NIPClassificator
Toward an evolvable neuroinformatics ontology ⋆

T. Förster, A.V.M. Herz, and R. Ritz

Institute for Theoretical Biology, Humboldt–University Berlin, Invalidenstr. 43, D–10 115 Berlin, Germany

Abstract

The Neuroinformatics Portal Pilot (NIP; `http://www.neuroinf.de`) is part of a larger effort to promote the exchange of neuroscience data, data–analysis tools, and modeling software. We develop the basic infrastructure needed for an optimal utilization of the many available resources on the web. Conceptual and technical problems have to be solved before the portal can serve the community in a most efficient way. Nevertheless it is operational already while still being further developed. In this contribution, we describe the design and planned usage of NIP's classification system.

Key words: Web portal; meta data; data sharing; classification

1 Introduction

Neuroscience aims at understanding the structure and function of neural systems including the human brain and has experienced a tremendous growth over the last few decades. The many facets of neural dynamics and computations, cognitive processes, and neurological disorders are reflected by a large number of highly specialized scientific disciplines involved in brain research. The rapid scientific progress has led to a constantly growing flood of data that are as diverse as gene–expression profiles, multi–electrode recordings, or functional brain images. Integrating all these data into one comprehensive body of knowledge has become a major challenge within neuroscience.

⋆ Supported by BMBF

The Neuroinformatics Portal Pilot (NIP; `http://www.neuroinf.de`) tries to establish a "warehouses" for data–analysis tools, modeling software and information on raw and processed data including publications. It tries to become a valuable community site for people and research groups in the field.

The overall rationale and objective for this portal has been described elsewhere [1,2] as well as the general usage of the site [3]. The exclusive focus of this contribution is on the most recent addition to the site, namely its highly improved classification system. The general design concepts as well as its intended usage are outlined now.

2 Why Classification is a Problem

One of the most common approaches to organizing content at any repository (here: web site) is to classify the individual content items by assigning them appropriate keywords to specify their subject matter (like for this article, see above). If there are no restrictions on the allowed keywords, however, this classification is of limited value because there is no way to infer the specific meaning of the terms used nor is possible to make sure that different people use the same terms to denote the same concepts.

2.1 Controlled Vocabularies

To deal with these issues it is common to use so–called *controlled vocabularies* meaning that keywords can only be chosen from a predefined set of terms the meaning of which is then well defined (in the context of the repository at least). The size of these sets can vary considerably (from just a few up to several thousand terms) depending on the application. Designing a scalable user interface to interactively work with these vocabularies (think of searching) easily becomes a major problem of its own. But no matter how complex the vocabulary, there are almost inevitably cases where there is no appropriate term to describe a resource of interest. Therefore, a combination of both approaches, controlled an free keyword classification, has often become the method of choice. It is possible, of course, to design rules according to which the controlled vocabulary can be extended and this is also what NIP does (more below).

2.2 Relationships Between Keywords

An additional benefit of using controlled vocabularies is that it is straight forward to establish relationships by just considering which terms to allow. Most often these relationships are only implicitly defined either based on common knowledge (e.g., *you know* that a cat is a mammal and a vertebrate) or on other conventions. This kind of approach basically works for all tree–structured relationships as long as the individual terms are unique. Given just one keyword it is possible to infer all its parents in the tree (the more general concepts) as well as its children (the more specific sub–categories). But this approach falls short as soon as more complex structures need to be taken into account (like synonyms, mapping of brain regions, method dependent definition of brain areas). Ideally one would like to be able to define arbitrary relationships between the terms or concepts in order to allow for a representation of any kind of knowledge in that domain. This is what NIP's classification system tries to achieve for the neuroinformatics community. How this is done is described in the following section.

3 NIP's Approach

At the Neuroinformatics Portal (NIP; `http://www.neuroinf.de`) any content item can be classified by relating it to an arbitrary number of existing keywords. In addition, every user can propose new keywords as needed but in order to do so a brief description defining the meaning of the new term as well as its relationships to the already existing ones have to be provided. Keyword submissions are then treated just like any other content submission and in case of acceptance the new term will be generally available to all (including its relations to other terms). That way it should become possible to build up and maintain an up–to–date keyword base to classify any content of interest to the community.

3.1 User Interface

It is to be expected that following the approach sketched above will sooner or later lead to a substantial amount of terms (on the order of thousands and above) which will make it hard for the user to directly interact with it via the web. It is just impracticable to put such an amount of terms in a selection box to choose from. Therefore specifying keywords (for classifying content as well as for formulating searches) is usually a two step process: First you enter the term that you expect to be a keyword, then in a second step

this term is looked up in the keyword base. If there is an exact match then this term is taken literally otherwise the system tries to infer related terms and it will ask you to select from those. Currently this inference is based on syntactic similarity only (using the algorithm by Ratcliff and Obershelp) but more powerful inference engines are to be included as NIP further develops.

3.2 Establishing Relationships Between Keywords

Since keywords themselves are treated as content items it is also possible to establish arbitrary relationships between different keywords. Examples are *is parent of/is child of* to establish tree structures or *is synonym of* to allow several terms to be used for the same concept (think of abbreviations or different languages). It is even possible to have the same term referring to different concepts, e.g., *neuron* can be used to generally specify a nerve cell or the simulation software package of that name. To the latter end, there are two keyword items with the same name but different identifiers, descriptions etc. related via *not the same*. Just as with the keywords themselves, every user of the portal can propose to add such a relationship between keywords. That way we hope it will become feasible to incorporate an increasing amount of knowledge into the system.

3.3 Inference

The advantages of following this kind of approach become obvious when considering questions like *What's related?* starting from an individual content item. Most often one would also like to know *... and why?* or *... how closely?* something else is considered related. To this end, NIP's classification system allows for introspection (answering *why*) as well as ranking of relationships (answering *how close*). The ranking is established by assigning appropriate weights to the relationships between keywords (e.g., 0.5 for *is parent of* or 1 for *is synonym of* or 0 for *not the same*) and then multiplying all weights along the shortest possible path from one content item to the other based on their classification.

How to best choose the weights and maybe even let the user manipulate these weights for certain tasks is an open question and the matter of ongoing research in the field of machine learning and artificial intelligence.

4 Interoperability

Establishing and maintaining a formal knowledge base for an entire scientific discipline is a formidable task and almost impossible to achieve within any individual project if there were no ways of sharing this knowledge between different applications and from different domains. It is one of the most promising visions in the context of the semantic web to make this kind of knowledge transfer available in a standardized way via the web. To this end, the world–wide–web consortium `http://www.w3.org` has issued a recommendation to establish the *Web Ontology Language OWL* as a semantic markup language for publishing and sharing ontologies on the world wide web. OWL is developed as a vocabulary extension of RDF (the Resource Description Framework) and is derived from the DAML+OIL web ontology language (which is the current semantic markup language for web resources).

NIP's classification system is designed to provide import and export support of ontologies using OWL. This way it shall become possible to easily share ontologies between different sites and applications in an evolvable but controlled manner.

5 Outlook

As NIP's classification system described here is currently a pilot study, expect it to be developed further. You are invited to browse the site and to join the portal community to actively contribute content, keywords, relationships, and therefore knowledge. The more you do so, the more useful the portal will be to the whole community.

References

[1] Organization for Economic Co–operation and Development, *Final report of the OECD Global Science Forum: Working Group on Neuroinformatics* (OECD, Paris, 2002).

[2] R. Ritz, R. Förster and A.V.M. Herz, Facilitating data and software sharing in the neurosciences – a neuroinformatics portal, in: R. Kötter, ed., *Neuroscience Databases: A Practical Guide* (Kluwer, Norwell, MA, 2003) 293–306.

[3] R. Ritz, R. Förster and A.V.M. Herz, Internetplattform Neuroinformatik: A Pilot Study for the OECD Neuroinformatics Portal, Neurocomputing, 52–54 (2003) 335–340.

Note to the referees for CNS:

In order to be as timely as possible with our contribution to the CNS meeting we describe here a feature of our portal that at the time of this writing is only available at our development server which is not publicly available. Incorporating it into the public site is scheduled for early February together with a major revision of the site's design and user interface.

This is also the reason why as of now there is no screen shot illustrating the user interface but this is planned for the final version.