

Sealife:

A Semantic Grid Browser for the Life Sciences applied to the study of Infectious Diseases

Sealife will develop a browser, which will link the existing Web to the currently emerging eScience infrastructure. The Sealife browser will be demonstrated within three application scenarios in infectious diseases relating to evidence-based medicine, literature mining, and molecular biology.

Objectives of the project

Problem or Context:

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

Project:

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. The browser will be evaluated in three applications relating to the study of infectious diseases

Project Description

Sealife will solve the following problems to achieve its objectives:

1. **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.
2. **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 automated mark-up modules for free text, which are based on textmining and natural language processing technologies.
3. **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.

The Sealife browser will be demonstrated within three application scenarios in evidence-based medicine, literature and 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.

To illustrate the power of this vision consider the following applications:



- **Evidence-based medicine:** Consider a clinician, who consults the national electronic library of infections to get trusted information on infections. The user visits the site and finds an interesting page on hepatitis and its treatment: "Ribavirin with or without alpha interferon for chronic hepatitis C". Using its background knowledge, the Sealife browser identifies hepatitis as disease and interferon as an immunologic factor. With this knowledge the browser automatically offers the user the ability to query the biomedical databases Ensembl and PDB to learn more.
- **Literature and Patent Mining:** Getting a quick overview over a field is often vital. In browsing a patent database, a researcher comes across the patent entitled "An improved infant formula is described which includes a phospholipids 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: "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.
- **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 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 some time of browsing, 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 browser now offers to perform a series of services on the protein sequences in the cart.

Expected Results & Impacts

Sealife builds on a number of relevant systems already developed by the partners:

- GoPubMed.org, an ontology-based literature search engine
- MyGrid, a Grid computing platform,
- Corese, a concept resource search engine,
- NeLI, the National electronic library of infectious diseases,
- Edinburgh Mouse Atlas.

These systems will be advanced through Sealife and will ensure a link to a user base. Additionally, Sealife has set up an advisory board with members from Pfizer, AstraZeneca, Unilever, and others. Dresden has spun-off Transinsight.com, which is dedicated to intelligent search for life sciences. Transinsight has secured seed funding by the German High-tech Gründerfonds and has obtained an award by the federal ministry for economic affairs.

Transinsight

Sealife

Project title: A semantic grid browser for the life sciences applied to the study of infectious diseases
Project co-ordinator: TU Dresden

Contact: Prof. Michael Schroeder

Tel: 0049 351 463 400 60

Fax: 0049 351 463 400 61

Email: ms@biotec.tu-dresden.de

Web: www.biotec.tu-dresden.de/sealife

Partners:

1. TU Dresden, Germany
2. Heriot-Watt University, Edinburgh, UK
3. City University, London, UK
4. University of Manchester, UK
5. Scionics GmbH, Dresden, Germany
6. Inria, Sophia-Antipolis, France

Timetable: from 4/2006 to 3/2009

Total cost: €2.6M

EC funding: €2.2M

Instrument: STREP

Project Identifier: FP6-2006-IST-027269

Keywords: Grid, semantic web, molecular biology, healthcare, bioinformatics