Computing and Informatics, Vol. 27, 2008, 37–52

SEBIO: A SEMANTIC BIOINFORMATICS PLATFORM FOR THE NEW E-SCIENCE

Juan Miguel Gómez, Damaris Fuentes Lorenzo,
Ángel García Crespo, Sung-Kook Han

University Carlos III
Computer Science Department
Av. Universidad 30-28911
Madrid, Spain
e-mail: {juanmiguel.gomez, damaris.fuentes, angel.garcia}@uc3m.es,
skhan@wonkwang.ac.kr

Revised manuscript received 11 January 2007

Abstract. Knowledge integration and exchange of data within and among organizations is a universally recognized need in bioinformatics and genomics research through the e-science field. The main problem looming over the lack of integration is the fact that the current Web is an environment primarily developed for human users and micro-array data resources lack widely accepted standards; this leads to a tremendous data heterogeneity. Using semantic technologies as a key technology for interoperation of various datasets enables knowledge integration of the vast amount of biological and biomedical data. In this paper, we aim at providing a semantically-enhanced bioinformatics platform (SEBIO), which handles these issues effectively. We will describe the problems arisen and the solutions applied so far. For that, the SEBIO approach is unfolded and its main components explained, to see in more detail how perfectly it copes with the aforementioned difficulties.

Keywords: Semantic web, e-science, bioinformatics

1 INTRODUCTION

In the last years, technological advances in high-throughput techniques and efficient data gathering methods, coupled with a world wide effort in computational biology, have resulted in a vast amount of life science data often available in distributed and heterogeneous repositories. These repositories contain interesting information such as sequence and structure data, annotations for biological data, results of complex and expensive computations, genetic sequences, and multiple bio-datasets.

However, the multiplicity and heterogeneity in the objectives, methods, representation, and platforms of these data sources and analysis tools have created an urgent and immediate need for research in resource integration and platform independent processing of investigative queries, involving heterogeneous data sources and analysis tools.

The Semantic Web and Semantic Web Services paradigm promise a new level of data and process integration that can be leveraged to develop novel high performance knowledge and process management systems for biological applications. Biomedical ontologies constitute a best-of-breed approach for addressing the aforementioned problems. The use of knowledge-oriented biomedical data integration would lead to achieving intelligent biomedical knowledge integration in a new semantic e-science, which will bring biomedical research to its full potential. In this paper, we aim at providing a Semantically-Enhanced BIOinformatics platform (SEBIO), which handles these issues effectively.

The paper is organized as follows. In Section 2, the problem statement is presented. We illustrate our motivation with the precise problem scenario and we use it to formulate the leading guidelines and solutions of our work. The gist of our work is discussed in Sections 3, 4, 5 and 6, where the SEBIO approach is unfolded. Finally, the rest of the paper presents our conclusions and future work.

2 PROBLEM SCENARIO

As discussed in [1], it is undeniable that, among the sciences, biology played a key role in the twentieth century. We have already selected many of the proverbial low hanging fruit of dominant mutations and simple diseases. At the same time, technological improvements in sequencing instrumentation and automated sample preparation have made it possible to create high throughput facilities for DNA sequencing, high throughput combinatorial chemistry for drug screening, high throughput proteomics, high throughput genomics, etc. In consequence, what was once a cottage industry marked by scarce expensive data obtained largely by the manual efforts of small groups of graduate students, post-docs and few technicians has become industrialized and data-rich, marked by factory scale sequencing organizations.

That role is likely to acquire further importance in the years to come. In the wake of the work of Watson and Crick [2] and the sequencing of the human genome, far-reaching discoveries are constantly being made. However, biomedical research is now information intensive; the volume and diversity of new data sources challenges current database technologies. The development and tuning of database technologies for biology and medicine will maintain and accelerate the current pace for innovation and discovery.

There are three main problems in this scope:

Semantic heterogeneity. Years ago, bioinformatics data sat in silos attached to specific applications. Then the Web came into the arena, bringing the hurly-burly of data becoming available across applications, departments and entities in general. However, throughout these developments, a particular underlying problem has remained unsolved: data resides in thousands of incompatible formats and cannot be systematically managed, integrated, unified or cleansed. To make matters worse, this incompatibility is not limited to the use of different data technologies or to the multiple different "flavors" of each technology, but also because of its incompatibility in terms of semantics.

Accessible resources. Biologists need software that is reliable and can deal with huge amounts of knowledge, as well as interfaces that facilitate human-machine interactions. Even though most of the needed information and analysis tools are accessible over the Web, they are designed for low-throughput human use and not for high-throughput automated use. Achieving the full potential of current search of biomedical information resources, fundamentally articles about a particular topic or subject, needs the ladder of IT to reach the higher branches.

Scientific interoperability and integration. Integration and exchange of data within and among organizations is a universally recognized need in bioinformatics and genomics research. By far the most obvious frustration of a life scientist today is the extreme difficulty in putting together information available from multiple distinct sources. A commonly noted obstacle for integration efforts in bioinformatics is that relevant information is widely distributed, both across the Internet and within individual organizations, and is found in a variety of storage formats, both traditional relational databases and non-traditional sources (e.g. text data sources in semi-structured text files or XML). Such knowledge integration is technically difficult for several reasons:

- The technologies on which different databases are based may differ and do not interoperate smoothly.
- The precise naming conventions for many scientific concepts (such as individual genes or proteins) in fast developing fields are often inconsistent and, so, mappings are required between different vocabularies.
- The precise underlying biological model for the data may be different and so to integrate this knowledge requires a common model of the concepts that are relevant and their allowable relations.
- As our understanding of a particular domain improves, not only will data change, but even database structures will evolve. Any users of the data source must be able to manage such knowledge source evolution.

Therefore, searching and integrating data from various sources has become a fundamental issue in bioinformatics research. As discussed in [3], such integration would permit to organize properly the data fostering the analysis and access of such knowledge to accomplish critical tasks and thus processing micro-array data to study protein function, and medical researchers in making detailed studies of protein structures to facilitate drug design.

2.1 Providing Semantics to Life Sciences

The Semantic Web term was coined in [4] to describe the evolution of a Web that consisted largely of documents for humans to read towards a new paradigm that included data and information for computers to manipulate. Ontologies [5] are its cornerstone technology, providing structured vocabularies that describe a formal specification of a shared conceptualization.

The fundamental aim of the Semantic Web is to provide a response to the ever-growing need for knowledge integration on the Web. The benefit of adding semantics is bridging nomenclature and terminological inconsistencies to comprehend underlying meaning in a unified manner. Semantics can be achieved by formally capturing the meaning of data, since a common data format will likely never be achieved, eventually leading to efficiently managing knowledge by establishing a common understanding [6].

On the other hand, with the explosion of online accessible bioinformatics literature, selection of the most suitable resources has become very important for further progress. Bioinformatics literature access relies heavily on the Web, but searching quality literature is hindered by the caveats of information overloading. Recently, the exchange of information on the Web has gained momentum with the raise of some socially-oriented collaborative trends. Together with current Semantic Web technologies and vocabularies that have gained momentum and proved useful, they can help overcome the significant shortcomings of information overload and foster sharing and collaboration through semantics.

Finally, since the current Web is an environment primarily developed for human users, the need of adding semantics to the Web becomes more critical as organizations rely on the service-oriented architecture paradigms to expose and on the data sources by means of Web Services. Semantic Web Services can be discovered, located and accessed since they provide formal means of leveraging different vocabularies and terminologies and foster mediation.

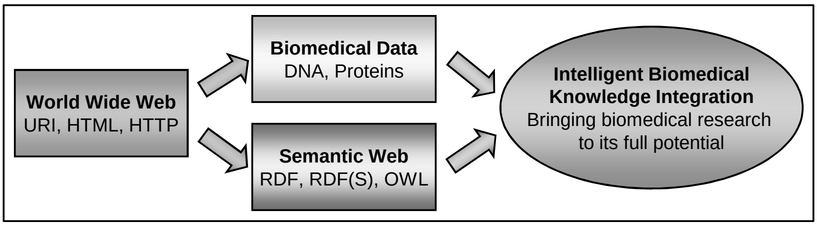

World Wide Web
URI, HTML, HTTP
Biomedical Data
DNA, Proteins
Semantic Web
RDF, RDF(S), OWL
Intelligent Biomedical Knowledge Integration
Bringing biomedical research to its full potential

Fig. 1. Intelligent biomedical knowledge integration

The breakthroughs of adding semantics to scientific knowledge is depicted in Figure 1; information integration is critical in bioinformatics and it could benefit from harnessing the potential of these new semantic approaches.

3 SEBIO ARCHITECTURE

We have divided SEBIO and its main goals into three main projects, which achieve all together the aforementioned goals. These three projects tackle a particular feature of the SEBIO platform each, namely: semantic knowledge integration, literature knowledge integration and finally, semantic web services knowledge integration. The three projects and the features covered are shown in Figure 2.

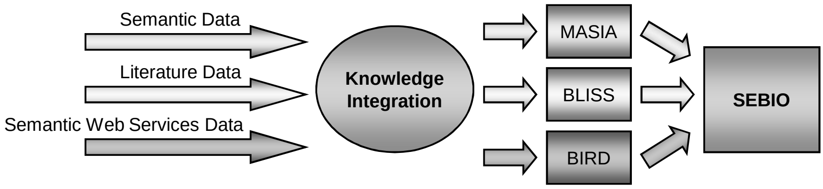


Fig. 2. SEBIO components

The Micro-Array Information and Knowledge Integration Semantics-based Architecture (MASIA) enables micro-array data sources integration. The Biomedical Literature Social Ranking System (BLISS) offers a wide range of documents and literature ranked in terms of interest about a number of topics. Finally, the Biomedical Information Integration and Discovery with Semantic Web Services (BIRD) aims at achieving fundamental integration for biomedical information sources, based on the integrated data achieved in MASIA.

4 MASIA: A MICRO-ARRAY INFORMATION AND KNOWLEDGE INTEGRATION SEMANTICS-BASED ARCHITECTURE

MASIA is a fully-fledged semantically-enhanced architecture for the integration of micro-array data sources. The breakthrough of MASIA is using semantics as a formal means of leveraging different vocabularies and terminologies and fostering of integration. The MASIA approach consists of a methodology to gather requirements, to collect and classify metadata and the different data schemas stemming from the data resources to be integrated, to construct a Unifying Information Model (UIM), to rationalize the data semantics and to utilize it. Finally, the MASIA approach focuses on a software architecture and its capability to enable integration.

4.1 Micro-Array Data Sources and Integration

A number of micro-array data sources scattered all over the world are providing arrays information. One of the most prominent efforts is the Stanford Micro-Array Database[1] (SMD), a micro-array experiments results database hoarding data from 62392 experiments. Another micro-array data source is the European Bioinformatics Institute (EBI) ArrayExpress[2], a public repository for micro-array data, complemented by the ArrayExpress Data Warehouse, which stores gene-indexed expression profiles from a particular subset of experiments in the repository. Also the Maxsd[3] project from the University of Manchester is a data warehouse and visualization environment for genomic expression data.

The lack of standardization in the data formats of these resources is hampering the potential exchange of array data and analysis. Various grass-roots open-source projects are attempting to facilitate the exchange and analysis of data produced with non-proprietary chips, the MGED ontology being one of the most important.

The MGED ontology is a conceptual model for micro-array experiments in support of MAGE v.1. The aim of MGED is to establish concepts, definitions, terms, and resources for standardized description of a micro-array experiment in support of MAGE v.1. Since the MGED has been recognized as a de-facto unifying terminology but most of the actors in the micro-array data sources scenario, it is a perfect gold standard candidate for being a common understanding model.

Finally, micro-array experiments are massive data gathering. Efforts to integrate different micro-array data from such experiments will always have to struggle with a large number of physically different data formats. While a common data format will likely never be achieved, the key to efficiently managing data is to establish a common understanding. This is accomplished by relating physical data schemas to concepts in an agreed-upon model; the Unifying Information Model (UIM). The UIM does not reflect any specific data model, but rather reflects the agreed-upon scientific view, scientific vocabulary and rules which will provide a common basis for understanding data.

4.2 MASIA Methodology

There are some problems which have to be faced when trying to use Semantic Information Management. Firstly, a fragmented data environment leads to business information quality problems. Also, information management is a key issue in a dynamic environment such as modern enterprise where application deployment, business process reengineering or possible restructuring of the data models leads to a burden of hard-coded scripts, data assets and proprietary definitions. Finally,

[1] SMD: `http://genome-www5.stanford.edu/`

[2] EBI ArrayExpress: `http://www.ebi.ac.uk/arrayexpress/`

[3] Maxd: `http://www.bioinf.man.ac.uk/microarray/maxd/index.html`

meaning and context of the knowledge must be captured and managed in a way that represents some long-term value for the enterprise.

How to bridge the gap between this situation and the Semantic Information Management level is defined by the Semantic Information Management methodology. This methodology is structured such that each stage adds value in its own right, while simultaneously progressing the enterprise towards the benefits of full semantic data integration.

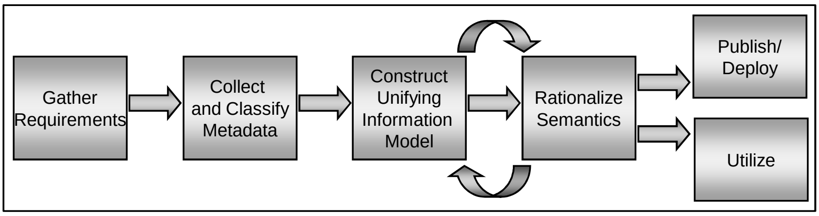

Gather Requirements
Collect and Classify Metadata
Construct Unifying Information Model
Rationalize Semantics
Publish/ Deploy
Utilize

Fig. 3. Methodology

Figure 3 depicts the steps in the methodology:

Gather Requirements: Establish the project scope, survey the relevant data sources and capture the organization's information requirements.

Collect and Classify metadata: Catalog data assets and collect metadata relevant to the organization and its use of data.

Construct Unifying Information Model: Capture the desired business worldview, a comprehensive vocabulary and business rules.

Rationalize the Data Semantics: Capture the meaning of data by mapping to the Information Model.

Publish/Deploy: Share the Information Model, metadata and semantics with relevant stakeholders; customize it to their specialized needs.

Utilize: Create processes to ensure utilization of architecture in achieving knowledge management, knowledge integration and knowledge quality.

4.3 MASIA Software Architecture

We propose a tailor-made value-adding technological solution which addresses the aforementioned challenges and solves the integration problem regarding to searching, finding, interacting and integrating heterogeneous sources by means of semantic technologies.

The MASIA architecture is related and composed by a number of components depicted in Figure 4:

Crawler: A software agent which browses the information sources in a methodical, automatic manner. It is a technology suitable for nearly any application that requires full-text search, expecially cross-platforms.

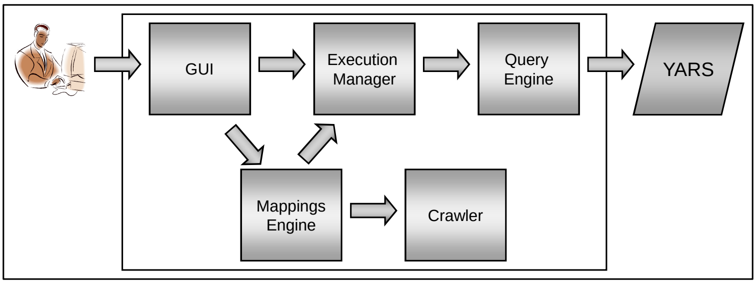


Fig. 4. The MASIA software architecture

Mappings Engine: A set of integrated tools for semantically mapping data schemas to such Unifying Information Model. The Mappings Engine is to enhance the semi-automatic mapping of schemas and concepts or categories of the UIM in order to alleviate the tedious process which requires human intervention. Since automatic mapping is envisaged as not recommendable due to semantic incompatibilities and ambiguities among the source schemas and data formats, it should bridge the gap between cos-efficient machine-learning mapping techniques and pure human interaction. The Mappings Engine takes the MGED ontology as a conceptual basis for the mappings from the various sources. It will then relate data schemas with the semantic structure of the ontology.

YARS: The YARS[4] (Yet Another RDF Store) system is a semantic data store that allows semantic querying and offers a higher abstraction layer to enable fast storage and retrieval of large amounts of RDF[5] (Resource Description Framework) while keeping a small footprint and a lightweight architecture approach. YARS deals with data and legacy integration.

GUI: The component that interacts with the user. It collects the users request and presents the results obtained. In our particular architecture, the GUI collects requests pertaining to search criteria, such as, for example, "descriptor". The GUI communicates with the Execution Manager component providing the user request and displays the results provided as a response from the *Execution Manager* component.

Query Engine: This component uses a query language to make queries into the YARS storage system. The semantics of the query is defined not by a precise rendering of a formal syntax, but by an interpretation of the most suitable results of the query. Since YARS enables SPARQL querying, due to pragmatic reasons this is the query language of our choice.

Execution Manager: It manages the different interactions among the components. Firstly, it communicates with the Mappings Engine to verify that the

[4] YARS: `http://sw.deri.org/yars`

[5] RDF: `http://www.w3.org/TR/rdf-primer/`

information extracted by the crawler are being correctly mapped on the MGED ontology as a Unifying Information Model (UIM) and finally stored into YARS with an RDF syntax. Secondly, it accepts the users search request through the GUI and hands them over the query engine, which, in turn, queries YARS to retrieve all RDF triples related with the particular search criteria. By retrieving a huge number of triples from all the integrated resources, the user benefits from a knowledge-aware search response which is mapped to the underlying terminology and unified criteria of the Unifying Information Model, with the added advantage that all resources can be tracked and identified separately, i.e., data provenance can be traced and assigned to a particular resource.

5 BLISS: A BIOMEDICAL LITERATURE SOCIAL RANKING SYSTEM

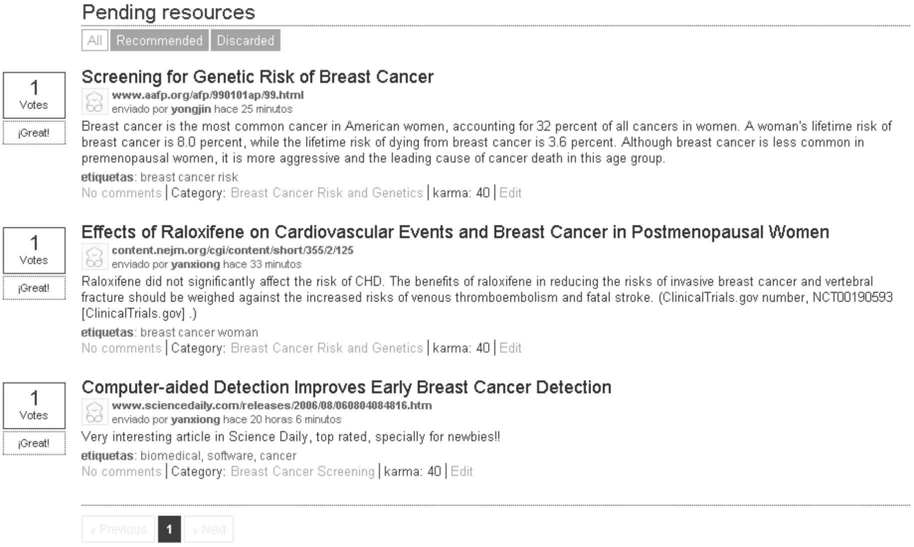
Biologists need software that is reliable and can deal with huge amounts of data as well as interfaces that facilitate the human-machine interactions. Most of the needed information and analysis tools are accessible over the Web. However, they are designed for low-throughput human use and not for high-throughput automated use.

In this section we present the BLISS system, a proof-of-concept implementation of a biological literature social ranking system used in the bioinformatics field. A screenshot of BLISS is depicted in Figure 5. The main features of the system are outlined as follows:

- The user (biologist, bioinformatician, medical, etc.) finds an article interesting and wants to communicate it to the community. For that, he selects that article (providing a URL as a pointer) and a category under which it is relevant (e.g. Yeast or Lung Cancer).
- Users who join the system can, provided their experience in the field, vote and hence rank the documents properly. The more votes an article gets, the higher it climbs up.
- Potential users can then be recommended and suggested a number of articles of particular importance for a number of topics, what, given the social nature of the approach, ensures the quality and feedback of the articles.

BLISS provides relevant metadata that can be harvested and used for intelligent collaborative discovery. For example, BLISS provides a labelled graph (based on a RDF representation) of resources being recommended under a common topic or with similar features. MEDLINE is a major repository of biomedical literature supported by the U.S. National Library of Medicine (NLM). It currently collects and maintains more than 15 million abstracts in the field of biology and medicine, and is incremented by additional thousands of new articles every day. PubMed[6]

[6] PubMed: `http://pubmed.gov/`



Pending resources

Screening for Genetic Risk of Breast Cancer
www.aafp.org/afp/990101ap/99.html
enviado por **yongjin** hace 25 minutos

Breast cancer is the most common cancer in American women, accounting for 32 percent of all cancers in women. A woman's lifetime risk of breast cancer is 8.0 percent, while the lifetime risk of dying from breast cancer is 3.6 percent. Although breast cancer is less common in premenopausal women, it is more aggressive and the leading cause of cancer death in this age group.

etiquetas: breast cancer risk
No comments | Category: Breast Cancer Risk and Genetics | karma: 40 | Edit

Effects of Raloxifene on Cardiovascular Events and Breast Cancer in Postmenopausal Women
content.nejm.org/cgi/content/short/355/2/125
enviado por **yanxiong** hace 33 minutos

Raloxifene did not significantly affect the risk of CHD. The benefits of raloxifene in reducing the risks of invasive breast cancer and vertebral fracture should be weighed against the increased risks of venous thromboembolism and fatal stroke. (ClinicalTrials.gov number, NCT00190593 [ClinicalTrials.gov] .)

etiquetas: breast cancer woman
No comments | Category: Breast Cancer Risk and Genetics | karma: 40 | Edit

Computer-aided Detection Improves Early Breast Cancer Detection
www.sciencedaily.com/releases/2006/08/060804084816.htm
enviado por **yanxiong** hace 20 horas 6 minutos

Very interesting article in Science Daily, top rated, specially for newbies!!

etiquetas: biomedical, software, cancer
No comments | Category: Breast Cancer Screening | karma: 40 | Edit

Fig. 5. A BLISS screenshot

is the most popular interface to access the MEDLINE database. If the articles searched by the biologists about osteoporosis are in MEDLINE, they will be found via a PubMed identifier link. They could then be certain that the article is of a certain quality (since it has been verified and recommended by a pool of users) and access it directly via the BLISS interface.

6 BIRD: BIOMEDICAL INFORMATION INTEGRATION AND DISCOVERY WITH SEMANTIC WEB SERVICES

The Biomedical Information and Integration Discovery with Semantic Web Services (BIRD) platform fosters the intelligent interaction between natural language user intentions and the existing Semantic Web Services execution environments. Our contribution is an overall solution, based on a fully-fledged architecture that transforms the user intentions into semantically-empowered goals that can be used to encompass interaction with a number of available Semantic Web Services architectures such as WSMX, OWL-S Virtual Machine and METEOR-S.

BIRD is a two-faced software agent designed to interact with human beings as a gateway or a man-in-the-middle towards Semantic Web Services execution environments. The main goal of the system is to help users express their needs in terms of knowledge retrieval and achieve knowledge integration by means of Semantic Web Services. BIRD allows users to state their needs via natural language or to go through a list of the most important terms, extracted from the Gene Ontology (GO). For this, BIRD makes use of ontology-driven data mining and of the data integration obtained from MASIA. BIRD first captures and gathers which are the terms and the user would like to search (e.g. Gene A, Protein Y) by using the aforementioned terms of the GO as a reference. Second, it builds up a "lightweight ontology". Finally, it looks for which "goal" from the "goal template" repository fits better with the search criteria and requirements from the user. A "goal" in Semantic Web Services technology refers to the aim a user expects to fulfil by the use of the service. Once BIRD has inferred the goals derived from the users' wishes, it sends them to the suitable Semantic Web Services execution environment, which will retrieve the outcome resulting of the integration of the applications being accessed.

Our approach is backed with a proof-of-concept implementation where the breakthrough and efficiency of integrating the biomedical publications database PubMed, the Database of Interacting Proteins (DIP) and the Munich Information Center for Protein Sequences (MIPS) has been tested.

Figure 6 depicts the main components of BIRD:

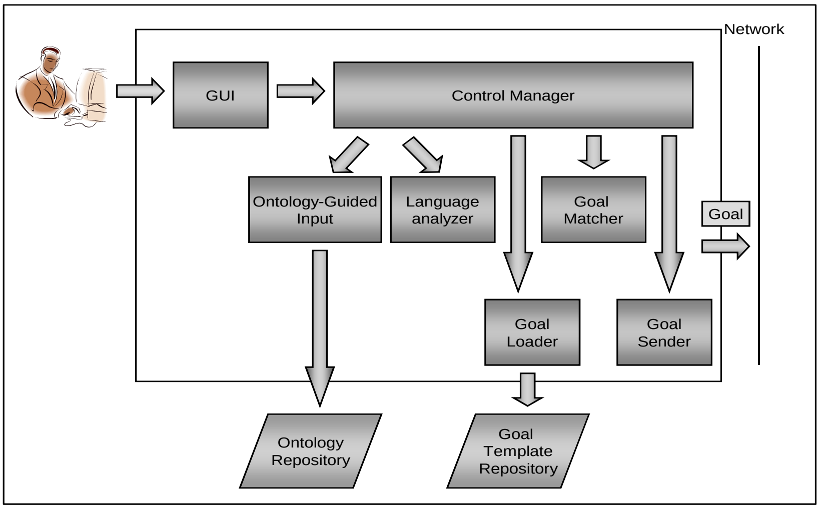


Fig. 6. The BIRD architecture

Language Analyzer: This analyzer is to filter and process the input introduced by the user in natural language and to determine the concepts (attributes and values) and relations included in it.

Goal Loader: This component looks for goal templates in the Goal Template Repository. Actually, the Goal Loader retrieves all the goal templates and transmits them to the Control Manager. Since in this version of BIRD there is no fixed Semantic Web Services execution environment, different types of goal repositories are taken into account. The repository is outside the architecture so that anybody may plug in his/her own goal repository.

Goal Matcher: The Goal Matcher compares the ontology elements obtained from the analysis of the user's wishes to the description of the goal templates extracted from the repository. From this matching, several goals are selected that are composed by the Control Manager in order to build up the sequence of execution.

Goal Sender: This component sends the different goals to the execution environment, which returns the results obtained from the execution of the services. The sending of goals is sequential, without taking into account any other workflow constructs.

GUI: It collects the users request and presents the results obtained to them.

Control Manager: It manages the different interactions among the components. First, it accepts the users input through the GUI. It can be either natural language text or a structured sentence written with the assistance of the Ontology-guided Input. If the input is in natural language, then it instructs the Language Analyzer to attempt the recognition of the major concepts in the text and communicates with the Goal Loader and the Goal Matcher to orchestrate the different goals that will be sent to the execution environment through the Goal Sender. Then, it communicates with the GUI so that the users receive a view of the selected goals and decides if they are correct and comply with their expectations. Finally, if the user approves them, they are sent sequentially.

One of the most important features of the system is its capability to interoperate with different Semantic Web Services execution environments.

7 RELATED WORK

Integration of heterogeneous data in life sciences is a growing and recognized challenge. As discussed in [7], several approaches for biological knowledge integration have been developed. Well-known approaches include rule-based links such as SRS [8] or [9], federated middleware frameworks, such as Kleisli system [10] or [11] as well as wrapper based solutions such as IBM Discovery Link [12]. However, these environments work with non-semantic approaches (i.e. XML data).

In parallel, progress has been made to organize biological knowledge in a conceptual way by developing ontologies and domain-specific vocabularies such as in [13] or [14]. With the emergence of the Semantic Web, the ontology-based approach to life science knowledge integration has become more ostensible. In this context, knowledge integration comprises problems like homogenizing the data model with schema integration, combining multiple database queries and answers, transforming and integrating the latter to construct knowledge based on underlying knowledge representation. However, the ontology-based approach can not solve the evolving concepts in biology and its best promise lies in specialized domains and environments where concepts and vocabularies can be well controlled [15].

A similar approach to the work presented has been followed in [7]. Their integration approach is based on the premise that relationships between biological entities can be represented as a complex network. The context dependency is achieved by a judicious use of distance measures on these networks. The biological entities and the distances between them are mapped for the purpose of visualization into the lower dimensional space suing the Sammons mapping. Finally, their system implementation is based on a multi-tier architecture using a native XML database and software tools for querying and visualizing complex biological networks. However, the forthcomings of the approach are hampered by the fact that they are stated at pure XML-level without taking into account particular semantics of the mappings and hence not being able to exploit the semantics inherent to the data formats.

Finally, even though we have not found any similar work such as BLISS, this work is related to existing efforts about social software and new distributive collaborative trends such as Digg[7] or Slashdot[8]. These works could also be enhanced with another type of search strategies as discussed in [16].

8 CONCLUSIONS AND FUTURE WORK

As the use of bioinformatics and biomedical research grows, the problem for searching, interacting and integrating relevant information is becoming increasingly a hurdle for the leverage of existing technologies. Currently, Semantic Web and Semantic Web Services, which have reached a certain level of maturity, offer an interesting alternative. Actually, these recent paradigms promise a new level of data and process integration that can foster the development of novel high-performance knowledge and process management systems for biological applications.

In this paper, we have proposed Semantically-Enhanced BIOinformatics Platform (SEBIO), a novel and trailblazer approach which aims at establishing the basis of fundamental research in the combination of the aforementioned approaches, handling effectively:

- conceptual models for biological data
- use of semantics to manage interoperation of biomedical datasets
- biomedical data engineering using ontologies
- support of ontologies for biological information retrieval and Web Services

To achieve these goals we have divided SEBIO into three main projects:

- The Micro-Array Information and Data Integration Semantics-based Architecture (MASIA), to enable micro-array data sources integration.
- The Biomedical Information Integration and Discovery with Semantic Web Services (BIRD), to achieve fundamental integration for biomedical information sources.

[7] Digg: `http://www.digg.com`

[8] Slashdot: `http://www.slashdot.com`

- The Biomedical Literature Social Ranking System (BLISS), which offers a wide range of documents and literature ranked in terms of interest about a number of topics.

As the use of new communication paradigms technologies on the Web grows and changes, the problem for finding and relating appropriate resources in order to achieve a particular goal will get more acute. The SEBIO approach is based on collaborative discovery and social semantic ranking system as a particular means to bridge the gap between provided metadata from both the service provider perspective and the current collaborative discovery techniques and initiatives on the Web for the benefit of semantic e-science.

The forthcomings of our approach are as follows: On the one hand, current technology can easily add some plug-in to improve and add the described functionality and benefit from the harvesting of more information on semantic e-science, such as the previously noted. On the other hand, a more accurate critical mass use of semantic collaborative discovery techniques can foster the effectiveness and efficiency of discovery, enhancing eventually the whole semantic scientifical resource discovery approach. Besides, our future work will focus on finding more use cases and real-world scenarios on semantic e-science to validate the efficiency of our approach and to determine the feasibility of the semantic match of lightweight ontologies and mappings in particular contexts.

REFERENCES

[1] Cohen, J.: Bioinformatics: An Introduction for Computer Scientists. ACM Computing Surveys, issue 36, 2004, pp. 122–158.

[2] Watson, J. D.—Crick, F. H. C.: Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature, Vol. 171, 1953, pp. 737–738.

[3] Ignacimuthu, S.: Basic Bionformatics. Alpha Science International, 2005.

[4] Berners-Lee, T.: Weaving the Web. Collins, 2000.

[5] Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag, 2002.

[6] Shadbolt, N.—Hall, W.—Berners-Lee, T.: The Semantic Web Revisited. IEEE Intelligent Systems, 2006.

[7] Gopalacharyulu, P.—Lindfors, E.—Bounsaythip, C.—Kivioja, T.—Yetukuri, L.—Hollmen, J.—Oresice, M.: Data Integration and Visualization System for Enabling Conceptual Biology. Journal of Bioinformatics, Oxford University Press, Vo. 21, 2005, pp. 177–185.

[8] Etzold, T.—Argos, P.: SRS: An Indexing and Retrieval Tool for Flat File Data Libraries. CABIOS, Vol. 9, 1993, pp. 49–57.

[9] Etzold, T.—Ulyanov, A.—Argos, P.: SRS: Information Retrieval System for Molecular Biology Data Banks. Methods Enzymol, Vol. 1, 1996, pp. 114–128.

[10] Davidson, S. B.—Overton, C. G.—Tannen,V.—Wong, L.: BioKleisli: A Digital Library for Biomedical Researchers. International Journal on Digital Libraries, Vol. 1, 1997, pp. 36–53.

[11] Chung, S. Y.—Wong, L.: Kleisli: A New Tool for Data Integration in Biology. Trends in Biotechnology, Vol. 17, 1999, pp. 351–355.

[12] Haas, L. M.—Schwarz, P. M.—Kodali, P.—Kotlar, E.—Swope, J. E. R. W. C.: Discoverylink: A System for Integrated Access to Life Sciences Data Sources. IBM Systems Journal, Vol. 40, 2001, pp. 489–511.

[13] Ashburner, M.—Ball, C.—Blake, J.—Botstein, D.—Butler, H.—Cherry, J.—Davis, A.—Dolinski, K.—Dwight, S.—Eppig, J.: Gene Ontology: Tool for the Unification of Biology. Nat. Genet., Vol. 25, 2000, pp. 25–29.

[14] Bard, J. B. L.—Rhee, S. Y.: Ontologies in Biology: Design, Applications and Future Challenges. Nat. Genet., Vol. 5, 2004, pp. 213–222.

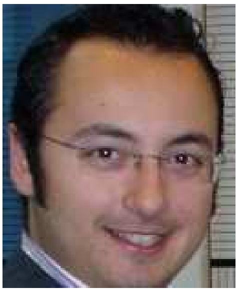
[15] Searls, D. B.: Data Integration: Challenges for Drug Discovery. Nature Review Drug Discovery, Vol. 4, 2005, pp. 122–158.

[16] Mahmood, A.: Object Replication Algorithms for World Wide Web. Computing and Informatics, Vol. 24, 2005, No. 4, pp. 371–390.

Ángel García Crespo is the head of the SofLab Group at the Computer Science Department in the Universidad Carlos III de Madrid and the head of the Institute for Promotion of Innovation Pedro Juan de Lastanosa. He holds a Ph. D. in industrial engineering from the Universidad Politécnica de Madrid (Award from the Instituto J. A. Artigas to the best thesis) and received an Executive MBA from the Instituto de Empresa. He has led and actively contributed to relevant European projects of the FP V and VI, and also in many business cooperations. He is the author of more than a hundred publications in conferences, journals and books, both Spanish and international.

Damaris Fuentes Lorenzo holds a B. Sc. in technical engineering in computer managements from the Universidad Carlos III de Madrid and a M. Sc. in computer engineering, development of the enterprise information system specialization. She is enrolled in a master about computer science and technology, specialized in software Engineering. She has took up several scholarships in the Departments of Telematics and Computer Science of the Universidad Carlos III de Madrid, involved in the latter as a research technician.Currently, she is working as a research assistant at IMDEA Networks.

Sung-Kook Han is a professor of computer engineering at Won Kwang University, Korea. He is a chair and advisory committee member of various academic societies: to name a few, Korea Information Society, The institute of Electronic Engineers of Korea, Korea Information Processing Society and The Korean Society for Cognitive Science. Currently he serves a Vice President of Korea Information Society and a board member of Korea Association of Semantic Information Technology. He is also a member of IEEE, ACM and AAAI. He was a committee chair of several international conferences. He received Ph. D. from the Inha University, Korea. He was visiting scholar at University of Pennsylvania, USA and visiting researcher at DERI, Austria. His main research area is artificial intelligence, ontology, semantic web and web services.



Juan Miguel Gómez is a Visiting Professor at the Computer Science Department in the Universidad Carlos III de Madrid. He holds a Ph. D. in computer science from the Digital Enterprise Research Institute (DERI) at National University of Ireland, Galway and received his M. Sc. in Telecommunications Engineering from the Universidad Politécnica de Madrid (UPM). He was involved in a number of EU FP V and VI research projects and was a member of the Semantic Web Services Initiative (SWSI). His research interests include the semantic web, semantic web services, business process modeling, B2B integration and, recently, Bioinformatics.