NIFSTD and NeuroLex: A Comprehensive Neuroscience Ontology Development Based on Multiple Biomedical Ontologies and Community Involvement

Fahim T. Imam, Stephen D. Larson, Jeffery S. Grethe, Amarnath Gupta, Anita Bandrowski, Maryann E. Martone

Neuroscience Information Framework, Center for Research in Biological Systems, University of California, San Diego, La Jolla, USA

Abstract. A core component of the Neuroscience Information Framework (NIF) project, the NIF Standard (NIFSTD) ontology, was envisioned as a set of modular ontologies that provide a comprehensive collection of terminologies to describe neuroscience relevant data and resources. We present here on the structure, design principles, community engagement and current state of NIFSTD that reuses multiple biomedical ontologies, applied for an effective semantic search mechanism.

Keywords: Multiple ontologies, modularity, ontology reuse, neuroscience ontology, semantic search

1 Introduction

The Neuroscience Information Framework (NIF)[1] project involves discovering and utilization of various neuroscience resources over the web. The end product is a semantic search engine and discovery portal consisting of a framework that describes resources and provides simultaneous access to multiple types of information, organized by various relevant categories.

Behind the scenes, NIFSTD [1] is a critical constituent for the NIF project in order to enable its effective concept-based search mechanism against a diverse collection of web based neuroscience data and resources. The overall ontology is assembled in a form that promotes reuse of multiple ontologies, easy extension and modification over the course of its evolution. NIFSTD relies on existing biomedical ontologies as the initial building blocks. These ontologies currently include CHEBI, Gene Ontology (GO), Protein Ontology (PRO), Ontology for Biomedical Investigations (OBI), and The Ontology of Phenotypic Qualities (PATO). Section 2 of this paper describes the basic structure and design principles of NIFSTD. The relation between NIFSTD and the NeuroLex wiki environment for collaborative development is the focus of Section 3. Current state and progress of NIFSTD is highlighted in Section 4 followed by conclusion and future work in Section 5.

2 NIFSTD Basic Structure and Design Principles

NIFSTD follows OBO Foundry [11] principles as long as they are reasonable for NIF's application purposes. While the principles promote developing highly interoperable and reusable reference ontologies in ideal cases, following some of them in a rigid manner has often proven to be too ambitious for day-to-day development. Some of the principles may become impractical for extending existing ontology modules especially when there is a deadline constraint imposed for those extensions. Gaining OBO Foundry community consensus for a production system is difficult as we often need to move quickly along with the project. Therefore, we rather favor a system whereby we start with minimal complexity as required and add more as the ontologies evolve over time towards perfection.

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

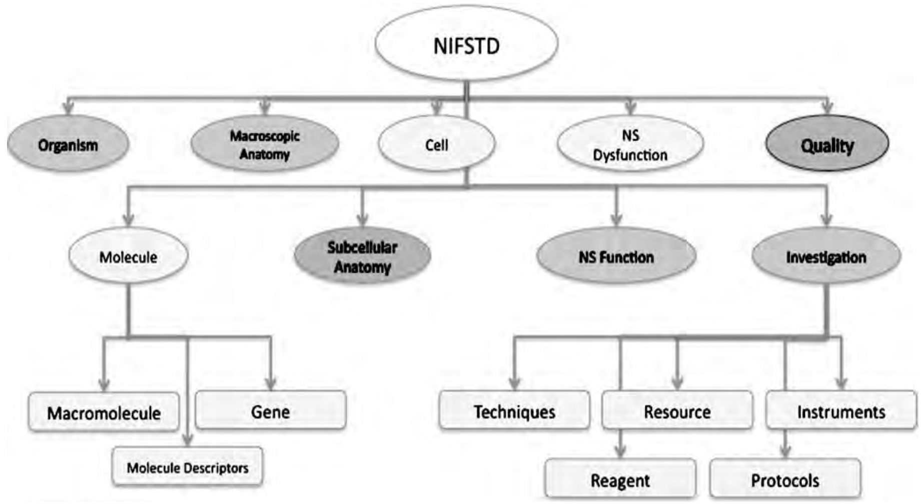

Figure 1. The semantic domains covered in the NIFSTD ontology. Separate OWL modules cover the domains specified within the ovals. The umbrella file http://purl.org/nif/ontology/nif.owl imports each of these modules when opened in Protégé. Each of the modules, in turn, may cover multiple sub-domains,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

Domain	External Source	Import/ Adapt	Module
Organism taxonomy	NCBI Taxonomy, GBIF, ITIS, IMSR, Jackson Labs mouse catalog	Adapt	NIF-Organism
Molecules	IUPHAR ion channels and receptors, Sequence Ontology (SO), ChEBI, and Protein Ontology (PRO); pending: NCBI Entrez Protein, NCBI RefSeq, NCBI Homologene, NIDA drug lists	Adapt IUPHAR, ChEBI; Import PRO, SO	NIF-Molecule NIF-Chemical
Sub-cellular	Sub-cellular Anatomy Ontology (SAO). Extracted cell parts and subcellular structures. Imported GO Cellular Component with mapping	Import	NIF-Subcellular
Cell	CCDB, NeuronDB, NeuroMorpho.org. Terminologies; pending: OBO Cell Ontology	Adapt	NIF-Cell
Gross Anatomy	NeuroNames extended by including terms from BIRN, SumsDB, BrainMap.org, etc; multi-scale representation of Nervous System Mac Macroscopic anatomy	Adapt	NIF-GrossAnatomy
Nervous system function	Sensory, Behavior, Cognition terms from NIF, BIRN, BrainMap.org, MeSH, and UMLS	Adapt	NIF-Function
Nervous system dysfunction	Nervous system disease from MeSH, NINDS terminology; Disease Ontology (DO)	Adapt/Import	NIF- Dysfunction
Phenotypic qualities	PATO Imported as part of the OBO foundry core	Import	NIF-Quality
Investigation: reagents	Overlaps with molecules above, especially RefSeq for mRNA	Import	NIF-Investigation
Investigation: instruments, protocols, plans	Based on Ontology for Biomedical Investigation (OBI) to include entities for biomaterial transformations, assays, data collection, data transformations.	Adapt	NIF-Investigation
Investigation: resource type	NIF, OBI, NITRC, Biomedical Resource Ontology (BRO)	Adapt	NIF-Resource
Biological Process	Gene Ontology's (GO) biological process in whole	Import	NIF-BioProcess
Cognitive Paradigm	Cognitive Paradigm Ontology (CogPO)	Import	NIF-Investigation

Table 1. Domains covered by the current NIFSTD along with the vocabularies imported from the external, community sources and the corresponding OWL modules.

Modularity. One of the key features of NIFSTD is that its ontologies are built in a modular fashion, each module covering orthogonal domain of Neuroscience concepts (Fig.1). These distinct domains include macroscopic anatomy, cell types, techniques, nervous system function, molecules etc. Following a powerful modularization ontology design pattern [3], NIFSTD promotes easy extendibility towards its evolution. It avoids duplication of efforts by conforming to standards that promote reuse. Each of the modules is standardized to the same upper level ontologies such as the Basic Formal Ontology (BFO) and OBO Relations Ontology (OBO-RO).

Representation Language. Expressed in W3C standard OWL Description Logic (OWL-DL) dialect, NIFSTD ensures computational decidability and supports automated reasoning via common DL reasoners such as Pellet and FACT++. The reasoners ensure the ontologies to be kept in a logically consistent state and allow computing inferred classification of asserted class hierarchies.

Reuse of Available Ontology Sources. Wherever possible, NIFSTD reuses existing available distilled knowledge sources, terminologies and ontologies. Community sources were culled from a variety of sources, ranging from fully structured ontologies to loosely structured controlled vocabularies. Table 1 illustrates various domains in NIFSTD that are either adapted or imported as a whole from various external, community sources. The process of importing or adapting a new ontology or vocabulary source varies depending upon its state [1].

- If a source already uses OWL, the OBO-RO and the BFO and is orthogonal to existing modules, the import simply involves adding an owl:import statement to the main ontology file (nif.owl).
- If an existing orthogonal ontology is in OWL but does not use the same foundational ontologies as NIFSTD, then an ontology-bridging module (explained later) is constructed declaring the deep level semantic equivalencies such as foundational objects and processes.
- If an external source is satisfiable by the above two rules but observed to be too large for NIF's scope of interests, a relevant subset is extracted as suggested by NIF's domain experts (e.g., CheBI, OBI etc. are adapted rather than imported as a whole). During our recent extensions, we have been following MIREOT principle [6] that allows extracting a required portion of a larger ontology.
- If the external source has not been represented in OWL, or does not use the same foundation as NIFSTD, then the terminology is adapted to OWL/RDF in the context of the NIFSTD foundational layer ontologies.

Single Inheritance. NIFSTD follows the simple inheritance principle for the hierarchy of *named* classes; i.e., an asserted *named* class can have only one named class as its superclass; however, a named class can have multiple *anonymous* superclasses. This principle promotes the named classes to be univocal and to avoid ambiguities. The classes with multiple superclass 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. Motivations behind this approach can be found in detail in [10].

Unique Identifiers and Annotation Properties. NIFSTD entities are identified by a unique identifier and accompanied by a variety of annotation properties derived from Dublin Core Metadata (DC) and the Simple Knowledge Organization System (SKOS) model. We reuse the same URI through MIREOTed extracted classes from the source, which allows us to avoid extra mapping annotations with the community sources, e.g., GO or PRO class identifiers remain unaltered. A tool based on MIREOT principles called OntoFOX [4] is proven to be useful in this regard.

Object Properties and Bridge Modules. NIFSTD object properties are mostly drawn from the OBO Relations Ontology (OBO-RO).

Fig. 2 illustrates some example object properties between various NIFSTD Classes.

The cross-domain relations are specified in separate *bridging* modules – modules that only contain logical restrictions and definitions on a required set of classes assigned between multiple modules [1]. The bridging modules allow the main domain modules – e.g., anatomy, cell type, etc. to remain independent of one another without the bridging modules (e.g., Fig. 3). They help to keep the modularity principles intact, and facilitate extensions for broader communities without worries about specific, NIF-centric views on how a class should be logically defined or restricted. These bridging modules can easily be excluded in order to focus on core modules and build separate bridging module that are appropriate for user specific domain.

Versioning. Various annotation properties are associated with versioning different levels of content within NIFSTD. These include creation and modification dates for each of the classes; file level versioning for each of the modules, annotations for retiring antiquated concept definitions, tracking former ontology graph position and replacement concepts. Refer to [1] for more details on NIFSTD versioning policies.

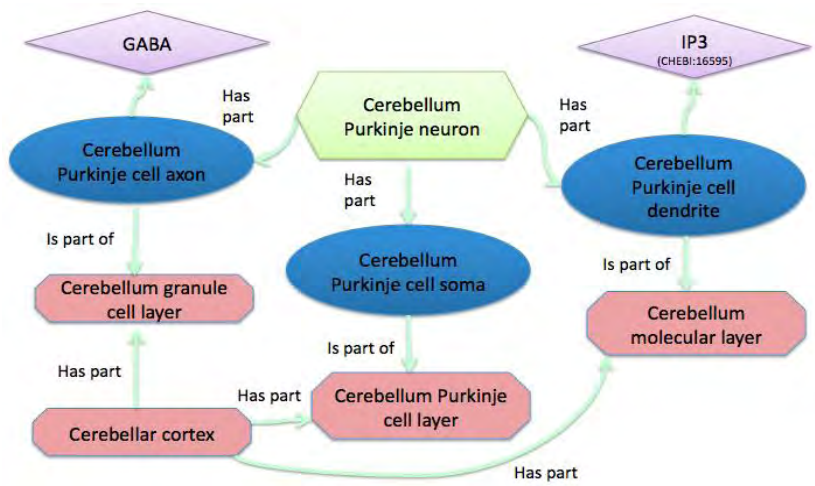

Figure 2. A typical knowledge model in NIFSTD. Both cross-modular and intra-modular classes are associated through object properties mostly drawn from the OBO Relations ontology (RO).

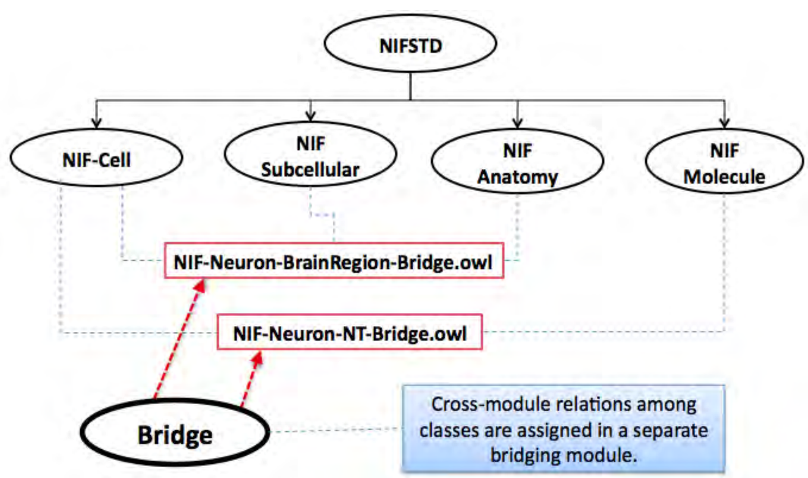

Figure 3. Two example bridging OWL modules in NIFSTD (rectangular boxes) that contain class property associations between multiple core modules.

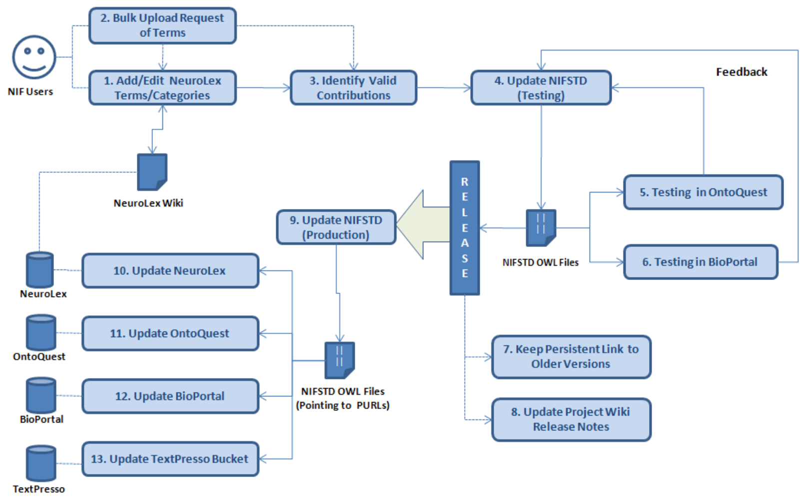

Figure 4. NIF workflow to transition knowledge between NeuroLex and NIFSTD.

Accessing NIFSTD Ontologies.[2] NIFSTD is primarily available in OWL format and can be loaded through Protégé [7] or similar OWL editing tools. It is also viewable through NCBO BioPortal [12]. Within NIF, NIFSTD is served through an ontology management system called OntoQuest [2]. OntoQuest generates an OWL-compliant relational schema and implements various ontological graph search algorithms for navigating, term searching and expanding query terms based on their logical restrictions for the NIF search portal. It also provides web services [8] to extract ontology contents. NIFSTD is also available in RDF and has a SPARQL endpoint [9].

3 NIFSTD and NeuroLex Wiki

One of the largest roadblocks that we encountered during our ontology development was the lack of tools for domain experts to view, edit and contribute their knowledge to NIFSTD. Existing editing tools were difficult to use or required expert knowledge to employ. We strive to balance the involvement of the neuroscience community for domain expertise and the knowledge engineering community for ontology expertise when constructing the NIFSTD. By combining open-source technologies related to semantic media wiki, NIF developed NeuroLex[3], a semantic wiki environment to promote easy collaboration between the neuroscience community and domain experts.

The NeuroLex Wiki. We envision NeuroLex 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 Wiki Facts. NeuroLex provides a bottom-up ontology development approach where multiple participants can edit the ontology instantly. Control of content is done after edits are made based on the merit of the content, rather than by the blessing of a few known individuals. Semantics is limited to what is convenient for the domain. Essentially, the NeuroLex approach is not a replacement for top-down construction, but critical to increase accessibility for non-ontologist domain experts.

NeuroLex has become critically important for the large corpus domain with no formal categories where the entities are unstable and unrestricted with no clear edges. NeuroLex is potentially necessary when participants are uncoordinated users, amateur users, or naive

[2] The NIF Standard Ontologies, http://ontology.neuinfo.org
[3] The NeuroLex Wiki Site, http://neurolex.org

cataloguers. NeuroLex provides various simple forms for structured knowledge where communities can contribute and verify their knowledge with ease. It also allows to generate a specific class hierarchy, or extraction of a specific portion of the ontology based on certain properties in a spreadsheet, without having to learn complicated ontology tools. The NIFSTD/NeuroLex development and curation workflow are depicted in Fig. 4.

NIFSTD vs. NeuroLex Properties. While the properties in NeuroLex are meant for easier interpretation, the restrictions in NIFSTD are usually based on rigorous OBO-RO standard 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 we use similar kind of 'macro' relation such as 'has_neurotransmitter' within NIFSTD, recognizing that these relations can be specified more rigorously. These 'macro' relations readily lend themselves into rigorous representations should they become necessary at a later date.

4 Current State and Progress

We use NIFSTD ontologies with the sense of being essentially useful for our semantic search interface. One of the most powerful features of having ontology is that it allows explicit knowledge of a domain to be *asserted* from which implicit logical consequences can be *inferred* using logical reasoners. The following example illustrates the strengths and usefulness of this feature for NIF search portal.

NIFSTD includes various neuron types with an asserted simple hierarchy within the NIF-Cell module (Fig. 5 is an example with five neuron types). However, 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 Fig. 6.

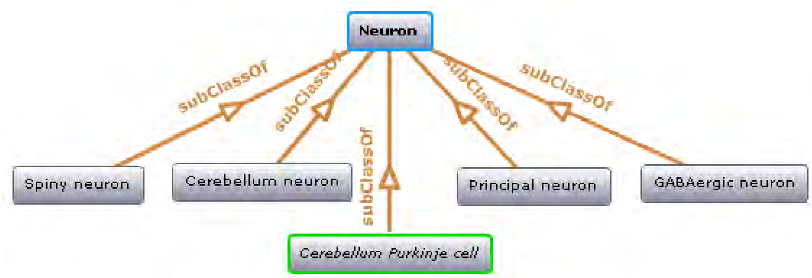

Figure 5. NIFSTD Cerebellum Purkinje cell is simply a subclass of a Neuron before invoking a reasoner along with asserted restrictions as specified in Fig. 5.

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 6. 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.

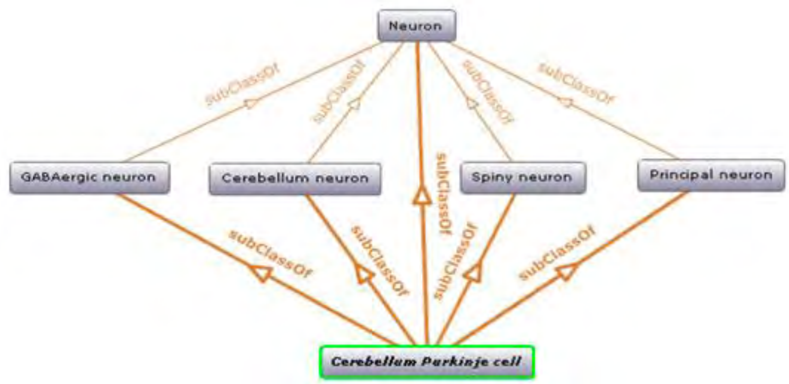

Figure 7. After invoking a reasoner NIFSTD Cerebellum Purkinje cell becomes a subclass of four different defined neuron types based on the restrictions specified in Figure 6.

When the NIF-Cell module along with the bridging modules is passed to a reasoner, the reasoner automatically computes for the asserted neuron types with restrictions (as indicated in Fig. 6) and produce a hierarchy where a neuron can have multiple inferred superclasses. 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 (Fig. 7) namely, GABAergic neuron, Cerebellum neuron, Spiny neuron and Principal neuron, based on the logical restrictions specified as in Fig. 6. Having the defined classes has enabled us to develop useful concept-based queries through the NIF search interface. For example, while searching for 'GABAergic neuron', the system 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 this 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.

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. A running list of defined concepts along with textual definitions can be found on the NIFSTD wiki page in [5]. 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, 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

The modularity principles along with the bridging modules allowed us to limit the complexity of the base ontologies and promoted easy extendibility. Also, it allowed us to rely on module-by-module reasoning for consistency checking, inferred subsumption and other reasoning tasks. We provide different inferred axioms saved in separate modules for the end-users convenience. We are still looking for a practical reasoning mechanism with more powerful inference engines that can scale with large ontologies like NIFSTD as a whole with all its individual modules imported together.

5 Conclusions and Future Work

The NIF project provides an example of practical ontology development and how it can be used to enhance search and data integration across diverse resources. Using the upper level BFO ontologies allowed us to promote a broad semantic interoperability between a large numbers of biomedical ontologies. We have defined a process to form complex semantics to various neuroscience concepts through NIFSTD and through NeuroLex collaborative environment. NIF encourages the use of community ontologies for resource providers, and as the project moves forward, we are using NIFSTD to build an increasingly rich knowledgebase for neuroscience that integrates with the larger life science community.

Acknowledgement

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

References

1. Bug, W.J., Ascoli, G.A., Grethe, J.S., Gupta, A., et al.: The NIFSTD and BIRNLex Vocabularies: Building Comprehensive Ontologies for Neuroscience. Neuroinformatics. 2008 Sep;6(3):175-94. PMID: 18975148 (2008)
2. Gupta, A., Bug, W., Marenco, L., Condit, C., et al.: Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF). Neuro-informatics. 2008 Sep;6(3):205-17. PMID: 18958629 (2008)
3. Ontology Design Patterns (ODPs) Public Catalog, http://odps.sourceforge.net
4. Xiang, Z., Courtot, M., Brinkman, RR., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. BMC Research Notes. 2010, 3:175. PMID: 20569493 (2010)
5. NIFSTD Defined Concept List, http://ontology.neuinfo.org/defined-types.html
6. Courtot, M., Gibson, F., Lister, A., Malone, J., Schober, D., Brinkman, R., Ruttenberg, A.: MIREOT: the Minimum Information to Reference an External Ontology Term. Available from Nature Precedings, http://dx.doi.org/10.1038/npre.2009.3576.1, (2009)
7. Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu
8. OntoQuest Web Services, http://ontology.neuinfo.org/ontoquest-service.html
9. NIFSTD SPARQL Endpoint, http://ontology.neuinfo.org/sparql-endpoint.html
10. Rector, A.: Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. Proc K-CAP (2003)
11. Smith, B., Ashburner, M., Rosse, C., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech. 1251-1255 (2007)
12. NIFSTD in NCBO BioPortal, http://bioportal.bioontology.org/ontologies/1084