Toward a New Paradigm for User Interaction on the Semantic Web to Support Life Sciences Investigation

Andrea Splendiani[1], Martin Kuiper[2] and Chris J Rawlings[1]

[1] Rothamsted Research, Biostatistics and Biomathematics Department, AL5 2JQ, Harpenden, UK
{andrea.splendiani, chris.rawlings}@bbsrc.ac.uk
[2] Systems Biology group, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway
kuiper@bio.ntnu.no

Abstract. Life Science information is increasingly available on the Semantic Web and this poses a demand for new tools and methodologies if it is to fulfill its potential to advance research in all areas of life sciences, including biomedicine. Life science information is obtained by traditionally distinct and varied scientific disciplines which explains why it is heterogeneous in its representation and in its semantics. The exploitation of this information by users relies only to a limited extent on well understood and shared formats, relations and metaphors; the interaction of users with biomedical information resources is an integral part of the definition and interpretation of the information that they provide. The need for this interactivity is reflected by current biomedical research practice. A range of life science software tools and methodologies focus on the analysis of biological networks and pathways. They provide interactive environments where relations among biological entities can be visualized and analyzed in the context of prior knowledge of biological systems or compared to experimental observations. Their representation of information, and the range of use cases they support, are broadly similar from what is provided by Semantic Web based life sciences information resources. Yet, despite similar motivation and efforts, there is still a disconnection between tools for the analysis of biological networks and Semantic Web knowledge bases. A few attempts to explore the synergies of a tighter integration of Semantic Web representation and Network analysis have been made, and this paper reports on the significance of this task, past experience and some idea for further developments.

Keywords: Semantic Web, Systems Biology, Life Sciences, User Interaction, Data Visualization.

1 Introduction

1.1 The Semantic Web in the Life Sciences

The Life Sciences has been one of the most enthusiastic adopters of Semantic Web technologies, and a key use case determining their evolution [1].

The first mention of the potential of these technologies in this area of research can be traced back to 2001 [2][3], a recent discussion can be found in [4]. Since then, a

considerable number of papers has been published reporting developments and experiences. Some reviews can be found in [5][6], a recent curated collection in [7]. The advantages that Semantic Web technologies bring to the integration of life sciences information stem from a few of its properties:

First, this information is heterogeneous and in constant evolution. The framework provided by RDF and SPARQL is effective for the development and for the fruition of integrated information resources, whose content can vary in extension and complexity not only in time, but on a per-application basis [8].

Second, the explicit declaration of semantics through ontologies, a central feature of the Semantic Web, fits the need for definitions of shared computable terminologies that is characteristic of the life sciences. This requirement has ancient roots, that can be traced in the definition of Anatomy by Aristotle, or in the use of terminologies for epidemiological statistics already in the 18th century [9]. Terminological systems such as UMLS [10] and SNOMED [11] are now at the cornerstone of modern biomedical information systems.

The last decade in particular has seen a revolution of biomedical research paradigms induced by the availability of information from biological systems on a multi-genomics scale. This paradigm switch, and the resulting need for computable description systems, has resulted in a strong emphasis on the development and integration of ontologies spanning several life sciences disciplines [12]. Finally, the web is assuming an increasing role in publishing scientific information: databases, scientific journals and the results of wiki based collaborative annotation efforts are available on the web, with an increasing level of semantic annotation [13][14][15][16][17].

For the reasons outlined above, a number of significant life sciences resources are already available on the Semantic Web. Some of the most important for research in systems biology are the collection of Open Biomedical Ontologies [18], PathwayCommons [19], a collection of models of biological systems represented in RDF[1], Uniprot [21], a comprehensive resource of information on proteins, and the Cell Cycle application ontology [22]. It should be noted that the information provided by these resources is all in the public domain.

1.2 Limits to the benefits of the Semantic Web Resources for the Life Sciences

Despite the availability of life science information represented in Semantic Web standards, and the promising features of the related technologies for data integration in this domain, practicing scientists are still far from being able to exploit the Semantic Web in their current research. One of the main reasons for this is the lack of tools and methodologies that exploit its full potential.

In their current incarnation, Semantic Web knowledge bases are constructed via the aggregation of public and domain-specific resources in centralized systems [23][24][25]. These knowledge bases offer a SPARQL interface for access to their content, sometimes complemented by web interfaces that make it easier to compose queries and visualize results [24].

This approach provides only a limited support to an effective user interaction, for a number of reasons:

First, after inspecting the results of a query, which corresponds to testing an hypothesis, the user may want to formulate a related query (corresponding to a related hypothesis) rather than follow links from its results. This implies that the user is forced, at least to some extent, to write SPARQL queries in order to interactively inspect the content of a knowledge base.

1 More precisely, via BioPAX [20], an exchange language that is based on an OWL ontology.

Second, the user doesn't know which relations will lead to the required information. Unlike other areas, that rely on a common understanding of entities and relations, different Semantic Web resources contained in the same knowledge base can be based on significantly different conceptualizations of a common domain.

As an example, both PathwayCommons and some of the OBO ontologies represent biological processes as Semantic Web resources. A typical query for both such resources would ask for all elements (proteins or genes) involved in a process. However, there is not a common “participation” relation in these systems, and the formulation of a query would require a detailed knowledge of both representations.

One solution to this problem would be to define common relations (such as “participation”) in the knowledge base.

A different solution, that doesn't require the definition of relations but exploits better user interaction with the knowledge base, will be presented later.

A third reason for the limited effectiveness of current Semantic Web knowledge bases in supporting user interaction is that in life sciences research, the exploration of data is often interactive, and based on the emergence of patterns or singularities: a visual representation of information that can highlight similarities and dissimilarities is essential. In fact, one of the major application of ontologies in biomedical research is in providing a functional characterization of patterns emerging from experimental data (Fig. 1).

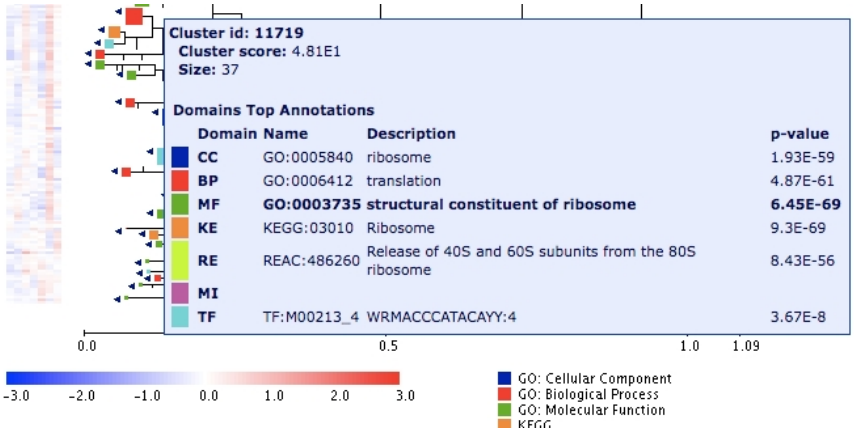

Fig. 1. Example of the use of ontologies to provide a functional characterization of patterns emerging from experimental data. On the left, gene expression values for different genes (rows) and conditions (columns) are clustered based on their profiles and visualized through a color code (bottom). To the right, a tree (partially visible) explains patterns found in these gene expression values. To each node in the tree a color code is associated that represents the most relevant annotation (as defined via ontologies) for each pattern. The pop-up box presents details and statistics for one of these patterns. Source: VisHic [26]

Finally, the information in which the user is interested is not necessarily what is represented in a knowledge base. In general, it is the information that can be derived from a knowledge base through some user assumptions that are more likely to be of interest in a specific context.

While the Semantic Web framework makes it possible to aggregate information first and to define its interpretation in an application context, this feature is rarely used to enable the users with the possibility of providing their context specific

interpretation of the information. An example of how this could be done is provided in the next session.

1.3 Enabling the Semantic Web in Life Science investigation.

One the most direct ways to advance the adoption of the Semantic Web in current research practice is to reformulate current biomedical investigation methodologies and tools so that they can profit from a Semantic Web based representation of information.

In the life sciences, software tools for the analysis of biological networks, such as Ondex [27] or Cytoscape [28] have emerged that provide feature-rich interactive environments for the analysis and visualization of information that is structured similarly to a Semantic Web representation. In the remainder of this paper we report some experiences in the direction of the integration of these tools and Semantic Web environments to support user interaction, and some idea for future developments. Given the variety of life sciences information, we will focus our attention on networks representing biological processes, or pathways.

2 Biological Network Analysis and the Semantic Web

Life science research is based on the observation of biological systems, and on the discovery and definition of their underlying mechanisms. This research proceeds by first individuating the elements that constitute a biological system, then by studying their relation or their organization in networks, and finally by considering the dynamic aspects of their interrelations. A large amount of research is focused on the second stage of this process, the study of biological networks, and several tools have been developed to support their integration and analysis [29].

We present in this section two of these tools that support the analysis of biological networks on a Semantic Web based knowledge representation. They combine information such as experimental data with classes and relations, and they provide interactive visual environment to explore Semantic Web information resources.

2.1 Ondex

Ondex is a suite for the integration and analysis of biological information coming from a variety of sources such as databases, experimental datasets, derived data and text-mining. Ondex is based on an internal representation (partially discussed in [30]) that is affine to a combination of RDF and OWL[2].

In this representation biological entities and their relations are represented as a graph where nodes and edges are decorated with attributes and associated to types. Types are organized in a hierarchy, characterized by their properties, and provide a support for inference.

This Semantic Annotation of biological entities supports user interaction tasks such as filtering of information, customization of its rendering and layout.

---

2 Ondex allows the possibility to import RDF data. The alignment of the Ondex data model to RDF and OWL is in process.

An example of a layout that is based on the semantic annotation of biological networks is presented in Fig. 2.

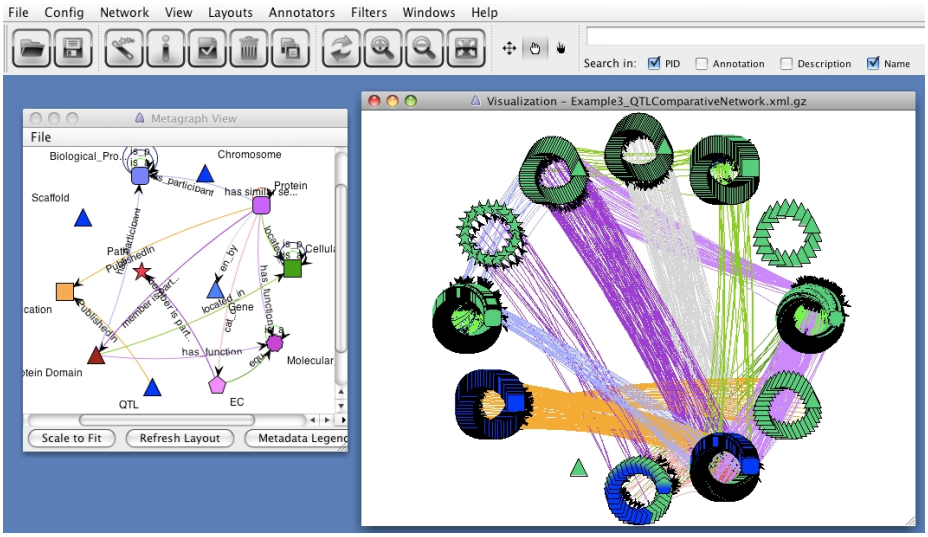

Fig. 2. Screenshot from Ondex that represent a biological network (right): layout, color coding and icons are based on the semantic categorization of its elements. To the left, a "metagraph view" shows classes and the relations among classes: it introduces color codes and icons that are used in the network visualization.

2.2 RDFScape

RDFScape is a Cytoscape plugin that explores the integration of Semantic Web resources into a widely used tool for the interactive visualization and analysis of biological networks [31].

One advantage of this integration is the use of an interactive environment to browse and query Semantic Web resources in an intuitive way. For instance, RDFScape allows interactive browsing of knowledge base contents (Fig. 3a) that can be customized depending on user needs: users can decide which resources to visualize based on their namespace, which color association to choose, or whether some elements should be rendered by multiple nodes. It also offers a means to define SPARQL queries in a visual and interactive way (Fig. 3b). A relevant feature of RDFScape is that it allows users to provide their interpretation of the observed knowledge base. This can be done via a combination of OWL axioms and inference rules that can be provided and modified interactively by users. As a result, users query, browse and visualize not the facts that are stated in a knowledge base, but deductions that are relevant for a particular investigation context. An example of this is reported in Fig. 4, for a more extended discussion we refer to [31]. It should be noted that the limits of the expressiveness of Semantic Web languages to support interpretations that are relevant to biomedical research are still to be investigated.

Finally, RDFScape also provides a mechanism to interactively associate elements in a Semantic Web knowledge base to experimental data represented in Cytoscape, and to define SPARQL queries that span the integrated information.

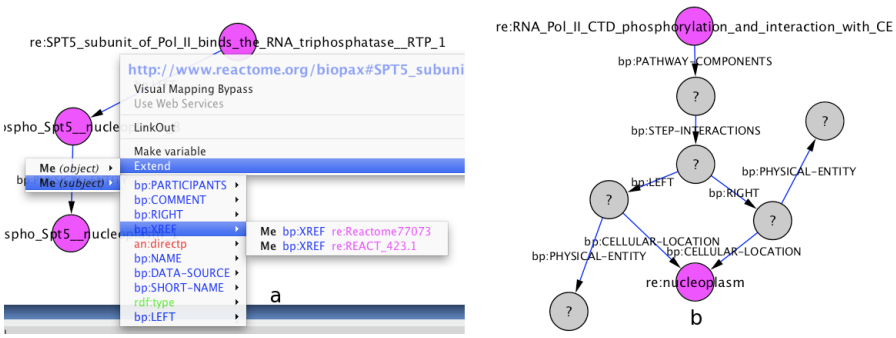

Fig. 3. a) Browsing the content of a knowledge base through an interactive right-selection system. b) A visual query (in this case all elements of a given process whose action takes place in a given cellular location). Visual queries are defined via the interactive browsing mechanism (note in (a) the the menu item “make variable”).

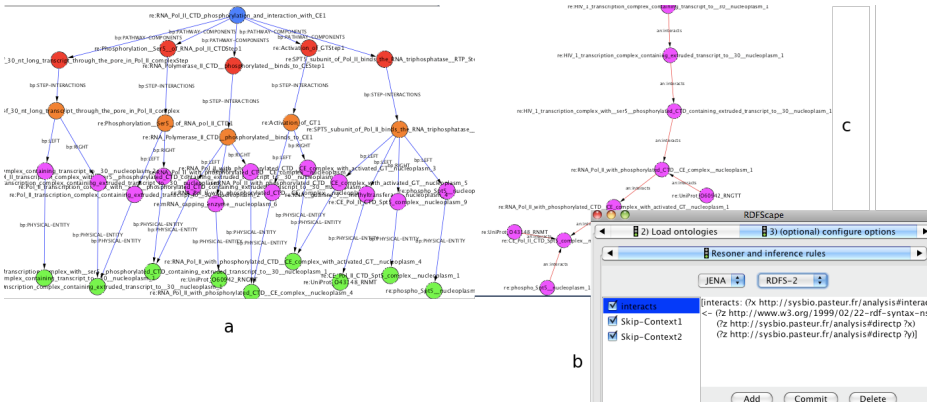

Fig. 4. Different rendering of a pathway, depending on the interpretation provided interactively by the user. a) Representation of a pathway as it is described in the original knowledge base. The pathway is represented by a blue node, pathway steps are colored in red, reactions in orange, the context of reactions in purple, and the interacting elements in green. b) Detail of the interpretation of a pathway provided by the user through inference rules. c) The pathway in (a) seen as an interaction network, given the interpretation provided in (c)[3].

3 The elements in purple in (c) correspond to the elements in green in (a).

3 New Paradigms for User Interaction on the Semantic Web in Life Sciences Investigation

Ondex and RDFScape are examples of how current biological network analysis systems can be brought to interact with Semantic Web based representation of Life Sciences knowledge in a mutually beneficial way. This is a first step in providing interactive systems that can exploit the full potential of the Semantic Web in current research.

New paradigms for user interactions and for the visualization of the information are required: what is needed are systems and methods that minimize the amount of information that the user has to process sequentially (reading), that make visually evident emerging patterns, and that support an interaction in which the user provides his know-how and his contextual interpretation of the data.

To take these ideas further, we can look for new paradigms in two directions.

First, we can research whether interaction techniques developed for Semantic Web systems in other domains can be exploited in the Life Science investigation context.

Alternatively we can then look for completely new paradigms, or metaphors, in different domains that share the same interaction needs and where the information has the same characteristics.

In the remainder of this section we will introduce two examples of potentially new interaction paradigms.

3.1 Visualizing the meaning of experimental data.

Many biological data analysis systems make use of ontologies to explain experimental data. A prototypical example of one of these systems is shown in Fig.1, where ontology terms are associated to patterns found in gene expression data. How this association is computed is out of the scope of this manuscript (we refer to [32] for an introduction). We focus here on how the resulting information is made available to the user. In the example considered, data patterns are annotated with ontology terms, via color coding.

This representation of the association of ontologies to data is suboptimal for two reasons: first, ontologies are *de facto* reduced to tags. The semantic characterizations of terms and their relations is not presented.

Second, the representation of ontology terms via color coding is perceptually suboptimal: a color code associated with a spatial representation can provide an intuitive representation for continuous information, as gradients map the change of continuous values in space (a common example is the representation of altitude on topographic maps). When rendering biological data or networks, the spatial aspect is important, as it highlights patterns or relevant parts of the network, but the color coding of annotations (discrete information) can only act as a legend, and introduce an indirect step into its interpretation.

As it is, this representation of ontologies to characterize experimental data is less efficient than simple tag clouds, commonly used in websites for a characterization of their content with a very limited semantic commitment.

We can then start with the inspiration of a simple tag cloud to imagine a method that would provide a more effective way to visualize this information, as exemplified in Fig. 5.

As previously discussed, one problem of the interaction of users with Semantic Web knowledge bases is that the sequence of selections, or of relations, that will lead to the sought information is hidden (we refer to Fig. 3 for an example of such interaction). The problem is that users can only see one step ahead and only where the next immediate selection will lead. We can imagine a system that extends simple tag clouds and that could compensate for this lack of foresight by associating to each possible selection a range of reachable classes. The distance from the center of the cloud could represent the number of interaction steps that information is away, while the size or the color of a single item could indicate the amount of individuals that are instances of each class. Selection of a specific tag (a proxy to a class) would lead to a different representation, or subset of information, which is relative to it.

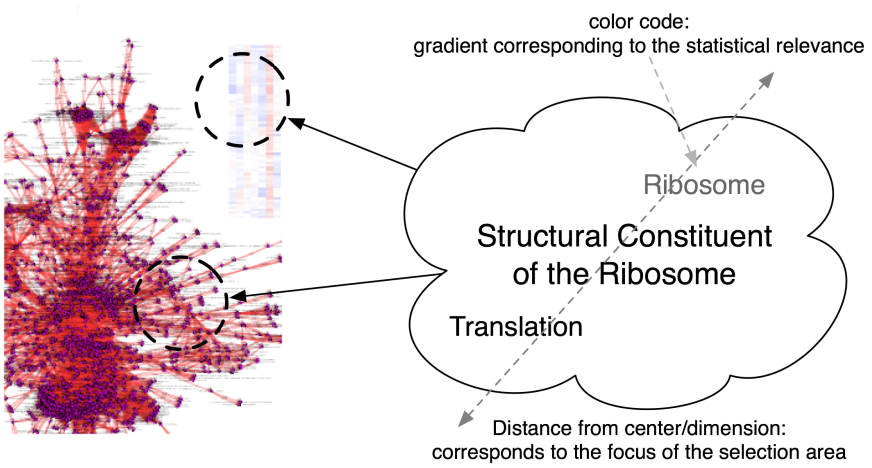

Fig.5. Example of a method to visualize information on the functional annotation of patterns in a dataset, or of clusters in a network (Fig. 2). This method takes inspiration from a simple tag cloud. Here distance from the center of the cloud is used to represent the specificity of the annotation with respect to the focus of the selection area (indicated by a radius around the pointer location). The statistical relevance of the association of elements to a given annotation can be encoded through a color gradient.

3.2 Less Is More (Watching a Marathon)

The estimated number of genes in *Homo sapiens* is about twenty three thousand. The number of participants in the New York City Marathon in 2008 was slightly below thirty-eight thousand. If we observe the response of a human cell to a given stimulus across its whole genome, we are dealing with an amount of information that is comparable to watching a marathon.

It can be argued that representing the behavior of an event involving all genes in a cell is more complex then representing the process of a marathon, for the complexity of the implications of gene regulation for subsequent expression and e.g. metabolic events. However, it is worth considering what makes the marathon an easy event to follow, and if some elements of it can be applied to the exploration of biological systems information.

In the case of the marathon, not all of the information is presented to the user: the focus is on professional runners, on notable events (perhaps a non professional in the front row?), and on aggregate values (first and last participant's arrival time, average arrival time...). This mix of detailed and synthetic information can be a basis for a new interaction paradigm. Users could focus on a set of entities they know of, for instance genes involved in biological pathway under investigation. In a network layout these could be assigned fixed position and they could function as an intuitive anchor for the interpretation of the remaining elements. For these remaining elements, only aggregate values, or detail of "outstanding" elements could be displayed (some of these ideas are sketched in Fig. 6).

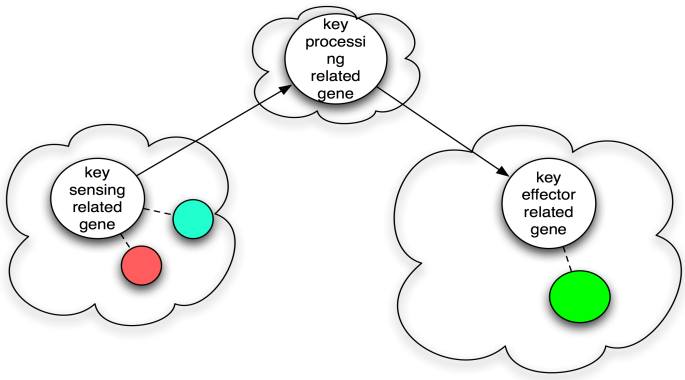

Fig.6. Schematic example of a possible rendering for a biological network that mixes detailed and synthetic information. In this simple example, three genes, corresponding to key genes of the sensing, processing and actuation parts of a signaling network are placed in fixed positions. They provide an anchor for an intuitive representation of the flow of signal in the pathway. The remaining genes are visualized as sub-networks (here shown as clouds) tightly clustered to the landmarks and whose size reflects a significant aggregate value, such as the overall variance in gene expression variation for the included genes. Only genes that perform significantly differently from the average in each sub-network are visualized in detail with a color code indicating their deviation from the average expression value.

4 Discussion: Beyond Biological Networks

Network based analysis tools can be extended to capitalize on the features of a Semantic Web based representation of biological information. This allows new paradigms for user interaction, where the visualization of information can intuitively convey a biological meaning, and where the relations among entities can be interactively explored and defined by the user.

While most of these tools have been developed in a biomedical context, they are essentially domain independent, and could be used as generic interactive system to explore the content of generic knowledge bases, even outside the Life Science area, such a web archives.

In this possible application domain, they present limitations and interesting features. Most of the limits stem from the lack of extensions that can handle data types that are not specific to the life sciences. Even simple examples such as images, and in general multimedia data, are not commonly found in life science Semantic Web knowledge bases.

A relevant limitation is in the lack of support for privacy, provenance and protection of the information. This is a very relevant topic in the life science domain: medical records pose evident privacy issue; the investments required to perform experiments induces limitations to the premature disclosure of derived data; the need to refer to trusted resources is an intrinsic part of science itself.

The use of network based analysis tools naturally introduces the possibility to compute analysis of the information they explore. For instance, algorithms for computing topological properties of networks such as “connectivity”, “betweeness” or “centrality” of their nodes could be used to determine the relevance of information in different contexts, such as the most influential papers in a citation network (a review of such algorithms can be found in [33]). This possibility, coupled with significant interactive and graphic capabilities, poses the basis for a new way to explore the Semantic Web.

References

1. W3C Semantic Web Health Care and Life Sciences (HCLS) Interest Group, http://www.w3.org/2001/sw/hcls/
2. Berners-Lee and T., Handler, J.: Publishing on the Semantic Web. Nature 410, 1032-1034 (2001)
3. Handler, J.: Communication. Science and the Semantic Web. Science 299, 520-521 (2003)
4. Antezana, E., Kuiper, M., Mironov, V.: Biological knowledge management: the emerging role of Semantic Web technologies. Brief. Bioinform. 10, 392-407 (2009)
5. Cheung, K.H., Prud'hommeaux, E., Wang, Y., Stephens, S.: Semantic Web for Health Care and Life Sciences: a review of the state of the art. Brief. Bioinform. 10, 111-113 (2009)
6. Ruttenberg, A., et al.: Life Sciences and the Semantic Web: the Neurocommons and beyond. Brief. Bioinform. 10, 193-204 (2009)
7. Semantic Web Applications and Tools for Life Sciences. BMC Bioinformatics 10, S10. Edited by Burger, A., Romano, P., Paschke, A., Splendiani, A. (2009)
8. Anwar, N. and Hunt, E.: Francisella tularensis novicida proteomic and transcriptomic data integration and annotation based on semantic web technologies. BMC Bioinformatics 10 (S10), S3 (2009)
9. Bodenreider, O. and Stevens, R.: Bio-ontologies: current trends and future directions. Brief. Bioinform. 7, 256–274 (2006)
10. Bodenrider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. N.A.R. 32 (Database issue), D267-70 (2004)
11. Cornet, R. and Keizer, N.: Forty years of SNOMED: a literature review. BMC medical informatics and decision making 8 (S1) (2008)
12. Smith, B. et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotech. 25, 1251-1255 (2007)
13. Mons, B. et al.: Calling on a million minds for community annotation in WikiProteins. Genome Biol. 9, R89 (2008)
14. Hoffmann, R.: A wiki for the life sciences where authorship matters. Nat. Genet. 40, 1047-51 (2008)
15. Cockerill, M.J. And Tracz, V.: Open access and the future of the scientific research article. J. Neurosci. 26, 10079-81 (2006)
16. Shotton, D. et al.: Adventures in semantic publishing: exemplar semantic enhancements of a research article. PloS Comput Biol. 5, e1000361 (2009)

17. Concept Web Alliance, http://conceptweblog.wordpress.com/
18. Open Biomedical Ontologies, http://www.obofoundry.org/
19. Pathway Commons, http://www.pathwaycommons.org
20. BioPAX : Biological Pathways Exchange, http:www.biopax.org
21. UniProt RDF, http://dev.isb-sib.ch/projects/uniprot-rdf/
22. Antezana, E. et al.: The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome Biol. 10, R58 (2009)
23. Ruttenberg, A. et al.: Advancing translational research with the Semantic Web. BMC Bioinformatics 8(S3), S2 (2007)
24. Antezana E. et al.: BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 10 (S10), S11 (2009)
25. Belleau, F. et al.: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 41, 706-716 (2008)
26. Krushevskaya, D. et al.: VisHiC – hierarchical functional enrichment analysis of microarray data. N.A.R. 37 (Web Server issue), w587-w592 (2009)
27. Köhler, J. et al.: Graph-based analysis and visualization of experimental results with Ondex. Bioinformatics 22, 1383-1390 (2006)
28. Shannon, P. et al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13, 2498-504 (2003)
29. Pavlopoulos, G.A. et al.: A survey of visualization tools for biological network analysis. BioData Mining 1, 12 (2008)
30. Taubert, J. et al.: The OXL format for the exchange of integrated datasets. Journal of Integrative Bioinformatics 4, 62 (2007)
31. Splendiani, A.: RDFScape: Semantic Web meets Systems Biology. BMC Bioinformatics 9 (S4), S6 (2008)
32. Khatri, P. and Draghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21(18), 3587-3595 (2005)
33. Huber, W. et al.: Graphs in molecular biology. BMC Bioinformatics 8 (S6), S8. (2007)