Neuroinform (2008) 6:149–160
DOI 10.1007/s12021-008-9024-z

The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience

Daniel Gardner · Huda Akil · Giorgio A. Ascoli · Douglas M. Bowden · William Bug · Duncan E. Donohue · David H. Goldberg · Bernice Grafstein · Jeffrey S. Grethe · Amarnath Gupta · Maryam Halavi · David N. Kennedy · Luis Marenco · Maryann E. Martone · Perry L. Miller · Hans-Michael Müller · Adrian Robert · Gordon M. Shepherd · Paul W. Sternberg · David C. Van Essen · Robert W. Williams

Published online: 23 October 2008

Abstract With support from the Institutes and Centers forming the NIH Blueprint for Neuroscience Research, we have designed and implemented a new initiative for integrating access to and use of Web-based neuroscience resources: the Neuroscience Information Framework. The Framework arises from the expressed need of the neuroscience community for neuroinformatic tools and resources to aid scientific inquiry, builds upon prior development of

D. Gardner (✉) · D. H. Goldberg · B. Grafstein · A. Robert
Laboratory of Neuroinformatics and Department of Physiology,
Weill Medical College, Cornell University,
1300 York Avenue,
New York, NY 10065, USA
e-mail: dan@med.cornell.edu

H. Akil
Molecular and Behavioral Neuroscience, University of Michigan,
Ann Arbor, MI 48109, USA

G. A. Ascoli · D. E. Donohue · M. Halavi
Center for Neural Informatics, Structure, and Plasticity
and Molecular Neuroscience Department,
Krasnow Institute for Advanced Study, George Mason University,
Fairfax, VA 22030, USA

D. M. Bowden
National Primate Research Center, University of Washington,
Seattle, WA 98195, USA

W. Bug · J. S. Grethe · M. E. Martone
Department of Neurosciences, University of California,
San Diego, CA 92093, USA

A. Gupta
San Diego Supercomputer Center, University of California,
San Diego, CA 92093, USA

D. N. Kennedy
Departments of Neurology and Radiology,
Harvard Medical School,
Boston, MA 02129, USA

L. Marenco · P. L. Miller · G. M. Shepherd
Department of Neurobiology and Yale Center for Medical
Informatics, School of Medicine, Yale University,
New Haven, CT 06510, USA

H.-M. Müller · P. W. Sternberg
Howard Hughes Medical Institute and Division of Biology,
California Institute of Technology,
Pasadena, CA 91125, USA

D. C. Van Essen
Department of Anatomy and Neurobiology, School of Medicine,
Washington University,
St. Louis, MO 63110, USA

R. W. Williams
Department of Anatomy and Neurobiology and Department
of Pediatrics, University of Tennessee Health Science Center,
Memphis, TN 38163, USA

neuroinformatics by the Human Brain Project and others, and directly derives from the Society for Neuroscience's Neuroscience Database Gateway. Partnered with the Society, its Neuroinformatics Committee, and volunteer consultant-collaborators, our multi-site consortium has developed: (1) a comprehensive, dynamic, inventory of Web-accessible neuroscience resources, (2) an extended and integrated terminology describing resources and contents, and (3) a framework accepting and aiding concept-based queries. Evolving instantiations of the Framework may be viewed at http://nif.nih.gov, http://neurogateway.org, and other sites as they come on line.

Keywords Neurodatabases · Data sharing · Terminologies · Portals

Introduction to This Special Issue of *Neuroinformatics*

This special issue of Neuroinformatics, edited by D. Gardner and M. Martone, informs the neuroscience and neuroinformatics communities of our plans and progress designing the Neuroscience Information Framework (NIF). We begin with this White Paper, which summarizes the project, briefly analyzes the present and future of neuroinformatics, introduces the work we have conducted under phases I and II of the Framework project, and discusses the challenges of serving the entire neuroscience community. Gardner et al. (2008) outline the rationale for, and the community-derived design of, the NIF core terminologies: a set of controlled-vocabulary terms for describing neuroscience data, the experiments that generate them, neuroscience Web resources, and their areas of interest. Müller et al. (2008) describe a parallel terminology effort, Textpresso, which marks up and provides new ways to search for an increasingly large fraction of the contemporary neuroscience literature. Bug et al. (2008) integrate NIF and other terminologies toward the NIFSTD, a standardized semantic framework and ontology bridging scales and areas. Gupta et al. (2008) describe the architecture, rationale and functions of the NIF information federation system, providing examples from the current release. Marenco et al. (2008a, b) present two enabling components, the NIF LinkOut Broker and a concept-based query interface. Finally, Halavi et al. (2008) use NeuroMorpho.Org, an integrated NIF repository for digitally reconstructed neurons, as an example of designing, creating, populating, and curating a neuroscience digital resource. With this issue, we all—as a team—offer to the neuroscience community and to the NIH our design for the Neuroscience Information Framework—and for its evolution.

Introduction to the Neuroscience Information Framework

The Neuroscience Information Framework Derives From, and Is Designed To Serve, the Neuroscience Community

The NIF is a new initiative for integrating access to—and thereby promoting use of—Web-based neuroscience resources. Working as a team, we and colleagues have designed and implemented the NIF under contract from the Institutes and Centers forming the US NIH Blueprint for Neuroscience Research.

In the initial phase, constrained by the enabling contract to exploratory work, we:

- Surveyed the web for neuroscience information resources: databases, literature, gene, tool, and material sites, and built an inventory,
- Developed terminologies to characterize and describe these resources and their contents,
- Convened expert terminology workshops,
- Converged on a feasible design for our initial release compatible with future extensions, and
- Prepared an initial version of this White paper.

Once extension to a technical implementation phase was approved by NIH, we:

- Constructed the Framework as a dynamic inventory of neuroscience data,
- Incorporated a user interface accepting and aiding concept-based queries that span resources across multiple levels of biological function, and
- Developed an underlying terminology for the Framework, brought together from multiple sources including Textpresso, other biomedical terminologies and ontologies, and a total of 18 neuroscience terminology workshop meetings.

All the above is being delivered to the NIH and offered under Open Source (OS) licensing to the neuroinformatics and neuroscience communities.

This is a US national project with contributions from beyond the authorship of this document. Figure 1 shows the paid and volunteer performance sites, emphasizing the geographic spread as well as the intellectual breadth of

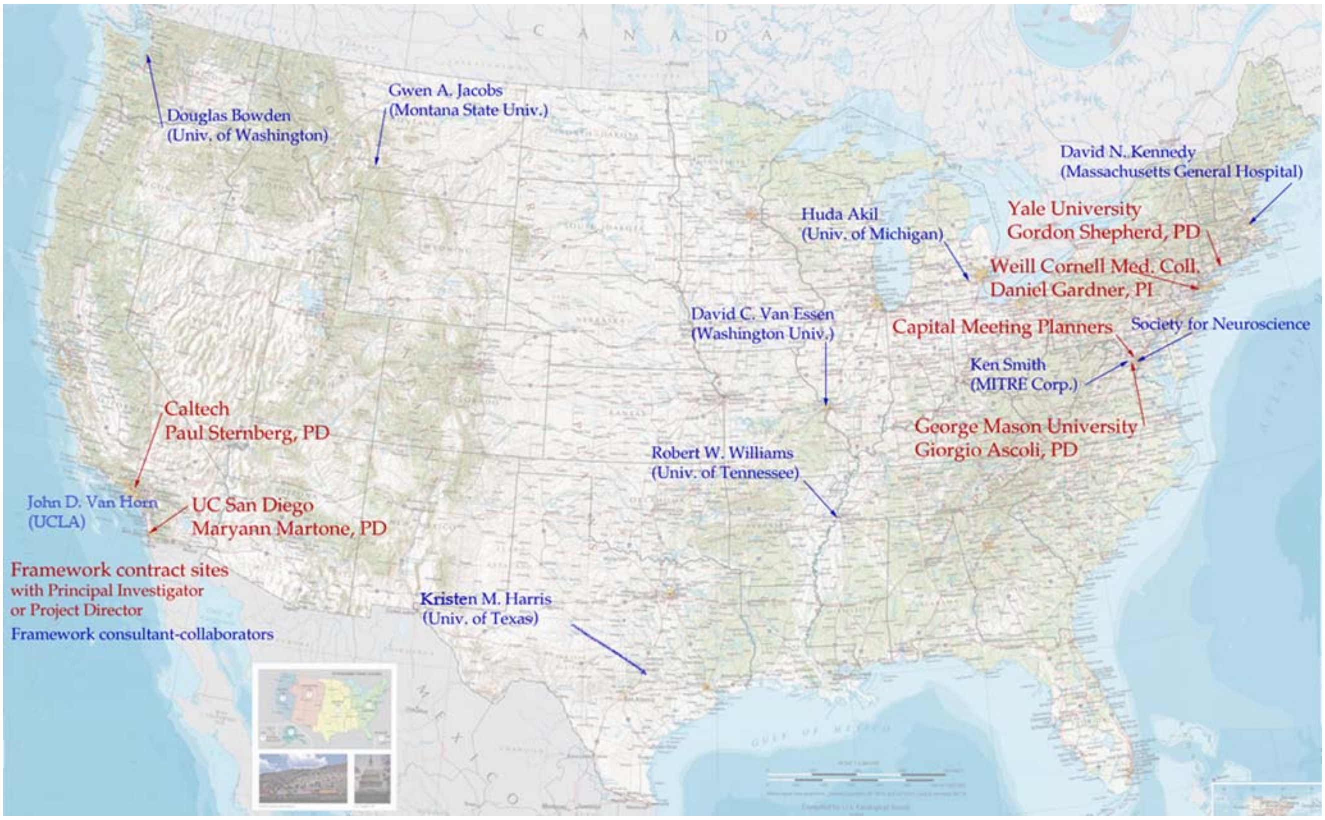


Fig. 1 Framework contributors include both contract sites and volunteer consultant-collaborators. An Appendix lists contributors in greater detail

neuroinformatic contributors to the Framework. An Appendix provides a more extensive list of participants.

The Neuroscience Information Framework Will Advance Neuroscience Research

The Framework is being designed to serve neuroscience investigators by:

1. Facilitating directed and intelligent access to data and findings,
2. Aiding integration, synthesis, and connectivity across related data and findings,
3. Stimulating new and enhanced development of neuroinformatic resources, and
4. Enabling new and enhanced analyses of data.

The Framework and its query tools are being designed to directly implement the first end and thereby enable informed investigators to achieve the second. The Framework, its components, and its satellites will support accessibility, interoperability, and integration; exploration and reasoning will continue to be performed by members of the research community.

We envision that Framework development will further advance neuroinformatics and links among neuroinformatics, bioinformatics, and the terminologies and ontologies relating them, supporting the third goal. The existence of the Framework will spur development of neuroinformatic resources in each of two ways. Many disease- technique- or preparation-focused communities may be reluctant to develop a database or other neuroinformatic resource. By offering a portal and entry point to be used by the entire neuroscience community, the Framework provides a much larger potential audience than a single community can muster. Larger numbers of viewers with broad expertise can add significant value to resources. As the Framework and its tools are Open Source, development will also be aided by making available modules useful for describing, archiving, and sharing data and findings. Framework terminologies, built with the support of many domains of neuroscience, will also aid development of a future semantic web of biomedical ontologies.

The fourth end is not a direct function of the Framework; rather, development of the Framework and easier access to data should spur development and utilization of analytic tools. The many tools indexed by the Internet Accessible Tool Resource, now accessible via the Framework, and the computational neuroinformatic resources at neuroanalysis.org provide two such examples.

The Neuroscience Information Framework is Designed to Advance the Mission and Goals of the NIH Blueprint for Neuroscience Research

The Blueprint "confronts challenges that transcend any single institute or center and serves the entire neuroscience community" and includes procedures that "focus on cross-cutting scientific issues." These summarize the goal and methodology of the Neuroscience Information Framework as well.

The Decade of the Brain (1990–1999; see http://www.loc.gov/loc/brain/) and the years beyond have continued to demonstrate the complexity of nervous systems, in their development, structure, function, and susceptibility to disease. Each individual technique, insight, scale of examination and depth of analysis, each individual disorder advances our understanding of neuroscience as a whole, informed by neuroscience as a whole. Neuroinformatics has served neuroscience well, but no neuroinformatic project has—until now—been designed to serve "the entire neuroscience community." New neuroinformatic tools and resources are needed to "focus on cross-cutting scientific issues" by facilitating access to data and findings that cut across traditional boundaries within neuroscience.

The Framework Will Enable New Paradigms for Neuroinformatics

The Neuroinformatic Ecosystem

Science is an ecosystem: its roots and soil are the experiments that support or disprove hypotheses, and the findings garnered from them. Its sun is the application and creativity of its investigators; their work tills and cultivates. Whether drip irrigation or heavy precipitation, the moisture needed for healthy growth is its funding. The product of all these is data—findings—and the goal is insight. The scientific ecosystem would fail without one other essential component: cross-fertilization. Science focuses on specific details, but gains significance in relation to the whole. Communication among scientists and between scientists and other interested individuals is necessary to relate, to inform, to explain, and to plan the conduct of science.

When techniques were few, direct observation by the unaided eye the only means of data acquisition, and the scale unitary, then words, numbers, and pictures were sufficient for scientific communication. As the scope and methods of science have expanded, and continue to expand, new and far more complex methods of communication and relation of results are needed for the scientific ecosystem to flourish. Bioinformatics is only the latest of these, a product of the fortuitous co-development of affordable computation and universal networking.

Neuroscience is among the most complex scientific activities the world has known. No other area uses more different techniques, develops more different models, explores across more scales: from Ångstrom units to populations. Just as no other contemporary area of science presents a more complex picture, so no other contemporary area of bioinformatics presents as many challenges as neuroinformatics. Our Neuroscience Information Framework is not, cannot be, a complete solution. It is, however, an essential first step towards an integrated ecosystem for neuroscience.

The Neuroinformatic Ecosystem Needs More Data, Better Access to Data, and Easier Re-use of Data

The amount of neuroscience data currently shared, although continuing to increase, is a tiny fraction of what exists and is potentially useful. To form a rich neuroinformatic ecosystem, what is needed is a greatly increased number of data and related resources, resources supporting many more techniques and areas, and a larger number of datasets for existing resources. This does not require significant technical breakthroughs: techniques exist or are being refined for receiving, archiving, describing, supplying, and displaying, and utilizing most types of data relevant to neuroscience. What is needed is recognition and commitment by many disparate neuroscience communities to annotate these data and make them freely and readily available both within their community and also to other domains of neuroscience.

Kennedy (2006) has identified *data sparseness* as a related important issue. If a resource is only sparsely populated with respect to the potentially available data, it loses both utility and credibility. If a researcher looks for data in an archive, fails to find it, and then discovers text partially describing the same data available through other means (e.g. Google, supplementary materials of papers, personal web pages of individual investigators), the archive is failing at a central task. The greater the fraction of the potentially available data of a given type that is accessible through a database, even if the absolute amount of data is small, the more likely that database is to become a useful, credible, and valued resource for those data.

Even in those areas where resources make data available, we find a notable continuum in the *utility* of the available data (Kennedy 2004). Data best suited to integration and re-analysis are the ones that neuroinformatic resources should leverage for development of links and terms. Sites that provide actual data have utility distinct from those that include statements about data, or figures displaying data, and have an essential role in the neuroinformatic ecosystem.

Interoperability is a Continuing Need

Potential utility and availability of web-accessible neuroscience data are not enough. Just as different components of a natural ecosystem interact in multiple and complex ways, so must components of the neuroinformatic ecosystem. We illustrate some of these interactions in Fig. 2, which represents interoperability of data, findings, and the resources that make them available, as a multidimensional set of vectors. For every dimension, distance from the origin gives increasing capacity for interoperability. Basic availability is indicated by the vertical axis, which spans closed data to data freely available via an open, public, resource. Use of standard open protocols and platform- and software-independence is indicated by the technical axis. From the Framework perspective, the domain and data compatibility axes are the most significant: these stress the need for common formats that permit data re-use beyond the immediate community that generated it, and the need for common or relatable descriptors for data, tools, methods, and materials that span different domains of neuroscience. The presence of the temporal axis serves as a reminder that the Framework itself, as well as the resources accessed through it, must incorporate methods for its graceful, scalable, evolution as datasets and resources multiply and techniques, our understanding of neuroscience, and the terminology used to characterize them evolve and expand.

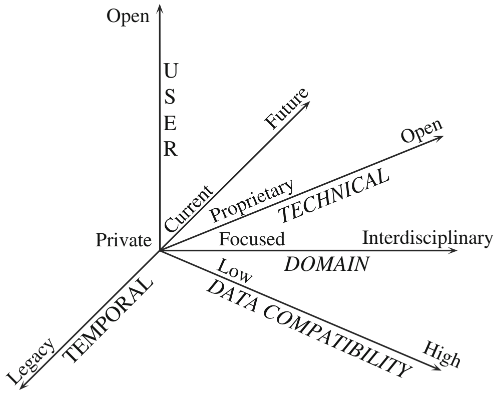


Fig. 2 Vector representation of interoperability dimensions for neuroinformatic resources. For each dimension, increasing interoperability is represented by distance from the origin. User interoperability is enhanced by open access to data, findings, or tools, and zero or minimal cost and licensing requirements. Technical interoperability measures openness of architecture and utility of standards for data format specification and for data and data model exchange. Domain interoperability includes the scope of a resource and the ease with which it interfaces with resources representing different subfields or domains of neuroscience. The data dimension measures relatedness of data and intersection of data models; the domain and data dimensions are thus non-orthogonal. Temporal interoperability reflects ease of migration and of incorporation of both future and legacy data (figure and legend modified from Gardner et al. 2001, © 2001 AMIA)

Methods for Post-Hoc Analysis are a Needed Component of the Ecosystem

The value of data for enabling multiscale integration via re-analysis, meta-analysis, or comparison depends upon both the availability of actual datasets themselves, the adoption of common or convertible data formats, and their characterization by metadata sufficient to permit post-hoc analysis. The Framework is designed to aid these, as well as to facilitate access to such data.

What is also needed, and must similarly be supported by the Framework, is the availability of analytic tools enabling the methods noted above. Such tools need to be robust, general, and characterized—just as data need to be characterized—using precise, neuroscience-aware descriptive terms. Such methods are now available for neuroimaging and some areas of neurophysiology, and need to be expanded, characterized, and made more widely available.

Foundations

The Framework Addresses Needs of the Neuroscience Community

Neuroscience investigators themselves have the greatest need for, and present the primary call for, intelligently directed access to data. As noted above, some of these data are not available outside the laboratory in which they were generated or recorded, others are available but not accessible to public search, and some are in existing web-accessible databases (see the data sparseness problem above). Neuroscientists welcome methods for describing and organizing their own data, and facilitating data sharing toward collaborative and citation-generating re-use of data (Gardner et al. 2003; Liu and Ascoli 2007). Investigators want their data to inform and be informed by others' data. Every database developer is familiar with requests from individual investigators for laboratory systems that organize data and potentially ready the data for sharing. Informatic systems for textual access are powerful and becoming more so, as illustrated by the report on Textpresso in this issue (Müller et al. 2008). However, as we note in a later section, access to and descriptions of datasets, images, tools, and syntheses transcend the capabilities of resources such as Google or PubMed.

The Framework Builds Upon Prior Development of Neuroinformatics

We acknowledge with gratitude but without explicit citation a very large and important body of neuroinformatics development, much of it funded by the NIH's Human Brain Project, that forms the necessary substrate for our Framework development (De Schutter et al. 2006; Koslow and Hirsch 2004). A representative set of projects that directly informed our work includes: SenseLab, Neurodatabase.org, the Internet Accessible Tool Registry (IATR), the Surface Management System Database (SumsDB), the Cell-Centered Database, GeneNetwork/WebQTL, and the Biomedical Informatics Research Network (BIRN) (Gardner 2004; Gardner et al. 2005; Kennedy and Haselgrove 2006; Marenco et al. 2005; Martone et al. 2005; Van Essen et al. 2005; Wang et al. 2003).

The Framework Derives from the Neuroscience Database Gateway

The Neuroscience Database Gateway (NDG) began in 2004 as a pilot project developed by the Society of Neuroscience to investigate the integration of federated neuroscience information on the Web (Gardner and Shepherd 2004). This task was initiated by the Society's Brain Information Group. It is now coordinated by the Society's standing Neuroinformatics Committee, supported through the Framework project, and located at http://ndg.sfn.org, hosted by the Yale Center for Medical Informatics.

This New White Paper Reflects Advances in Neuroinformatics

We here report significant advances in the state of the field presented in an earlier neuroinformatics White Paper, a project of the Society for Neuroscience Brain Information Group led by Floyd Bloom. That paper, available at: http://web.sfn.org/index.cfm?pagename=NDG_whitepapers, highlighted information infrastructure needs of neuroscience research and offered three specific and highly relevant goals for the proposed White Paper and the other three objectives as well: an inventory of neuroscience databases, creation of a database portal, and to "promote broader and more integratable information infrastructural tools to place…neuroscience data in the public domain."

We note the close alignment between these goals, those of the subsequent Neuroinformatics Committee, and the Framework project, as well as our adoption of Open Source. We additionally note that the earlier work's authors included team members Huda Akil, Douglas Bowden, Daniel Gardner, Gwen A. Jacobs, Luis Marenco, Maryann Martone, Gordon Shepherd, David Van Essen, and Robert W. Williams.

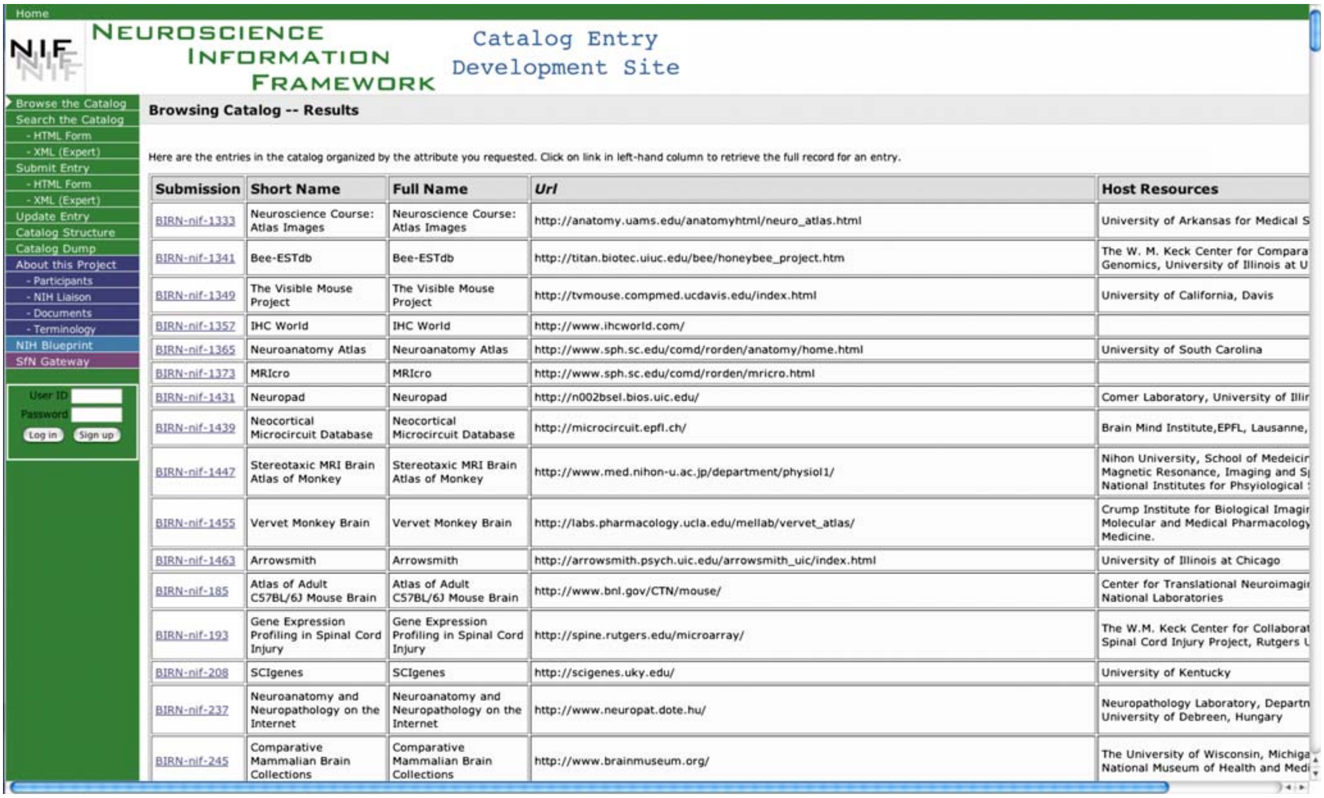
Challenges for Framework Development

The Framework Project Began with an Inventory of Web Neuroscience Databases and Related Resources

To provide a representative sample of web-accessible neuroinformatic resources, and a testbed for syntactic and semantic tags distinguishing among available Web-based neuroinformatic resources, the Framework established a test site at http://neurogateway.org. Figure 3 shows one view of this working development site. We emphasize that this is not the Framework: the other reports in this special issue describe multiple facets of the current NIF (Bug et al. 2008; Gardner et al. 2008; Gupta et al. 2008, Halavi et al. 2008; Marenco et al. 2008a, b; Müller et al. 2008).

The Framework can incorporate only the data or knowledge that are made available; it can integrate these only if sufficient metadata are provided. We note above that in spite of the vigorous development of neuroinformatics, and the many techniques for data collation, archiving, annotation, and distribution developed over the last decade, the amount of neuroscience data available is only a small fraction of the total. The solution depends upon commitments from both data providers across neuroscience and funding agencies to encourage the open archiving and sharing of data. We have also noted that it is important to distinguish between available data—publicly accessible, often via a web archive—and potentially-available data—residing locally in a laboratory or Department willing to share, but not web-accessible or lacking essential metadata (Kennedy 2004). For an example leveraging the Framework component NeuroMorpho.Org see Halavi et al. (2008) in this issue.

Inventoried resources differ in their potential for interoperability. Global neuroscience web resources include experimental, clinical, and translational neurodatabases, knowledge bases, atlases, genetic/genomic and material resources, and tool and modeling sites for processing, analysis, or simulation of brain data. This diversity of sites spans multiple biological scales, techniques, and data



Browsing Catalog -- Results

Here are the entries in the catalog organized by the attribute you requested. Click on link in left-hand column to retrieve the full record for an entry.

Submission	Short Name	Full Name	*Url*	Host Resources
BIRN-nif-1333	Neuroscience Course: Atlas Images	Neuroscience Course: Atlas Images	http://anatomy.uams.edu/anatomyhtml/neuro_atlas.html	University of Arkansas for Medical S
BIRN-nif-1341	Bee-ESTdb	Bee-ESTdb	http://titan.biotec.uiuc.edu/bee/honeybee_project.htm	The W. M. Keck Center for Compara Genomics, University of Illinois at U
BIRN-nif-1349	The Visible Mouse Project	The Visible Mouse Project	http://tvmouse.compmed.ucdavis.edu/index.html	University of California, Davis
BIRN-nif-1357	IHC World	IHC World	http://www.ihcworld.com/	
BIRN-nif-1365	Neuroanatomy Atlas	Neuroanatomy Atlas	http://www.sph.sc.edu/comd/rorden/anatomy/home.html	University of South Carolina
BIRN-nif-1373	MRIcro	MRIcro	http://www.sph.sc.edu/comd/rorden/mricro.html	
BIRN-nif-1431	Neuropad	Neuropad	http://n002bsel.bios.uic.edu/	Comer Laboratory, University of Illir
BIRN-nif-1439	Neocortical Microcircuit Database	Neocortical Microcircuit Database	http://microcircuit.epfl.ch/	Brain Mind Institute,EPFL, Lausanne,
BIRN-nif-1447	Stereotaxic MRI Brain Atlas of Monkey	Stereotaxic MRI Brain Atlas of Monkey	http://www.med.nihon-u.ac.jp/department/physiol1/	Nihon University, School of Medeicir Magnetic Resonance, Imaging and S National Institutes for Phsyiological
BIRN-nif-1455	Vervet Monkey Brain	Vervet Monkey Brain	http://labs.pharmacology.ucla.edu/mellab/vervet_atlas/	Crump Institute for Biological Imagir Molecular and Medical Pharmacology Medicine.
BIRN-nif-1463	Arrowsmith	Arrowsmith	http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html	University of Illinois at Chicago
BIRN-nif-185	Atlas of Adult C57BL/6J Mouse Brain	Atlas of Adult C57BL/6J Mouse Brain	http://www.bnl.gov/CTN/mouse/	Center for Translational Neuroimagi National Laboratories
BIRN-nif-193	Gene Expression Profiling in Spinal Cord Injury	Gene Expression Profiling in Spinal Cord Injury	http://spine.rutgers.edu/microarray/	The W.M. Keck Center for Collaborat Spinal Cord Injury Project, Rutgers U
BIRN-nif-208	SCIgenes	SCIgenes	http://scigenes.uky.edu/	University of Kentucky
BIRN-nif-237	Neuroanatomy and Neuropathology on the Internet	Neuroanatomy and Neuropathology on the Internet	http://www.neuropat.dote.hu/	Neuropathology Laboratory, Departn University of Debreen, Hungary
BIRN-nif-245	Comparative Mammalian Brain Collections	Comparative Mammalian Brain Collections	http://www.brainmuseum.org/	The University of Wisconsin, Michiga National Museum of Health and Medi

Fig. 3 This working development site was established initially to assemble an inventory towards assessing the state of the neuroinformatic ecosystem; later uses included testing 'detector' controlled vocabularies

models, serving communities of neuroscientists with specific conventions, individual terminologies, and distinct foci. The potential for interoperability among resources depends upon design decisions and practices of the inventoried resources, including data model, user interface, and adoption of standard formats and terminologies. Some resources are accessible only via a proprietary or specialized interface, some allow browsing but not query, some allow query using non-intuitive indices or descriptors. Some do not provide sufficient metadata to allow their data or findings to be integrated or analyzed. Some tool sites do not clearly indicate the scope or applicability of their tools, provide verification, or facilitate pipelining.

Disparate neuroscience resources have areas of intersection that allow their findings to be compared and extended. The breadth of contemporary neuroscience ensures that the neuroinformatic resources accessed via the framework will be disparate, but like neuroscience itself these will have areas of intersection that allow findings to be related or extended. Such areas of intersection cannot be predicted in advance; they depend upon both what questions are being asked and how new findings enable connections to be bridged across previously-disparate subfields. The potential for intersection depends upon the scope and type of data or finding in each resource (or the applicability of tools in each toolkit). Identifying such areas was a key goal of Framework design, and we believe, as described below, that common or relatable terminologies, whether *detectors* describing resources as a whole or *selectors* that narrowly specify a cell type, gene, antibody, or protocol, will aid such connectivity.

Framework Design Must Facilitate Maintenance, Expansion, Extension, and Evolution

Neuroscience continues to grow and evolve and this is the greatest challenge to the Framework stability. Here we lay out specific features of this challenge; in the section on Framework design we briefly outline the reasons why Open Source development best meets this challenge.

The Framework must be a stable, reliable, yet extendable resource. This key requirement needs careful planning to accommodate extension of our initial version-1 Framework—NIFv1. Were NIFv1 to be merely a static software system that would require little to no extension or bug-fixing, then the requirements would be minimal. Instead, both the technology required to create a functional and effective Framework and the inevitable expansion of the domain of neuroscience requires long-term support, maintenance, and evolution. We envision that this evolution will also encompass specialization so that groups will be able to tailor the Open Source Framework for their sub-community or special use. Both design methodology and community agreements should ensure that this diversity is accommodated and these additions and extensions are fed back into the Framework in general.

Framework Open Design Specifications

This section presents design choices for a dynamic, scalable Framework capable of degrees of integration from multiple sources. In particular, we detail our adoption of Open Source, suggest that Open Source design and broad scope will aid efficient access to and use of data, and briefly discuss the needs of and solutions toward interoperable and adoptable terminologies.

Overall planning for the technical implementation was agreed upon at a meeting of the Principal Investigator, Project Directors (with P. Miller representing G.M. Shepherd), and selected team members at Caltech on 16 and 17 April, 2007, following NIH approval of the development phase. Also at that meeting, the team selected the goals that were possible given the time and resources available, made a list and detailed plan for development beyond NIFv1, and agreed to remain a consortium for future work. The other reports in this special issue detail the NIFv1 Framework development agreed upon at that time, and carried out in the following year.

Framework Design Combines Specific Technical Choices and Broad Community Support

Open data, access and exchange, via open source and platform, aid Framework-enabled open discovery for neuroscience. Perhaps the most important design principle we have adopted for the Framework is openness. The original NIH proposal for Framework development specified transfer of copyright to the U.S. government. At the insistence of the P.I., this was modified to allow the NIF consortium to substitute Open Source (OS) development. The goal of the Framework is open access to data, facilitating open discovery throughout and across neuroscience and bridging neuroscience with complementary areas of biomedicine. Open Source development methodology supports the informatic ecosystem just as the Framework is designed to aid the neuroinformatic ecosystem. Open Source is implemented through release of all code, terminology, and algorithms under a copyright license that permits unlimited re-use, adoption, and extension of the material, requiring only the continued incorporation of the OS license permitting such use. The Framework is offered under BSD and MIT compatible OS licenses (http://opensource.org/licenses).

In practical terms, this means that the Framework is available to any group that wishes to establish a mirror site, focused subset, or extension of the Framework, or to modify it for a complementary purpose. As we detail below, we also believe that Open Source development will significantly reduce maintenance and versioning costs by promoting multi-site and multi-organization replication and adoption of the Framework and related tools.

Framework Design is Projected to Reduce Costs and Enhance Benefits of Data and Knowledge

We envision the NIF as not only a resource in itself, but as a nucleus and an exemplar to aid bioinformatic development across neuroscience and potentially to linked fields of biomedicine. We project that the Framework will not only promote data sharing and utilization in neuroscience, but also reduce the cost/benefit ratio for data acquisition and utilization, in each of several ways. These include providing Open Source neuroinformatic tools and code that others can leverage, as well as stimulating development by others. Some of these reduce costs that other groups would have to expend to develop resources centered upon their subfields of neuroscience. Others increase the benefit of such development by expanding audience, utility, and opportunities to collaborate and to leverage findings outside the immediate subfield.

Framework inventory and content-aware queries will disseminate and relate neuroscience data and knowledge. We justify our commitment to Framework development—including the many contributions of time, code, tools, insights, and findings from neurobiological and neuroinformatic investigators—by projecting that access via the Framework will increase the distribution, utility, and significance of data and other findings. The content-based query tool will enable more investigators to ask more questions, and will make more easily available the resources capable of providing answers. Just as a paper with a greater number of citations increases the value and therefore decreases the cost/benefit ratio of data contained within, so Framework-enabled examination, coordination, and possible re-analysis of data does the same.

Framework availability and scope will spur development of additional neuroinformatic resources. As noted in the Introduction, we believe that the existence of a single Framework query point for a very wide range of Web-based neuroscience will itself encourage the growth of the neuroinformatic ecosystem. The potential is great for additional communities in neuroscience, whether centered on specific areas of function, disease, technique, or preparation, to develop terminologies and methods for making available data, findings, or tools useful for their domain and beyond. By providing a portal and query point to the entire neuroscience community, the Framework expands the potential audience, increasing exposure of the site's contents and offering the possibility for collaborations and informative links to related areas. This can motivate communities to support the neuroinformatic ecosystem and thereby reduce the data sparseness problem.

Framework Terminology Integrates Multiple Streams

The NIFv1 Framework and content-based query tool development include multiple neuroscience terminology thrusts, detailed in Gardner et al. (2008) Bug et al. (2008), and Müller et al. (2008) in this volume. Good design also favors adoption of existing terminologies, both to ease integration of neuroscience knowledge with that of other fields and also to reduce the magnitude of lexical development. We recognize that interoperability and efficiency would both be aided by our adoption of terms taken from existing standards, subject to relevance for neuroscience and availability under Open Source licensing. Obvious choices include BIRNLex and the NCBI taxonomy. We also acknowledge the first neuroscience-centric keyword development, established more than a decade ago by Framework team member Bernice Grafstein. The Framework adoption of XML for future terminology representation, and parallel Human Brain project efforts to place Framework terms in BrainML format, allow incorporation of other XML-based terminologies in whole or in part using the namespace feature of XML.

Implementation and Core Functionality of the NIFv1

We have implemented NIFv1 as a Web resource available to any neuroscientist user with a contemporary Web-accessible computer; all functionality is available on any platform and operating system compatible with current Java. Supporting this goal required adherence to standards permitting current use and future evolution, and of course administrative tools aiding content management and update of the system. The NIFv1 was developed following standard commercial-grade techniques for Web-accessible code development, tracking, and testing. Delivered under a non-contaminating Open Source license, it includes software components and terminologies needed to establish a Web-based Framework application on any contemporary multi-processor or multi-core Unix server with gigabyte (GB) or better memory and 250 GB or larger disc, standard Open Source gnu compilers and library, Java 1.5, MySQL or PostgreSQL database, and Apache web server components including Tomcat.

Details of Framework design and implementation are provided in the accompanying papers, especially Gupta et al. (2008). An overview of major system components of the NIF is shown in Fig. 4. Implementation of the system delivering core NIFv1 functionality includes four main modules. At the top level of Fig. 4 are the NIFv1 interfaces: the NIFv1 Query Interfaces supporting neuroscientist users and administrative interfaces, including those for registering and maintaining entries specifying interoperable NIF resources. At the middle level in Fig. 4 are the NIF Database Resource Directory, the NIF Database Mediator, and the NIF Document Archive. Additional NIFv1 components include NeuroMorpho.Org as well as multi-tiered back-end data resources and NIFv1 services which provide specific functionality.

Why Nothing Else Does What the Framework Will Do

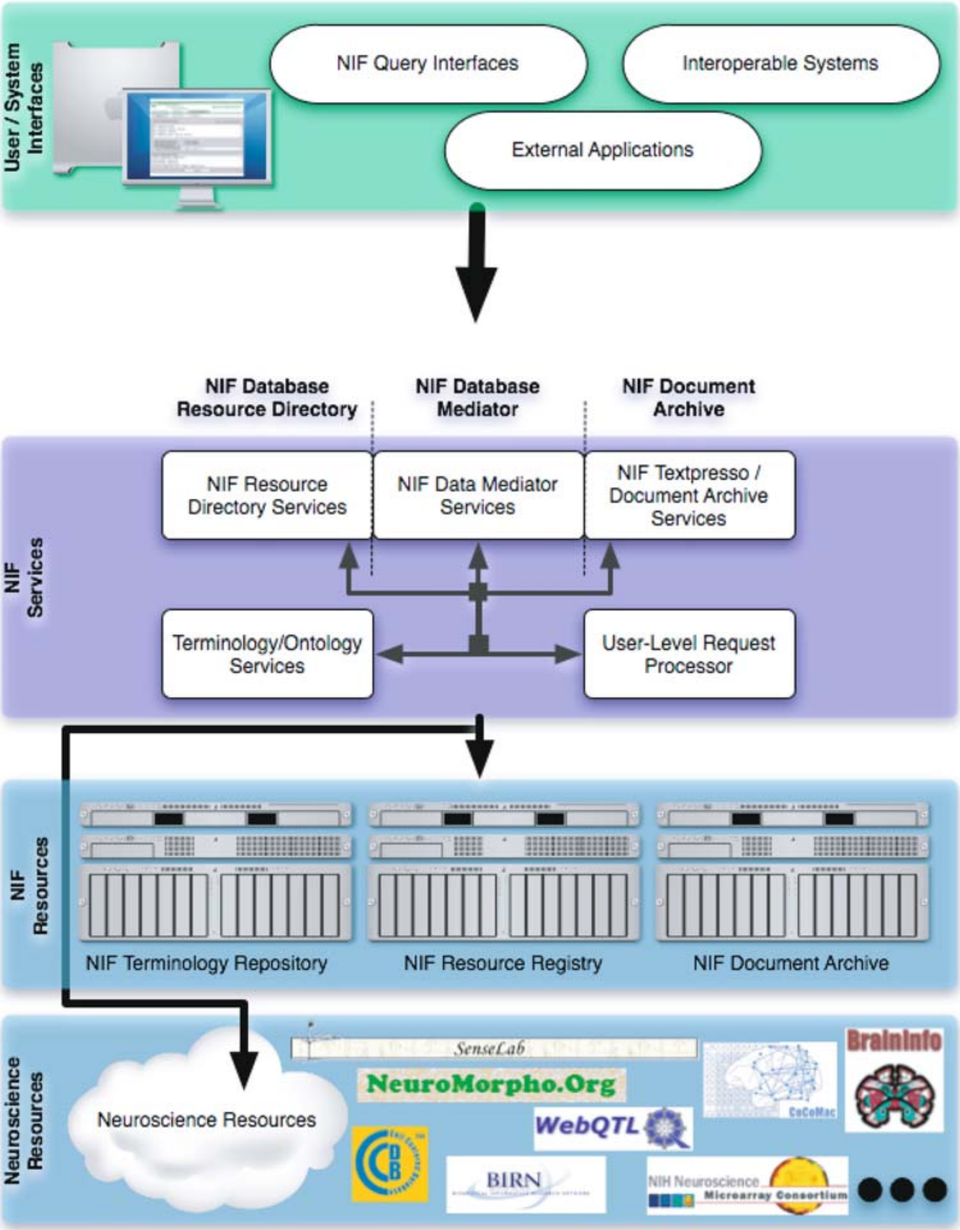
The Framework is Neuroscience-Specific and Neuroscience-Generated

Neuroscience does not at present have a central, general source for relevant data. Geneticists, structural biologists, and molecular biologists have universally-accessed databases that emphasize gene and protein sequence and structure data (e.g., NCBI Entrez, PDB, and others). Because there is no site that directly addresses their needs, neuroscientists by default make use of a variety of search engines (e.g., Google, Google Scholar, and PubMed) that are largely literature-oriented.

We are designing NIFv1 to change this. The Framework presents neuroscientists with a single starting point for their searches, one that can be a portal that students start using at the dawn of their training and continue to utilize as their primary access to multiple and complex sets of data available from a growing number of neuroscience-specific databases. No other site or tool is comparable because this approach has never before been attempted for neuroscience. This will not echo material available through other sources, but will complement it.

- The Framework is focused on neuroscience, with access to resources that individually address key specific areas or techniques, that supply data in addition to knowledge, and that in aggregate span the breadth of neuroscience.
- The Framework derives from the neuroscience community itself; many of the authors are developers but we are all in addition neuroscientists and users.
- The Framework has the Society for Neuroscience as a resource (Kennedy 2007). Three SfN Presidents have said: 'The Society for Neuroscience strongly supports the joint effort by members of the Society's Neuroinformatics Committee to spearhead establishment of a Neuroscience Information Framework',

Fig. 4 Overview of the NIFv1 implementation core architecture



'Development of the NIF has benefited and will continue to benefit greatly from the volunteer contributions from SfN membership, particularly from members of the NeuroInformatics Committee' and: 'this partnership with the SfN is pivotal, because the SfN can promote the power of the NIF in presentations, courses, on its web site and even provide a venue for training and demonstrations. The goal is to fully integrate neuroinformatics into the daily life of the average neuroscientist, and none of the existing databases, search engines or entities have ever succeeded in doing that.'

- The Framework builds on a broad series of neuroscience expert terminology workshops. These workshops are to our knowledge the only coordinated unified efforts to assemble working *neuroscientist-users* representing focused communities within the breadth of neuroscience and derive collegial consensus terminologies broadly characterizing the questions they ask, the data they collect, and the techniques they use (Gardner et al. 2008).
- The Framework allows users to specify both the types of resource to query and whether data or literature references are required; this capability may in the future

be expanded to allow synthesizing information from multiple sources and ranking by value.

NIF Functionalities Relative to Other Tools

We offer comparisons to popular search tools:

Google: Compared to Google, the Framework enables neuroscientists by offering content-based queries, access to data, and a focus on neuroscience:

- Framework neuroscience concept-based queries, provide a more comprehensive, yet focused search result than Google and thereby reduce the number of false negative results. Unlike Google, the Framework allows users to clarify, specify, or modify search terms, reducing the number of false positive items in the response, and so increasing the signal to noise ratio.
- Google indexes existing Web pages. However, many neuroscience datasets are contained in databases accessible only via query interfaces, and only presented dynamically (often not in HTML or PDF) in response to an ad-hoc query. This provision of data, rather than text describing data or pictures showing a static representation of some feature of data, further distinguishes many Framework-accessible resources from those that Google can find.
- Unlike Google, the Framework specifically references neuroscience resources that are known to provide meaningful, useful data or other information. This is because the Framework only links to Web resources that members of the Framework team have visited and approved as relevant and reliable.

Entrez-PubMed: Compared to Entrez, the Framework again enables neuroscientist users by its focus on neuroscience and its use of content-based queries:

- The NIF is a portal to a rapidly growing body of neuroscience information on the web, much as Entrez provides a portal to a curated set of biomedical resources, largely built around genomics and proteomics (although expanding to other areas). Though Entrez does provide combined searching against documents plus data repositories, it does so in a manner that can't fully tap the conceptual inter-relatedness of the individual elements. Indexing all NIF entities with the NIF terminology/ontology specifically enriched for concepts relevant to neuroscientists makes it possible to provide a much more contextually-relevant and thorough correlated concept analysis to drive query resolution and to organize query results.
- As a literature service, PubMed provides somewhat better focus than Google by, (1) limiting citations to documents related to biomedicine, (2) enabling users to narrow their searches by language, species, age, type of document, etc., (3) utilizing Boolean logic, and (4) indexing literature citations using MeSH; however, it remains largely a search-by-key-word service. Thus, it is vulnerable to both false negatives and false positives when users' terminology differs from that used for indexing.

Information Sharing Statement

Lector, si monumentum requiris, Circumspice.

Acknowledgements This project has been funded in whole or in part through the NIH Blueprint for Neuroscience Research with Federal funds from the National Institute on Drug Abuse, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN271200577531C. The Neuroscience Information Framework team gratefully acknowledges the support of volunteer consultant-collaborators and friends, and The Society for Neuroscience. Early development of the SfN Neuroscience Database Gateway was supported by the Society for Neuroscience by means of a generous gift from Paul Allen and Jody Patton and by contract (NIH Order No. 263-MD-409125-1) from NIMH, NINDS, and NIDA. We thank C. Wren for the epitaph adaptable as the Information Sharing Statement. The Advisory Committee consists of Huda Akil, Giorgio Ascoli, Daniel Gardner, Bernice Grafstein, Maryann E. Martone, Gordon Shepherd, Paul Sternberg, David C. Van Essen, and Robert W. Williams.

Appendix

The Framework Team

The Framework Team includes many individuals, representing many nodes of a collegial network for neuroinformatic development.

The Contractor for Phases I and II, described in this White paper and the special issue it introduces, is Weill Medical College of Cornell University, Daniel Gardner, PI, and subcontractors (with the PD at each) are:

- Yale University (Gordon Shepherd, PD)
- Caltech (Paul Sternberg, PD)
- University of California, San Diego (Maryann Martone, PD)
- George Mason University (Giorgio Ascoli, PD), and
- Capital Meeting Planners Inc

Team members supported via Framework Contractor or Subcontractor sites include: Giorgio A. Ascoli, Vadim Astakhov, William Bug, Fabien Campagne, Mark Ellisman, Ronit Gadagkar, Daniel Gardner, Bernice Grafstein, Jeffrey Grethe, Amaranth Gupta, Erdem Kurul, Luis Marenco, Maryann E. Martone, Perry L. Miller, Hans-Michael Müller, Thien Nguyen, Xufei Qian, Adrian Robert, Ruggero

Scorcioni, Gordon M. Shepherd, Paul W. Sternberg, Willy Woong, and Ilya Zaslavsky

The team also includes a set of consultant-collaborators. None received direct support from the Framework project; each is pleased to make available, towards supporting the neuroinformatic ecosystem, code, products, or expertise that aid Framework development:

- The Society for Neuroscience
- Huda Akil, Univ. of Michigan Med School
- Douglas Bowden, Univ. of Washington
- Kristen M. Harris, Univ. of Texas at Austin
- Gwen A. Jacobs, Montana State Univ.
- David N. Kennedy, Massachusetts General Hospital
- Ken Smith, MITRE Corp.
- David C. Van Essen, Washington Univ.
- John D. Van Horn, UCLA
- Robert W. Williams, Univ. of Tennessee

As this work was being submitted for publication, the team learned of the sudden and untimely death of our valued colleague William Bug. Untiring in his vision, enthusiasm for the project, and ability to bridge communities of biomedicine, he will be greatly missed. In his honor we echo his invariable signoff from hundreds of inspiring e-mails: Cheers, Bill.

References

Bug, W., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A., et al. (2008). The NIFSTD and BIRNLex vocabularies: Building comprehensive ontologies for neuroscience. *Neuroinformatics*, doi:10.1007/s12021-008-9032-z.

De Schutter, E., Ascoli, G. A., & Kennedy, D. N. (2006). On the future of the Human Brain project. *Neuroinformatics*, *6*, 129–130. doi:10.1385/NI:4:2:129.

Gardner, D. (2004). Neurodatabase.org: Networking the microelectrode. *Nature Neuroscience*, *7*(5), 486–487. doi:10.1038/nn0504-486.

Gardner, D., Abato, M., Knuth, K. H., & Robert, A. (2005). Neuroinformatics for neurophysiology: The role, design and use of databases. In S. H. Koslow & S. Subramaniam (Eds.), *Databasing the brain: From data to knowledge (Neuroinformatics)* (pp. 47–67). New York: Wiley.

Gardner, D., Goldberg, D. H., Grafstein, B., Robert, A., & Gardner, E. P. (2008). Terminology for neuroscience data discovery: multi-tree syntax and investigator-derived semantics. *Neuroinformatics*, doi:10.1007/s12021-008-9029-7.

Gardner, D., Knuth, K. H., Abato, M., Edre, S. M., White, T., DeBellis, R., et al. (2001). Common data model for neuroscience data and data model interchange. *Journal of the American Medical Informatics Association*, *8*, 17–31.

Gardner, D., & Shepherd, G. M. (2004). A gateway to the future of neuroinformatics. *Neuroinformatics*, *2*, 271–274. doi:10.1385/NI:2:3:271.

Gardner, D., Toga, A. W., Ascoli, G. A., Beatty, J., Brinkley, J. F., Dale, A. M., et al. (2003). Towards effective and rewarding data sharing. *Neuroinformatics*, *1*, 289–295. doi:10.1385/NI:1:3:289.

Gupta, A., Bug, W., Marenco, L., Qian, X., Condit, C., Rangarajan, A., et al. (2008). Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics*, doi:10.1007/s12021-008-9033-y.

Halavi, M., Polavaram, S., Donohue, D. E., Hamilton, G., Hoyt, J. Smith, K. P., et al. (2008). NeuroMorpho.Org implementation of digital neuroscience: dense coverage and integration with the NIF. *Neuroinformatics*, doi:10.1007/s12021-008-9030-1.

Kennedy, D. N. (2004). Barriers to the socialization of information. *Neuroinformatics*, *2*, 367–368. doi:10.1385/NI:2:4:367.

Kennedy, D. N. (2006). Where's the beef? Missing data in the information age. *Neuroinformatics*, *4*, 271–274. doi:10.1385/NI:4:4:271.

Kennedy, D. N. (2007). Neuroinformatics and the Society for Neuroscience. *Neuroinformatics*, *5*, 141–142. doi:10.1007/s12021-007-0014-3.

Kennedy, D. N., & Haselgrove, C. (2006). The internet analysis tools registry: A public resource for image analysis. *Neuroinformatics*, *4*, 263–270. doi:10.1385/NI:4:3:263.

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*, 267–270. doi:10.1385/NI:2:3:267.

Liu, Y., & Ascoli, G. A. (2007). Value added by data sharing: Long-term potentiation of neuroscience research. *Neuroinformatics*, *5*, 143–145. doi:10.1007/s12021-007-0009-0.

Marenco, L., Ascoli, G. A., Martone, M. E., Shepherd, G. M., & Miller, P. L. (2008a). The NIF LinkOut broker: A web resource to facilitate federated data integration using NCBI Identifiers. *Neuroinformatics*, this issue.

Marenco, L., Crasto, C. J., Liu, N., Migliore, M., Liu, J., Morse, T. M., et al. (2005). SenseLab: A decade of experience with multilevel, multidisciplinary neuroscience databases. In S. H. Koslow & S. Subramaniam (Eds.), *Databasing the brain: From data to knowledge (Neuroinformatics)* (pp. 343–347). New York: Wiley.

Marenco, L., Li, Y., Martone, M. E., Sternberg, P. W., Shepherd, G. M., & Miller, P. L. (2008b). Issues in the design of a pilot concept-based query interface for the Neuroinformatics Information Framework. *Neuroinformatics*, doi:10.1007/s12021-008-9035-9.

Martone, M. E., Peltier, S. T., & Ellisman, M. H. (2005). Building grid-based resources for neurosciences. In S. H. Koslow & S. Subramaniam (Eds.), *Databasing the brain: From data to knowledge (Neuroinformatics)* (pp. 111–121). New York: Wiley.

Müller, H.-M., Rangarajan, A., Teal, T. K., & Sternberg, P. W. (2008). Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics*, doi:10.1007/s12021-008-9031-0.

Van Essen, D. C., Harwell, J., Hanlon, D., & Dickson, J. (2005). Surface-based atlases and a database of cortical structure and function. In S. H. Koslow & S. Subramaniam (Eds.), *Databasing the brain: From data to knowledge (Neuroinformatics)* (pp. 369–388). New York: Wiley.

Wang, J., Williams, R. W., & Manly, K. F. (2003). WebQTL: Web-based complex trait analysis. *Neuroinformatics*, *1*, 299–308. doi:10.1385/NI:1:4:299.