Internetplattform Neuroinformatik

A Pilot Study for the OECD Neuroinformatics Portal

Raphael Ritz, Rainer Förster, and Andreas Herz
Innovationskolleg Theoretische Biologie, Humboldt–Universität zu Berlin
Invalidenstr. 43, 10115 Berlin, Germany

Abstract - Following the open source philosophy, more and more neuroscientists are willing to share their primary data as well as custom software with the scientific community. To facilitate this interaction, we want to establish a website that links publicly available resources (like experimental data, numerical tools, computer models) and contains extensive annotation. Special emphasis is given to high quality standards of the linked databases and software tools and to integrate both users and providers in this process. The site also contains information about research groups, ongoing activities, links to re- and preprint servers and other interesting resources. In addition, we are currently exploring web services to see where they could advance the field of neuroinformatics.

I. 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 major deficit in a recent progress report [1]. To overcome the problem, the Working Group has issued a proposal to create the "Neuroinformatics Portal", an internet based global knowledge management system 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 the portal. This project is hosted by the Innovationskolleg Theoretische Biologie at the Humboldt–Universität zu Berlin.

Already in 1999, the European Union established the *Thematic Network Computational Neuroscience and Neuroinformatics*. The primary objective of the thematic network is to promote the fields of computational neuroscience and neuroinformatics and to start building a European neuroinformatics infrastructure. One of the initiatives within the thematic network was to establish a website that contains a database of ongoing activities and research groups in the fields of neuroinformatics and computational neuroscience. This project is hosted by the Laboratory of Theoretical Neurobiology at the University of Antwerp and started the website `http://www.neuroinf.org`. It provides the international neuroinformatics community with information and access to courses and workshops and hosts the comp–neuro mailing list.

By the end of 2000, the two initiatives decided to join forces and to develop a *Joint Neuroinformatics Internet Portal*. Here now we describe the contribution of the Berlin group.

II. THE TASK

Our current task is to contribute to the enhancement of the `www.neuroinf.org` site to become 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, like, e.g., *who is doing what and where* or services like news and bulletin boards, threaded discussions and the like. The current state of our project can be checked at `his.biologie.hu-berlin.de`.

III. THE PROBLEMS

All major community sites on the internet are facing the following severe problems:

1. How to get all the relevant information in and updated?
2. How to keep the displayed information consistent and up–to–date?
3. How to assure a certain quality?
4. How to structure the site and allow for rapid and goal–directed searching?

In the long run we also need to address additional issues:

5. How to establish interoberability with other websites?
6. How to provide computational services through the web?

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 approaches that we have taken or are currently considering.

IV. A SOLUTION

The way we envision to solve these problems is the following; let's consider each in turn:

A. Solution to 1.: Get the Community involved

The best way to get at all the information necessary for making the site useful is to get the community involved. Ideally, everybody who would like to should be able to provide whatever information he or she wants to share, including personal information (who is doing what where) as well as information about software tools, data sets or any other things 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 that the site won't be outdated soon. Therefore, the software implementing the site makes it easy to contribute.

B. Solution to 2.: Build a Dynamic Site

Having the users providing information is one thing, 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 bits and pieces of information need to be stored and updated in a database. Most particular webpages are not hard–coded using HTML but they are dynamically generated from the database on request.

C. Solution to 3.: Define Workflow and Roles

If everyone would be able to enter or upload and provide anything without any further control, the site could easily be overwhelmed by irrelevant or inflammatory items. There needs to be a way of assuring a certain quality of the information published. Therefore we implemented a web–based review process very similar in spirit to the traditional scientific review process (but it is not anonymous).

In practice this means that we first want to know who you are, before you are allowed to enter anything into our site. Therefore one needs to register before submitting data. Any data submitted for view will be reviewed electronically before it becomes public.

D. Solution to 4.: Classification

The fourth problem listed above shall be overcome by appropriate classification of each item at the time of entry (i.e., by the member who submits it). The most obvious way of classifying, namely through a number of freely chosen keywords, would not work because it would not allow to automatically group and relate different entries. We therefore implemented a predefined classification scheme based on a controled vocabulary. The scheme consist 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. Through selection from each menu a string of six keywords gets constructed. This is the basic classification step. Each item can be classified using multiple such strings. Based on this classification it now becomes possible to selectively group the content either for specific listings within the portal directly or in response to a (guided) search of the site.

E. Solution to 5.: Expose Metadata

At the heard of automated data sharing between different web sites is the consequent and standardized usage of structured data about data or metadata for short. In 1994 a group of experts on this issue met in Dublin, Ohio and developed what is now refered to as the *Dublin Core*: a simple yet effective element set for describing a wide range of networked resources. The Dublin Core standard comprises fifteen elements, the semantics of which have been established through consensus by an international, cross–disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields of scholarship.

The Dublin Core element set is outlined in Table 1. Each element is optional and may be repeated. Each element also has a limited set of qualifiers, attributes that may be used to further refine (not extend) the meaning of the element. The Dublin Core Metadata Initiative [2] has defined standard ways to "qualify" elements with various types of qualifiers. A set of recommended qualifiers conforming to DCMI "best practice" is available, with a formal registry in process [2].

The crucial step in enabling interoperability between different web sites now is a standardized way to expose

TABLE I
Dublin Core Metadata Element Set

Name	Definition
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content of the resource
Subject	The topic of content of the resource.
Description	An account of the content of the resource.
Publisher	An entity for making the resource available.
Contributor	An entity responsible for making contributions to the content of the resource.
Date	A date associated with an event in the life cycle of the resource.
Type	The nature or genre of the content of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A reference to a resource from which the present resource is derived.
Language	A language of the intellectual content of the resource.
Relation	A reference to a related resource.
Coverage	The extend or scope of the content of the resource.
Rights	Information about rights held in and over the resource.

and access the metadata. One way how to do so has been outlined by the *Open Archives Initiative* [3] who developed a *Metadata Harvesting Protocol* [4] for this porpuse. Generaly they destinguish between *data providers* who administer systems that support the OAI protocol as a means of exposing metadata about the content in their systems and *service providers* who issue OAI protocol requests to the systems of data providers and use the returned metadata as a basis for building value–added services. From a more general perspective this is just one particular example of a *web service* which will be introduced now.

F. Solution to 6.: Web Services

The next step in the development of the World Wide Web is to connect computational capabilities (effectively software and CPU time) to the web. The efforts in this context are now typically refered to as *Web Services* [5]. Web services are a standard for interfaces between applications and content services on the internet. The key design principle is that one web service can be called from another web service to use the methods provided by the first as if it were a locally integrated modul even though the services may reside at different sites and be implemented on different software platforms (operating systems, programming languages, etc.). It works because the interface to each and every service is designed according to a well defined standard enabling the service to be called from a third party without knowing the underlying software infrastructutre.

This implies a general solution to the problem of integrating different (even proprietary) systems. The impact of this new evolving distributed computing paradigm is potentially huge and might drastically change the way we will do computations in the future.

V. IMPLEMENTATION

A. Dynamic Site

The site is implemented using open source and original, self–developed software. We use the web–application server Zope [6] together with its content management framework [7] behind an Apache web server [8]. Zope and all its components are programmed in Python [9] 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 [10] and a relational database where appropriate.

To improve the general accessibility we plan to globally mirror the site. These mirror servers will be synchronized using Zope's Enterprise Objects [11].

B. Exposing Metadata

Metadata encoding based on Dublin Core can be done in several different sytaxes, including HTML and RDF/XML. The Resource Description Framework (RDF; [12]) allows multiple metadata schemes to be read by humans as well as parsed by machines. It uses the eXtensible Markup Language (XML; [13]) to express structure thereby allowing metadata communities to define the actual semantics. This decentralized approach recognizes that no one scheme is appropriate for all situations, and further that schemes need a linking mechanism independent of a central authority to aid description, identification, understanding, usability, and/or exchange.

RDF allows multiple objects to be described without specifying the detail required. The underlying glue, XML, simply requires that all namespaces be defined and once defined, they can be used to the extent needed by the provider of the metadata.

For example:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/
          22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description
     rdf:about="http://www.somewhere.de">
    <dc:creator>Raphael Ritz</dc:creator>
    <dc:title>Internetplattform Neuroinformatik
    </dc:title>
    <dc:description>Users guide for the
     neuroinformatics portal.</dc:description>
    <dc:date>2002-01-24</dc:date>
  </rdf:Description>
</rdf:RDF>
```

could describe a users guide to our portal. With XML and RDF, Dublin Core can now be mixed with other metadata vocabularies. For example, the simple Dublin Core description above might be used alongside other vocabularies such as vCard [14] that can describe the author's affiliation and contact information, or a more specialised "neuroinformatics" vocabulary that described the scientific scope in greater detail.

C. Web Service Standards

The success of the internet is mainly based on standardized protocols (rather than any particular implementations) enabling different technologies to talk to each other.

To enable now the integration of different applications, web service standards have been developed (by Microsoft, IBM, Sun and others in the XML standard protocol commission) and have been submitted to the W3C [15] for recommendation.

The base protocol is called SOAP (Simple Object Access Protocol; [16]). It enables the core transmission of objects between different sites over HTTP or SMTP (i.e., the web or e–mail). It is a W3C recommendation (i.e., web standard), it is supported by all major software vendors and the open source community, and there are implementations for Java, Apache, WebSphere, Visual Basic, etc. already available.

To standardize the formal description of interfaces WSDL (Web Services Description Language; [17]) has been developed. The description has to include all the information necessary (supported protocols, address and port number, procedures and functions defined including their supported input and output formats).

To combine different web services in one work flow WSFL (Web Services Flow Language; [18]) has been developed. The WSFL enables developers to easily create, execute, and combine web services into complex applications.

To find a particular web service, a registration system called UDDI (Universal Discovery, Description, and Integration; [19]) has been established. The UDDI directory can be accessed manually (by software developers) as well as by applications at runtime.

VI. OUTLOOK

Given the motivation described at the beginning of this article the portal introduced here is only a first step toward the more ambitious goal of inducing a paradigm shift within the neurosciences toward data and software sharing.

While changing the attitude people have with respect to "their" data is partly a cultural challenge there are also further technical challenges: To really 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 all these data, databases and software available. But the real task on top of this is to make these databases and software tools interoperable. In the future it should be possible to seamlessly search through all of these databases and then apply whatever kind of analysis tool you like to whatever data set you have found that way gaining new insights into nervous system function.

Supported by BMBF (German Ministery for Education and Research).

References

[1] Final report of the OECD Megascience Forum: working group on biological informatics, January 1999
http://www1.oecd.org/dsti/sti/s_t/ms/prod/BIFINREP.pdf

[2] DCMI: Dublin Core Metadata Initiative
http://www.dublincore.org/

[3] OAI: Open Archives Initiative
http://www.openarchives.org/

[4] Metadata Harvesting Protocol; developed by the Open Archives Initiative
http://www.openarchives.org/OAI_protocol/
openarchivesprotocol.html

[5] A portal for web services http://www.webservices.org/

[6] Zope: A free, open source web–application server
http://www.zope.org

[7] CMF: Zope's Content Management Framework
`http://cmf.zope.org`

[8] Apache: A free, open source web server
`http://www.apache.org`

[9] Python: A free, open source scripting language
`http://www.python.org`

[10] ZODB: A free, open source object database management system
`http://www.zope.org//Wikis/ZODB/StandaloneZODB`)

[11] ZEO: Zope's Enterprise Objects; a replicated storage server
`http://www.zope.org/Products/ZEO`).

[12] RDF: Resource Description Framework
http://www.w3.org/RDF/

[13] XML: eXtensible Markup Language
http://www.w3.org/XML/

[14] vCard: an electronic business card by the *Internet Mail Consortium*
http://www.imc.org/

[15] W3C: the World Wide Web Consortium
http://www.w3.org/

[16] SOAP: Simple Object Access Protocol
http://www.w3.org/TR/SOAP

[17] WSDL: Web Services Description Language
http://www.w3.org/TR/wsdl

[18] WSFL: Web Services Flow Language
http://www-106.ibm.com/developerworks/webservices/library/ws-ref7/

[19] UDDI: Universal Discovery, Description, and Integration
http://www.uddi.org/