Formal Ontology in Information Systems
M. Donnelly and G. Guizzardi (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-084-0-15

Maturation of Neuroscience Information Framework: An Ontology Driven Information System for Neuroscience

Fahim T. IMAM, Stephen LARSON, Anita BANDROWSKI, Jeffrey S. Grethe, Amarnath GUPTA, and Maryann E. MARTONE
University of California, San Diego, United States

Abstract. The numbers of available neuroscience resources (databases, tools, materials and networks) on the web have, and continue to expand; particularly in light of newly implemented data sharing policies required by funding agencies and journals. However, the nature of dense, multi-faceted neuroscience data and the design of classic search engine systems makes efficient, reliable, and relevant discovery of such resources a significant challenge. This challenge is especially pertinent for online databases, whose dynamic content is largely opaque to contemporary search engines. The Neuroscience Information Framework[1] (NIF) was initiated to address this problem of finding and utilizing neuroscience-relevant resources. The NIF provides simultaneous, concept-based search across multiple data sources allowing neuroscientists to connect with available resources, including the deep content of experimental data in online databases. Searching the NIF portal is semantically enhanced through the utilization of a comprehensive ontology, the Neuroscience Information Framework Standard (NIFSTD), developed internally and with community involvement at NeuroLex.org. The NIFSTD also provides the foundation for a standard semantic resources description framework to facilitate navigation across resources, as well as integration and interoperability of available neuroscience data. Since the first production release in 2008, NIF has grown significantly in content and functionality, particularly with respect to the ontologies and ontology-services driving the system. This paper presents NIF as a comprehensive information system that is decisively driven by its ontologies.

Keywords. Ontology, semantic search, neuroscience ontologies

Introduction

An initiative of the NIH Blueprint for Neuroscience Research, the Neuroscience Information Framework[1] (NIF) project is targeted towards advancing Neuroscience by enabling discovery and access to public research data and tools worldwide through an open source, semantically enhanced networked environment. The ultimate end product of NIF is a semantic search engine and knowledge discovery portal that provides federated access to a vast amount of Neuroscience data and resources over the web. Now entering its fourth year of production-release, NIF has matured into the largest source of neuroscience-relevant information on the web.

[1] The Neuroscience Information Framework (NIF), http://neuinfo.org

One of the critical components for the overall NIF system, NIF Standardized Ontologies (NIFSTD) [1] provides a comprehensive collection of Neuroscience relevant concepts (60,000+) along with their synonyms and conceptual relationships. NIFSTD covers major domains in neuroscience, including diseases, brain anatomy, cell types, subcellular anatomy, small molecules, techniques and resources descriptors. The conceptual knowledge models defined within the NIFSTD ontologies are materialized through an ontological query processing engine called OntoQuest [2, 6], which enables an effective concept-based search over heterogeneous types of web-accessible information entities for the NIF's production system.

Unlike traditional, generic search engines like Google, NIF provides a powerful domain-specific query processing mechanisms that allow the query strings to be searched against their ontological semantics rather than their exact lexical matches. In this paper we present the integral components of the NIF system, the NIFSTD ontologies, the NeuroLex Wiki, and the OntoQuest system, and how they are utilized to enhance the overall NIF platform.

1. NIF System

The core objective of NIF was to address the problem of finding neuroscience-relevant resources. NIF provides simultaneous search across multiple information sources to connect neuroscientists to available resources. These sources include: (1) NIF Registry: A human-curated registry of neuroscience-relevant resources; (2) NIF Literature: A collection of neuroscience relevant corpora; (3) NIF Database Federation: A federation of independent databases registered to the NIF, allowing for direct search and discovery of database content, often referred to as the “hidden web”. Semantic search through the NIF portal is enhanced through the utilization of NIFSTD. OntoQuest enhances the search by providing an ontology-based query formulation, source selection, term expansion, and finally better ranking on the search results.

1.1. NIF Data Federation

Although the trend of representing web resources through RDF like resource description formalisms are increasingly practiced to achieve interoperability, the vast majority of resources are still residing in relational databases, or simply in html documents. Through the NIF data federation, NIF have established strategies that would address the most common types of resources with biomedical interests. Thus, for this phase of the NIF, we elected to focus our data federation efforts on these types. NIF queries get translated to queries in the native forms of the databases we federate. NIF contents are also transformed in RDF and have its SPARQL endpoint. NIF will soon implement its RDF query interface for its federated resources. For a portal such as NIF, the challenge is to provide a simplified view of a complex resource that is understandable to a user coming through the NIF portal. To accomplish this, the NIF curators work with the native resource to ensure that the local semantics of the data are expressed correctly through NIF by augmenting the model where necessary and mapping the data to NIFSTD ontologies.

Before the current era of Semantic web, the common practices for the vast majority of web-based resource providers simply overlooked the idea of structuring their data models that could be machine processable, reusable and interoperable. NIF curators are

taking up the challenge of filtering these resources so that they could become available under the hood of a common, interoperable semantic layer. The NIF data federation currently contains one of the largest unified collections of neuroscience-relevant data available through the web, with over 50 independent data sources integrated through the NIF indices. NIF is closely following the movements such as Open Data, Linked Data, and the Web of Data that could integrate data regardless of their sources.

1.2. NIF Resource Registry

To aid the neuroscience community to discover useful digital resources, such as academic databases, software, funding etc. NIF has developed a large digital catalog of resources related to neuroscience. NIF has developed a simple OWL ontology module dedicated to cataloging digital resources called the resource ontology in collaboration with the BRO (biomedical resource ontology). So far NIF curators have assigned a vast amount of digital resources to one or more of the Resource ontology categories. The resource category classes are mostly derived from NIF-Investigation and the Ontology for Biomedical Investigation (OBI) that can serve as 'resource descriptors'. The top level resource categories are: Data, Funding, Job, Material, People, Services, Software, and Training. These categories comprise with human readable definitions and labels targeted for curators , along with sub-categorizations and synonyms to assist the search systems to locate the right data.

1.3. Growth of NIF Contents and Outreach

Since the first release in 2008, NIF has grown significantly in content and community building. Currently, NIF provides access to the largest collection of neuroscience relevant data on the web, all from a single interface. The chart on the left in Figure 1 illustrates the growth of federated data resources within NIF since June, 2008. The chart on the right illustrates the utilization growth in visits per month across NIF holdings, including NIF search portal and NeuroLex. Currently NIF search portal has ~6000 visits per month, and NeuroLex has over 15,000 visits per month. Also, it should be noted that a significant number of NIF users are successfully finding their desired keywords from the NIF ontologies. For example, based on the recent google analytics report (March 19-25, 2012), out of total 1,823 search events, 846 were auto complete search (i.e., terms exist in NIFSTD), and 85 of them were advanced ontological queries.

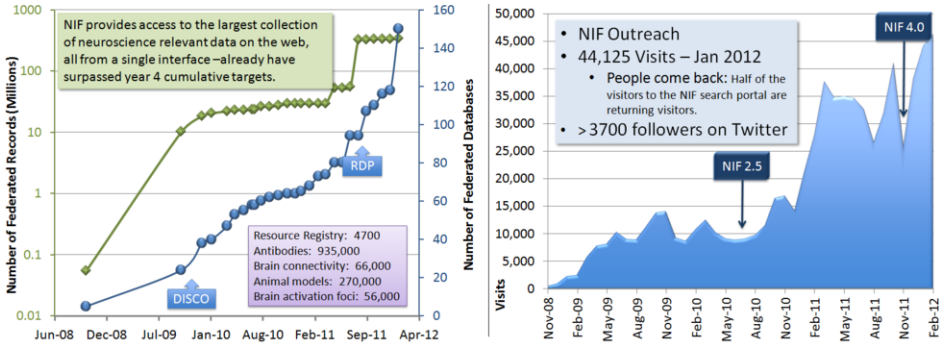

Figure 1. On the left, the increase of NIF contents in terms of the number of federated records and databases. On the right, the increase of community outreach in terms of the number of visitors to the NIF portal.

2. The NIFSTD Ontologies

NIFSTD is a set of modular ontologies where each module covers a distinct orthogonal domain of Neuroscience [1] (Figure 2). The modules in NIFSTD are expressed in OWL-DL [11] which affirms the balance between expressivity and computations decidability, and supports automated reasoning via common DL reasoners e.g., Pellet, FACT++. Acknowledging the fact that information is most often incomplete, the open-world assumption in OWL has been a suitable approach to represent the Neuroscience domain in NIF which also allows its ontologies to be deployed incrementally. NIFSTD is loadable through Protégé [13] ontology editor and has a SPARQL endpoint [15]. It is also available on the web through NCBO BioPortal [16].

Wherever possible, NIFSTD reuses terms and their classification schemes from existing, standard Biomedical knowledge sources. These sources include fully structured ontologies to loosely structured controlled vocabularies or lexicons following standard nomenclatures. Depending on the state of the sources, they are either adapted into an OWL ontology, extracted a relevant portion using MIREOT principles [8], or simply imported as whole. Domains covered by the current NIFSTD along with the vocabularies imported from the external, community sources and the corresponding OWL modules can be found in [5]. Each class in NIFSTD is assigned with a unique identifier accompanied with a variety of annotation properties. These annotation properties were mostly drawn from SKOS and Dublin Core Metadata model.

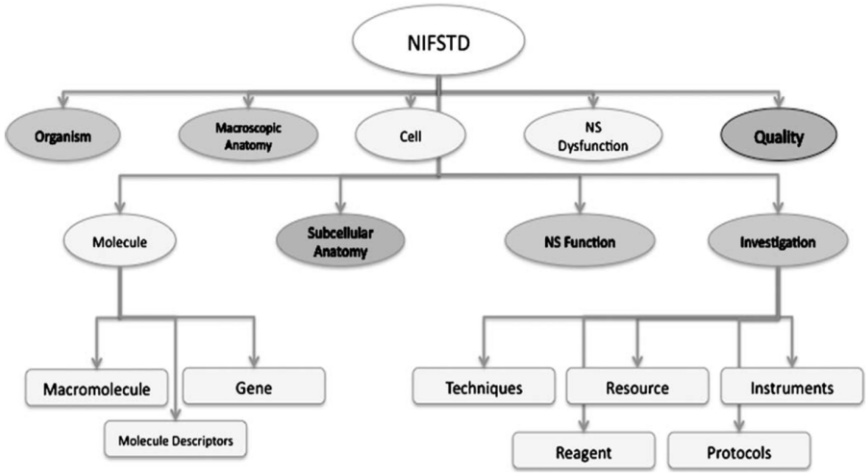

Figure 2. The semantic domains covered in the NIFSTD ontology. Separate OWL modules cover the domains specified within the ovals. The umbrella file http://purl.org/nif/ontology/nif.owl imports each of these modules when opened in Protégé. Each of the modules, in turn, may cover multiple sub-domains, some of which are shown in the rectangular boxes [1].

As one of the core principles, NIFSTD included the Basic Formal Ontology (BFO) [18], the most widely used upper ontology in biomedical communities, to represent the upper level semantic layer of its different orthogonal modules. The upper level classes in NIFSTD modules are normalized under the appropriate BFO entities. Since the beginning, NIF always recognized the significance of a formal ontology. As the number of biomedical ontologies increases, a formal ontology like BFO plays an important role to promote semantic interoperability for data integration [4]. BFO provides a logical basis to categorize the domain-independent high level classes, such

as the entities, characteristics and processes. This kind of categorization helped us to avoid erroneous and ambiguous ontological assertions, and was necessary to develop a large-scale ontology like NIFSTD.

NIFSTD closely follows the OBO Foundry best practices [4] as long as they are practical for day to day development. NIFSTD includes PATO (obofoundry.org/wiki/index.php/PATO:Main_Page) to describe the phenotypic qualities of the NIFSTD entities. The relational properties in NIFSTD (Example: Figure 3) are derived from standard OBO-Relations Ontology (RO). The relations that are domain-independent and exist as universally true within the classes of a specific module, those relations are kept integrated together within the same module. The relations between entities that could vary based on specific application, or require domain-dependent viewpoint, those relations are kept in a separate module called a *bridging module*. A bridging module would typically incorporate relational properties between multiple distinct modules. This kind of isolation enables NIFSTD to keep its modularity principles intact. The core modules are hence easily extendible and re-usable without the need for any modification. New bridging modules can be developed should a user desire a customized application of their own domain.

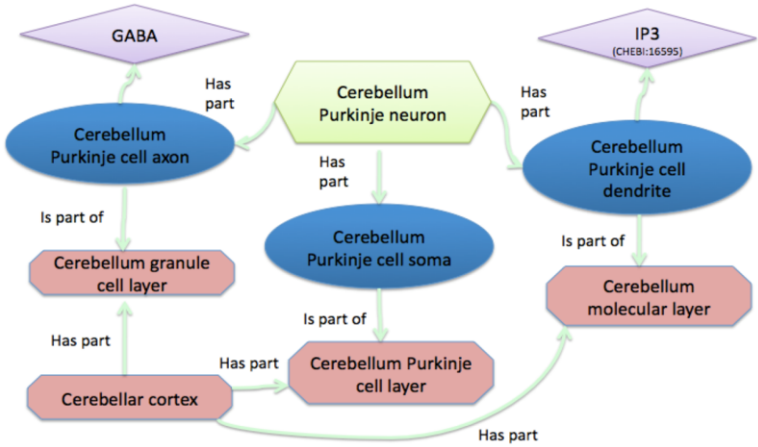

Figure 3. An exemplary knowledge model in NIFSTD. Both cross-modular and intra-modular classes are associated through object properties drawn from the OBO Relations ontology (RO). Different color/ shape of the boxes indicate that the classes belong to different modules. The cross-domain relations are typically kept in a separate bridging module in NIFSTD.

NIFSTD follows the simple inheritance principle [3] for its hierarchy of named classes; i.e., an asserted named class can have only one named class as its super-class; however, a named class can have multiple anonymous super-classes. The classes with multiple super-class are derivable via automated classification on defined NIFSTD classes with necessary and sufficient conditions. This approach saves a great deal of manual labor and minimizes human errors inherent in maintaining multiple hierarchies. Also, this approach provides logical and intuitive reason as to how a class may exist in multiple, different hierarchies.

Since the first release, the NIFSTD ontologies have undergone extensive revision and refinements over the course of its evolution during the last couple of years. These include simplified structural changes to its import hierarchies, simplifying the 'back-end' modules that comprises the common entities shared by all of the NIFSTD modules,

enforced modularization principles, and refactoring the modules under more appropriate BFO classes. Also, NIFSTD included new modules extracted from standard Biomedical ontologies such as the Gene Ontology (GO), Protein Ontology (PRO), ChEBI, and Human Disease Ontology (DOID). NIFSTD core contents have also been rapidly enhanced from NeuroLex contributions.

3. The NeuroLex Wiki

One of the largest roadblocks that NIF encountered during its ontology development phase was the lack of tools for domain experts to view, edit and contribute their knowledge to formal ontologies like NIFSTD. Existing ontology tools were difficult to use or required expert knowledge to employ. NIF strives to balance the involvement of the neuroscience community for domain expertise and the knowledge engineering community for ontology expertise when constructing its ontologies. By combining several open source technologies related to semantic media wikis, NIF created NeuroLex[2], a semantic wiki for the neuroscience community and domain experts.

The initial contents of the NeuroLex were derived from NIFSTD which established its neuroscience centric semantic framework and enabled the semantic relationships among its category pages. NIFSTD OWL classes are were automatically transformed into category pages containing a simplified, human readable class descriptions. The category pages are editable and readily available to access, annotate or enhance by the community or domain experts. Additions of new categories and enhancements to the NeuroLex contents are regularly transformed into NIFTSD in formal OWL-DL expressions. While the properties in NeuroLex are meant for easier interpretation, the restrictions in NIFSTD are more rigorous and based on standard OBO-RO relations. For example, the property 'soma located in' is translated as `'Neuron X' has_part some ('Soma' and (part_of some 'Brain region Y'))` in NIFSTD. Sometimes a similar kind of 'macro' relation such as 'has_neurotransmitter' are used in NIFSTD, recognizing that these relations can be specified more rigorously. These 'macro' relations readily lend themselves into more rigorous representations using OWL 2.0 [12] property chain composition, should they become necessary at a later date.

NIF considers NeuroLex.org as the main entry point for the broader community to access, annotate, edit and enhance the core NIFSTD content. The peer-reviewed contributions in the media wiki are later implanted in formal OWL modules. It should be noted that NIF is not charged with development of new modules but relies on community for new contents. Therefore, the NeuroLex wiki has proven to be ideal for NIF's current scope. For example, it has proven to be effective and helpful in the area of neuronal cell types where NIF is working with a group of neuroscientists to create a comprehensive list of neurons and their properties. NeuroLex category pages are linked with NIF Search interface where users can quickly view a descriptive ontological details about a search term.

NeuroLex can be viewed as a full-fledged information management system that provides a bottom-up ontology development approach where multiple participants can edit the ontology instantly. Semantics in NeuroLex are limited to what is convenient for the domain experts. Essentially, the NeuroLex approach is not a replacement for top-down construction, but critical to increase accessibility for non-ontologist domain

[2] The NeuoroLex Wiki, http://neurolex.org

experts. NeuroLex provides various simple forms for structured knowledge where communities can contribute and verify their knowledge with ease. It also allows the generation of specific class hierarchies, or extraction of a specific portion of the ontology contents based on certain properties in a spreadsheet, without having to learn complicated ontology tools.

4. The OntoQuest System

Within NIF, NIFSTD is served through a powerful ontology management system, called OntoQuest. OntoQuest views the ontology as a graph and performs graph-like operations (e.g., finding the k-neighborhood). The NIFSTD is an OWL/RDF graph currently containing 60,000+ terms and about 25 edge labels. In OntoQuest, the subclassOf property induces a spanning DAG over all ontology nodes. Periodically, the major release of NIFSTD gets loaded in OntoQuest repository. However, we have recently established an incremental update mechanism to allow day-to-day updates of NIFSTD to be reflected more frequently in OntoQuest.

As the ontologies used in Neuroscience and relevant biomedical domain are usually massive in size, it was necessary for OntoQuest to utilize database technologies in order to provide a scalable query processing mechanisms. OntoQuest provides efficient retrieval of knowledge and information that are semantically structured through ontologies.

Currently, OntoQuest can process full-fledged ontologies represented in OWL or OBO formalisms as well as simpler taxonomies expressed in RDFS. OntoQuest is built on top of a relational schema motivated by the IODT system from IBM [17] that shreds OWL and OBO ontologies into this relational schema. It is important to note that OntoQuest is not a reasoner; rather, it provides an efficient navigation and query facility on ontologies by treating them as directed graphs. Details on OntoQuest graph model can be found in [7].

OntoQuest utilizes Jena library APIs to parse OWL ontologies. However, OntoQuest has its own customized algorithms for generating the mappings between the OWL ontologies and the backend schemas. OntoQuest stores all distinguished relationships permitted by OWL (e.g., `subclass-of, allValuesFrom, disjoint`) in separate tables, while all user-defined relation names are stored in a quad-store. Each class and property definitions from the ontologies are mapped into appropriate relational table. Using an advanced encoding and indexing algorithm [9], the DAG-structure of the ontological class hierarchies are stored in such a way that allows an efficient computations on transitive relations such as class subsumption and partonomic relations among the classes.

The current version of OntoQuest allows the expression of property chain rules advocated by the proposed OWL 2.0 standard [12]. This enables non-recursive first order rules like `(A subclass-of B), (C part-of B)®(C part-of A)`. When such rules are specified, OntoQuest may be instructed to materialize them or to evaluate them during query processing.

OntoQuest contains its own query processing engine to support ontological queries which is integrated with NIF search portal providing automated query expansion of the terms asserted in NIFSTD ontologies. It also provides a collection of web services to extract specific ontological contents [14]. Finally, OntoQuest accepts bridge ontologies

that provide mappings between multiple existing ontologies through rules specified in OWL. This is particularly important for NIF as most biomedical ontologies in the OBO consortium[3] significantly cross reference each other. NIFSTD currently have several modules containing extensive inter-ontology mappings.

4.1 OntoQuest Operations

OntoQuest implements various useful operations on its ontological graphs. Table 1. lists the functions along with their operations provided by OntoQuest. The NIF system extensively utilizes these functions trough its interface. Refer to [7] for details on the specific operations. It is important to note that the OntoQuest ontology repository does not store instances; rather, it only stores the classes and interclass properties. The instance data is stored in the various databases, and are duly mapped to ontological concepts where possible. Therefore the instantiate operator in OntoQuest calls data access operations.

Table 1. OntoQuest operators manipulate the ontology graph stored in the ontology repository. Node and edge IDs are unique in the system over all stored ontologies.

Function	Operation
`scangraph(p)`	Performs a scan operation over the edges that are evaluated to satisfy predicate p.
`selectNodeLabels(p)`	Selects a set of node labels satisfying predicate p.
`selectNodes(p)`	Selects a subset of nodes based on predicate p.
`selectEdges(p)`	Selects a subset of edges based on predicate p.
`project(pat)`	Projects a set of subgraphs that satisfy a graph pattern pat.
`label(g)`	Accepts a graph g and returns a copy of it by replacing the node-ids by node labels, and edge-ids by edge-labels.
`merge(g1,g2)`	Performs a node and edge union of graphs g1, g2.
`flattenPropTree(pLabel)`	Accept a property label pLabel and return a set of subproperties of label pLabel.
`flattenQuality(qValue)`	Accept a quality value qValue and return a set of subdomain quality names under qValue.
`induce(N)`	Given a node set N, returns the graph induced by N
`reachable(n1, n2, ei)`	Whether node n2 is reachable n1 by traversing edges satisfying regular expression ei.
`getTransitiveAncestors(n, Label, k)`	Get k levels of ancestors of node n, by following the transitive edge label Label.
`getTransitiveDescendants (n, Label, k)`	Get k levels of descendants of node n, by following the transitive edge label Label.
`neighbors(N, k, ei, ex)`	Given nodes N, returns the k-neighborhoods of each node in N, such that the edges satisfy the regular expression ei, and do not satisfy the regular expression ex.
`LCA(N, Label)`	Find the least common ancestor of node set N by traversing the transitive edge label Label.
`dagPath(n1, n2, Label)`	Find the paths connecting nodes n1 to n2 along transitive edge label Label.
`centerpiece(N)`	Given a node set N, compute the centerpiece subgraph intervening the nodes in N.
`unfoldPropertyChain(chID)`	Compute and materialize derived edges by unfolding OWL2 property chains identified by property chain ID chID.
`instantiate(N)`	Find instances of the nodes N, where N can be only class nodes.

[3] OBO Consortium, www.obofoundry.org

4.2 OntoQuest Query Processing

The NIF system uses a query language inspired by current search engines like Google. In this language, the simplest option is to ask a keyword query, but one can optionally add predicates on metadata and data attributes, specify return structures, and make references to ontologies. A detailed treatment of the full query specification with all options, and the evaluation strategy is beyond the scope of this paper. Simpler constructs are presented here but focus on how the ontologies are utilized.

Table 2: Typical ontological query expansions in NIF through OntoQuest.

Example Query Type	**Ontological Expansion**
A single term query for Hippocampus and its synonyms	`synonyms(Hippocampus)`, expands to Hippocampus OR "Cornu ammonis" OR "Ammon's horn" OR "hippocampus proper".
A conjunctive query with 3 terms	`transcription AND gene AND pathway`
A 6-term AND/OR query with one term expanded into synonyms	`(gene) AND (pathway) AND (regulation OR "biological regulation") AND (transcription) AND (recombinant)`
A conjunctive query with 2 terms, where a user chooses to select the subclasses of the 2nd term	`synonyms(zebrafish AND descendants(promoter,subclassOf)))`, zebrafish gets expanded by synonym search and the second term transitively expands to all subclasses of promoter as well as their synonyms.
A single term query for an anatomical structure where a user chooses to select all of the anatomical parts of the term along with synonyms	`synonyms(descendants(Hippocampus,partOf))`, expands to all parts of hippocampus and all their synonyms through the ontology. All parts are joined as an "OR" operation.
A conjunctive query with 2 terms, where a user chooses to select all the equivalent terms for the 2nd term	`synonyms(Hippocampus) AND equivalent(synonyms(memory))`, the second term uses the ontology to find all terms that are equivalent to the term memory by ontological assertion, along with synonyms.
A conjunctive query with 2 terms, where a user is interested in a specific subclasses for both of the terms	`synonyms(x:descendants(neuron,subclassOf) where x.neurotransmitter='GABA') AND synonyms(gene where gene. name='IGF')`, x is an internal variable.
A query to seek all subclasses of neuron whose soma location is in any transitive part of the hippocampus	`synonyms(x:descendants(neuron,subclassOf) where x.soma.location = descendants (Hippocampus, partOf))`
A query to seek a conceptual term that is semantically equivalent to a collection of terms rather than a single term.	`'GABAergic neuron' AND equivalent ('GABAergic neuron')`, the term gets recognized as ontologically *equivalent* to any neuron that has GABA as a neurotransmitter and therefore expands to a list of inferred neuron types.

A keyword query in NIF is an Boolean expression with wildcards (PARK* stands for PARKIN, PARK2,…). The basic query generation processes involves the following.

1. The query is first sent to a query analysis unit to identify terms that are known to the ontology sources.
2. The analyzed query goes through query expansion unit that uses the ontology to find synonyms and related terms stored with the ontology.

3. The terms in the expanded query are looked up from an inverted index to locate candidate records that are in different data stores (graph store, relation store etc.)
4. The Boolean conditions are then evaluated to generate a candidate list of data items (of heterogeneous types) that form the result.

Here, a candidate term refers to the actual term that matched with ontologies, and the candidate record refers to the containing data structure that has the candidate term. Table 2 presents a set of typical queries along with their expansions. It is important to note that, to the best of our knowledge, none of the traditional search engines provide this kind of query expansion mechanism. Readers interested in performance evaluation of typical NIF queries are referred to [7].

4.3 OntoQuest and NIF Search Interface

NIF is essentially an application system built upon the heterogeneous data management infrastructure that utilizes OntoQuest. NIF hides the complexity of the query language within the elements of the user interface. Specifically, user presents only keyword and ontological keyword queries; the ontology-based expansion and predicate search happens by user interaction.

The NIF search engine takes the user's keyword query and in the most common case, performs an ontological search to retrieve conceptual terms that closely match the terms in the ontology, and if desired, the neighborhood of these ontological terms. This process of exploring the ontology to find related terms is performed interactively. When the user settles on the final query terms, the keyword module uses the index to locate sources that have the data or web documents satisfying the keywords. Once the data sources are located, the source query wrapper module transforms the query into queries against all sources and broadcasts these transformed queries. The process of transformation converts the query keywords into SQL (or HTTP calls and so on) for structured data sources, XML requests, search against the web index and so forth. If the user's search terms are not found in the ontology, OntoQuest allows the query to be posted directly against the sources as a string search.

5. NIF Semantic Search

One of the most powerful aspects of ontologies is that they allow explicit knowledge of a domain to be asserted from which the implicit, inferred knowledge can be automatically derived as logical consequences. NIFSTD is designed to capitalize this ontological feature to enhance NIF's semantic search mechanism. The key feature of the current NIFSTD is the inclusion and enrichment of various cross-domain bridging modules. These modules contain necessary restrictions along with a set of defined classes to infer useful classification of neurons and molecules. The following list illustrates some of the defined concepts in NIFSTD and their classification schemes:

- Neurons by their soma location in different brain regions - e.g., Hippocampal neuron, Cerebellum neuron, Retinal neuron etc.
- Neurons by their neurotransmitter - e.g., GABAergic neuron, Glutamatergic neuron, Cholinergic neuron
- Neurons by their circuit roles - e.g., Intrinsic neuron, Projection neuron

- Neurons by their morphology - e.g., Spiny neuron
- Neurons by their molecular constituents - e.g., Pervalbumin neuron, Calretinin neuron
- Classification of molecules and chemicals by their molecular roles - e.g., Drug of abuse, Neurotransmitter, Calcium binding protein

A list of defined concepts along with their textual definitions can be found on the NIFSTD wiki page in [5]. The following example illustrates the strengths and usefulness of this feature for our NIF system. NIF has various neuron types with an asserted simple single hierarchy within the NIF-Cell module (Figure 4 is an example with five neuron types).

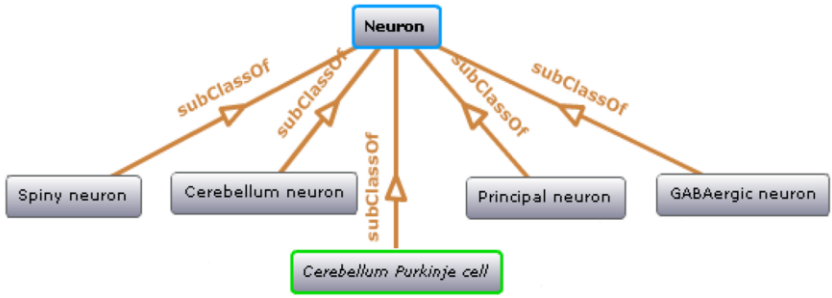

Figure 4. NIFSTD Cerebellum Purkinje cell is simply a subclass of a Neuron before invoking a reasoner along with asserted restrictions as specified in Fig. 5.

We assert various logical necessary restrictions about these neurons in a bridging module where we also specify various *defined* neuron types with necessary and sufficient conditions as illustrated in Figure 5.

Class name	Asserted defining (necessary & sufficient) expression
Cerebellum neuron	Is a 'Neuron' whose soma lies in any part of the 'Cerebellum' or 'Cerebellar cortex'
Principal neuron	Is a 'Neuron' which has 'Projection neuron role', i.e., a neuron whose axon projects out of the brain region in which its soma lies
GABAergic neuron	Is a 'Neuron' that uses 'GABA' as a neurotransmitter

Class name	Asserted necessary conditions
Cerebellum Purkinje cell	1. Is a 'Neuron' 2. Its soma lies within 'Purkinje cell layer of cerebellar cortex' 3. It has 'Projection neuron role' 4. It has 'GABA' as a neurotransmitter 5. It has 'Spiny dendrite quality'

Figure 5. Typical NIFSTD asserted restrictions for various neuron types. The first table in the figure defines three neuron types with logical necessary and sufficient conditions. The second table lists a set of necessary restrictions for Cerebellum Purkinje cell. All these restrictions written in a readable format here is expressed in OWL DL language in actual NIFSTD.

When the NIF-Cell module, along with the bridging modules are passed to a reasoner, the reasoner automatically computes for the asserted neuron types with restrictions (as indicated in Figure 5) and produce a hierarchy where a neuron can have multiple

inferred super-classes. In this example, although we did not explicitly state that Cerebellum Purkinje cell is anything other than a simple neuron, the reasoner identified that the neuron is an inferred subclass of four different defined neurons (Figure 6) namely, GABAergic neuron, Cerebellum neuron, Spiny neuron and Principal neuron, based on the logical restrictions specified as in Figure 5.

Having the defined ontological classes have enabled NIF to formulate useful concept-based queries. For example, while searching for 'GABAergic neuron', the NIF query expansion through OntoQuest recognizes the term as 'defined' from the ontology, and looks for any neuron that has GABA as a neurotransmitter (instead of the lexical match of the search term) and enhances the query over those inferred list of neurons. Searching these defined terms in a Google search would essentially exclude all the GABAergic neurons unless they are explicitly listed within the search.

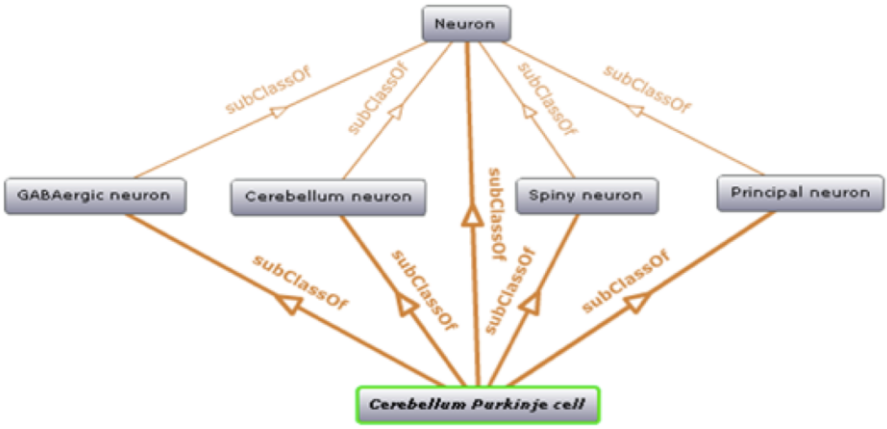

Figure 6. After invoking a reasoner NIFSTD Cerebellum Purkinje cell becomes a subclass of four different defined neuron types based on the restrictions specified in Figure 5.

To further its mission to deliver a concept-based interface for neuroscience, OntoQuest is being extended to support the rules for defining certain classes to include quantitative definitions as well as logical ones. Such concepts include quantities such as 'age maturity categories for common organisms, concepts such as "dementia" which may be defined as a range of scores on a set of test batteries and qualitative assessments of expression studies. These standards allow a researcher to issue a query through the NIF keyword interface for "increased gene expression" in "adults" for "drugs of abuse". For example, many native databases report age results for individual organisms. However, when searching from a keyword-based interface, users will not be easily able to enter a set of age ranges into the query without specialized forms.

Furthermore, looking for the expression of gene X in adult animals across species would require that the user enter age ranges for all organisms, clearly impractical. Thus, for concepts like maturity stages and expression levels, NIF maps the quantitative values to qualitative categories like "Adult" or "increases expression". For example, NIF defines an adult mouse as a mouse `>=40` days of age. For sources within the integrated view, we will provide the age provided by the data source but also an additional tag that maps this range to "Adult". For gene expression levels from microarray or other types of assays, NIF annotates a change relative to some control at a significance level of `p>=0.05` as "increased" or "decreased" expression. These

standards are applied when the translation of results is straightforward and the original values as recorded in the source database are always provided. When annotation of results requires interpretation because the meaning of values recorded in the source database is unclear, we do not apply such standards. For example, in the Gene to Brain Region view, each of the sources provides assessment on gene expression levels within a brain region in a different way (See Figure 7).

Database	Gene Symbol	Gene Name	Brain Structure...	Expression	Assay	Age	Organism
MGI	Grm1	glutamate receptor, metabotropic 1	midbrain	TRUE	RNA in situ	embryonic day 14.5	mouse, laboratory
GENSAT	Grm1	glutamate receptor, metabotropic 1	Midbrain	weak signal	BAC-Cre recombinase driver	Adult	Mouse
GENSAT	Grm1	glutamate receptor, metabotropic 1	Midbrain	moderate to strong signal	BAC-Cre recombinase driver	P7	Mouse
GENSAT	Grm1	glutamate receptor, metabotropic 1	Midbrain	moderate to strong signal	BAC-Cre recombinase driver	Adult	Mouse
ABA	Grm1	glutamate receptor, metabotropic 1	Midbrain	30	RNA in situ hybridization	Adult	Mouse
MGI	Grm1	glutamate receptor, metabotropic 1	midbrain	TRUE	RT-PCR	postnatal	mouse, laboratory

Figure 7. Results for "GRM1 and Midbrain" from the 3 expression atlases.

In the Allen Brain gene expression atlas, based on analysis of in situ hybridization images through the mouse brain, gene expression levels are expressed in a numerical scale from 0-100 where expression is normalized compared to controls. Higher numbers indicate an increased likelihood that the gene is expressed within that region, but there is no way for NIF to translate the numbers into discrete categories that would accurately reflect the intent of the resource providers. Similarly, the GENSAT resource provides assessments of labeling patterns for the BACS transgenics per brain region using a qualitative rating of low, moderate and high staining. However, the staining intensity does not necessarily correlate with the expression levels of gene, as it reflects the number of GFP's inserted into the cell. Thus, in both cases, NIF retains the assessments provided by the source databases without additional annotations.

6. Conclusion

In this paper we have presented the NIF as an ontology-driven information system, a system where the data resources are imported into the system by the domain experts, annotated by the domain experts to connect the data to its standardized formal ontologies or other data sources through the ontologies, and finally, all the data mapped to the ontologies can be queried in a federated manner. The NIF system enables multi-model data federation. It uses an ontologically enhanced system catalog, an ontological data index, an association index to facilitate cross-model data mapping, and an algorithm for ontology-based keyword queries with ranking.

Since the launch of the NIF in September 2008, it has grown significantly in content and functionality. During this rapid period of growth, NIF has been working to develop an understanding of the current state of biomedical resources and established an ontology-based infrastructure, procedures and guidelines for maximizing their utility.

The NIF project provides an example of practical ontology development and how ontologies can be used within hybrid information systems to enhance search and data integration across diverse resources.

Acknowledgement. Supported by a contract from the NIH Neuroscience Blueprint HHSN271200800035C via NIDA.

References

[1] W.J. Bug, G.A. Ascoli, J.S. Grethe, A. Gupta, M.E.Martone et al., The NIFSTD and BIRNLex Vocabularies: Building Comprehensive Ontologies for Neuroscience, *Neuroinformatics* 6(3) (2008), 175-94

[2] A. Gupta, W.J. Bug, L. Marenco, C. Condit, M. E. Martone et al., Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF), *Neuroinformatics* 6(3) (2008), 205-17

[3] A. Rector, Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL, *Proc K-CAP* (2003)

[4] B. Smith, M. Ashburner, C. Rosse, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat Biotech* (2007), 1251-1255

[5] F.T. Imam, S.D. Larson S., J.S. Grethe, A. Gupta, A. Bandrowski, M.E. Martone, NIFSTD and NeuroLex: A Comprehensive Neuroscience Ontology Development Based on Multiple Biomedical Ontologies and Community Involvement, *Proc. of Intl. Conf. on Biomedical Ontologies (ICBO), Buffalo, NY* (2011)

[6] L. Chen, M.E. Martone, A. Gupta, L. Fong, M. Wong-Barnum, Ontoquest: exploring ontological data made easy, *Proc. 31st Int. Conf. on Very Large Database (VLDB)* (2006), 1183–1186

[7] A. Gupta, C. Condit, X. Qian, An ontology-enhanced information system for heterogeneous biological information, *BioDB* (2011)

[8] M. Courtot, F. Gibson, A. Lister, J. Malone, D. Schober, R. Brinkman, A. Ruttenberg, MIREOT: the Minimum Information to Reference an External Ontology Term. *Nature Precedings* (2009), http://dx.doi.org/10.1038/npre.2009.3576.1

[9] L. Chen, A. Gupta, M.E. Kurul, Stack-based algorithms for pattern matching on DAGs, *Proc. 31st Int. Conf. on Very Large Databases (VLDB), Stockholm* (2005), 493–504

[10] Ontology Design Patterns (ODPs) Public Catalog, http://odps.sourceforge.net

[11] Web Ontology Language (OWL), http://www.w3.org/2001/sw/wiki/OWL

[12] OWL2 Web Ontology Language Primer, http://www.w3.org/TR/2009/REC-owl2-primer-20091027/

[13] Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu

[14] OntoQuest Web Services, http://ontology.neuinfo.org/ontoquestservice.html

[15] NIFSTD SPARQL Endpoint, http://ontology.neuinfo.org/sparqlendpoint.html

[16] NIFSTD in NCBO BioPortal, http://bioportal.bioontology.org/ontologies/1084

[17] J. Mei, L. Ma, Y. Pan, Ontology query answering on databases, *Proc. of Int. Semantic Web Conf.* (2006), 445–458.

[18] P. Grenon and B.Smith, SNAP and SPAN: Towards Dynamic Spatial Ontology, *Spatial Cognition and Computation* 4: 1 (2004), 69-103