A Semantic Web for Bioinformatics: Goals, Tools, Systems, Applications

Mid June, 2007

Department of Computer Science,
University of Pise, Italy

Why Semantic Web

Biological information: an underused resource

- biomedical research activities produce an increasing quantity of new data and new data types
  - EMBL 86 (March 2006) 69,783,593 sequences and 126,401,347,060 bases (+41,05% for sequences and +48,47% for bases in one year)
  - ArrayExpress data on 1,187 experiments, accounting for 800 Gb (Dec 21, 2005)
  - 858 molecular biology databases (NAR 2006), 1,300 SRS libraries, more than 1,000 molecular biology web servers (NAR 2006)
- secondary databases, which are of the highest quality for their good and extended annotation and quality control, and specialized databases, e.g. by gene, organism, disease, mutation, are created by small groups or even by single researchers
- this leads to a distributed environment where data sources have heterogeneous DBMS, data structures, query methods and information
- data integration is needed to achieve a wider view of all available information, to automatically carry out analysis involving more databases and software and to perform large scale analysis. Only a tight integration of data and analysis tools can lead to a real data mining.

Why Semantic Web (ii)

Data integration

- Integration of data and processes needs stability
  - deep knowledge of the domain, well defined information and data, leading to standardization of schemas and formats, clear definition of goals.
- Integration fears heterogeneous data and systems
  - uncertain domain knowledge, highly specialized and quickly evolving information, lacking of predefined, clear goals, originality of procedures and processes.

Data integration in biology

- pre-analysis and reorganization of the data is very difficult, because data and related knowledge change very quickly,
- complexity of information makes it difficult to design data models which can be valid for different domains and over time,
- goals and needs of researchers evolve very quickly according to new theories and discoveries, this leading to new procedures and processes.

Why Semantic Web (iii)

Biological data integration

- Integration methods based on syntactical tools are inadequate
  - explicit cross-references,
  - implicit links (e.g., through names of biological entities),
  - common contents (by using common vocabularies, reference lists and lexicons).

- Semantic integration methods seem more adequate
  - common data models,
  - reference ontologies,
  - metadata descriptions.

Why Semantic Web (iv)

Accessing biological knowledge in a distributed, heterogeneous environment

Moving from an interactive to an automated approach for data integration requires new technologies and tools.

Some starting assumptions

- XML schemas for the creation of the models of the information,
- XML based languages for data representation and exchange,
- Web Services for the interoperability of software
- computerised workflows for the definition and execution of analysis processes
- Workflow enactment portals for utilization of automated processes

Little has been practically made for supporting semantic integration.

Why Semantic Web (v)

Overcoming heterogeneity by adding semantics

- Defining common ontologies and applying them to software tools and databases may be seen as a first attempt to organize and better define the information
- Associating the huge amount of information included in existing databases and information sources with concepts defined in ontologies is the most demanding task
- Developing metadata for biological information on the basis of the Semantic Web standards and tools and describing all information sources, not only databases, à la Resourceome,
- Developing and applying searching tools that are able to make the best use of this additional information.
- Developing tools for expanding integration to unstructured information source (e.g. literature) whose contents must be integrated with structured data in order to achieve the best results.

Preliminary list of topics

- **Goals**
  - Roles and uses of ontologies in knowledge discovery, text analysis and data mining
  - Expected results of adoption of Semantic Web tools in Bioinformatics
- **Standards, Technologies, Tools**
  - Semantic Web standards (RDF, OWL, …)
  - RDF Schemas and Query systems
  - Biomedical Ontologies and related tools
  - Formal approaches to large biomedical controlled terminologies and vocabularies
- **Systems**
  - RDF repositories and query systems for life sciences
  - Semantically aware biomedical Web Services
  - Semantic Biological Data Integration Systems
- **Existing and perspective applications**
  - Case studies, use cases, and scenarios
  - Semantic Web applications in life sciences
- **Network Standards, Technologies, Tools, Applications in Bioinformatics**

Chairs

- Workshop Chairs
  - Paolo Romano, IST – Genoa, Italy
  - Michael Schroeder, Tech Univ Dresden, Germany
  - Nicola Cannata, University of Camerino, Italy
  - Oreste Signore, W3C, Pise, Italy
- Local Organizing Chair
  - Roberto Marangoni, Univ. Pise, Italy
- Scientific Committee
  - Under definition
    (if interested, propose you name!)

Foreseen deadlines

- September 29, 2006: Scientific Committee formed, Web site available, First announcement released
- **November 6, 2006: Call for papers launched**
- February 16, 2007: Opening and invited lectures defined
- March 16, 2007: Tutorials submission
  - Acceptation communication: March 30, 2007
  - Tutorials documentation available: May 4, 2007
- **March 23, 2007: Oral communication submission**
  - Acceptation communication: April 13, 2007
  - Final version available: May 4, 2007
- April 20, 2007: Posters submission
- **April 27, 2007: Early registration**

- **June 12-15, 2007: Workshop and Tutorials**

Pise

Pise is one of the most beautiful and renowned Italian towns, especially famous for the Torre Pendente (Leaning Tower), the Duomo (Cathedral) and the Battistero (Baptesime Church), but also for its scientific tradition, from Galilei's time, up to current research at the University of Pisa, the National Research Council and the High School (Scuola Normale).

It is located in Tuscany, near to many wonderful towns, like Florence, Siena, Lucca, Pistoia and Arezzo, and to nice natural places, like shores on the Tirreno Sea (Versilia, Viareggio, Tirrenia) and the country (Apuane Alps, Chianti region, Maremma).

It is easily reachable by plane, since the Pise Airport "Galileo Galilei" is well connected to many International Airports with flights by many lines, including low cost companies, and by car and by train.

Pise in June

In Tuscany, there is a tradition of assigning a special month to each town, so we have the famous "Maggio Fiorentino" (May in Florence").

June is the month devoted to Pise (Giugno Pisano). During this month, it is a tradition that many events are organized in the town, especially near to the mid of the month, when the main events are devoted to San Ranieri (luminaria, gioco del ponte).

June is also a wonderful time for tourism in Italy. The water of the sea is no more cold, while the weather is not so hot as in July and August. Also, in the countryside the weather is sunny.

We are looking forward to meeting you at NETTAB 2007 in Pise!

http://www.nettab.org/2007/

info@nettab.org

paolo.romano@istge.it

Paolo Romano, Michael Schroeder, Nicola Cannata, Oreste Signore