nature neuroscience

e-Neuroscience: challenges and triumphs in integrating distributed data from molecules to brains

Maryann E Martone, Amarnath Gupta & Mark H Ellisman

Imaging, from magnetic resonance imaging (MRI) to localization of specific macromolecules by microscopies, has been one of the driving forces behind neuroinformatics efforts of the past decade. Many web-accessible resources have been created, ranging from simple data collections to highly structured databases. Although many challenges remain in adapting neuroscience to the new electronic forum envisioned by neuroinformatics proponents, these efforts have succeeded in formalizing the requirements for effective data sharing and data integration across multiple sources. In this perspective, we discuss the importance of spatial systems and ontologies for proper modeling of neuroscience data and their use in a large-scale data integration effort, the Biomedical Informatics Research Network (BIRN).

The molecular biology explosion in the 1980s and 1990s was driven not only by technical advances in our ability to rapidly acquire large amounts of sequence information, but also by the concurrent revolution in computer science and information technology. It is unlikely that the Human Genome Project would have been undertaken if the millions of base pairs were to be published in a text book. Two main developments fueled the immensely valuable effort to assemble entire genomes: the availability of ubiquitous and highly interconnected computational resources, and the availability of DNA sequences in a machine-readable form, which enabled computational algorithms to be applied to them.

Experimental advances have given the individual neuroscientist an increasingly powerful arsenal for obtaining data across multiple scales, from the level of molecules to entire nervous systems. Unlike the relatively simple genomic data that can be viewed as a sequence of four letters, the objects of study in neuroscience are considerably larger and more complex—cellular architectures, connectivity, physiological and behavioral data—making the primary data equally large and complex. Moreover, it is clear that difficult neuroscience problems like mapping gene expression in the whole brain and understanding Parkinson's disease are too large to be accomplished unless the research of multiple groups working across disciplines can be combined. The obvious prerequisites in combining multiple groups' research are twofold: to provide an infrastructure where they can share their data and analyses, and to ensure the information produced by any group is in a form that another group can use. But how do we accomplish these goals?

Government agencies recognized over a decade ago that a significant effort was required to develop the information infrastructure and computational tools to make these data generally available to the scientific community and to begin to integrate brain data into unifying models of the nervous system. Pioneering initiatives like the Human Brain Project (www.nimh.nih.gov/neuroinformatics), led by the National Institutes of Health (NIH), support development of computational algorithms for visualization, analysis and modeling of neuroscience data and databases to provide access to neural data. During the same decade, the Neuroscience Thrust of the National Science Foundation (NSF) supported the National Partnerships for Advanced Computational Infrastructure (NPACI; www.npaci.edu) initiative, which brought together computational neuroscientists and leading computer scientists from nearly 50 universities to collaborate in shaping advances in database technologies, grid computing and advanced networking research so they would better fit the requirements of advancing neuroinformatics and biomedical research.

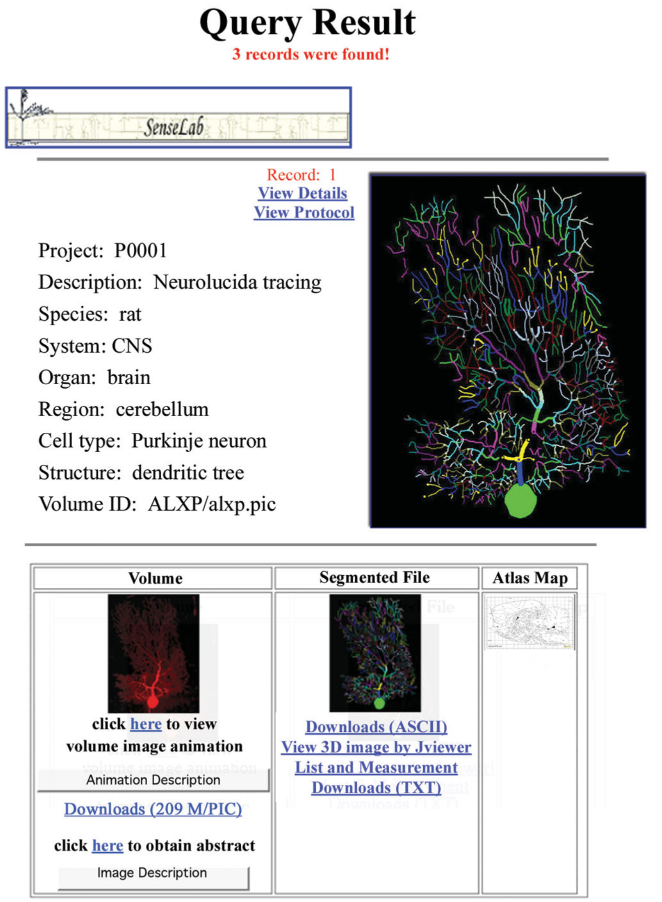
The challenges involved in creating informatics resources for complex neuroscience data are cogently laid out in an excellent commentary by Kotter[1]. These range from sociological hurdles involved in sharing hard-won data that has not yet been fully utilized, to technological problems involved in representing, storing and accessing large amounts of non-standardized, complex data. Nevertheless, neuroinformatics research has succeeded in formalizing many key issues involved in data sharing and data exchange and presenting partial solutions that will go a long way toward realizing the promise of a 'global scientific forum'[2]. Here we describe our vision of how this ideal—but immensely difficult—goal can be achieved, as well as our efforts in developing the Cell-Centered Database (CCDB) and the Biomedical Informatics Research Network (BIRN).

Creating informatics resources for neuroimaging data

The major impetus behind neuroinformatics efforts is the belief that biological data have value beyond the purpose for which they were originally acquired[2]. To quote from the NIH data-sharing policy guidelines: "…sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measure-

Maryann E. Martone is in the Department of Neurosciences, National Center for Microscopy and Imaging Research and The Center for Research in Biological Systems, The University of California San Diego, La Jolla, California 92093-0608, USA. Amarnath Gupta is at the San Diego Supercomputer Center, The University of California San Diego, La Jolla, CA 92093-0505, USA. Mark H. Ellisman is affiliated with all of the above institutions.
e-mail: mark@ncmir.ucsd.edu

Published online 27 April 2004; doi:10.1038/nn1229



Query Result

3 records were found!

Record: 1
View Details
View Protocol

Project: P0001
Description: Neurolucida tracing
Species: rat
System: CNS
Organ: brain
Region: cerebellum
Cell type: Purkinje neuron
Structure: dendritic tree
Volume ID: ALXP/alxp.pic

Volume	Segmented File	Atlas Map
click here to view volume image animation Animation Description Downloads (209 M/PIC) click here to obtain abstract Image Description	Downloads (ASCII) View 3D image by Jviewer List and Measurement Downloads (TXT)	

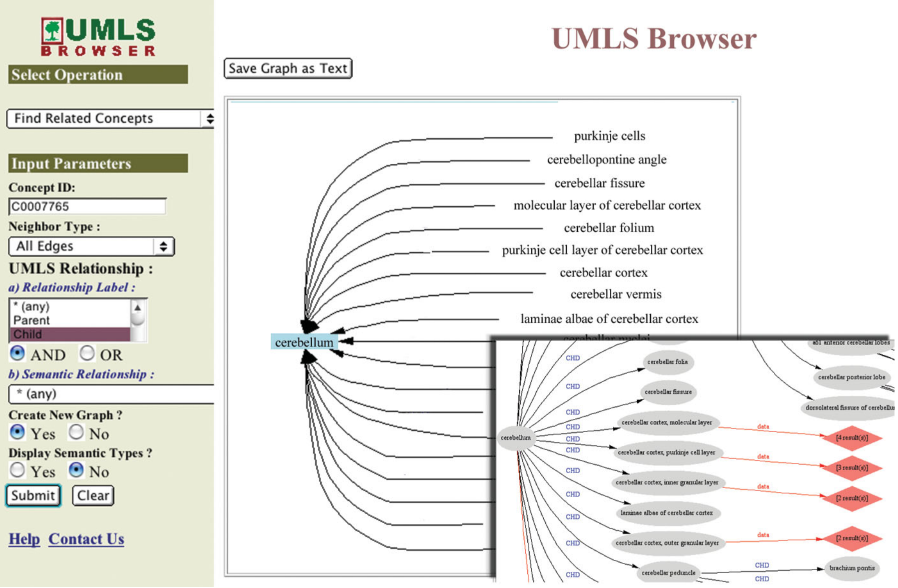
Figure 1 A query result from the CCDB shows a dynamically generated view of the type of data available for any given dataset. This particular data set represents an optical section series of a filled neuron imaged with confocal microscopy (Volume). The dendritic branching structure was traced using Neurolucida (Segmented Object). A higher-resolution view of the traced structure is shown at the upper right. The location of the cell in a brain atlas is also available (Map Image) and the coordinates are stored in the database. Users may also issue a query against the Yale Senselab databases on neuronal properties[11,32] to find additional information about the cell type returned from the query.

ment, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined". Science has always been in the business of sharing information through publications and presentations, but as primary datasets get larger and analytical methods more complex and less standardized, the printed record becomes increasingly unsatisfactory. Imaging in particular, from MRI-based methods to microscopy, remains a major tool by which scientists investigate biological systems, but this is poorly served by current publication formats.

Because image-based data are rich in content, large in size and laborious to obtain, they have been a prime driver for neuroinformatics efforts to create suitable databases and analysis tools to make them more broadly accessible to the scientific community. In many recent surveys of neuroimaging resources[3,4], the term "database" is used broadly to refer to any organized data collection. Creating a collection of images and offering them on a web site is certainly useful and far better than storing data in a drawer or on a shelf. But the problem with simple data collections is that they do not lend the data to any automation—no software can organize, manipulate or search the data because the information is not in a format that can readily be processed by computer. To achieve this, we must go beyond simple data collections to more structured data representations. Here, we adopt the definition that a structured database is an instantiation of a 'data model'. A data model is a mathematically expressible description of the data, including the structure of the data, the relationships among different pieces of data, and the operations allowed on the data.

The difference between a data collection and a structured database can be illustrated by the following example. A researcher is interested in comparing the branching patterns of Purkinje neurons from different species. She collects a series of filled and immunolabeled neurons and uses a program like Neurolucida (Microbrightfield, Inc.) to extract a model of the dendritic branching pattern. Feeling generous, she decides to make these data available to others who may be interested in using this data for computational modeling studies or to perform additional analyses. How she chooses to share these data dramatically impacts how accessible and useful they are. She can create a simple data collection site by posting the Neurolucida files along with some descriptive information on how they were produced and perhaps a viewer to allow the visualization of the branching structure. In this case, a potential user specifically interested in Purkinje cells from adult mouse cerebellum may find the site by searching the web using key words "Purkinje Cell", wading through the thousands of hits, and browsing through the posted data to find suitable examples. If there are too many data sets to browse manually, the user may search the site for keywords such as "adult mouse". In this case, he might hit adult mouse Purkinje cells, adult rat Purkinje cells and Purkinje cells labeled with a mouse monoclonal antibody, because the search engine knows nothing about the concepts species and age. Suppose instead, the resource creator decides to go a little further and stores the descriptive information in a simple database with the fields "filename, species, age". In this case, the user can immediately "find all mouse Purkinje cells from animals aged > 1 month" because the structure of the database specifies that each image has a species, which in turn has an age, and the query engine knows where to look to find the values and what operations can be performed on them (for example, age is a numerical value). The richer the data model, the more targeted the queries and the more benefit to the user. For example, if the data resource stores both experimental and analytic results, the user could "find all Purkinje cells labeled with Lucifer Yellow with more than one primary dendrite". On the other hand, the more complex the data representation, the more information technology expertise is required to establish and maintain the resource[5].

Most existing databases for complex data types enable the user to select data sets based on descriptive information stored with the data ("meta data"). For example, using the GENSAT database[6] (www.gensat.org; see also p. 483 of this issue[7]), it is possible to query a set of gene expression images based on annotations regarding its localization pattern ("localized versus widely expressed"). However, relatively few databases have investigated more rigorous modeling of complex imaging data, so that its content is exposed to direct query. Such limited data modeling perhaps contributes to the perception noted by Kotter[1] that "databases are … mere data repositories that do not necessarily create insights". This viewpoint is perhaps unnecessarily dismissive of the power of proper information management; the ability to locate relevant data when following a train of inquiry certainly has the power to stimulate hypothesis formation. Ultimately, however, as the amount of data continues to increase, neuroinformatics must explore data representations that allow neuroscientists to "mine" computationally the content of imaging data sets for the purposes of "…non-trivial extraction of implicit, previously unknown, and potentially useful information from neuroscience data"[1]. To achieve this goal, computer and neuroscientists are working together to represent complex imaging data so that it can be queried and used in computation while providing a rich set of methodological descriptors so that the data can be understood and interpreted by others. Finally, data must be situated in a larger framework so that it can be related to data taken by others, either at the same or different scale.


UMLS BROWSER
Select Operation
Input Parameters
Help
Contact Us
UMLS Browser
Save Graph as Text
purkinje cells
cerebellopontine angle
cerebellar fissure
molecular layer of cerebellar cortex
cerebellar folium
purkinje cell layer of cerebellar cortex
cerebellar cortex
cerebellar vermis
laminae albae of cerebellar cortex
cerebellar nuclei
cerebellum
CHD
cerebellar folia
cerebellar fissure
cerebellar cortex, molecular layer
cerebellar cortex, purkinje cell layer
cerebellar cortex, inner granular layer
laminae albae of cerebellar cortex
cerebellar cortex, outer granular layer
cerebellar peduncle
data
[4 result(s)]
[3 result(s)]
[2 result(s)]
[2 result(s)]
a61 anterior cerebellar lobes
cerebellar posterior lobe
dorsolateral fissure of cerebellu
brachium pontis

Find Related Concepts

Concept ID: C0007765

Neighbor Type : All Edges

UMLS Relationship :

a) Relationship Label : * (any) Parent [Child]

● AND ○ OR

b) Semantic Relationship : * (any)

Create New Graph ? ● Yes ○ No

Display Semantic Types ? ○ Yes ● No

Submit Clear

Figure 2 Portion of UMLS showing concepts related through the "child (is_a)" relationship shown using a graphical browsing tool developed by BIRN. Users may search and browse through the UMLS and may also perform some simple graph queries (*e.g.*, "Compute shortest distance between two concepts" using the forms on the left). Lower-right, the relationships in the UMLS are being used to query distributed multi-scale database sources through the BIRN mediator. The diamonds indicate concepts for which data were found. The connected ovals show concepts and relationships contained in the UMLS.

The Cell-Centered Database

The Cell Centered Database[8,9] (CCDB; www.ncmir.ucsd.edu/CCDB) serves as a development platform for pushing the capabilities of database systems for storing and analyzing large, three-dimensional (3D) imaging datasets. The CCDB is a web-accessible database, providing 3D structural and protein localization data derived from light and electron microscopy to the scientific community. It was designed around the process of 3D reconstruction from 2D micrographs, capturing key steps in the process from experiment to analysis. The types of imaging data stored in the CCDB range from large-scale maps of protein distributions, taken by laser-scanning multi-photon or confocal microscopy across multiple brain regions, to 3D reconstruction of individual cells, subcellular structures and organelles obtained using electron tomographic methods. Electron tomography is a powerful technique for 3D reconstruction at electron-microscopic resolution[10]. It is conceptually similar to medical imaging techniques such as CAT scans, in that it derives a 3D volume from a series of 2D projections through a structure. In this case, the structures are contained in sections prepared for electron microscopy, and projections are obtained as the sections are tilted through a limited angular range.

The CCDB schema contains more than 50 tables for descriptive information about the experiment, subject, tissue processing, imaging method and analysis details, and it keeps track of the different types of primary and derived data for a single set of images (**Fig. 1**). The CCDB also allows the results of morphometric analysis (for example, surface area) to be stored for any object segmented from a 3D dataset. The schema is generic for 3D reconstruction and can accommodate cellular data regardless of tissue of origin. However, it contains several features specialized for neuronal data, and the bulk of data currently available for download derive from the nervous system.

The CCDB is exploring the instantiation of more sophisticated data models based on mathematical modeling of the types of data that typically result from imaging experiments. For example, the CCDB has many instances of filled neurons imaged with confocal microscopy, and, as described in the hypothetical case above, many of these have had their dendritic branching patterns traced using Neurolucida (**Fig. 1**). These branching patterns can be viewed as an instance of a tree data structure, where a node represents a dendritic branching point, and an edge between two nodes represents a branch. Each branch has attributes such as diameter and length, and the model can impose constraints (*e.g.*, the diameter of higher-order branches cannot be more than that of lower-order branches). Because the set of operations on a tree is well understood in computer science, this models a single neuron well enough to enable questions like "Find the diameter distribution of the third-order branches of those Purkinje neurons that have more than one primary branch". Unlike searching on descriptive attributes, which requires access to an explicit representation in the schema, a user can potentially query for any property that can be computed from a tree structure.

Data integration of distributed multi-scale data sources

Assuming that neuroscientists agree to make their primary data available through well-structured web repositories, will that be sufficient to achieve the type of large-scale collaboration and data integration that we seek? As has been well acknowledged, placing data into shared data repositories is only a part of the battle; they must also be integrated into a body of cross-accessible knowledge, where results from disparate data sets can be accessed and understood on the basis of a common understanding of neural systems[1,2,11,12]. As the emerging cyberinfrastructure removes limits on the physical location of data and resources, the need for information to be gathered into single, centralized repositories (data warehouses) is decreasing. However, the advantage of data warehouses is that the meaning and interrelationships among different pieces of data are represented in the data structure. The disadvantage is that these repositories are often too rigid to accommodate the addition of new types of data, and they can be difficult to manage as they grow more complex[13].

Both neuroscientists and computer scientists are increasingly turning to more distributed architectures, where independent, distributed data resources can participate in larger, collaborative virtual data federations[13]. The federation approach is attractive for many reasons, not the least of which is that it maintains the independence of indi-

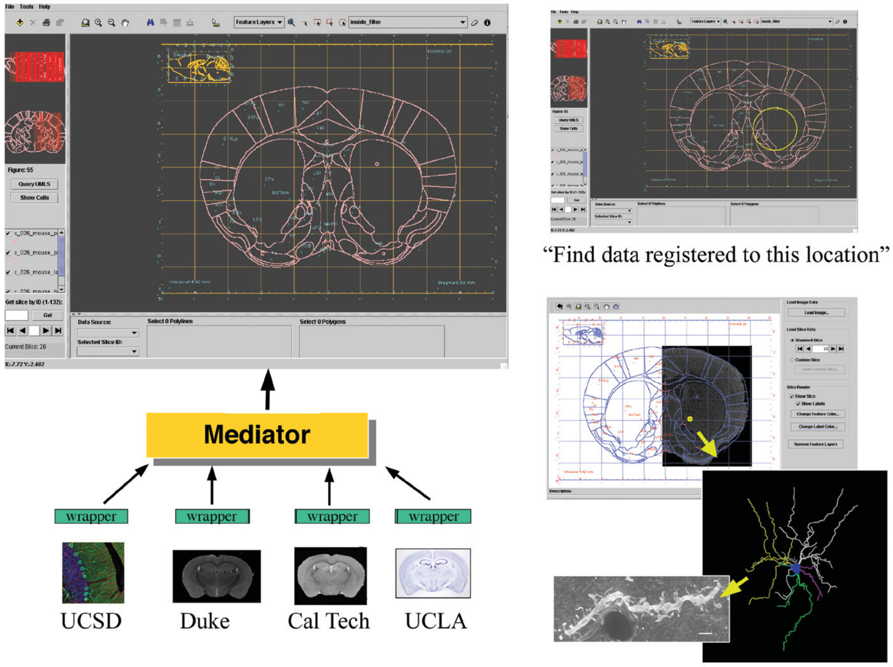


Figure 3 The Smart Atlas tool is being developed as a graphical interface and spatial query tool for distributed, spatially registered multi-scale imaging data in the Mouse BIRN. The Smart Atlas is a java-based GIS tool currently built on top of a commercially available brain atlas[31] but for use with any vector-based atlas[15]. Through the interface, users can query for data registered to a particular location (upper right), and navigate through multiple levels of resolution from the tissue level to cellular and subcellular data (lower right). A model of the dendritic branching pattern of a medium spiny neuron from the mouse caudoputamen is shown in the lower right. To the left of this image is a reconstruction of a portion of spiny dendrite from electron tomography (Scale bar, 1 μm). The Smart Atlas works through the BIRN mediator (lower left) to retrieve spatially registered data from Mouse BIRN participants. Medium spiny neuron and spiny dendrite images courtesy of D. Price and M. Terada from the National Center for Microscopy and Imaging Research.

vidual database efforts. Scientists can design a specialized database encapsulating their particular area of expertise, and maintain control of the primary data, while still making it available to other researchers. Data integration over distributed data sources is a major research topic in computer science[12] and has begun to receive attention in the neuroscience community as well[14]. On the computer science side, researchers are grappling with challenges such as database interoperability across heterogeneous platforms and integration of the complex data types characteristic of scientific data. On the neuroscience side, neuroscientists are working to find ways of providing 'content and context'[14–16] so that neuroscience data can be reliably compared by both man and machine.

One of the well-recognized roadblocks to the creation of shared data resources in the biological sciences concerns reconciling semantic differences between sources[12]. Scientific terminology, even in circumscribed fields like neuroanatomy, is vast, non-standard and confusing[17,18]. Anatomical entities may have multiple names, such as neostriatum and caudoputamen; the same term may have multiple meanings, for example, spine (vertebral spine) versus spine (dendritic spine), and, worse, the same term may be defined differently by different scientists (*e.g.*, basal ganglia). Such semantic ambiguity presents considerable frustration even to experienced neuroscientists; to a machine, it can be all but intractable. One solution is to develop a set of standard terms that neuroscientists agree to use when describing their data ('controlled vocabularies'). Groups are also working to develop so-called 'meta languages' that can be used to describe the content of neuroscience data and databases in a standardized way[19]. However, many neuroscientists instinctively balk at words like "standard" and "controlled". We must also recognize that some of the ambiguity in scientific terminology reflects genuine confusion and disagreement and cannot be easily solved by limiting the terms used[17,20].

One solution explored in the CCDB and BIRN projects and by others[14] is the use of ontologies to provide the necessary knowledge structures for interrelated concepts contained in distributed resources. An ontology is essentially a set of terms and the relationships among them (**Fig. 2**), and provides one means for communities to formalize an understanding of a field[18]. These relationships may be simple "is a" and "structural part of" relationships. For example, Purkinje cell is a neuron, neuron has part nucleus, or may be more complex[21]. The Neuronames project has been building ontologies for neuroanatomy for many years and now includes over 12,000 terms[18]. Neuronames is available as one of the source vocabularies for the Unified Medical Language System (UMLS), a large knowledge source for the biomedical sciences maintained by the National Library of Medicine[22]. The UMLS does not develop the vocabularies, but rather curates and maintains them[18]. Each concept entering the UMLS is mapped to existing concepts and assigned a unique identifier (ID). All synonymous terms are assigned the same ID (*e.g.*, Purkinje cell, cerebellar Purkinje cell and Purkinje's corpuscles share C0034143). Conversely, even if two terms share the same name, they are distinguishable by their unique IDs. In the example given above, spine (vertebral spine) = C0037949, whelreas spine (dendritic spine) = C0872341.

Each CCDB concept is mapped to its corresponding UMLS ID. As part of the BIRN project (see below), we are creating a custom ontology to house terms not contained in UMLS. Any time a term is added, it must be explicitly defined, related to existing terms in the UMLS and assigned a unique ID. In this way, the BIRN community can participate in extending existing ontologies. While mapping to the ontology might be considered the imposition of a controlled vocabulary, the use of numerical identifiers to establish meaning is actually quite flexible. Their use does not require that researchers agree on a term, only that they make their definition explicit and assign the appropriate ID. For example, ontologies could in theory distinguish between two definitions of basal ganglia by researchers A and B by assigning each a unique ID. The use of unique identifiers contained in shared knowledge bases to communicate about database content is a powerful means by which information is made machine-readable without restricting variety of viewpoint. IDs are readily searchable, and any piece of data identified by a UMLS concept ID can be related to any other similarly identified data[14] either through direct equality or through relationships defined in UMLS. The challenges involved in using ontologies for promoting data integration include the need to develop more comprehensive ontologies for most neuroscience domains than are currently available and to explore the extent to which ontologies can be used to represent more complex neuroscience theories with their attendant uncertainties and conflicts[21].

Although language will continue to be a major means by which scientific data are annotated and shared, neuroscientists are increasingly recognizing the power of spatial coordinate systems as a means of communicating about the brain[4,16,23,24]. Computer-based atlases and associated tools for warping and registration are providing the means to express the location of anatomical features or signals in terms of a specified coordinate system. A standardized spatial framework for reporting brain-derived data is a more explicit means for comparing data across different experiments and laboratories[24] and provides the means to build up integrated views of brain features by accumulating data taken at the same location across multiple experiments[23]. Although there may not be consensus among neuroscientists about the identity of a brain area that creates a signal, its location in terms of spatial coordinates is at least quantifiable. The creation of probabilistic atlases provides the means to represent the variability present in individual brains after registration, as well as the confidence limits for anatomical boundaries and signal distributions[4]. The expression of brain data in terms of spatial coordinates allows it to be transformed easily into additional coordinate systems that may provide additional information (for example, cortical flat maps or alternative parcellation schemes[4,25]). Finally, mapping data to a spatial coordinate system allows the user to query brain data on the basis of spatial attributes, for example, "find all anatomical regions within 1 mm of a given signal or activation pattern" (**Fig. 3**).

Several mature neuroinformatics projects are creating atlases for spatially registered data. These projects include the Mouse Atlas Project at the University of California, Los Angeles (www.loni.ucla.edu/MAP) for the C57BL/J6 adult mouse[15], the EMAP and EMAGE projects from the University of Edinburgh (http://genex.hgu.mrc.ac.uk/intro.html) for spatiotemporal mapping of gene expression and structural data in the embryonic mouse and human[20] and the CARET project from David Van Essen's group at Washington University for cortical maps of several species[25] (http://brainmap.wustl.edu/resources). An important result of these projects is the creation of software tools for spatial warping and alignment of image data so that outside researchers can spatially register their own data to these frameworks.

Part of the challenge of using coordinate systems to describe data will be to develop such systems for finer levels of resolution than are represented in a whole brain atlas. For example, Bjaalie and colleagues have introduced a 3D coordinate system for the rat pontine nuclei[23,24] that can be used to compare histological and immunocytochemical data across subjects. As part of the CCDB project, we are developing coordinate systems for individual neuron types, so that subcellular data can be placed in a spatial context relative to the other cellular components.

Data mediation in the BIRN

The BIRN project (www.nbirn.net) was launched by the NIH in September 2001 to build a persistent and robust infrastructure for data sharing and collaboration on a large scale. Neuroimaging was chosen as the pioneering application for BIRN's infrastructure development because of the large size and rich content of 3D imaging data and because a considerable amount of neuroinformatics expertise and tools have been developed through initiatives like the Human Brain Project. While significant technological and sociological hurdles to data sharing and data integration remain, the NIH recognized that they were well-enough understood for biomedical scientists to engage more significantly with those charged with developing the tools to create global scientific communities organized around distributed, shared resources. The BIRN uses the power of so-called 'grids' to provide scientists the means to pool expertise and data to address large-scale problems in neurological disease. The Grid is defined as an infrastructure for the integrated, collaborative use of computational resources, networks, databases and scientific instruments owned and managed by multiple organizations[26]. The tools of the grid movement in computer science (also referred to by the NSF as 'cyberinfrastructure') facilitate the smooth interoperation of these heterogeneous resources through the development of specialized software layers ('middleware') that sit between the resource and the user[26]. Projects like the BIRN and Telescience[27] are developing simple interfaces through which neuroscientists are able to use grid resources or services through high-bandwidth connections established between distributed file systems, processing pipelines and workflows, databases and computational resources.

Three neuroimaging BIRN test projects were established by groups working across 12 universities in the United States: Human Morphometry BIRN, using structural MRI to investigate possible links between depression and Alzheimer disease, Function BIRN, using functional MRI to investigate schizophrenia, and Mouse BIRN, using MRI-based and microscopy-based methods for multi-scale investigations of animal models of human neurological disease. The BIRN Coordinating Center (BIRN-CC) was established to support large-scale collaborations among test sites by coordinating the development and deployment of the grid architecture and providing a framework for integration of data as well as the interoperation of a broad range of existing software tools.

At the heart of the BIRN is the ability to integrate data taken at different sites, whether at the same or different scales, for the purposes of creating larger subject pools (Morph and Function BIRN) and for creating multi-scale views of mouse models of disease (Mouse BIRN). Each of the BIRN participants has an existing database or is creating one to house their contributions to the BIRN project. All participants have agreed to map their data to shared spatial and ontological knowledge sources like the UMLS. The BIRN data integration framework builds upon the NPACI tools for federation of distributed multi-scale neuroscience data and the CCDB project for modeling complex neuroscience data. It uses a mediator as the agent of communication between the user and the distributed databases. The user sends a query to the mediator, which then breaks it up and accesses the appropriate databases to retrieve components of the answer.

The mediator exploits expert knowledge contained in ontologies or spatial coordinate systems as the necessary 'glue' to bridge the diversity of data present in the BIRN databases, a system which we call knowledge-guided or model-based mediation[28–30]. Here, connections between database elements do not have to be direct, but may be inferred through reasoning operations performed on knowledge sources registered to the mediator at time of query (**Fig. 2**). As a simple example, a scientist posing a query to the mediator for information on mouse cerebellum would retrieve gross anatomical information on the cerebellum from a database established at Duke University for MRI brain volumes and information on Purkinje cell structure from the CCDB. The mediator would perform this connection because it accessed an ontology of brain anatomy with the relationships "Cerebellum has a cerebellar cortex; cerebellar cortex has a Purkinje cell layer; Purkinje cell layer has a Purkinje cell". From these relationships, the mediator infers that "Cerebellum has a Purkinje cell" and retrieves relevant cell data from the CCDB. This capability is critical for the type of multi-scale data integration central to the Mouse BIRN project.

Users can also interface with the BIRN data grid through the Smart Atlas, a GIS- based brain atlas query tool and spatial database (**Fig. 3**). In a GIS (Geographical Information System) system, spatial features and associated attributes are stored in a coordinate system, in this case established by brain landmarks in a stereotaxic atlas[31]. The Smart Atlas makes use of this common coordinate system to bring together multi-scale imaging data on the mouse brain stored at each of the BIRN sites. Through the Smart Atlas, researchers can query multi-scale image data based on their location in the brain (**Fig. 3**) and also issue spatial queries about relationships between anatomical features.

Conclusions

The neuroinformatics efforts of the past decade have highlighted both the necessity of better information management and the promise of the new global electronic forum for scientific inquiry and exchange. In recognition of information technology's critical role as neuroscience moves forward, the Society for Neuroscience organized a series of meetings by the Bioinformatics Group (BIG) to examine the information needs of the Society and to consider its role in promoting and developing such resources. As a direct outcome of these meetings, the Society is planning to host a neuroscience database gateway to make its members aware of existing tools and databases and encourage their use.

Sydney Brenner was quoted at the recent conference on Digital Biology: "We now have unprecedented ability to collect data about nature…but there is now a crisis developing in biology, in that completely unstructured information does not enhance understanding" (www.bisti.nih.gov/2003meeting). The power of informatics infrastructure to build integrated models across vast amounts of neuroscience data will require that neuroscientists not only be willing to store data in a database, but also ensure they are properly modeled and mapped to shared knowledge sources like ontologies and spatial coordinate systems. The neuroscience community as a whole will have to commit to supporting and contributing to ontology development and spatial systems, not just for whole-brain anatomy, but also for other disciplines and scales of resolution. The good news is that the approaches to data integration are sufficiently powerful that they respect the flexible and dynamic nature of scientific inquiry. Regardless of whether scientists are engaged in high-throughput, large-scale data acquisition efforts like GENSAT, or whether they are working as individuals, providing the necessary 'hooks' will ensure that their data do not become isolated islands but can be readily integrated into a larger corpus of knowledge by whatever system is engineered to make use of them, now or in the future.

ACKNOWLEDGMENTS
This work was supported by NIH grants from the National Center for Research Resources (RR04050, RR08605) and the Human Brain Project (DA016602), and funded by the National Institute on Drug Abuse, the National Institute of Biomedical Imaging and Bioengineering and the National Institute of Mental Health. Further funding was provided by NSF grants supporting the National Partnership for Advanced Computational Infrastructure (NSF-ASC 97-5249 and MCB-9728338). The BIRN is supported by NIH grants RR08605-08S1 (BIRN-CC) and RR043050-S2 (Mouse BIRN).

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/

1. Kotter, R. Neuroscience databases: tools for exploring brain structure-function relationships. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 1111–1120 (2001).
2. Koslow, S.H. Opinion: sharing primary data: a threat or asset to discovery? *Nat. Rev. Neurosci.* **3**, 311–313 (2002).
3. Toga, A. Neuroimage databases: the good, the bad and the ugly. *Nat. Rev. Neurosci.* **3**, 302–308 (2002).
4. Van Essen, D.C. Windows on the brain: the emerging role of atlases and databases in neuroscience. *Curr. Opin. Neurobiol.* **12**, 574–579 (2002).
5. Bug, W. & Nissanov, J. A guide to building image-centric databases. *Neuroinformatics* **1**, 359–378 (2003).
6. Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925 (2003).
7. Heintz, N. Gene Expression Nervous System Atlas (GENSAT). *Nat. Neurosci.* **7**, 483 (2004).
8. Martone, M.E. *et al.* A cell-centered database for electron tomographic data. *J. Struct. Biol.* **138**, 145–155 (2002).
9. Martone, M.E. *et al.* The cell centered database: a database for multiscale structural and protein localization data from light and electron microscopy. *Neuroinformatics* **1**, 379–395 (2003).
10. Medalia, O. et *al.* Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**, 1209–1213 (2002).
11. Miller, P.L. *et al.* Integration of multidisciplinary sensory data: a pilot model of the human brain project approach. *J. Am. Med. Inform. Assoc.* **8**, 34–48. (2001).
12. Lacroix, Z. Issues to address while designing a biological information system. in *Bioinformatics: Managing Scientific Data* (eds. Lacroix, Z. & Critchlow, T.) 75–108 (Morgan Kaufmann, San Francisco, 2003).
13. Eckman, B. A pracitioner's guide to data management and data integration in bioinformatics. in *Bioinformatics: Managing Scientific Data* (eds. Lacroix, Z. & Chritchlow, T.) 35–73 (Morgan Kaufmann, New York, 2003).
14. Marenco, L. *et al.* Achieving evolvable web-database bioscience applications using the EAV/CR framework: recent advances. *J. Am. Med. Inform. Assoc.* **10**, 444–53 (2003).
15. MacKenzie-Graham, A. *et al.* The informatics of a C57BL/6J mouse brain atlas. *Neuroinformatics* **1**, 397–410 (2003).
16. Fox, P.T. & Lancaster, J.L. Opinion: mapping context and content: the BrainMap model. *Nat. Rev. Neurosci.* **3**, 319–321 (2002).
17. Bota, M., Dong, H.W. & Swanson, L.W. From gene networks to brain networks. *Nat. Neurosci.* **6**, 795–799 (2003).
18. Bowden, D.M. & Dubach, M.F. NeuroNames 2002. *Neuroinformatics* **1**, 43–59 (2002).
19. Gardner, D., Abato, M., Knuth, K.H., DeBellis, R. & Erde, S.M. Dynamic publication model for neurophysiology databases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 1229–1247 (2001).
20. Baldock, R.A. *et al.* EMAP and EMAGE. *Neuroinformatics* **1**, 309–326 (2003).
21. Gupta, A., Ludascher, B., Grethe, J.S. & Martone, M.E. Towards a formalization of disease-specific ontologies for neuroinformatics. *Neural. Net.* **16**, 1277–1292 (2003).
22. Humphreys, B.L., Lindberg, D.A., Schoolman, H.M. & Barnett, G.O. The Unified Medical Language System: an informatics research collaboration. *J. Am. Med. Inform. Assoc.* **5**, 1–11 (1998).
23. Bjaalie, J.G. Opinion: localization in the brain: new solutions emerging. *Nat. Rev. Neurosci.* **3**, 322–325 (2002).
24. Brevik, A., Leergaard, T.B., Svanevik, M. & Bjaalie, J.G. Three-dimensional computerised atlas of the rat brain stem precerebellar system: approaches for mapping, visualization, and comparison of spatial distribution data. *Anat. Embryol. (Berl.)* **204**, 319–332 (2001).
25. Van Essen, D.C. *et al.* An integrated software suite for surface-based analyses of cerebral cortex. *J. Am. Med. Inform. Assoc.* **8**, 443–459 (2001).
26. Foster, I. The grid: a new infrastructure for 21st century science. *Physics Today* **55**, 42–47 (2002).
27. Peltier, S.T. *et al.* The Telescience Portal for advanced tomography applications. *J. Parallel Distrib. Comput.* **63**, 539–550 (2003).
28. Ludascher, B., Gupta, A. & Martone, M.E. A model-based mediator system for scientific data management. in *Bioinformatics: Managing Scientific Data* (eds. Lacroix, Z. & Critchlow, T.) 335–370 (Morgan Kaufmann, San Francisco, 2003).
29. Gupta, A., Ludaescher, B. & Martone, M.E. Knowledge-based integration of neuroscience data sources. in *Proc. 12th Int. Conf. Scientific Statist. Database Management IEEE Comput. Soc.* (2000).
30. Ludaescher, B., Gupta, A. & Martone, M.E. Model-based mediation with domain maps. in *Proc. 17th Int. Conf. Data Eng. IEEE Comput. Soc.* (2001).
31. Paxinos, G. & Franklin, K.B.J. *The Mouse Brain in Stereotaxic Coordinates* (Academic Press, San Diego, 2000).
32. Shepherd, G.M. *et al.* The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci.* **21**, 460–468 (1998).