

Approaches to neuroscience data integration

Kei-Hoi Cheung, Ernest Lim, Matthias Samwald, Huajun Chen, Luis Marenco, Matthew E. Holford, Thomas M. Morse, Pradeep Mutalik, Gordon M. Shepherd and Perry L. Miller

Submitted: 23rd February 2009; Received (in revised form): 8th May 2009

Abstract

As the number of neuroscience databases increases, the need for neuroscience data integration grows. This paper reviews and compares several approaches, including the Neuroscience Database Gateway (NDG), Neuroscience Information Framework (NIF) and Entrez Neuron, which enable neuroscience database annotation and integration. These approaches cover a range of activities spanning from registry, discovery and integration of a wide variety of neuroscience data sources. They also provide different user interfaces for browsing, querying and displaying query results. In Entrez Neuron, for example, four different facets or tree views (neuron, neuronal property, gene and drug) are used to hierarchically organize concepts that can be used for querying a collection of ontologies. The facets are also used to define the structure of the query results.

Keywords: data integration; neuroinformatics; ontology; semantic web; user interface

INTRODUCTION

In an article titled ‘Data-sharing in an information age’ [1], Insel *et al.* [2] made the following prediction: ‘as we emerge from the “decade of the brain”, we are entering a decade for which data-sharing will be the currency for progress in neuroscience’. The growth in the number and diversity of neuroscience databases on the Internet (or the Web) is consistent with this statement. As the number of neuroscience databases continues to grow, the problem of database interoperability is becoming prominent in neuroinformatics [3]. Also, as described in [4], the neuroinformatic ecosystem involves not only data, but also access to and re-use of data.

APPROACHES TO DISCOVERY AND INTEGRATION OF NEUROSCIENCE DATABASES

Data accessibility and re-use cannot be achieved without locating the data sources first. One common method for discovering Web-accessible data sources is to use a Web search engine such as Google to perform a keyword search. This approach often lacks the specificity and sensitivity the user wants. For example, if the user enters the keyword ‘neuron’, numerous hits are returned. It is tedious and time-consuming for the user to sift through a large number of hits to find the relevant ones. Some of these hits are not related to ‘neuron’ in the way expected by neuroscientists. For example, one of these unrelated sites is hosted by

Corresponding author. Kei-Hoi Cheung, Yale Center for Medical Informatics, 300 George Street, Suite 501, New Haven, CT 06511, USA. Tel: 1-203-737-5783; Fax: 1-203-737-5708; E-mail: kei.cheung@yale.edu

Kei-Hoi Cheung is an associate professor at the Center for Medical Informatics, Yale University School of Medicine.

Ernest Lim is a programmer at the Center for Medical Informatics, Yale University School of Medicine.

Matthias Samwald is a postdoctoral researcher at DERI Galway, Ireland and the Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria.

Huajun Chen is an associate professor at the College of Computer Science, Zhejiang University, Hangzhou, China.

Luis Marenco is an assistant professor at the Center for Medical Informatics, Yale University School of Medicine.

Matthew E. Holford is a programmer of the Center for Medical Informatics, Yale University School of Medicine and Department of Biostatistics, Yale University School of Public Health.

Thomas M. Morse is an associate research scientist in the Department of Neurobiology, Yale University School of Medicine.

Pradeep Mutalik is an associate research scientist at the Center for Medical Informatics, Yale University School of Medicine.

Gordon M. Shepherd is a professor in the Department of Neurobiology, Yale University School of Medicine.

Perry L. Miller is a professor and director of the Center for Medical Informatics, Yale University School of Medicine.

a company named 'NEURON' that specializes in manufacturing and selling magnetic and smart card readers and encoders. However, if the user chooses a very specific search term (e.g. 'hippocampal CA1 non-pyramidal neuron'), the search may not return any hits. Part of the problem is that generic Web search engines such as Google do not index the resources according to the needs of domain experts, such as neuroscientists.

NEUROSCIENCE DATABASE GATEWAY

To address the non-specific keyword search problem, the Neuroscience Database Gateway (NDG) (<http://ndg.sfn.org/>) has been created as a resource through the Society for Neuroscience (SfN) (<http://www.sfn.org/>). NDG provides a registry of neuroscience databases annotated with controlled keywords. Nearly 200 databases are currently listed in NDG. These databases span different neuroscience subdomains such as neurophysiology (e.g. SenseLab [5, 6]), neuroanatomy (e.g. BAMS) and neuroimaging (e.g. CCDB [7]). They also cover a wide variety of types of data for different species (e.g. mouse, rat and human). The types of data include images (e.g. CCDB [7]), brain atlases (e.g. Allen Brain Atlas) and neurological diseases (e.g. Alzforum [8]). Each NDG-registered database is annotated with keywords derived from a standard vocabulary/terminology that is curated and approved by a neuroscience user committee. Categories of keywords (e.g. database access, species and clinical conditions) are provided to help the user browse the databases within the registry. In addition, NDG provides a structured search interface for the user to query the registry based on controlled keywords. Figure 1a shows the home page of NDG. The search page can be accessed by clicking the 'Search' button as shown in the figure. Figure 1b shows the search interface and an example search involving selection of two keywords: 'Experimental data' and 'Alzheimer's disease' in the 'DB Category' field and 'Clinical conditions' field, respectively. Each keyword was selected from a popup list of controlled terms (the popup lists are also shown in Figure 1b). The search returns a list of matching databases shown in Figure 1c. Figure 1d shows the detailed description (annotation) of one of the matching databases 'Whole Brain Atlas' that contains normal and abnormal brain images (belonging to the category of experimental data)

including those that are related to the study of Alzheimer's disease. Such a controlled keyword search approach has the potential to allow a more precise and accurate annotation and discovery of neuroscience resources compared to non-specific text-based document indexing and retrieval (e.g. Google).

While NDG is one of the first attempts in terms of enabling neuroscience database discovery by annotating and indexing (at a high abstract level) databases based on controlled keywords, other efforts are underway to make use of standard terminologies and ontologies to support both 'shallow' and 'deep' interoperability among neuroscience resources.

THE NEUROSCIENCE INFORMATION FRAMEWORK (NIF)

One major effort supporting integrative neuroscience research is the Neuroscience Information Framework (NIF) initiative funded by the National Institutes of Health, which is currently being refined by a multi-institutional consortium. The overarching goal of the NIF is to be a one-stop shop for neuroscience. A central element in the NIF is an ontology called 'NIFSTD' (which stands for 'NIF standardized ontology') [9]. This ontology represents an integration/alignment of multiple ontologies (e.g. BirnLex and Gene Ontology) that have been developed for use in various biomedical/neuroscientific domains. It is available in the Web Ontology Language (OWL) format [7], which is a standard ontology format used by the Semantic Web [10]. The OWL format allows machine-based querying and reasoning. The NIFSTD allows the neuroscientist to perform searches based on neuroscience-related concepts and concept relationships. The NIFSTD utilizes the OBO's Relation Ontology [11] for specifying relationships between entities.

The NIF has three main components: NIF resource registry, NIF database mediator and NIF document archive. All are currently still undergoing ongoing development and refinement. The *NIF resource registry* uses the NIFSTD to annotate neuroscience databases and tools in a fashion similar to the NDG. The *NIF database mediator* uses the NIFSTD to facilitate query mapping between multiple databases. The *NIF document archive* uses a text-mining tool called *Textpresso* [12] for storing and accessing the neuroscience literature. Like the

(A) NDG Home Page

NDG
Neuroscience Database Gateway

A Gateway to Neuroscience Resources on the Web

Welcome to the **SIN** Neuroscience Database Gateway!

Databases are of growing importance in neuroscience, as in many other biomedical research fields. The Neuroscience Database Gateway is a new resource for SIN members, aimed at promoting awareness and facilitating access to relevant neuroscience databases.

The SIN Neuroscience Database Gateway provides links to five main types of database. These can be seen by making the appropriate selection in the sidebar or below.

- Databases of experimental data
- Knowledge bases
- Software tools for neuroscience
- Bioinformatics resources
- Providers of research materials
- All neuroscience databases

The gateway also includes basic search capabilities, using the **Linked Fields** option.

The Neuroscience Database Gateway (NDG) began in 2004 as a pilot project developed by the Society's **Brain Information**

(B) Search: Databases

Parameters to retrieve	Select	Conditions
0 Id	<input type="checkbox"/>	
1 Name	<input checked="" type="checkbox"/>	
2 Description	<input type="checkbox"/>	
3 Notes	<input type="checkbox"/>	
4 DB Category	<input type="checkbox"/>	Keywords Experimental data
5 Access	<input type="checkbox"/>	Keywords
6 URL	<input type="checkbox"/>	Keywords
7 Project contact	<input type="checkbox"/>	Keywords
8 DB contact	<input type="checkbox"/>	Keywords
9 Categories	<input type="checkbox"/>	Keywords
10 Institution	<input type="checkbox"/>	Keywords
11 Clinical conditions	<input type="checkbox"/>	Keywords Alzheimer's disease
12 Species	<input type="checkbox"/>	Keywords
13 Supporting Agencies	<input type="checkbox"/>	Keywords
14 Principal Investigator	<input type="checkbox"/>	Keywords
15 Frameworks	<input type="checkbox"/>	Keywords
16 Law status	<input type="checkbox"/>	Keywords

(C) List of Databases

SN Name
1 BIRN
2 IBVD
3 IONI Image Database
4 NeuroImage
5 BCDB
6 Brede
7 MorphEIRN
8 NBD
9 NINDS NIMH Microarray Enterprise Manager
10 Whole Brain Atlas
11 National Brain Databank
12 RGD
13 PDB
14 Human Protein Reference Database
15 GENBANK
16 NIH Neuroscience Microarray Consortium
17 Array Express-European Bioinformatics Institute
18 SMD
19 Alzheimer's Research Center
20 National Institute of Neurological Disorders and Stroke
21 BrainPharm
22 Coriell Cell Repositories
23 NeuroMorpho.Org
24 OASIS
25 GeneNetwork
26 BRAINnet

(D) Annotation of Whole Brain Atlas

Databases	
Name	Whole Brain Atlas
Description	Database of Normal and Abnormal Brain Images
Notes	Database of normal and abnormal brain images.
DB Category	Knowledge show other Tools show other Experimental data show other
Access	Public show other
URL	http://www.med.harvard.edu/AANLIB/home.html
Project contact	Keith A. Johnson (keith@zwh.harvard.edu)
DB contact	John A. Becker (kajohnson@partners.org)
Categories	Anatomy show other Atlas show other
Institution	Massachusetts General Hospital
Clinical conditions	Alzheimer's disease show other Huntington's disease show other Multiple Sclerosis show other Clinical conditions: other show other Parkinson's disease show other Stroke show other Tumor show other
Species	Human show other
Supporting Agencies	Other show other

Figure 1: Screen shots illustrating the steps involved in performing a keyword search through NDG: (A) NDG home page, (B) form-based search interface that includes selection of controlled keywords, (C) list of databases whose annotation contains the selected keywords and (D) annotation of one of the matching databases.

NDG, the NIF resource registry contains information about a wide range of different types of databases and other Web-based resources relevant to neuroscience. These resources are indexed with a high level of abstraction using NIFSTD's ontological concepts. Such high level concepts are used to support 'shallow' interoperability. To search and/or integrate the content of the resources themselves, one has to use the NIF database mediator. The mediator allows automated searching of the contents of a set of mediated databases whose internal vocabularies have been mapped to the general and specific concepts in the NIFSTD ontology.

A prototype Web interface called 'CBQI' (Concept Based Query Interface) [13] (whose components are shown in Figure 2) was developed in an early version of the NIF to demonstrate how the Textpresso and NIFSTD are used to allow: (i) text

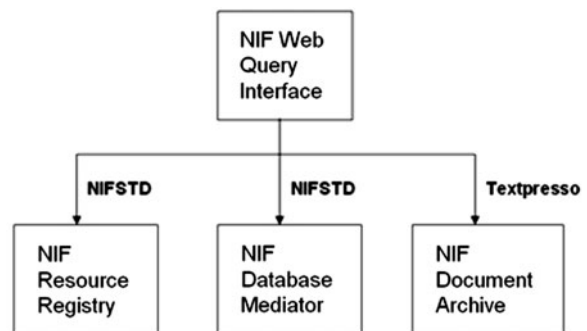


Figure 2: Components of the NIF Web Query Interface.

documents to be retrieved through the Textpresso text-search engine and (ii) concept-based queries to be issued against multiple databases through the NIF resource registry and the NIF database mediator.

In this prototype implementation, a keyword entered by the user is mapped to the corresponding concept in the NIFSTD ontology. Through the NIF database mediator, the NIFSTD concept is mapped to the corresponding items in the underlying databases including SenseLab, CCDB, SumsDB (<http://sumsdb.wustl.edu:8081/sums>) and NeuroMorpho.org (<http://neuromorpho.org/neuroMorpho>). The process of mapping all the relevant terms in a database to the equivalent concepts in NIFSTD is a manual task. As a result, the expansion of the NIF database mediator will be slow compared to the population of the other two NIF resources. Thus the full power of the concept-based approach can only be achieved incrementally over a relatively extended period of time. With these mappings, queries are issued against the local databases. The results returned from the local databases are displayed separately to the user. As described in [13], this query interface was limited

in that concept-to-database mappings were only partially implemented. It also points out that some meaningful display of a subset of the NIFSTD ontology should ideally be implemented to better help the user choose the appropriate concept(s) for querying the databases within a particular context.

Figure 3a shows how a search is formulated via the prototype CBQI search interface (form). The interface has four components, reflecting the four major steps involved in formulating a search. The first step is labeled 'Search for Keywords'. Here the user has entered the text term 'purkinje' for this simple example search. After entering this term, the user clicks on the 'Search for Keywords' button. This results in a search of the NIFSTD ontology for any concepts (keywords) that match the text word 'purkinje'. A list of the concepts found is then displayed in the box labeled 'Select Keywords'. In this case three concepts are displayed. The user may then highlight one or more of

(A) Search Page: The interface is titled 'Integrated Search of the NIF Prototype'. It has four main sections: 1. Search for Keywords (input field with 'purkinje'), 2. Select Keywords (list: Purkinje Cell, Purkinje cell layer of cerebellar cortex, Purkinje neuron), 3. Compose Query (checkboxes for '1. Purkinje neuron', 'AND selected keywords', 'OR selected keywords'), and 4. Retrieve Information (checkboxes for 'Neuroscience Web resources', 'Neuroscience literature (through Textpresso)', 'NIF federated databases').

(B) NIF Database Mediator Form: Shows search results for 'Purkinje neuron'. It lists resources like 'anatomicaldetail obytab', 'neurondb neuronal transmitter', 'neurondb neuronal current', and 'SenseLab_NIF'. Each resource has a 'Retrieve Data' button and a list of selectable fields.

(C) Query Results: Results from table 'neurondb neuronal current' for 'Cerebellar purkinje cell'. The table has columns: Neuron, preferred, Ion channel, Neuronal compartment, and NOTES.

Neuron	preferred	Ion channel	Neuronal compartment	NOTES
Cerebellar purkinje cell	-1	I Potassium	Distal equivalent dendrite	Macropatch clamp and intracellular recordings in guinea pigs suggested that the pattern of Ca2+ spike firing in the dendrites of Purkinje cells is dynamically modulated by a highly aminopyridine-sensitive K+ current, and probably also by a Ca2+-activated potassium current (>294<)
Cerebellar purkinje cell		I Na,t	Distal equivalent dendrite	
Cerebellar purkinje cell	-1	I T low threshold	Soma	Suggested.
Cerebellar purkinje cell	-1	I K,Ca	Soma	(>86<; reviewed in Linas and Walton, 1990).
Cerebellar purkinje cell	-1	I L high threshold	Proximal equivalent dendrite	(>36<; reviewed in Linas and Walton, 1990).
Cerebellar			Distal	Intradendritic recordings show

Figure 3: The CBQI Web interface: (A) search page, (B) NIF Database Mediator form with the option of selection query output fields from each database, (C) query results from one of the database tables.

those concepts and click ‘Select’. The selected keywords are then copied into the ‘Compose Query’ box. Figure 3b shows how search results are displayed for the NIF Database Mediator. This screen displays different databases that contain potentially relevant data and allows the user to launch a search directly into any one of those databases to retrieve that data. From left to right, we see the names of (i) the database, (ii) a table in that database and (iii) fields within that table which may contain relevant information. Each table may have up to two buttons, one (a ‘Web link out’ button labeled with the name of the database) that links to a specific page for the search concept (in this case ‘Purkinje neuron’) in the resource, and another (‘Retrieve Data’) that retrieves information directly from the resource’s back-end database. Note that the search term ‘Purkinje neuron’ has been translated to its corresponding term in each database: e.g. Purkinje neuron (in CCDB—Cell Centered Database ([7]), and Cerebellar purkinje cell (in SenseLab [5, 6]). Database term translations are performed via the NIF Mediator using mappings between those terms and concepts in the NIFSTD ontology. For each database, the user is given the option of indicating (via checkboxes) which data fields he would like retrieved from each database (by default all fields are selected). For example, for the NeuronDB neuronal current table, if the user clicks on the ‘Retrieve Data’ button he is taken to a new (pop-up) screen (Figure 3c) containing data about neuronal currents that have been identified in various compartments of the purkinje cell. The advantage of this query output format is that the data can be inspected in a generic tabular format, and could for example be copied and pasted into a spreadsheet (or into a local database) for integrated analysis with data from other sources.

ENTREZ NEURON: A CASE STUDY FOR SEMANTIC WEB INTEGRATION AND VISUALIZATION

A further search enhancement is illustrated by Entrez Neuron. This is a prototype Web application (<http://neuroweb3.med.yale.edu:8080>) developed by our group, which provides an intuitive interface for the user to issue concept-based queries against multiple neuroscience ontologies. *Entrez*

Neuron differs from *CBQI* in a number of ways including the following.

- (i) While *CBQI* features query mediation using the NIF database mediator, *Entrez Neuron* represents a warehouse approach in which multiple databases are converted into OWL ontologies that are loaded into a single triplestore (Oracle). Concept-based queries are issued against a single OWL repository without query mapping. There are tradeoffs between the centralized (warehouse) and the decentralized (mediated) querying approaches. The former generally tends to give better query performance as integrated queries are executed locally, while the latter ensures data currency as the latest version of the datasets and their links can be accessed at query runtime.
- (ii) While *CBQI* allows the user to issue a query based on a single concept or a Boolean combination of concepts, *Entrez Neuron* allows the user to express a query that retrieves data based on relationships among multiple concepts.
- (iii) While *CBQI* hides the ontology structure of NIFSTD from the user, *Entrez Neuron* provides visualization of different views of the ontology collection to help guide the user to formulate concept-based queries.
- (iv) *Entrez Neuron* currently does not support literature search.

Derived loosely from the approach of *Entrez Gene* (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>), the prototype *Entrez Neuron* system supports neuron-centric data integration. Its target users are neuroscientists spanning multiple subdisciplines including neurophysiology, neuroanatomy, neuro-molecular biology, neuroimaging, neuropharmacology and neurobiochemistry. It provides an interactive Web interface that allows neuroscientists to retrieve different types of information about neurons across multiple ontologies implemented using OWL and stored using Oracle 11g that supports SPARQL and native OWL inferencing. These ontologies constitute the SenseLab ontology collection that includes conversion of several SenseLab databases into OWL ontologies (<http://www.w3.org/TR/hcls-senselab/>) and their links to other biomedical ontologies. While genes are one of the fundamental units that biologists use in their research studies, neurons are one of the most

important elements that neurobiologists explore. *Entrez Neuron* provides an interface that allows queries to be issued based on concepts displayed in different hierarchically structured views (or facets) derived from the ontologies. There are currently four facets involving the following concepts: *neuron*, *neuronal property*, *gene* and *drug*. These concepts and their relationships are depicted in Figure 4. As shown in the figure, a *neuronal property* is located in neuronal compartment of a *neuron*; a *gene* encodes a molecular component that enables a *neuronal property*; and a *drug* binds to a *receptor*.

Each of the facets and its associated data ontologies is described below.

- *Neuron facet*. This facet uses the hierarchically-related brain regions to organize neurons. The hierarchy of brain regions is obtained by merging the hierarchies of brain regions from NeuronDB (<http://senselab.med.yale.edu/neurondb/>) and the Subcellular Anatomy Ontology (SAO) (<http://ccdb.ucsd.edu/CCDBWebSite/sao.html>) that is used for image annotation at the subcellular level. The information about neurons includes fluorescent-labeled images of neurons obtained from SAO, neuronal data (what types of neuronal properties are located in what neuronal compartments) from NeuronDB, model descriptions (e.g. what types of neuronal properties are involved in the model) from ModelDB (<http://senselab.med.yale.edu/modeldb/>), and drug-receptor bindings from the Psychoactive Drug Screening Program (PDSP) Ki database (<http://pdsp.med.unc.edu/indexR.html>).
- *Neuronal properties facet*. This facet allows the user to browse different types of neuronal membrane properties including receptors, neurotransmitters and ionic currents. These properties and their hierarchical organization are obtained from

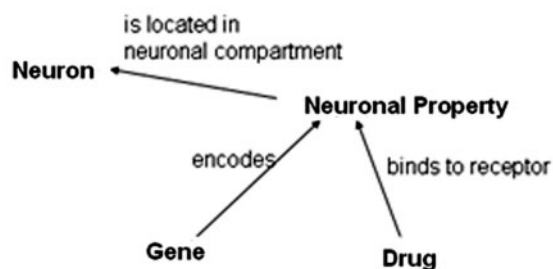


Figure 4: Relationships between different entity types.

SenseLab's CellPropDB (<http://senselab.med.yale.edu/CellPropDB/CellCxSearch.asp>). The information related to these neuronal properties includes descriptions of models that include the neuronal properties, drug-receptor bindings and neuron compartment localization information, which are obtained from ModelDB, PDSP and NeuronDB, respectively.

- *Drug facet*. This facet allows the user to browse drugs based on different classes of drug actions. Both the drugs and the drug action classes are obtained from a subset of the Chemical Entities of Biological Interest (CheBI) [14]. The information about the drug includes drug-receptor bindings from PDSP and neuronal information from NeuronDB.
- *Gene facet*. This facet allows the user to browse gene-related information (e.g. gene products) based on functional categories that are derived from the molecular functions defined in the Gene Ontology. The gene-related information includes neuronal data obtained from NeuronDB.

These facets allow the user to browse different types of concepts organized in ways that are familiar to neuroscientists. They also represent different levels of granularity. Brain regions constitute the top layer. The next level consists of different types of neurons located in different brain regions. Below the level of neurons, there are cell membrane properties including neurotransmitters, receptors, and channels located in different neuronal compartments. Finally, there is the molecular level involving molecular entities like genes and drug molecules that interact with neuronal cell membrane properties. Such molecular interactions may relate to synaptic activities such as neurotransmission. Such multi-level organization not only allows the user to access information at any given level, but it also permits a drill-down approach to accessing different layers of information systematically.

These four facets represent those presently being explored on a prototype basis. Further facets can be introduced as the granularity of user demand requires. The Neuron facet for example might be enhanced to include the interregional circuits that form the major pathways and systems of the brain, and the microcircuits that carry out the processing within each region. The Neuronal Properties facet could include the levels of integrative activity that occur within the dendritic trees and dendritic

branches of a neuron. Both of these facets could retrieve information on physiological recordings and neural correlates of behavior. Current steps in those directions are illustrated in NeuronDB, which provides the identification of neuron properties in relation to canonical compartments of the neuron, and in ModelDB, which provides generation of the physiological recordings by simulations based on those properties. Entrez Neuron can provide a framework for incorporating these and other levels of organization and function into hierarchies,

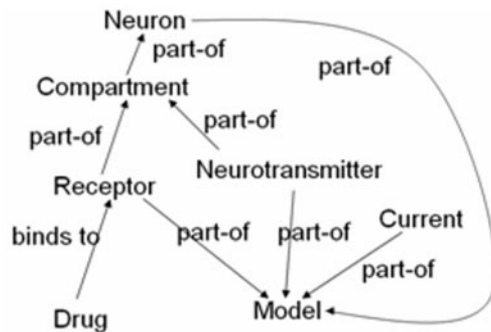


Figure 5: Combined ontological components of NeuronDB, ModelDB and PDSP.

where desired by the corresponding research community.

The facets not only facilitate concept browsing, but they also allow related types of information to be retrieved from different sources. Figure 5 shows the types of information retrieved from NeuronDB, ModelDB and PDSP ontologies. The network graph shown in Figure 5 represents a semantic neighborhood view of the following concepts: *Neurons*, *Neuronal Property* and *Drug*. It consists of the following types of information (with some overlap) obtained from NeuronDB, ModelDB and PDSP:

- (i) NeuronDB: *Neuron*, *Compartment*, *Receptor*, *Neurotransmitter* and *Current*.
- (ii) ModelDB: *Model*, *Receptor*, *Neurotransmitter* and *Current*.
- (iii) PDSP: *Drug* and *Receptor*.

The Web interface of *Entrez Neuron* was implemented using Ajax (<http://en.wikipedia.org/wiki/AJAX>) to provide responsiveness and interactivity. Figure 6 shows the Web-based query interface of Entrez Neuron. The interface consists of three panels: (i) search panel, (ii) facet tree panel and

Figure 6: Entrez Neuron query interface.

(iii) query results panel. As shown in the figure, the user has chosen three different terms (CA1 pyramidal neuron, GABA-A receptor and diazepam) from their corresponding facets (neuron facet, neuronal property facet and drug facet). For example, the user clicks on the browse button next to the Drug search box for browsing the drug facet tree and selecting a search term from the tree. In this case, the user expanded the tree and selected the term 'diazepam'. The selected term is automatically entered in the Drug search box. Then the user presses the 'GO' button in the search panel to issue the query. The query results are displayed in the query results panel. As shown in the query results, there are two data sources (NeuronDB and PDSP) containing information about the chosen drug, neuron and neuronal property. The query results returned from each data source can be seen by clicking the corresponding tab. The query results are scrollable and presented in text format that is readable to neuroscientists (e.g. 'Diazepam binds to GABA-A . . .' for the NeuronDB tab and 'CA1 pyramidal neuron with GABA-A receptor in Dam' for the PDSP tab). The text output corresponds naturally to the semantic web structure (e.g. *subject, property, object*) shown in Figure 5.

The query interface supports two search types: *exact search* versus *fuzzy search*. While the former only allows exact matches of the search term, the latter allows the search to be broadened by including matches of the terms that have the same parent as the search term. Such a fuzzy search employs inferencing based on the parent-child relationship. For the example query shown in Figure 6, if the fuzzy search type were chosen, the query results will include results from matches of 'CA3 pyramidal neuron' that has the same parent ('Hippocampus') as 'CA1 pyramidal neuron'. The fuzzy search results are displayed in separate tabs in the query results panel.

SUMMARY

We have reviewed several approaches to the problem of registering, discovering and integrating neuroscience databases. Domain independent and text-based search engines such as Google are inadequate in terms of meeting the specific but diverse needs posed by neuroscientists. More focused search strategies are needed. NDG and NIF are representative approaches to implementing such search

strategies. For ontologically-based querying and integration, CBQI and Entrez Neuron give demonstration of how this can be achieved. They also highlight the importance of an intuitive interface in enabling the neuroscientist to issue complex integrated queries without a steep learning curve.

FUTURE DIRECTIONS

We have identified several future directions for Entrez Neuron.

- We will incorporate additional data sources into Entrez Neuron to cover a broader spectrum of neuroscience research needs. For example, there are a significant number of neuroscience resources (e.g. databases) that have been registered through NIF (more resources will be registered in the future). These resources provide diverse types of neuroscience information ranging from molecular data (e.g. genes and pathways), imaging data (MRI brain images) to disease related data (e.g. Alzheimer's and Parkinson's). In addition to the NIF registry, there are semantic web portals that have converted and integrated different biomedical databases into the semantic web format, including the BioPortal [15], Neurocommons [16], SWAN [17] and the HCLS KB (<http://www.w3.org/TR/hcls-kb/>). Incorporating new data sources may call for new facets (trees). The identification of new facets could be facilitated by discovering the hierarchical relationships encoded using the Semantic Web.
- While the Semantic Web offers a global identifier scheme (URI), there has been a proliferation of synonymous URI's, which hinders data integration. To address this, efforts such as the Shared Names initiative (http://sharedname.org/page/Main_Page) have begun to normalize URI's in the biomedical context. In addition, NIF has created NeuroLex (http://neurolex.org/wiki/Main_Page) for providing a dynamic and standard lexicon of neuroscience concepts to improve the way that neuroscientists communicate about their data. It will be beneficial for Entrez Neuron to be synergistic with such community efforts.
- While the current version of Entrez Neuron uses a centralized approach (data warehouse) to support data integration, we will explore the use of the Semantic Web in a query federation scenario

where semantic web data are queried at local sites by query mediators. To facilitate query mediation, we will explore how to describe semantic web data resources in such a way that these resources can be discovered by the semantic web-based mediator automatically.

Key Points

- Several approaches to neuroscience data integration are discussed.
- An informatics infrastructure is needed for neuroscience data representation, annotation and integration.
- Informaticians and neuroscientists need to work together in order to come up with the appropriate solution to the neuroscience data integration problem.

Acknowledgements

The authors thank Maryann Martone and Willy WaiHo Wong for providing the OWL version of the Subcellular Anatomy Ontology (SAO) and assisting with their integration into Entrez Neuron. They also thank Melliyal Annamalai and Alan Wu from Oracle Corporation for their technical assistance.

FUNDING

National Institutes of Health grants P01 DC04732, R01 DA021253 and U24 NS051869, The Fidelity Foundation.

References

1. Stark C, Breitkreutz B-J, Reguly T, *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:535–9.
2. Insel TR, Volkow ND, Li TK, *et al.* Data-sharing in an information age. *Neuroscience Networks* 2003;**1**:e17.
3. Marenco L, Nadkarni P, Martone M, *et al.* Interoperability across neuroscience databases. *Methods Mol Biol* 2007;**401**: 23–36.
4. Gardner D, Akil H, Ascoli GA, *et al.* The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 2008;**6**:149–60.
5. Crasto CJ, Marenco LN, Liu N, *et al.* SenseLab: new developments in disseminating neuroscience information. *Brief Bioinform* 2007;**8**:150–62.
6. Marenco L, Tosches N, Crasto C, *et al.* Achieving evolvable web-database bioscience applications using the EAV/CR framework: recent advances. *J Am Med Inform Assoc* 2003;**10**:444–53.
7. Martone ME, Tran J, Wong WW, *et al.* The cell centered database project: an update on building community resources for managing and sharing 3D imaging data. *J Struct Biol* 2007;**161**:220–31.
8. Clark T, Kinoshita J. Alzforum and SWAN: the present and future of scientific web communities. *Brief Bioinform* 2007;**8**:163–71.
9. Bug WJ, Ascoli GA, Grethe JS, *et al.* The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* 2008;**6**:175–94.
10. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Scientific American* 2001;**284**:34–43.
11. Smith B, Ashburner M, Rosse C, *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–5.
12. Müller H-M, Rangarajan A, Teal TK, *et al.* Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics* 2008;**6**: 195–204.
13. Marenco L, Li Y, Martone ME, *et al.* Issues in the design of a pilot concept-based query interface for the neuroinformatics information framework. *Neuroinformatics* 2008;**6**:229–39.
14. Degtyarenko K, Matos Pd, Ennis M, *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;**36**:D344–50.
15. Musen M, Shah N, Noy N, *et al.* BioPortal: ontologies and data resources with the click of a mouse. *AIMIA Annu Symp Proc*. Philadelphia: Hanley & Belfus, Inc, 2008; 1223–4.
16. Ruttenberg A, Rees JA, Samwald M, *et al.* Life sciences on the Semantic Web: the Neurocommons and beyond. *Brief Bioinform* 2009;**10**:193–204.
17. Ciccarese P, Wu E, Wong G, *et al.* The SWAN biomedical discourse ontology. *J Biomed Inform* 2008;**41**:739–51.