A Pilot Study for the OECD Neuroinformatics Internet Portal and Related Activities

Raphael Ritz, Rainer Foerster, and Andreas Herz

Institute for Theoretical Biology

Humboldt-University Berlin, Germany

r.ritz@biologie.hu-berlin.de

BACKGROUND AND MOTIVATION

Novel experimental and computational techniques have led to major transitions in the neurosciences, all the way from the molecular to the system level. At the same time, more and more scientists now share their experimental data, analysis tools and computer models and have thus started a new research culture reflecting the open source philosophy. However, many of the data and computer programs already publicly available are not known to the general neuroscience community and information is often difficult to locate.

The Working Group on Neuroinformatics of the OECD Megascience Forum has identified this as a major deficit. To overcome the problem, the Working Group has issued a proposal to create Neuroinformatics Portals, i.e., internet based global knowledge management systems for all data relating to nervous system structure and function. In November 2000 the German Federal Ministry of Education and Research started a three year pilot project to help in jump-starting portals. This project is hosted by the Institute for Theoretical Biology at the Humboldt-University Berlin.

TASK

Our long-term goal is to contribute to a global internet portal for the entire field of neuroscience with a particular emphasis on facilitating the exchange of data and software but also providing various other kinds of information, such as "who is doing what and where" or services like news, bulletin boards and threaded discussions. The current state of our project can be checked at http://www.neuroinf.de

We do not plan, however, to serve as a primary data repository where people can upload and publish raw data sets for instance. All we are interested in is the collection and dissemination of so-called metadata, meaning structured data about data. People can use the portal to advertise their resources which are already available somewhere else on the internet. The potential benefit of doing so is to become more visible to the community.

CHALLENGES

All major community sites on the internet are facing the following challenges:

1. How to get all the relevant information into the underlying database and how to get it updated?
2. How to make the website reflect this information in a consistent manner?
3. How to assure a certain quality?
4. How to structure the site and allow for rapid and goal-directed searches?
5. How to establish interoperability with other websites?

None of these problems can be solved by the traditional static HTML-programming approach as no one person or team is able to keep up with all the additions and changes necessary once a site has grown sufficiently. Solving these problems requires a well-designed structure of the portal and efficient organization of data manipulation. The following section describes some of the approaches that we have taken or are currently considering.

SOLUTIONS

Getting the Data: Community Involvement

The best way to get at all the information necessary for making the site useful is to get the community involved. Ideally, everybody should be able to provide whatever information he or she wants to share, including individual information (who is doing what where) as well as information about software tools, data sets or any other topics of potential interest to the community. If the person who provides a certain piece of information can also be made responsible for keeping it up-to-date, chances are high that the site will not be outdated soon. In addition, we are taking efforts to deploy automatic techniques to detect if there are updates on the original sites.

Up-to-date Pages: a Dynamic Site

Having the users provide information is one important issue, having the actual display of the site reflecting the 'current knowledge' is another. To achieve this, the site has to be dynamic, which means that all the different pieces of information need to be stored and updated in a database. Most web pages of our site are therefore not hard-coded using HTML but - on request - they are dynamically generated from the database.

Quality Control: Define Workflow

If everyone were able to provide content without any editorial control, the site might easily be overwhelmed by irrelevant or inflammatory items. There needs to be a way of assuring a certain quality of the information published. We therefore include a workflow engine to support a web-based review process very similar in spirit to the traditional scientific review process - but it is not anonymous.

Any data submitted for view will be checked by one of the site reviewers or managers for consistency and completeness before it is made public. This means that we are not peer-reviewing the proposed resource according to scientific standards. This is left to the community by commenting on it. We are just checking the (meta-) data provided through the user interfaces of the portal.

Structuring the Site: Classification

The fourth challenge listed above shall be overcome by appropriate classification of each item at the time of entry (i.e., by the member who submits it). To this end we implemented a predefined classification scheme based on a controlled vocabulary. The scheme consists of six hierarchically organized categories to select the species, the anatomical structure, the neural system, the phenomenon or behavior under study, the experimental condition, and the method used.

Interoperability: Collect and Expose Standard Metadata

At the heart of automated data sharing between different web sites is the consequent and standardized usage of structured data about data. Such data about data are also called metadata. The de-facto standard for the most basic metadata is the "Dublin Core" element set, but more specialized sets of elements including rules for their potential values (controlled vocabularies) still need to be developed.

One crucial step in enabling interoperability between different web sites is a standardized way to expose and access the metadata. The "Open Archives Initiative" (OAI) developed a "Protocol for Metadata Harvesting" for this purpose. We propose to follow this approach in order to establish interoperability at the level of metadata at least between all relevant sites within the neurosciences.

TECHNICAL IMPLEMENTATION

The site is implemented as a so-called dynamic site using open source and original, self-developed software. We use the web-application server "Zope" together with its content management framework (CMF) and "Plone" behind an "Apache" web server on a "Linux" server. Zope and all its components are programmed in "Python" except for a few performance critical parts which are implemented using C.

The databases used are Zope's own Zope Object DataBase (ZODB), a persistent and transactional object database that can also be used on its own and a relational database where appropriate.

RELATED ACTIVITIES

A Spin-off from the Portal Pilot Study: LabTools

LabTools are a collection of software modules to support the sharing of data from research laboratories through the web. LabTools enable individual researchers (or the head of the laboratory) to control in detail who is allowed to access which particular information or data.

This solves one of the most prominent current problems in data sharing via the web in that it enables individual laboratories to make their data accessible through the web in principle but without loosing control about who is able to view or download which data. If desired, LabTools can be setup to support the submission of descriptive metadata about individual content items to the neuroinformatics portal at http://www.neuroinf.de described above.

Technically, LabTools are a realization of a customized framework on top of Plone which in turn is built on Zope, a general object publishing environment, and its content management framework (CMF) the same framework we use to run the portal. All these software components are distributed under an open source license and can be obtained free of charge. The most up-to-date information about the project can be obtained from http://www.neuroinf.de/LabTools.

System Requirements: LabTools will install on any of the platforms that Zope supports: Windows, Mac OSX, Linux, most Unixes and Solaris. Windows 2000 requires writing to the registry in order to install the software, this may require more rights than standard users have but an administrator should experience no problems.

Requirements on the Server Side: A higher performance computer will obviously make LabTools/Plone perform better. At least 600 Mhz and 64 MB of RAM is recommended. For a base installation of LabTools and Plone about 50MB of hard drive space is required. One must also account for the object database (and potentially others) and the file repository which can grow to almost any size depending upon the amount of data stored.

Requirements on the Client Side: LabTools only require a web browser to access the server. If users want to login, cookies must be enabled. JavaScript is not required but will provide a richer user experience. For best user experience a recent browser is recommended but also with older browsers the system is fully functional. It might just look different from the originally intended view.

Obtaining and installing the Software: All components needed for LabTools can be downloaded easily from the web. Zope is available at http://www.zope.org/Products. The distribution includes Python so there is no need to get it separately. The CMF can be obtained from http://cmf.zope.org/download and Plone is hosted at http://plone.org/download. For Windows and Mac users the Plone community provides standalone installers. The standalone installers include everything one needs to get Plone up and running. This is the best choice for people that are new to Plone and Zope. To get and install LabTools itself follow the

instructions on http://www.neuroinf.de/LabTools. It should not take more than about half an hour to get the full system up and running from scratch.

Selected Activities from Third Parties

If the "International Neuroinformatics Coordinating Facility" as proposed by the "Working Group on Neuroinformatics" of the "OECD Global Science Forum" gets established, this will boost international collaboration in the field. Already today, countries are making arrangements for national neuroinformatics nodes and related activities. And it is now, that collaborations between those early initiatives can be very effective and most helpful. This is the more so because of the common goal to establish global interoperability standards for the field of neuroinformatics.

Here we mention just two concrete examples for such collaborations that we are currently involved in: Together with the Polish neuroinformatics initiative we develop a web service interface for remotely querying neuroinformatics portal servers in an automated, unsupervised way. Together with Italy, we work on bridging the gab from current web technology to the GRID world in order to make distributed computing and data storage available to the community as easily and conveniently as possible.

OUTLOOK

While changing the attitude many scientists have with respect to 'their' data is partly a cultural challenge there are also technical challenges: To realize the potential of neuroinformatics to re-connect and integrate neuroscience researchers and their work, we need to implement a connected network of databases and tools that offers a systematic coverage of neuroscience, in all its genetic, molecular, cellular, local circuit, systems, and behavioral aspects, and in all the species of central interest for neuroscientists.

The portal currently helps in disseminating information about these data, databases and available software. But the real task on top of this is to make these databases and software tools interoperable. Ideally, it will be possible to seamlessly search through a large set of databases from the molecular to the organism level and to apply the rich repertoire of analysis and modeling tools - first steps are taken but obviously there is still a long way to reach this goal.

Acknowledgment: Supported by the German Federal Ministry for Education and Research (BMBF).