Announcement recognizes that for "advanced information technologies to be put to wide use by the neuroscience community, they should be generalizable, scalable, extensible, and interoperable" and that this requires "technologies for data synthesis and integration and electronic collaboration" including "federation of related databases and tools".

We here offer proposals—based on our resources (neurodatabases, user tools, and computational aids) and our schemas and methods for interoperability (BrainML, BrainMetaL, and GENIE)—targeted toward development and acceptance of standards and practices aiding data sharing and interoperability among neuroinformatic resources. In a pendant poster, Robert et al tomorrow offer guidelines to aid development of the internal structure of efficient, maintainable, persistent neuroinformatic resources.

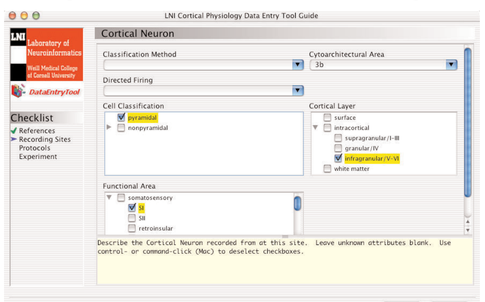
Incentives Promote Data Sharing

Data sharing will also be aided by technological and societal incentive, protective, and reward structures. Such protections and incentives can be designed into individual neuroinformatic resources; we describe two offered at neurodatabase.org.



1. Explicitly Recognize Data Submitters Rights

To access databases at neurodatabase.org requires agreement with a statement on appropriate use of data, including a requirement to acknowledge sources and a reminder that extensive re-use of data may favor a collaboration. Such steps toward protection of data can reassure potential submitters and thus increase data sharing.

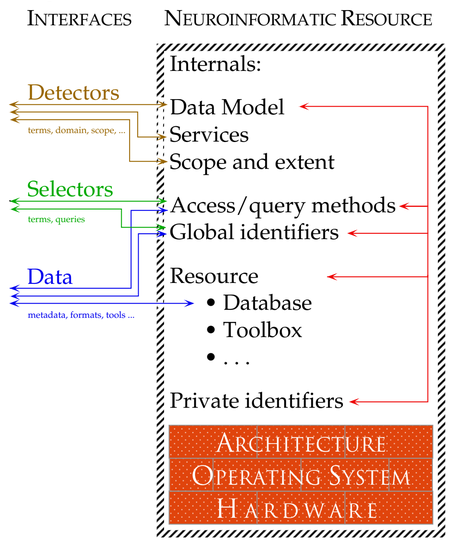


2. Minimize the Effort to Share Data

Ease of data submission and perceived benefits should match those provided by conventional journal publication. User-friendly tools such as our wizard mode DataEntryTool provide one such incentive.

Interfaces Implement Interoperability

We propose the use of compatible *interfaces* to implement interoperability between, and federation of, resources. Interfaces shield users from resource internals, while making available in compatible formats both structure and descriptors of input queries and output data.



3. Interfaces Mediate Open Access

Compatible interfaces expose *detectors* for identifying resources, *selectors* for specific search, and *data* served by the resource. Ideally, such interfaces should be either standardized or traceable to a defined schema. Interfaces map without exposing the internal structure or architecture of resources and can complement, rather than replace, native access methods.

	Gene	Protein	Cell	Organ	Subject
Sequence	comparable	comparable			relatable
Expression			relatable	relatable	relatable
Structure		same	relatable	varies	varies
Sensor data			largely same	technique-dependent	technique-dependent
Image		same	technique-dependent	varies	varies
Chart note					unique

4. Interoperability Interfaces Need To Relate Data of Disparate Types Across Multiple Levels

To support interoperability, such interfaces must accommodate and allow relating neuroscience data of many types (*rows*) obtained from systems at many levels (*columns*). The chart summarizes identical, comparable, relatable, variable, and unique combinations of type and level.

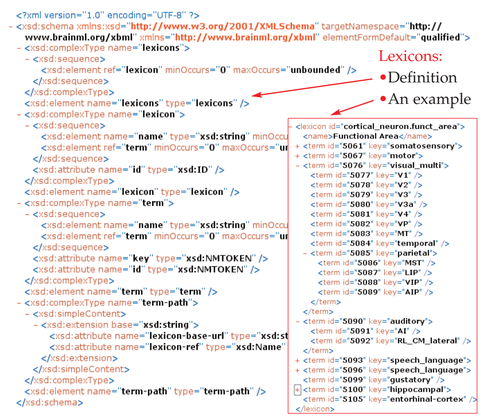
BrainML: An Extensible Interoperability Interface To Aid Data Sharing

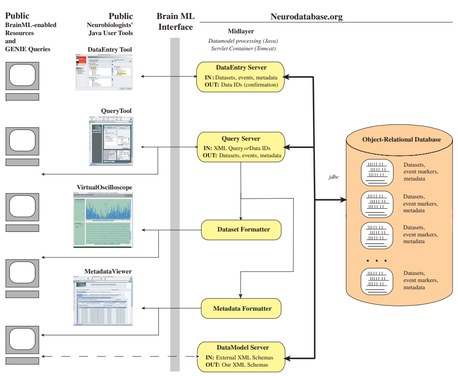
Such interfaces require development of, and standardization upon, not only formats but also a rich semantics for descriptors that cover both neuroscience anatomy and data types, as well as an appropriate range of experimental techniques, procedures, and practices. In addition, the complexity and diversity of experimental neuroscience requires as well a defined syntax as well as semantics.

In support of these goals, and to serve as a generalizable yet potent interface, we offer our BrainML, an XML-derived expandable multilevel data description language suite. BrainML is built on the metalanguage BrainMetaL, a substrate defining basic types and structures (see Xiao et al, these meetings 2002). To support multiple specialized levels and fields within neuroscience, and links to related areas of biomedicine, BrainML is designed as a suite of compatible, BrainMetaL-derived, XML Schema-defined, structures and ontologies.

Syntax and semantics can enable targeted searches that rely on more than simple term matching (especially important for trees or hierarchies of terms), differentiating detectors (common to most or all entries in a focused database) from selectors (descriptors used to select records of interest), and parsing and specifying datasets.

In BrainML, queries are XML-defined. Hierarchic lexicons map trees of descriptors to attributes, enabling both broad and focused searches as well as extensibility. With further development, attribute values will be mappable to source definitions.





5. BrainML Interface Aids Data Exchange

An implementation of the model of Fig. 3, BrainML is the interface for neurodatabase.org, transmitting XML-wrapped datasets, metadata, and queries via controlled-vocabulary selectors between Java2 user tools and database midlayer servers and formatters. The DataModel Server is not yet operational.

Molluscan Neuron:

identifier	
species	ganglion
coordinates	appearance
connects-to	relative_to
innervation	
RP	spont_AP
receptive_field_modality	
receptive_field_location	
motor_behavior	
releases	expresses
homolog-of	

Cortical Neuron:

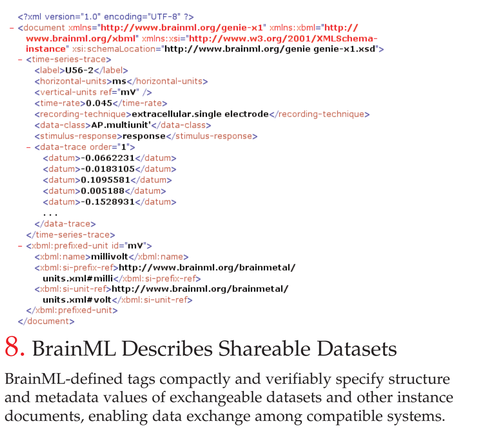
cell_class	classif_method
subject_taxonomy	
coordinates	coord_scheme
funct_area	cytoarch_area
cortical_layer	
directed_firing	
receptive_field_modality	
receptive_field_location	
motor_behavior	
releases	expresses

6. Specific Preparations or Techniques Require Targeted Sets of Descriptive Attributes and Values

Molluscan and mammalian cortical neurons illustrate how sets of descriptive attributes may vary. Some common attributes share the same set of values, some need need distinct values (not shown), and some attributes describe only one of the cell types, not the other.

7. BrainMetaL Lexicons Provide Structure for BrainML Controlled-Vocabulary Descriptor Trees

BrainMetaL defines a structure for lexicons: hierarchic attribute-value trees of terms: descriptors modifying specified attributes. BrainML and GENIE use this structure to specify focused sets of controlled vocabularies for biophysics and for neuroscience.



8. BrainML Describes Shareable Datasets

BrainML-defined tags compactly and verifiably specify structure and metadata values of exchangeable datasets and other instance documents, enabling data exchange among compatible systems.

GENIE: Generalized Extensible Neuroscience Internet Examiner

- What
  - *Peer-to-peer structured* neuroscience data sharing
  - Designed for ease of use and rapid deployment
  - Enhanced by BrainML
- Why
  - The new data sharing mandate, with projected expansion of its scope
  - Data are heterogeneous and distributed
  - Can be generalized to link beyond neuroscience
  - Many datasets are stored as flat files, not in databases
  - Permits both free and restricted distribution
- How
  - Peer-to-peer architecture empowers local archives
  - System uses platform-independent XML, standard protocols and common open-source software
  - Local archives serve searchable metadata detectors and selectors
  - System self-organizes web of GENIE-empowered sites
  - Unstructured variants enable rapid deployment

9. GENIE for Interoperable Peer-to-Peer Data Sharing

For some preparations, techniques, and communities, nformal or mediated networks for sharing data peer-to-peer will supplement centralized databases. To support interoperability between centralized and distributed resources, we have begun development of GENIE: the Generalized Extensible Neuroscience Internet Examiner, a peer-to-peer self-organizing system for serving and sharing neuroscience data by individual or community databases or flat file servers. Although GENIE is BrainML-based, it has not escaped our attention that unstructured variants may enable rapid deployment.

```
<html>
<head>
   <title>NeuroPeer Database</title>
   <meta name="detectors"
   content="genie,visual_multi,optical.neurointrinsic,
   nonhuman.primate, neuroinformatix" scheme="brainml">
   <meta name="robots"content="genie,neuroinformatix,brainml">
</head>
. . .
```

10. Sample GENIE Header Enables Search

This sample HTML header for the `/genie` file served by a local archive enables search agents to recognize a GENIE-compatible server with specific content and defined data model. The local server can use any of HTML, Java, or Javascript. Other local server files specify genie-hosts, databases, or data models.

URLs :

- BrainML, GENIE & BrainMetaL: `brainml.org`
- Our cortical database: `neurodatabase.org`
- Towards data sharing: `datasharing.net`

Acknowledgments:

Human Brain Project Neuroinformatics research funded by NIMH and NINDS via MH/NS57153, and by NIMH SBIR MH60538.