Book Title
Book Editors
IOS Press, 2003

Sealife: A Semantic Grid Browser for the Life Sciences Applied to the Study of Infectious Diseases

MichaelSchroeder [a,1], Albert Burger [b], Patty Kostkova [c], Robert Stevens [d], Bianca Habermann [e], and Rose Dieng-Kuntz [f]

[a] *TU Dresden, Germany*
[b] *Hariot-Watt University, Edinburgh, UK*
[c] *City University, London, UK*
[d] *University of Manchester, UK*
[e] *Scionics, Dresden, Germany*
[f] *INRIA, Sophia-Antipolis, France*

Abstract. The objective of Sealife is the conception and realisation of a semanticgGrid browser for the life sciences, which will link the existing Web to the currently emerging eScience infrastructure. The SeaLife Browser will allow users to automatically link a host of Web servers and Web/Grid services to the Web content he/she is visiting. This will be accomplished using eScience's growing number of Web/Grid Services and its XML-based standards and ontologies. The browser will identify terms in the pages being browsed through the background knowledge held in ontologies. Through the use of Semantic Hyperlinks, which link identified ontology terms to servers and services, the SeaLife Browser will offer a new dimension of context-based information integration.

In this paper, we give an overview over the different components of the browser and their interplay. This SeaLife Browser will be demonstrated within three application scenarios in evidence-based medicine, literature & patent mining, and molecular biology, all relating to the study of infectious diseases. The three applications vertically integrate the molecule/cell, the tissue/organ and the patient/population level by covering the analysis of high-throughput screening data for endocytosis (the molecular entry pathway into the cell), the expression of proteins in the spatial context of tissue and organs, and a high-level library on infectious diseases designed for clinicians and their patients.

For more information see http://www.biote.ctu-dresden.de/sealife

Keywords. Grid computing, bioinformatics, ehealth, semantic web, textmining, ontologies

1. Introduction

Currently, much effort is being spent on creating a new computational and data infrastructure to facilitate eScience, the cooperation of geographically distributed organisa-

[1]Correspondance to: Michael Schroeder, Biotec, TU Dresden, ms@biotec.tu-dresden.de, +49 351 46340060

tions, which transparently integrate their computational and data resources at a structural and semantic level. Progress has been made with standards for grid computing and semantic representations for life science data with many projects creating a host of grid-enabled services for the life sciences.

How can the researcher in the lab benefit from this new infra-structure to science? A technology is needed to transparently bring such services to the desks of the scientists. The Web started with a browser and a handful of Web pages. The vision of eScience with an underlying Grid and Semantic Web will only take off with the development of a Semantic Grid browser that gives a user easy access to Grid and e-Science resources.

The Sealife project is filling this gap by developing such a semantic grid browser. These browsers will operate on top of the existing Web, but they introduce an additional semantic level, thus implementing a Semantic Web.

Using ontologies as background knowledge the browsers can automatically identify entities such as protein and gene names, molecular processes, diseases, types of tissue, etc. and the relationships between them, in any Web document, they collect these entities and then apply further analyses to them using applicable Web and Grid services.

If the user points the mouse at a Semantic Hyperlink the SeaLife Browser offers a definition of the encountered term, the application of services relevant to the term, and to add the term to a shopping cart. After browsing through various pages and adding various terms to the shopping cart, the user decides to check out. The SeaLife Browser presents the contents of the shopping cart including the list of items collected, the type of the identified terms, and the sources where they were collected by the user.

The SeaLife Browser offers to apply additional services considering combinations of terms. For example, if the user collected a set of proteins, then the browser will offer to apply a tool to compare the proteins' sequences against each other, to create a multiple sequence alignment, or to query the literature for co-occurances of the two proteins. The user can save the current state of the shopping cart and return at a later stage to continue the semantic exploration.

To summarise, the SeaLife Browser links the existing Web to the new eScience grid infrastructure paving the way for a future generation Web for the life sciences.

2. Case studies

To illustrate the power of this vision consider the following applications in the context of infectious diseases. The applications vertically integrate the molecular/cell, tissue/organ, and patient/population layers by covering high-level information stemming from the national library of infectious diseases to detailed studies of high-throughput screening data for endocytosis, the entry pathway into the cell.

- **Evidence-based medicine**: Consider a clinician, who consults the national electronic library of infections to get curated and trusted information on infections. The user visits the site and finds an interesting page on hipatitis and its treatment: "Ribavirin with or without alpha interferon for chronic hepatitis C". Using its background knowledge, the SeaLife Browser identifies hipatitis as a disease and interferon as a cytokine and immunologic factor. With this knowledge the browser automatically offers the user the ability to query Ensmbl in order to learn more about the genetics related to hipatitis and the Protein Databank to look at struc-

tures of cytokines. The browser also offers the opportunity to explore the literature further. Via the ontology the browser can either refine searches and look for interferon type I e..g., or generalise and search for liver diseases etc.

- **Literature and Patent Mining**: Getting a quick overview of a field is vital for companies to stay ahead of their competitors. In browsing a patent database, a researcher comes across the patent entitled “An improved infant formula is described which includes a phospholipid supplement in order to more closely resemble the composition of human milk.”. The SeaLife Browser identifies the term “phospholipid metabolism” and offers the following definition to the user: "The chemical reactions and physical changes involving phospholipids, any lipid containing phosphoric acid as a mono- or diester.". It also identifies human in its taxonomy. The user decides that this is relevant and wishes to learn more about phospholipids. The SeaLife Browser automatically offers the service of showing all human proteins in the UniProt database, which are involved in phospholipid metabolism.

  As another example, consider a biologist who wishes to know which enzymes are inhibited by levamisole. The researcher visits a traditional literature search engine such as PubMed to find relevant literature. PubMed returns over 100 articles for the query levamisole inhibitor. While the first articles already mention the enzyme alkaline phosphatase, there is only one article, which is ranked very low and which mentions phosphfructokinase. It is unlikely that the user would find this article. With the SeaLife Browser, the situation is different. With the background knowledge the browser identifies terms such as phosphofructokinase or alkaline phosphatase in the PubMed result page. With the ontology the browser can infer that both terms are enzymes. It can now offer to the user literature services which categorises the abstracts by enzyme activity, thus giving a direct overview of all the results and more directly answering the researcher’s question.
- **Molecular Biology**: Consider a biologist, who encountered the statement “Rabaptin-5 interacts with the small GTPase Rab5 and is an essential component of the fusion machinery for targeting endocytic vesicles to early endosomes”. The SeaLife Browser identifies “Rabaptin-5” and “Rab5” as protein names, “endocytosis” as biological process, and “early endosome” as cellular component. When the user moves the mouse over "Rab5", the SeaLife Browser offers to search sequence databases for Rab5 proteins. At the same time, it offers to move the protein sequence of Rab5 to a shopping cart. After browsing for some time, the user decides to visit his/her shopping cart and takes a look at the proteins he/she has collected in the web-session. The SeaLife Browser now offers to perform a series of services on the protein sequences in the cart.

  One simple analysis includes a domain search of the collected sequences. For Rab5, this results in the identification of a GTPase domain. A multiple sequence alignment is displayed from the domain database, which gives the user an idea about conserved residues of the respective domain. In cases where a known three-dimensional structure exists for the domain, the user invokes a molecular display tool to visualize the possible fold of his/her protein.

  Querying several databases for related information about his/her protein, like the online mendelian inheritance in man (OMIM), expression databases like SAGE, or the protein structure database (PDB) give the user information about links to

diseases, expression levels of his/her protein of interest in several tissues, as well as available structural data at one mouse-click. More sophisticated analysis tools allow the user to perform a sequence database search with his/her protein sequence in order to retrieve possible related sequence to his/her protein of interest or perform fold recognition with his/her collected items in the shopping cart, with which he/she could get an idea about the function of the proteins.

3. Aims and Objectives

In these scenarios, there is an obvious reliance on the computer having some notion of domain semantics–what is the relationship between symbols in the language of biomedicine? To achieve the above vision, the following semantic problems need to be solved:

- Ontologies: Design and integration of ontologies and associated infrastructure, which can serve as background knowledge for a Semantic Grid browser geared towards life science applications ranging from the molecular level to the person level.
- Concept Mapping: Bridging the gap between the free text on the current Web and the ontology-based mark-up for the Semantic Web and Grid by developing an automated mark-up modules for free text, which are based on text-mining and natural language processing technologies.
- Service Composition: Bridging the gap between the ontologies of the Semantic Web and the services of the Grid by linking suitable ontology mark-up to applicable services and by supporting the interactive creation of such mappings for complex services.

3.1. State-of-the-art

Current work in the Semantic Web has concentrated upon the development of languages and infra-structure. Few real Semantic Web applications have been made to date. Biology is, however, already well placed to create a Semantic Web for Life Sciences with its large Web presence and growing use of ontologies. There is as yet, no transparent, user facing, easy browser for a Semantic Web or Grid.

Stein's vision of a bioinformatics Nation[1] to bring together the distributed and heterogeneous resources of the bioinformatics community, will rely on such infra-structures as suggested by the semantic Web and Grid. In order to deliver a SeaLife Browser for biologists, target data and services are needed and bioinformatics is already well placed to do this. Many Web and Grid services are now available, with some delivering data formatted according to XML schema descriptions. Such efforts can be seen in ^{my}Grid [2], HeatlhGrid [3] and the Biomedical Informatics Research Network (BIRN)[4]. These projects, and others, bring together virtual organisations of computers, data, pro-

[1] Lincoln Stein. Creating a bioinformatics nation: A web-services model will allow biological data to be fully exploited. Nature 417(119); 9 May 2002

[2] `http://www.mygrid.org.uk`

[3] `http://www.healthgrid.org`

[4] `http://www.nbirn.net`

grammes, instruments and users to collaborate to perform *in silico* analyses and health care–that is, they form Grids in the bio-health domain.

The bioinformatics domain, in particular, is already deploying much of the necessary infra-structure for these projects. Ontologies are already widely used in describing and analysing biological data. Foremost in these is the Gene Ontology[5] and others in the Open Bio-Ontologies consortium. These provide a common language for describing, amongst other features, molecular function, biological processes and location of gene products; sequence features; description of microarray and proteomic experiments. This means large bodies of semantically marked up data already exist that could be explored by a SeaLife Browser. A growing number of these bio-ontologies are in the OWL format and the OBO formats can be represented in OWL, suggesting that this markup is in a form accessible to the proposed SeaLife Browser.

As well as data, many Web and Grid Services mean that there is now programmatic access to many bioinformatics tools. EuroGrid, for example, the Bio GRID part of EuroGrid [6] developed an access portal for biomolecular modeling resources. The Semantic Grid, like the Web itself, relies on standards such as HTML and HTTP for the original Web. The UniGrids project [7] developed standard access mechanism over both Glogus and Unicore that are compliant with the Open Grid Services Architecture (OGSA). Many of the services in the projects above use Unigrid output to support these computationally intensive services across a Grid of computational resources such as DEISA (Distributed European Infrastructure for Supercomputing Applications)[8]. DEISA is a consortium of leading national supercomputing centres in Europe that intends to jointly build and operate a distributed terascale supercomputing facility. It is such networks of computational power that will support computationally intensive analyses over eScience and eHealth data.

As well as Grid Services, projects such as myGrid, use Web Services, which are envisaged to come together with Grid Services to provide a unified access style. There are already well over 2 000 Web Services available in bioinformatics including sequence searches with BLAST (also available through EuroGrid), the major databases such as the Ensembl database for genetic and disease information, MSD for protein structures, PubMed for biomedical literature, Kegg for metabolic pathways, INTERPRO for sequence profiles, the Emboss suite and many others. Together with Grid Services, these offer programmatic access to bioinformatics tools never seen before.

Once available, these 1 000's of services need to be able to be discovered and deployed, by humans as well as computers. There is a large effort to develop frameworks for semantically describing services through ontologies. Foremost amongst these is the Web Services Modelling Ontology (WSMO) and its markup form WSML. This creates a template to describe the inputs, outputs, pre- and post-conditions and tasks supported by Web and Grid services. This kind of markup, extended to the biological domain, will enable a SeaLife Browser, and other applications, to discover tools appropriately from semantic markup in a page. Projects such as myGrid and bioMOBY [9] have already ex-

[5] http://www.geneontology.org
[6] http://www.eurogrid.org
[7] http://www.unigrids.org
[8] http://www.deisa.org
[9] http://www.biomoby.org

plored the use of semantic markup to aid discovery and composition of Web Services in bioinformatics.

Intimately linked into the role of ontologies in Semantic Grids is text-mining. Pages and services need to be marked up with semantic descriptions provided by the background knowledge provided by ontologies. In addition, the knowledge captured by ontologies need to be collected from data resources, as well as human experts. These are the twin roles of text-mining within Semantic Grids. Through techniques such as stemming and part of speech tagging, coupled with co-location, text-mining can deliver the terms, their synonyms and relationships that need to be used within ontologies. Once deployed, a SeaLife Browser needs to identify places within pages to be accessed for pages not yet marked up and this will be achieved with text-mining techniques. Finally, the semantic markup of pages will be driven by text-mining tools – too much data exist for human curation to be relied upon.

text-mining is already a widely used technique within bioinformatics. Entity recognition, to create dictionaries of gene and gene product names is well explored. For example, the goal of the BioMinT[10] project aims to develop a generic text-mining tool that (1) interprets diverse types of query, (2) retrieves relevant documents from the biological literature, (3) extracts the required information, and (4) outputs the result as a database slot filler or as a structured report. The E-bioSci platform[11] offers access to retrieval of full texts and facts from the vast natural language knowledge base formed by the collected biomedical literature, databases from sequences to images. Such tools will be an invaluable component for generating both ontologies and markup using those ontologies. Such efforts are already underway in projects such as MMTx[12], that is mapping terms in documents to the UMLS metathesaurus. This provides, in essence, a nascent Semantic Web, but without the browsing tools to exploit the marked up documents.

text-mining can also serve to annotate contents in formats such as the Resource Description Framework (RDF). Haystack[13] is such an RDF browser, which allows a user to view RDF stores and to personalise data, by placing links where they feel the need and configuring the user interface, through links, buttons and actions, to do the job they wish in a particular context. It attempts to enable users to work with information, not applications. Instead of having barriers between, for instance, calendar, email, browser and word processor, metadata allows information to be used in any suitable context of work.

These technologies come together in several bioinformatics projects. The myGrid project has developed a set of middleware components that support the e-Scientist in performing and managing *in silico* experiments in biology. Web and Grid Services provide access to distributed resources, while workflow techniques enable the orchestration of these resources to perform experiments. The myGrid middleware is a toolkit of core components for forming, executing, managing and sharing discovery experiments. The components are intended to be adopted in a "pick and mix" way by developers and tool builders to produce end userapplications.

This state of the art reveals that biomedical informatics is already making great use of Semantic Grid and Web infra-structure and technology. Many have concentrated on

[10] http://www.biomint.org
[11] http://www.e-biosci.org
[12] http://mmtx.nlm.nih.gov/docs.shtml
[13] http://haystack.lcs.mit.edu/

data generation, data description and other aspects of the domain. A few have brought elements of this infra-structure together in applications. At no point, however, is there the equivalent of a browser that may be used to look at arbitrary resources and exploit the semantic content to perform eScience. Thus Sealife takes the next step towards implementing the vision of eScience for the life sciences.

4. The Sealife Components and their Interplay

As mentioned in Section 3, to implement the vision of Sealife, three problems need to be solved: ontology design and evolution, concept mapping to link the web to the ontology terms, and service composition to apply relevant services.

4.1. Ontologies

At heart, an ontology is a structured set of vocabulary terms and their definitions that captures a community's understanding of its domain. the idea is to create a shared understanding of the symbols (terms) used to communicate in that domain. Thus, the Gene Ontology creates an agreed set of vocabulary terms for describing the major attributes of gene products. However, it is not only a facilitator for human communication. By capturing this knowledge in a knowledge representation language with strict semantics, it is possible to enable machines to manipulate these symbols through the semantics of the language.

The Web Ontology Language (OWL) is the WorldWideWebConsortium's recommendation for representing ontologies for the Semantic Web. OWL has a strict semantics and its description logic version (OWL-DL) can be used for reasoning over the ontology and its instances. Many bio-ontologies, however, are represented in a more simple language that describes a directed acyclic graph (DAG). This allows only minimal machine usage, but it is directly transformable to OWL. The large number of ontologies in this form (all those in the Open Biomedical Ontologies collelction) offer a potentially vast background knowledge for the SeaLife Browser. Medical ontologies are available in a variety of representations. Some are open and some of these can be mapped into OWL with ease. Others, such as the Medical Subject Headings (MeSH) are a simple thesaurus design for informaiton retrieval and are not really automatically transformaable to OWL. Nevertheless, there is a large amount of biomedical ontology already extant for SeaLife Browser.

Protégé is the most widely used ontology development environment. Its OWL plugin offers a GUI style interface for building and using OWL ontologies. protege's wide range of plugins make it a rich environment. SWOOP, however, offers a much lighter development environment, but has considerable debugging facilities. Outside the OWL world, DAGEdit and OBOEdit are the most widely used tools in bio-ontologies. the former produces the DAG format of the OBO collection. OBOedit, a later development than DAGEdit, offers a richer environment with more modelling constructs.

Protégé, being a more robust and wide-ranging environment than the others, captures more of the principles for building ontologies. These can be split into two broad areas: First those that represent a software engineering approach and second, those that embody philosophical principles within ontology. The first are guidelines of requirements/scope;

knowledge elicitation; design, conceptualisation; encoding; testing/evaluation; publication. these phases map onto a typical software engineering process and many tools and Protégé plugins exist for these stages. Philosophical aspects of ontology building represent the debate on what an ontology can and should represent; styles of building; writing definitions; etc.

One development principle not mentioned is that of re-using ontologies. As already mentioned, many ontologies exist in biomedicine. Once transformed to a common representation and thus a common language semantics, they must be either merged into one or mapped to one another. This is because ontologies can overlap etc. and these overlaps must be recognised and accomodated. A number of such integrations efforts exist within biomedical ontologies. One example is Xspan.[14] This uses a cross-species ontology of anatomy from embryo stages to adult form. The terms from the various species have to be mapped and Xspan have developed the COBrA tool to facilitate this mapping.

4.2. Text-mining

The concepts of the ontologies have to be linked to text in web pages. This task is far from trivial as the concepts will occur in wide variations. The following problems need to be addressed:

- Information content of words: Consider the term *alkaline phosphatase activity* from the GeneOntology. A query on the literature database PubMed for *alkaline phosphatase* leads to more than two times more results than *alkaline phosphatase activity* and to more than ten times more results than *"alkaline phosphatase activity"*. This is particularly striking as the word *activity* is not very informative, as nearly one third of GeneOntology terms end in *activity*.
- Insertions and deletions of words: An ontology term may consist of several words, which are separated by inserted words in free text. For example, the text *...at a higher rate than freshly isolated monocytes upon activation...* should match the GeneOntoogy concept *monocyte activation* and the text *...large family of transcription factors that bind to ...* should match the term *transcription factor binding*.
- Stemming: Words such as *binding* and *binds* have to reduced to the stem *bind*.
- Sentence splitting: text-mining has to identify sentences as units. This is not trivial, as a dot separates two sentences, but it occurs also in abbreviations such as *ca.*, *etc.*, *C. elegans*.
- Special characters: Often ontology terms contain special characters such as slashes, commas, brackets, dashes, etc., which have to be treated appropriatedly. For example, the slash in the term *chromatin assembly/disassembly*, the slash acts as delimiter between two tokens, while in *Arp2/3 complex* the slash is no delimiter.
- Ambiguous concepts: Sometimes ontology concepts are not formulated unambiguously. for example, the term *small-molecule carrier or transporter* should have to match both *small-molecule carrier* and *small-molecule transporter*.

Sealife's text-mining module addresses these problems and thus maps concepts to text in the web pages.

[14] `http://www.xspan.org`

4.3. Service Composition

Once terms have been identified in the SeaLife Browser, they are linked to other resources. A user can, for instance, put a sequence into their Sealife cart. This could be submitted to a service or series of services to perform an analysis. In many cases, more than one service will be used. The following issues will have to be addressed:

- Services will have to be discovered. Many thousands of services now exist. Currently, these are only described by their name and these are not necessarily informative. Efforts to semantically describe these services will reduce this barrier for both people and machines. What should be described? the following are some axes of description:
  - Input
  - Output;
  - Task performed by the service;
  - Service name;
  - Algorithm used;
  - etc.
- Once discovered, how are the services to be composed? Here the following issues are revealed:
  - In many cases, bioinformatics services are implicitly typed. A service takes an input of string and gives an output of string. There is often much structure within one of these strings (for instance, a Uniprot record). Services are needed to locally impose some type on these strings in order to compose them.
  - A minority of services have input and output in some structured XML document. Again, a variety of XML schema exist, so typing services are still needed. Nevertheless, the XML syntax of such input/output documents makes this process easier.
  - A variety of typical type operations are needed in order to compose services: Access, coercion; etc.

An open system such as ^{my}Grid brings more of these problems than a closed system. In a closed system, it is easier to impose a type system, but it does place a barrier to third party services joining the system. SeaLife Browser will of necessity be open, so poorly typed services will be endemic. Composition of services will be part of the SeaLife Browser solution.

5. Existing Prototypes

To realise the Sealife browser, there are already the following key components.

5.1. Ontology editors, evolution and design

In creating the ontology background knowledge for the SeaLife Browser, it will be necessary to transform ontologies into the OWL format. The Gene Ontology Next Generation

(GONG) project[15] developed a methodology to migrate from the simple DAG used by GO to rich descriptions that are possible in OWL. After transforming the DAG to OWL, a simple mapping, the source ontology is already useable in the SeaLife Browser. it is possible, however, to make the source ontology even more useful by migrating it towards a property based description. Many of the OBO ontologies have much of their definition implicit in the class name or term. For instance, "glucose biosynthesis" is a chemical term followed by a process name. These can be made explicit in OWL and in GONG. The mapping is done using regular expressions to match term styles and then generate the implicit OWL definition. When combined with appropriate supporting ontologies, OWL and a reasoner are able to find many implicit subsumption relationships (on average, one in ten classses had a missing subsumption relationship). GONG is now available as a Protégé plugin and will form a component of the SeaLife Browser infra-structure.

An example of the use of a richer GeneOntology resulting from GONG was to guide annotation. By inter-linking the three ontologies of function, process and location, it is possible for a machine to know what molecular functions, for instance, are involved in a particular biological process. by reasoning using the relationships between GeneOntology classes, it is possible to narrow the range of other ontological terms to those sensible to be used (for instance in annotation).

It is clear that a similar process can be used in the SeaLife Browser. Rich interlinked ontologies, together with a reasoner can be used to guide a user through the web of science delivered by SeaLife Browser.

5.2. GoPubMed

GoPubMed[16] is an ontology-based literature search engine, which has indexed over 15 000 000 PubMed abstracts with GeneOntology terms. This system will be a key component underlying Sealife's text-mining module. GoPubMed - as shown in Fig. 1 allows users to explore PubMed search results with the GeneOntology. GoPubMed submits a user's query to PubMed, retrieves the recommended articles, extracts GeneOntology terms from these abstracts, and then displays the part of the GeneOntology covering the extracted GeneOntology terms. This induced ontology can then be used to display articles from the result set, which mention a specific GeneOntology term, including its synonyms or children. With this approach, GoPubMed goes beyond classical search and allows users to answer questions.

Consider the following example: A researcher wants to know which enzymes are inhibited by levamisole. A keyword search for *levamisole inhibitor* produces well over 100 hits in PubMed. With GoPubMed these hits can be systematically be explored for enzyme activities. As shown in Fig. 1, the user can click on *molecular function* and then *catalytic activity*, which reveals that the result set contains *cyclases*, *transferases*, *isomerases*, *hydrolases*, *lyases*, *small protein conjugating enzyme activity*, and *oxidoreductases*. Following the most frequently mentioned enzymes, the user learns that many papers mention *alkaline phosphatase*. The user can also find less obvious facts, such as a single paper on *phosphofructokinase activity* listed among the transferases, which indeed confirms that levamisole inhibits tumor phosphofructokinase.

[15] `http://gong.man.ac.uk`
[16] `http://www.gopubmed.org`

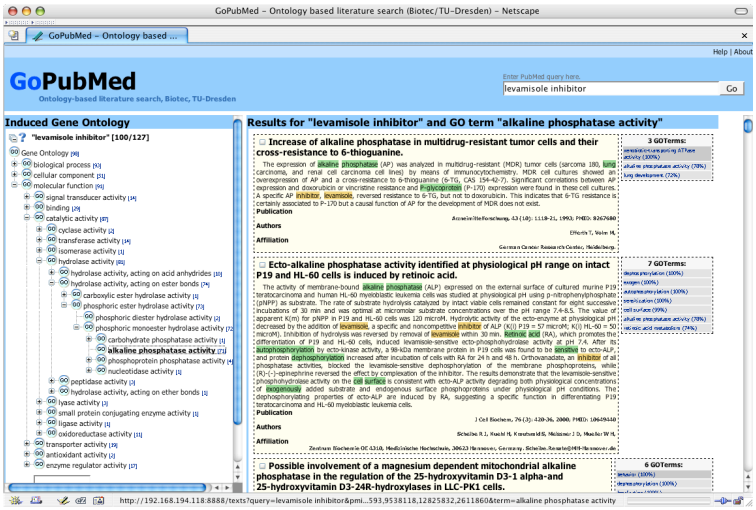

Figure 1. User interface of GoPubMed. On the left, part of the GeneOntology relevant to the query is shown and on the right the abstracts for a selected GeneOntology term. Clicking on a term in the tree, the papers, which have been annotated with this term, are displayed.

5.3. myGrid

myGrid offers a range of Grid enabled services in a *pick and mix* style. An applicaiton builder can take a variety of these services and use them within their application. The SeaLife Browser will be such an application and could use the following services:

- A workflow enactor, Freefluo, can take a workflow described in the XSCFL language and run it against external, distributed services.
- myGrid is capable of using any third party Web Service. A registry of services can be used like a library catalogue to find and then retrieve Web Services.
- SOAPLab is an application for automatically wrapping command line applications as Web services. In a discipline such as bioinformatics, with many legacy programmes and use of command-line as a rapid development route, such an application is vital.
- A notification service, with a subscription mechanism, can be used to notify users and applications in the change of status of services.
- A provenance service records the Web of science generated from workflows. this is itself a Semantic Web that the user can browse via SeaLife Browser.

These services and more can offer support for the activity envisaged within SeaLife Browser.

6. Conclusion

The SeaLife Browser will make eScience's web servers and services available to the bench scientists by using text-mining to identify ontology terms in free text and by linking the ontology terms to applicable services. The SeaLife Browser thus introduces

the novel concept of semantic hyperlinks, which are generated on the fly and use the browser's background knowledge to dynamically link web pages to relevant services. The technical key challenges of the system are the design of ontologies, text-mining for concept mapping and service composition. For all three aspects, there are existing systems and results such as the ontology editor GONG, the ontology-based literature search engine GoPubMed, and the bioinformatics grid system ^{my}Grid. These will form the backdrop for the realisation of the SeaLife Browser, which will be applied to the study of infectious diseases ranging from the patient and clinician exemplified by the National electronic Library of Infectious diseases[17] to molecular biologists studying endocytosis.

Acknowledgements

Funding by the EU project FP6-2006-IST-027269 is kindly acknowledged.

[17] http://www.neli.org.uk