Notes on Neuroscience Ontologies

Dan Gardner

Ontologies and Neuroinformatics

Words and phrases often have multiple meanings, and individual domains of neuroscience use them in different ways. Also, many similar terms are not synonymous. Either context or structure can be used to extract significance, to disambiguate similar terms and relate different ones. Tools that do this by representing knowledge about words and meanings are called ontologies.

How can ontologies serve neuroinformatics? Whether we need neuroscience ontologies, and if so, how they are to be designed, depends in part on how we select:

- ***level:*** what level or levels of knowledge the ontologies will specify
- ***scope:*** how much of neuroscience is to be covered
- ***methodology:*** what tools for ontology development are used
- ***implementers:*** who specifies, designs, maintains, and extends the ontologies

Selections should be guided by what is needed, what is implementable, and how well the results can be used by colleagues, including both other neuroinformatics developers and the body of neuroscientists who generate and use data.

Levels and Scope

We can focus discussion by grouping these into discrete but useful categories:

- ***general:*** top-level ontologies that include definitions of parts of speech or XML semantics, using generic parsers
- ***comprehensive discipline-specific:*** knowledge representation spanning neuroscience. The magnitude of this task can be appreciated by asking whether all 1414 pages of Kandel and Schwartz could be reduced to a series of statements in the sentential calculus, or in XML.
- ***focused domain-specific:*** knowledge representation for one domain of neuroscience, such as an ontology to relate mammalian anatomical descriptors.
- ***functional ontology:*** not knowledge representation, but specification of a controlled or reduced syntax and semantics, designed for parsing and interfacing descriptive neurobiological attributes of XML-derived schemes.

Methodology and Implementers

These are interdependent:

- Neuroscientists are needed to select and encapsulate relevant concepts.
- Both neuroscientists and ontology specialists together should construct definitions.
- Language and neuroscience domain specialists should select or develop tools to parse, interface, and interconvert terms, mediate queries, and federate databases.
- Neuroinformatics developers and graphic user interface (GUI) specialists, with input from neuroscientists who generate and use data, should design GUIs that build upon ontologies without requiring awareness of their technical detail.
- The neuroinformatics community should define methods for updating and extending ontologies in response to developments in research methodology and our understanding of neuroscience.

We should in addition track ongoing development of universal ontologies such as OIL and tools for ontology development or knowledge parsing, including mediators such as MIX.

THE CASE FOR FUNCTIONAL ONTOLOGIES

Where descriptors are free text with ad-hoc terminology, a discipline-specific or domain-specific ontology can parse and translate each descriptor into a standard set of identifiers and relations. For neuroinformatics, my group favors a simpler scheme, a functional ontology we are developing called the Common Data Model. Here, context and structure are supplied by a domain- and discipline-specific data model with controlled-vocabulary values to defined attributes.

The newest Common Data Model encapsulates an implicit ontology that is transparent to users and implementable by parsers. It is being implemented as BrainML—a markup language to be defined using XML Schema and relying on a metalanguage we are also developing called BrainMetaL. The model derives most neuroscience terms from one of five broad categories, a scheme we informally call *quintessence*:

- ***data:*** datasets, wrappers for data, and quasi-data from simulations
- ***entities:*** expanded from recording sites to include anatomy, molecules, and more
- ***models:*** simulation parameter sets, hypotheses, and diagnoses
- ***methods:*** protocols, experimental conditions, and simulation engines
- ***references:*** publications and authors

Each category is a hierarchy, allowing specialized descriptors to be derived from more general ones. Each component of the data model utilizes multiple specific descriptive attributes. Most attribute values are arranged in a controlled-vocabulary (CV) hierarchy. Although attributes and hierarchies together provide a set of implicit is-a and has-a operators that form a functional ontology, neuroscientist users need only select appropriate terms from intuitive GUI tools.

Implementation and maintenance are straightforward. Although each CV hierarchy is a simple tree, the use of multiple attributes appears to be able to serve the same knowledge concepts as a multi-inheritance directed graph. The use of uniform CV hierarchies coordinated via forms also reduces or eliminates the need for moderators to review the submitter-supplied terminology. However, for such an implicit ontology to continue to serve neuroscience, it cannot remain static. There must be continued monitoring and upgrading of each CV hierarchy (easy) and if needed extensions to the data model (hard).

Although BrainML is projected as a community effort designed to fit a large fraction of contemporary neuroscience, it is designed to permit interconversion with other XML-based scientific markup languages.

Laboratory of Neuroinformatics
Department of Physiology and Biophysics
Weill Medical College of Cornell University
New York, NY 10021
USA

dan@aplysia.med.cornell.edu

APPENDIX

Two examples—the patient record and a neuroscience conference—show the difficulty of implementing usable discipline-specific ontologies:

1. Consider the tasks required to parse the ad-hoc descriptors that appear in patient records, and the complex multidimensional semantic nets and ontologies now under development for classification of diseases, clinical data types, and subjectivity of chart entries. Word appearance alone is clearly inadequate for this task, impelling the use of contextual ontologies.

2. As an example of the need for context in vocabulary development, and the complexity of a comprehensive neuroscience ontology, consider the following announcement for a meeting on "neural signaling". Here, this term covers a wide range of molecular intra- and inter-cellular processes including vesicle endocytosis, synaptic plasticity, retrograde messengers, signal transduction pathways, growth cones, and biogenic amine uptake. You'd have to know each speakers' work to select the appropriate meaning. Moreover, the term as used here doesn't include such other common meanings of 'signaling' as action potential timing, and related topics including signal sequences controlling protein processing.

```
COLLOQUIUM ON "NEURAL SIGNALING"

... Sackler Colloquium of the National Academy of Sciences
entitled "Neural Signaling", will be held ... February 16-17,
2001 ... co-organized by Solomon H. Snyder and Richard L. Huganir.

Recent advances in molecular neuroscience have revealed links in
signaling mechanisms between and within cells in systems as diverse
as neurotransmission, olfaction and nerve growth.  ... from
elementary synaptic events to Alzheimer's Disease and drug
addiction.

The colloquium sessions will cover the following topics:

1. Inter- and Intracellular Signaling in the Nervous System
   Pietro DeCamilli(Yale), Richard L. Huganir (Johns Hopkins), Roger
   Nicoll (UCSF), Moo Ming Poo (UC-San Diego)

2. Inter- and Intracellular Signaling in the Nervous System
(continued)
   Lily Jan (UCSF), Michael Greenberg (Harvard),
   Paul Greengard (Rockefeller), Cori Bargmann (UCSF)

3. Signaling in the Developing Nervous System
   Richard Axel (Columbia), Marc Tessier-Lavigne (UCSF),
   Corey Goodman (UC-Berkeley), Carla Schatz (Harvard)

4. Drugs and Disease and Signaling in the Nervous System
   Eric Nestler (Yale), Marc Caron (Duke), Dennis Choi
   (Washington U.), Dennis Selkoe (Harvard)
```