neuroinformatics, knowledge bases

Grzegorz BLINOWSKI[*], Piotr DURKA[**],
Artur SPASIŃSKI[***]

INTER-NEURO: FROM CHAOS TO NEUROINFORMATICS KNOWLEDGE BASE

Almost unlimited possibilities of sharing neuroinformatics resources, opened by the Internet, create an almost unlimited number of issues. Growing amount of available data, combined with the lack of reliable and large enough metainformation resources, limits the proliferation and reliability of this media. In this paper we propose a solution, which may help in an efficient sharing of neuroinformatics resources, by means of a network of vortals dedicated to particular and well defined topics. These vortals are responsible for collection of high quality resources in their particular fields. They are interconnected in a way transparent to the user, using a low level interface for interchanging queries. For a user this means that a query entered in one of the vortals will return relevant results found also in the other vortals of the Network. We also describe technical details and pilot implementation; metainformation is based upon Open Archives/ Dublin Core standards, and interchange of queries on XML/SOAP.

1. INTRODUCTION

Internet becomes the major media for interchange of scientific information records. It allows for an instant publication of scientific results for a potentially unlimited audience. It overcomes the greatest drawback of the traditional, “paper” publishing scheme, which is the growing average time from submission to publication. Another extremely important feature is the possibility of publishing not only text and figures found in the paper publications, but also datasets, software, models and all relevant files. According to the advocates of Reproducible Research, “an article about computational science in a scientific publication is **not** the scholarship itself; it is merely **advertising** of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures” [1][7].

Therefore, sharing algorithms and datasets is a must for any coherent progress in biomedical sciences. However, the sole availability of all these information does not yet imply its usability, because the difficulty of finding relevant information is proportional to

[*] Warsaw University of Technology, Institute of Computer Science, Nowowiejska 15/19, 00-665 Warsaw, Poland, e-mail: GBlinowski@ii.pw.edu.pl

[**] Warsaw University, Institute of Experimental Physics, Laboratory of Medical Physics, Hoża 69, 00-681 Warszawa, Poland, e-mail: durka@fuw.edu.pl

[***] CC Otwarte Systemy Komputerowe sp. z o.o., ul. Rakowiecka 36, 02-532 Warszawa, http://www.cc.com.pl/, e-mail: Artur.Spasinski@cc.com.pl

the volume of potentially accessible files. Relatively huge contribution of an information noise is trivially exemplified by a results returned by the search engines. On the other hand, more specific terms may return no hits if relevant resources are hidden in specialized databases.

This problem could be in theory solved by a centralized storage of resources, as e.g. one of the largest open biomedical repositories physionet.org [3].

However, even this initiative, with an impressive history dating back to the times of interchange of computer tapes, cannot efficiently cover all the important topics, growing quickly on the borders of established fields—like the neuroinformatics and other biomedical sciences. Therefore, specialized repositories, databases and Web pages are being dedicated to many interrelated topics. But there are no mechanisms which would allow relating and efficiently searching these resources. In this paper we propose such a solution, based exclusively on Open Source tools, and with freely available source code.

2. INTER-NEURO

Fragmentation of neuroinformatics resources, resulting from the spontaneous creation of services dedicated to narrow topics, is per se a positive phenomenon. Owing to a more or less well defined scope, such initiatives can gather relevant and high quality information. However, the main strength of the interdisciplinary research lies in combining knowledge, techniques and algorithms from sometimes distant fields. Therefore, to allow for an efficient use of these resources, we need an efficient and simple way to search and relate them. Firstly, it requires a certain amount of metainformation about the available items. Secondly, we need an efficient algorithm to search all these resources, in a simplest possible way from the user's point of view. Two following sections present propositions of such solutions, based upon Open Archives/Dublin Core standards for metainformation and XML/SOAP for interchange of queries.

3. METAINFORMATION SCHEME

3.1. SEMANTIC AWARE SEARCH

The major difference between semantic-aware search (i.e. search with meta information) and “ordinary” search - such that is provided by typical Internet-wide search engines (for example Google) is that the former indexes not only simple keyword data by also the meaning of the data. As an example consider search for a (any) book by “Mark Twain”: typing this phrase in a search engines' search text input will yield a lot of documents about the author, possibly also documents about other people who happen to have the same name, among thousands of documents there will be also literary works of Mark Twain – however there is no simple way to separate them from other returned content. The problem of finding precise and relevant content is solved with semantic-aware search:

like ordinary search engine the search engine indexes documents but unlike ordinary ones it is aware of metainformation, which includes data about: author, creator (note, that the author is not necessarily the creator (or owner) of the document), title, major keywords, references, etc. All the meta information content is stored together with plain whole text keyword index.

The choice of meta information is not trivial, fortunately standards exist which regulate naming and scope of meta information attributes. On of the most popular standards in this field is the Dublin Core standard (DC). The DC specification is developed and maintained by "The Dublin Core Metadata Initiative" (DCMI) an "open forum engaged in the development of inter operable on line metadata standards that support a broad range of purposes and business models.". The full specification of the DC standard may be found in [8]. Here we will summarize only the most important elements of the DC metadata:

- Type - "The nature or genre of the content of the resource" – this may be a text (paper, article, preprint); a software item (i.e. a description of a freeware or commercial software piece); a dataset (i.e. an experiment collected time series in a well know format)
- Title - "A name given to the resource", e.g. in case of a paper – its title
- Identifier - "An unambiguous reference to the resource within a given context", the identifier does not have to have a sensible meaning to a human being – it is simply a unique token identifying the resource – e.g. an URL
- Creator - "An entity primarily responsible for making the content of the resource" - i.e. - a person, an organization, or a service
- Description - "An account of the content of the resource" - abstract, table of contents, reference, etc.
- Subject - "The topic of the content of the resource " - keywords, key phrases, classification codes that describe the resource.

DC defines also a handful of other attributes which include: time&date information, information about the publisher, more data about the content itself, etc.

Having adopted and agreed upon the DC standard as an universal way of description we can build a sophisticated distributed search mechanism around it. With meta information standardized there is no longer an issue of "what to search for?" only an issue of "how to search?" (technically) remains.

3.2. INTERCHANGE OF QUERIES - BUILDING THE DISTRIBUTED SEARCH

In the Inter-neuro initiative we have adopted the SOAP/RDF XML [9] based standards for describing queries and results, in consequence the HTTP protocol [2] is used for transporting the query and the response over the network.

The search service is build around the distributed P2P paradigm – each actor (i.e. portal or web site taking part in the initiative) is both a client and a server – it is able both: to formulate and send the queries as well as listen for search requests and to answer them.

The rationale for using SOAP/RDF/XML is the following:

- SOAP/XML is portable and both platform and system independent
- SOAP/XML and SOAP over http are a de-facto standards for building distributed applications
- SOAP is simple - there is no heavyweight software required to generate and parse it
- There is a multitude of XML parsers and tools available (both commercial and open-source) so building software compatible with our format should not be a technical problem

The distributed query is executed as follows (the process is shown on figure 1):

(1) User enters the query – i.e. he connects to one of the initiative sites (e.g. http://eeg.pl/) chooses "advanced search", enters the search phrase(es), marks the "external search" check box, and clicks the search button

(2) query is translated into universal format (SOAP/XML) and sent to all participating sites

(3) each site executes local query

(4) each site returns results

(5) results and aggregated and displayed to the user

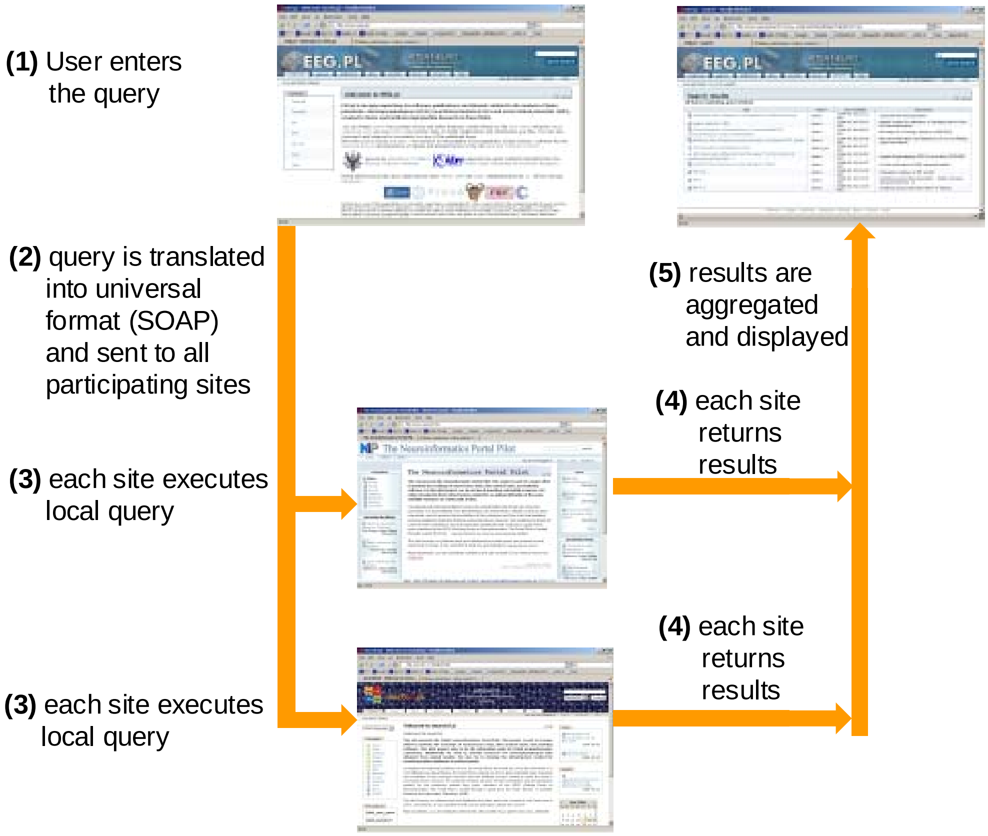
(1) User enters the query
(2) query is translated into universal format (SOAP) and sent to all participating sites
(3) each site executes local query
(3) each site executes local query
(5) results are aggregated and displayed
(4) each site returns results
(4) each site returns results
EEG.PL
EEG.PL
The Neuroinformatics Portal Pilot

Fig. 1. Distributed searching with Inter-neuro

4. TECHNICAL DETAILS – QUERY AND RESULTS

SOAP (Simple Object Access Protocol) [6] is a stateless, message exchange paradigm based on XML. In simpler terms – SOAP is a mechanism similar to RPC (Remote Procedure Call) based on open standards: the remote object access (or a “procedure call”) is express purely in XML notation, the same applies to returned results. A SOAP message consists of an outermost envelope, an optional header and body. From the logical point of view the body consists of a remote objects' (or procedures') identifier and parameters. The SOAP standard describes how parameters should be represented, serialized and encoded. SOAP defines both a method for encoding simple types (strings, integers, etc) as well as complex types such as arrays and structures. In case of the remote search employed in Inter-neuro a relatively simple query is used: only string type parameters representing DC attributes are passed - see fig. 2.

```
<?xml version="1.0"?>
<SOAP-ENV:Envelope xmlns:SOAP-
ENV="http://schemas.xmlsoap.org/soap/envelope/" ...>
  <SOAP-ENV:Body>
    <NeuroQuery>
       <QueryTypexsi:type="xsd:string">Software </QueryType>
      <FullTextQuery xsi:type="xsd:string">
        some pattern here
       </FullTextQuery>
      <search xsi:type="SOAP-ENC:Array"
       SOAP-ENC:arrayType="ns1:searchcrit[3]">
        <item>
          <pname>DC:creator</pname>
          <pvalue>regexp</pvalue>
        </item>
        <item>
          <pname>DC:title</pname>
          <pvalue>regexp</pvalue>
        </item>
        <item>…</item>
      </search>
      <searchlogic xsi:type="xsd:string"> AND </searchlogic>

      <datebeg xsi:type="xsd:string"> 2002-01-01</datebeg>
      <dateend xsi:type="xsd:string"> 2004-01-01</dateend>
    </NeuroQuery>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Indicates the DC object type: Software ; Dataset ; etc.

This component is for "full-text" search

The third component specifies universal "by-DC-attribute" search

and so on for other DC attributes

either AND or OR

Date conditions further limit the search scope

Fig. 2. The SOAP query

The result is generated as and RDF record serialized (encoded) in SOAP response - see fig. 3. RDF - Resource Description Framework [5] is a language for representing information about resources in the World Wide Web. RDF similarly to SOAP is based on XML. It is particularly intended for representing metadata about Web resources, such as the title, author, and modification date of a Web page. RDF is intended for situations in which information needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning - hence it is ideally suited for our application.

```
<?xml version="1.0"?>
<SOAP-ENV:Envelope xmlns:SOAP-
ENV="http://schemas.xmlsoap.org/soap/envelope/"
      ...>

<SOAP-ENV:Body>
      <NeuroQueryResponse>
            <statusxsi:type="xsd:int">0
      </status>
            <resultsxsi:type="SOAP-ENC:Array"
             SOAP-ENC:arrayType="ns1:res[3]">
                  <item
      rdf:about="http://www.eeg.pl/somepaper">
                   <title>A fine paper
                     on EEG</title>
                    <dc:date>2003-06-23</dc:date>
                    <dc:title>Analysis of EEG
                     signas</dc:title>
                    <dc:description>some info
                     here</dc:description>
                  ...
                  </item>
             ...
            </results>
      </NeuroQueryResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

general status, e.g. 0 - OK, <0 - error
because more than one record may be returned an SOAP array is used here
First result tuple

more attributes here

more result tuples

Fig. 3. The RDF response

5. THE IMPLEMENTATION

Our implementation is based on the Zope/CMS/Plone [4] free application server / content management / portal engine. Although Zope/Plone provides some mechanisms for distributed communication between different sites (RPC-over-XML) it currently lacks SOAP/RDF support as such. We have used ZOPE's template mechanisms and programming capabilities to develop a distributed search component. The software is written in Python (a default development language for ZOPE, in which the whole system is actually written) and freely available as ZOPE package (technically ZOPE “product”). Software components are freely available at http://eeg.pl.

6. CONCLUSIONS

We presented a possible solution to the major problem of information noise, which sometimes overweights advantages of the Internet in scientific communication. Our solution lies in between the two extremes of the absolute centralization and a complete decentralization. Disadvantages of one central repository of information are obvious, but, on the other hand, Semantic Web and super-intelligent software agents, creating structure from the chaos, are still rather buzzwords than reality. We propose a humble compromise. As in the presented example of the Inter-neuro initiative, relevant information should be gathered in specialized repositories of possibly well defined scope. Owing to this specialization, these relatively small services can assure the quality and proper annotation of resources. Seamless integration of these small repositories into a significant knowledge base can be effectuated by the paradigm presented in this paper. More technical details and a complete software implementation of this solution are freely available from http://eeg.pl.

BIBLIOGRAPHY

[1] BUCKHEIT, J.B. and DONOHO, D.L. (1995) Wavelab and reproducible research.in A. Antoniadis (ed) Wavelets and statistics, Springer-Verlag 1995 (also: http://citeseer.ist.psu.edu/buckheit95wavelab.html)

[2] FIELDING R., GETTYS J. et al., RFC 2616 - Hypertext Transfer Protocol – HTTP/1.1, Jun. 1999, http://www.rfc-editor.org

[3] GOLDBERGER AL, Amaral LAN, et al., PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 (also: http://circ.ahajournals.org/cgi/content/full/101/23/e215); 2000 (June 13).

[4] LATTEIER A., et al., The Zope Book (2.6 edition), 2003, http://zope.org/Documentation/Books/ZopeBook/2_6Edition/ZopeBook-2_6.pdf

[5] MANOLA F., MILLER E. (eds), RDF Primer (W3C Recommendation), 10 Feb. 2004, http://www.w3.org/TR/rdf-primer/

[6] MITRA N. (ed.), SOAP Version 1.2 Part 0: Primer (W3C Recommendation), 24 Jun. 2003, http://www.w3.org/TR/2003/REC-soap12-part0-20030624/

[7] SCHWAB, M.; KARRENBACH, N.; CLAERBOUT, J. Making scientific computations reproducible in Computing in Science & Engineering, Vol. 2, Issue6, Nov.-Dec. 2000, p.61-67 (also: http://sep.stanford.edu/research/redoc/cip.html)

[8] WEIBEL S., KUNZ J., et al., Dublin Core Metadata for Resource Discovery, http://www.ietf.org/rfc/rfc2413.txt

[9] W3 Consortium, Extensible Markup Language (XML), 2004, http://www.w3.org/XML/#intro

This work was partially supported by Polish State Committee for Scientific Research grant #4438/IA/115/2003.