Biomedicine meets Semantic Web

Juan Bernabé Moreno
University of Granada, Department of Computer Science and Artificial Intelligence

Abstract— **Bio-sciences have a lot of potential to be the best fit to profit from the power of the semantic web. This work explains how ontologies have become a main stream within the biomedical research, and discusses the main scenarios ontologies are used in. Finally, after summarizing the conclusion, a brief outlook is provided**

***Index Terms*— Semantic Web, Data sources, Integration, Biomedicine, Life Sciences**

I. INTRODUCTION

The amount of available knowledge doubles every 5 years but this increase is even faster in the Life Sciences and this is the reason why the digital revolution has become a cornerstone in the day-to-day work in Life Sciences.

In 1953 Watson and Crick discovered the structure of DNA. Almost 47 years later, the first draft of the human genome is published which was considered as one of the most significant scientific landmarks of all time, comparable with the invention of the wheel or the splitting of the atom [11]

The biomedicine nowadays is the result of a very fast evolution within the last years, becoming a data-driven internet-based science. Date driven because of the masses of data that high-throughput experiments produce on a daily basis: more than 16k million DNA bases, more than 25.000 protein structures with an average of ca. 400 residues, more than 130k annotated protein sequences –SWISSPROT[12]- and more than 850k protein sequences –TrEMBL[13]-, more than 14 million scientific articles available –PubMed[13]-), etc. Internet-driven because of the incipient development of the telemedicine, the shift from a preventive to a curative medicine supported by information technologies, the online collaboration platforms that break down the communication barriers, etc

Current information state and the underlying problems

The biggest challenge and effort driver in the IT resources available in the life sciences is the integration of existing disparate data sources.

The same investment is done over and over by several projects because of emerging requirements that make the integration with legacy systems very tedious

The integration and recycling of information sources lead to the creation of several data bases

From the developer point of view this situation obviously presents drawbacks:

1) High redundancy among the resources implies wasting of development capacity
2) Parallel scanning of resources is almost impossible. Plenty of queries to be executed in order to retrieve the available information
3) Relevant dependencies or contradictions between data remain hidden, because the information is spread out among different data bases
4) Users to get used to different data models and different interfaces, which requires ramp up time
5) Need for custom development to integrate each and every data source
6) Suboptimal data exploiting: small data bases stay unexploited. Adding of new data sources requires manual intervention for discovery and biding, which means the model does not scale.

There have been approaches to overcome such problems (see section III.B)

The result of theses approaches was in some cases a good short-term solution, but not sustainable. The fact of adding a new data source implied remapping with other databases schema or with a central data schema and eventually building a new connector, which lead to a cost explosion and to a poor designed home-grown system

The problem starts right after a new data source project goes live: the developed data source becomes difficult to access from outside because of the lack of a semantic basis and application context

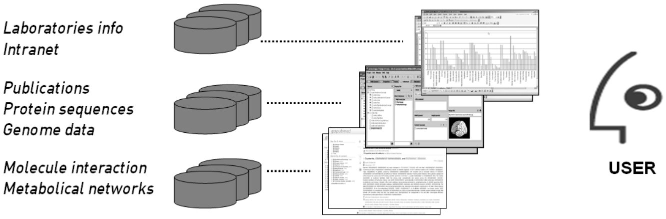

Figure 1 Before Semantic Web in place

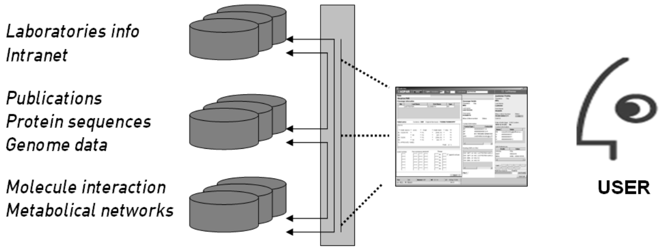

Figure 2 Semantic Web enabled scenario

II. The Semantic Web

As per W3C [1], semantic web is about giving well-defined meaning to the information to enable that people and computers work in cooperation.

The semantic web extends the ordinary web in two major aspects:

- The information is expressed in special machine-targeted language (instead of a wide range of natural languages for the human consumption)
- The data is formally and semantically interlinked, whereas the web is a set of informally interlinked information

The Semantic Web to demolish communication barriers

Semantic Web shouldn't be seen as a new technology, but rather as a completely new idea to organize the knowledge. The benefits of structuring the knowledge in the scientific community have shifted the semantic web into a key role rather than an unnecessary overhead.

Older artificially created communication barriers between the scientific community members can be demolished in a way that the entire community will be able to profit from any small contribution of any individual.

III. Bioinformatics as perfect candidate

As indicated before, the amount of data produced by the different researches and the number of algorithms created to workout this information is huge, and the trend is making them available over the internet for the community. Additionally, people are more and more recognizing the advantages of adhering to open standards and it is no longer unusual that researchers from the bioinformatics community foster the creation of new data standards, as they are required. XML as publishing format is pushing out the proprietary free-text position based formats.

The number of data sources providing valuable information over the internet is exponentially increasing, and their integration –jointly querying- and interoperation has become the first concern.

On the other hand, the upcoming internet applications are more and more allowing the information retrieval over the internet by moving to service invocation architectures (REST and web services) [15]

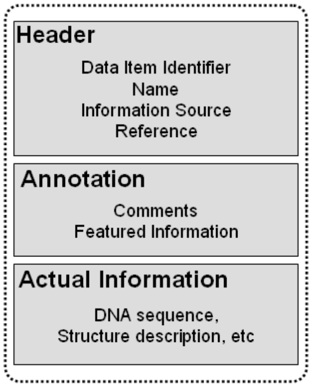

Figure 3 Typical biomedical content information item

A. *Biomedicine Data Sources*

Even if most data bases can be accessed by a web interface, ftp and email accessing methods are still mostly supported

The underlying data models differ substantially from each other, ranging from full-fledged object-oriented data bases to file-based models.

The content information is commonly given as per the diagram shown in Figure 3, with header, annotation and actual information

The majority of data bases can be accessed by queries that retrieve information based on the occurrence of certain text (or stemmed text) within a data item –what's known as full-text search- or within certain predefined fields (i.e.: abstract of an article). Boolean searching operations (and, or, not) are commonly supported as well as wildcards –the entire support of regular expressions is hardly ever supported-. The interface given to the user is usually form-based but sometimes a console access is also given to allow the usage of a data querying language DQL usually with certain restrictions. As mentioned above, programmatic access is usually supported by means of web services, REST, etc. The results set is usually a (paged) set of the entries matching the query than can be reduced by refining the query

Very well-known data bases are *GenBank* and *DDBJ* for nucleotide sequences, *SWISS-PROT* and *ENZYME* for protein sequences, *BLOCKS* and *PROSITE* for protein families and *PDB* and *MMDB* for 3D macromolecular structures.

B. *Biomedicine Data Base Interoperation*

Since the information is available over internet and scattered into different data bases, there have been several approaches to address the integration and interoperation concern:

1) *Link-driven federation*

Most of the existing web interfaces that offer multi-database querying are based on this mechanism. The data source system is often implemented by using files and specialized retrieval packages and the integration is done by means of cross-reference indexes between the data items. The querying processing time is low and the interface is easy to use, but the syntax-based ad-hoc linkage doesn't address the heterogeneities in the terminology used by the disparate data sources. This mechanism presents serious scaling problems, i.e. incorporating a new data source requires re-indexing the system.

The best known systems running this method are SRS from Lion Bioscience/Biowisdom[16] and Entrez[17]

2) Data warehouses

Central databases that keep a copy of data from different sources into central schema (e.g..: Atlas)[18]

3) Query optimizers

They are basically applications that enable the user to create queries to different data sources in a comfortable way. They often rely on view integration, where the different schemas are integrated to form a global one, which is queried in a high level language (e.g.: Discovery Link)[19]. Other examples are BioKleisli [21], K2 [22], TINet [20], P/FDM [24], TAMBIS [23], etc

4) Middleware frameworks

Intended to query different data models by means of different interfaces.

IV. Ontologies

In a more and more data intensive world the computers play a key role in helping people manage the information explosion.

Ontologies have become the cornerstone to structure the complex knowledge domains and establish standards

Ontologies have been defined as "a way to express formally a shared understanding of information" [25]

A. Ontologies between Philosophy and Computer Science

As indicated in [2] the fact that ontologies is a plural raises the major difference between the philosophical and computer science approach to the term.

A philosophical ontology would encompass the whole of the universe, but computer scientists allow the existence of multiple, overlapping ontologies, each focused on a particular domain.

Indeed an understanding of the ontology of a particular domain may be crucial to any understanding of the domain. The combination of ontologies, and communication between them, is therefore, a major issue within computer science, although such issues are problematic with the philosophical use of the term. At the limit, an ontology that perfectly expresses one persons understanding of the world is useless for anyone else with a different view of the world. Communication between ontologies is necessary to avoid this type of solipsism.

B. Ontologies usage in biomedical use cases

Biomedicine has profited from ontology-driven technologies in many ways [26]:

1) Reference for naming things

The motivation behind it is establishing a set of controlled terms for labeling entities in databases and data sets.

It will ensure the consensus between people on the name to be given to certain entity, and the consensus between people and machines to identify and name things. The immediate consequence of this consensus is the fact that computers can help researches to make sense of massive data available to perform analysis on. The challenging side is the variety of synonymous terms and polysemy or lexical ambiguity, defined as "the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings" in [27]

The biggest effort driver is the unification of disparate data that are labeled differently in different data sources. Thus, where the ontology adds value is in "fixing" the terminology so that people can label medical entities in a consistent way. Additionally, synonymy, acronyms and abbreviations can augment the ontologies.

The most popular example is the Gene Ontology [28] Their entities have *is-a* and *part-of* relations to other entities, providing the basis for representing biological knowledge. These relations support the creation of computer reasoning applications, which can infer subsumption (*is-a* relations) or composition (part-of relations) between entities.

The ad-hoc usage of the GO ontology allows the querying of many Model Organism Databases (MODs) thanks to the disambiguation of meanings.

Looking beyond the mere terminology fixing, the GO is used as a basis for the term extraction and better information retrieval on the life sciences documents (i.e. the GoPubMed project)

Ontologies are commonly used to provide a common way of describe the patient information in health records (see US National Library of Medicine UMLS) [30]

The description of audio visual information is also address by the usage of ontologies to provide names for anatomy, pathology and observations in images (i.e. Open Microscopy Environment –OME-)

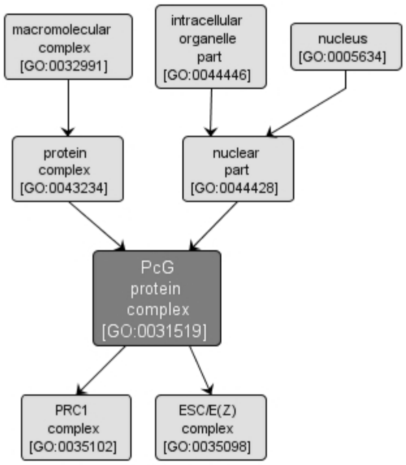
macromolecular complex [GO:0032991]
intracellular organelle part [GO:0044446]
nucleus [GO:0005634]
protein complex [GO:0043234]
nuclear part [GO:0044428]
PcG protein complex [GO:0031519]
PRC1 complex [GO:0035102]
ESC/E(Z) complex [GO:0035098]

Figure 4 A chromatin-associated multiprotein complex containing Polycomb Group proteins. GO View

2) Representation of encyclopedic knowledge

The second natural step to capture and represent knowledge is by means of rich relationships between the entities of a domain. The textual description of complex knowledge gives the humans the possibility to access this knowledge, but not the machines. Using well-defined, univocal, standardized relationships to structure and make explicit the knowledge enable the access to machines and humans.

The Foundation Model of Anatomy (FMA)[31] is a very popular example that specifies canonical knowledge for the anatomy domain (entities and relationships). It is the result of the anatomist and knowledge engineers collaboration and unlike other ontologies, it has not been created for a particular application, but with the long-term goal of providing digital accessible encyclopedic knowledge for anatomy

3) Information model specification

Specifying information and data models using ontologies instead of UML provides several advantages:

- Explicit specification of the terms used to express information in the biomedical domain
- Augmented capabilities like explicit relationship making among data types and automatic reasoning – subsumption and composition-.
- Complex structures visualization capabilities (like the ones offered by Protégé [32]
- Publishing of information model in the semantic web (if standards have been adhered –like OWL-) [33]

Ontologies for this purpose have been adopted in the microarrays world. Microarray is the term standing for modern bio-molecular analysis systems used to generate molecular level biomarkers for a variety of biological states and medical diseases. The creation of large amounts of microarray data and the creation of databases to enable sharing of these data quickly raised the need for standards in describing microarray experiments and results. The MGED ontology aims at providing a common terminology and information schema for annotating microarray experimental results, resolving ambiguity situations on how microarray experiments are described and providing a mechanism for query expansion exploiting subsumption relationships [34]

Another remarkable example is the Ontology of Biomedical Investigation which is a more generic approach targeting the description of biological and medical investigation [35]

4) Specification of data exchange format

The emerging of multiple data based containing related biomedical information requires a mechanism to specify the standard exchange format. The ontological capabilities for structuring information are being more and more used.

The BioPax organization has been working for years in defining a standard for representing metabolic, biochemical, transcription regulation, protein synthesis and signal transduction pathways [36]. It has already be taken as standard for the leading pathways resources like Kyoto Encyclopedia of Genes and Genomes [37], BioCyc [38] or Reactome [39]

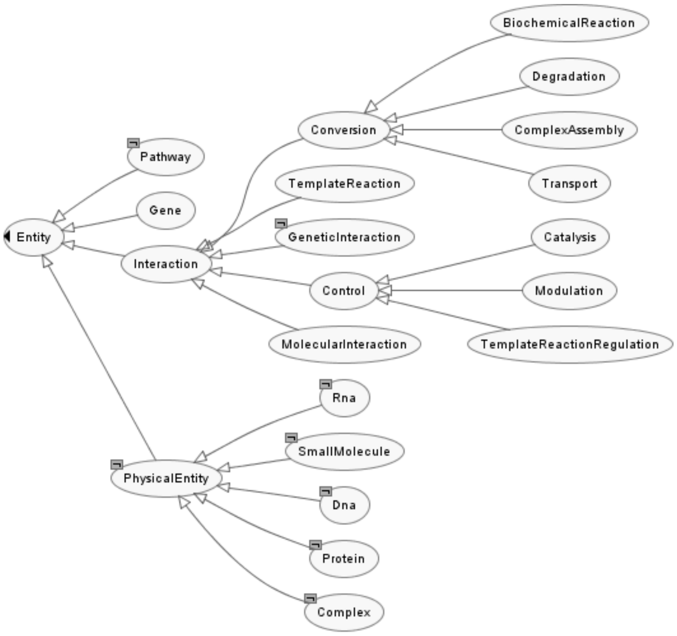

Figure 5 BioPax ontology browsed with Protégé

5) Semantic based Information Integration

The integration scenario of heterogeneous yet related data sources requires manual ad-hoc processing currently based in syntactic-based methods (e.g.: linking object with the same name facing polysemy, acronyms, abbreviations and synonymy related issues). Specifying the semantics of data in a variety of databases can enable researchers to integrate heterogeneous data across different databases. Linking entities in different data sources based on shared characteristics supported by an ontology that provides a common declarative foundation to describe biomedical content has proven to be a better approach. The additional ontological reasoning capabilities can support the linking process and resolve ambiguity and at the end of the day facilitate the integration and validation of disparate information.

The TAMBIS project implements an ontology driven integration middleware [23]

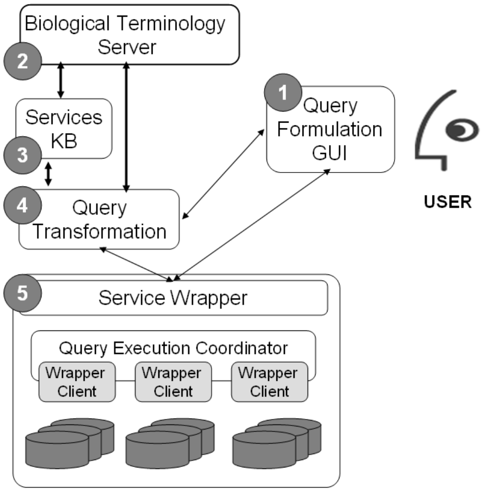

Figure 6 Information flow in TAMBIS

1- The user interacts with Query Formulation Dialogues, expressing queries in terms of the biological model. The dialogues are driven by the content of the model, guiding the user towards sensible queries. The query is then passed to the transformation process, which may require further user input to refine and instantiate the query.
2- The Terminology Server provides services for reasoning about concept models, answering questions like: What can I say about Proteins? Or what are the parents of concept X? It communicates with other modules through a well-defined interface
3- The Services Knowledge Base links the biological ontology with the sources and their schemas. This information is used by the transformation process to determine which source should be used.
4- Query Transformation takes the conceptual source-independent queries and rewrites to produce executable query plans. To do this it requires knowledge about the biological sources and the services they offer Information about particular user preferences - say favourite databases or analysis methods - may also be incorporated by the query planner. The query plans are then passed to the wrappers.
5- The Wrapper Service coordinates the execution of the query and sends each component to the appropriate source. Results are collected and returned to the user.

6) *Computer reasoning with data*

The competitive advantage of representing the knowledge by means of ontologies is the possibility to exploit knowledge by means of computer reasoning or the capability of making inferences based in the knowledge contained in the ontology, the contextual information and the asserted facts. For a scientist the panorama looks like a huge amount of well-structured information and a set of tools to analyze this information and allow for drawing meaningful inferences.

This steps means shifting from the mere information retrieval to the meaning of information mindsets. Typically, when a researcher is formulating hypothesis, it’s extremely difficult to verify that the data available support this hypothesis and if no, to figure out where the inconsistencies are. The need for tools capable of querying and interpreting the information at hand is becoming more and more incipient.

The HyBrow [40] or Hypothesis Browser allows for evaluating alternative hypotheses applying biological knowledge to integrated biological data –such gene expression, protein interactions and annotations-.

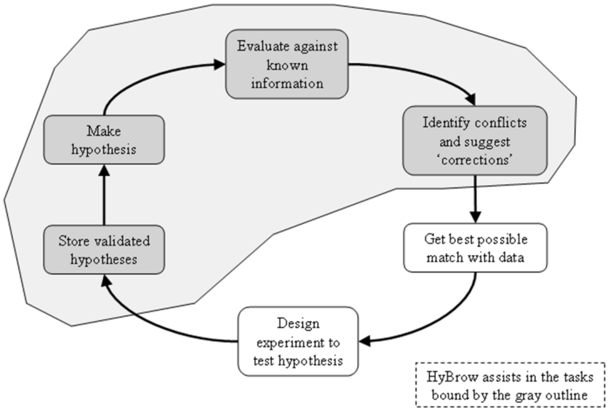
Evaluate against known information
Make hypothesis
Identify conflicts and suggest 'corrections'
Store validated hypotheses
Get best possible match with data
Design experiment to test hypothesis
HyBrow assists in the tasks bound by the gray outline

Figure 7 "Understanding cycle" proposed in the HyBrow project

As extensively discussed in [10], in the recent years, several formalisms have been proposed for modeling biochemical processes [4][5][6] or quantitatively [7][8] The tools that are being developed integrate a graphical user interface and a simulator, but only a very small subset manage to provide truly reasoning capabilities on the processes. For example, the Biocham [9] has been on the design of a biochemical rule language and a query language of the model in temporal logic, that are intended to be used by biologists. Biocham is a language and a programming environment for modeling biochemical systems, making simulations, and querying such models in temporal logic, composed by: a rule-based language for modeling biochemical systems, a simple simulator, a powerful query language based on Computation Tree Logic CTL and several interfaces for automatic evaluation of CTL queries.

V. Conclusion and Research Directions

This work explains the challenges the Life Sciences are faced with, and how the semantic web technologies are being used to address the major integration problems.

Ontologies, the semantic web cornerstone, are being adopted to solve a wide rage of problems, among them, the establishing of controlled vocabularies, the representation of knowledge, the specification of information models overcoming certain limitations of the classic UML, the specification of exchange formats to transfer knowledge between distributed systems, the integration capabilities empowered by the usage of semantics, or the usage of automatic reasoning techniques to discover or infer hidden knowledge inherent to the data model.

The uptake of this technique is so widespread, that many institutions start owning the development of a particular ontology. As an immediate result of it, the number of reference ontologies in biomedicine is constantly increasing.

The future will bring more formalism and therefore better analytical possibilities. Collaboration platforms for the community development of ontologies will also be a big area of research, as well as knowledge sharing for ontologies re-usage and expansion. Another point to be address in the near future is the mapping and overlapping of ontologies.

REFERENCES

[1] Berners-Lee, T., Hendler, J. and Lassila, O. "The Semantic Web," Scientific American, 284(5), 2001, pp. 34-43.

[2] Parry, D. ACM International Conference Proceeding Series; Vol. 54 Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32, New Zealand 2004

[3] De Roure, D.; Frey, J.; Michaelides, D.; Page, K. Collaborative Technologies and Systems, 2006. CTS 2006. International Symposium on Volume , Issue , 14-17 May 2006 Page(s): 411 – 418

[4] Regev, A., Silverman, W., and Shapiro, E. (2001a). *Representation and simulation of biochemical processes using the pi-calculus process algebra*. In Proceedings of the Pacific Symposium of Biocomputing,

[5] Nagasaki, M., Onami, S., Miyano, S., and Kitano, H. (2000). Biocalculus: Its concept, and an application for molecular interaction. In Currents in Computational Molecular Biology., volume 30 of Frontiers Science Series.

[6] Eker, S., Knapp, M., Laderoute, K., Lincoln, P., Meseguer, J., and Sönmez, M. K. (2002b). *Pathway logic: Symbolic analysis of biological signaling.* In Proceedings of the seventh Pacific Symposium on Biocomputing

[7] Matsuno, H., Doi, A., Nagasaki, M., and Miyano, S. (2000). *Hybrid Petri net representation of gene regulatory network.* In Pacific Symposium on Biocomputing

[8] Hofestädt, R. and Thelen, S. (1998). Quantitative modeling of biochemical networks. *In Silico Biology*

[9] Chabrier-Rivier, N., Fages, F., and Soliman, S. (2004). The biochemical abstract machine BIOCHAM. In Danos, V. and Schächter, V., editors, CMSB'04: Proceedings of the second Workshop on Computational Methods in Systems Biology, Lecture Notes in Computer Science. Springer-Verlag.

[10] State-of-the-art in Bioinformatics, Reasoning on the Web with Rules and Semantics (REWERSE Project), Deliverable A2-D1, available at http://rewerse.net/deliverables/a2-d1.pdf (accessed March 2009)

[11] Available at http://news.bbc.co.uk/2/hi/science/nature/805803.stm (accessed March 2009)

[12] Available at http://www.expasy.ch/sprot/ (accessed March

[13] Available at http://www.trembl.org/ (accessed March 2009)

[14] Available at http://www.ncbi.nlm.nih.gov/pubmed/ (accessed March 2009)

[15] Fielding, Roy T.; Taylor, Richard N. (2002-05), "Principled Design of the Modern Web Architecture" ACM Transactions on Internet Technology (TOIT) (New York: Association for Computing Machinery)

[16] Available at http://www.biowisdom.com/solutions/srs/ (accessed March 2009)

[17] Available at http://www.ncbi.nlm.nih.gov/sites/gquery (accessed March 2009)

[18] Available at http://bioinformatics.ubc.ca/atlas (accessed March 2009)

[19] Available at https://www3.ibm.com/solutions/lifesciences/solutions/discoverylink.html (accessed March 2009)

[20] Eckman, B., Kosky, A., and Laroco, L. (2001). Extending traditional query-based integration approaches for functional characterization of post-genomic data.

[21] Davidson, S., Overton, C., Tannen, V., and Wong, L. (1997). Biokleisli: A digital library for biomedical researchers. Journal of Digital Libraries.

[22] Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C., and Stoeckert, C. (2001). K2/kleisli and gus: Experiments in integrated access to genomic data sources. IBM Systems Journal, Issue on Deep computing for the life sciences.

[23] Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., and Brass, A. (2001). Transparent access to multiple bioinformatics information sources. IBM Systems Journal, Issue on Deep computing for the life sciences

[24] Kemp, G., Angelopoulos, N., and Gray, P. (2000). A schema-based approach to building a bioinformatics database federation. In Proceedings of the IEEE Inter- national Symposium on Bioinformatics and Biomedical Engineering

[25] N. Noy, et al., Creatring Semtric Web contentss with Protege!-2000, IEEE Intelligent Systems 16 (2) (2001 March/April).

[26] Smith B, Nigam S. Ontologies in biomedicine. How to make use of them (tutorial) Universtity of Buffalo, USA (available at http://bioontology.org/wiki/images/d/d2/ISMB_2007_Handout.pdf)

[27] D. Slomin. R. Tengi, WordNet, Princeton University Cognitive Science Lab., 1-003

[28] Gene Ontology, available at http://www.geneontology.org/ (accessed March 2009)

[29] GoPubMed project, available at http://www.gopubmed.com/ (accessed March 2009)

[30] US National Library of Medicine, available at http://www.nlm.nih.gov/ (accessed March 2009)

[31] Foundation Model of Anatomy, available at http://sig.biostr.washington.edu/projects/fm/FME/aboutFME.html (accessed March 2009)

[32] Protégé. Stanford University, available at http://protege.stanford.edu (accessed March 2009)

[33] OWL Web Ontology Language, available at http://www.w3.org/TR/owl-features/ (accessed March 2009)

[34] MGED, available at http://mged.sourceforge.net/ontologies/index.php (accessed March 2009)

[35] Ontology of Biomedical Investigation, available at http://obi-ontology.org/page/Main_Page (accessed March 2009)

[36] Biological Pathways Exchange, available at http://www.biopax.org/ (accessed March 2009)

[37] Genes and Genomes, available at http://www.genome.jp/kegg

[38] BioCyc, available at http://biocyc.org

[39] Reactome, available at http://www.reactome.org/

[40] Hypotheses Browser (HyBrow), available at http://www.hybrow.org/ (accessed March 2009)