Inf Syst Front (2009) 11:471–480
DOI 10.1007/s10796-009-9159-y

Knowledge management in biomedical libraries: A semantic web approach

Damaris Fuentes-Lorenzo · Jorge Morato · Juan Miguel Gómez

Published online: 12 February 2009

Abstract In recent years, technological advances in high-throughput techniques and efficient data gathering methods, coupled with a world-wide effort in computational biology, have resulted in an enormous amount of life science data available in repositories devoted to biomedical literature. These repositories lack the ability to attain an effective and accurate search. Using semantic technologies as the key for interoperation enables searching and processing of biomedical literature in a more efficient way. However, emerging semantic applications take for granted specific knowledge that biomedical researchers may not have. This paper presents design principles for easy-to-use biomedical semantic applications by means of ontology-based annotations and faceted search. The proposed approach is backed with a usable prototype that shows the breakthroughs of adding these principles to a biomedical digital library where identifying and searching information are critical aspects for non-semantic Web experts.

Keywords Semantic web · Annotation · Faceted browsing · Digital libraries · Biomedical data · Usability

D. Fuentes-Lorenzo (✉)
IMDEA Networks,
Av. del Mar Mediterráneo 22,
28918 Madrid, Spain
e-mail: damaris.fuentes@imdea.org

J. Morato · J. M. Gómez
Carlos III University,
Av. de la Universidad 30,
28911 Madrid, Spain

J. Morato
e-mail: jorgeluis.morato@uc3m.es

J. M. Gómez
e-mail: juanmiguel.gomez@uc3m.es

1 Introduction

As discussed in Cohen (2004), it is undeniable that biology and medicine played a key role in the twentieth century. Over the last 15 years, a dramatic transformation in the practice of life sciences research has been witnessed. One of the central factors promoting the importance of biology is its relationship with medicine. Fundamental progress in medicine depends on elucidating some of the mysteries that occur in the biological sciences. However, biomedical research is now information intensive; the volume and diversity of new data, as shown in Jurisica and Glasgow (2006), challenge traditional database technologies.

The Semantic Web has emerged as an attempt to provide machine-processable metadata to the ever increasing information resources on the Web. As its creator Tim Berners-Lee said (Berners-Lee et al. 2001), the Semantic Web "is not a separate Web, but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation".

Therefore, the breakthrough of adding semantic metadata to biomedical applications is providing a new level of data and process integration that can be leveraged to develop novel high-performance data and process management systems. The two-pronged use of ontologies allows humans to grasp the meaning of any element and, secondly, allows machines to have formal semantics to support reasoning.

The remainder of the paper is organized as follows. Section 2 describes the motivation statement for the research. Section 3 identifies the problems which arise when semantics or Semantic Web techniques are applied to traditional applications. Section 4 mentions and compares different related works that try to overcome those problems. In Section 5, the solution is presented, implemented in *BioSem*, a prototype for semantic social-oriented management of biomedical literature. Conclusions are discussed in Section 6 and finally, future work is presented in Section 7.

2 Motivation

When the Web appeared into the arena, it made biomedical literature available across applications (such as Medline[1] or PubMed[2]), departments and entities in general. However, throughout these developments, a particular underlying problem has remained unsolved: articles are stored in a weak structure (or not structured at all), a disadvantage that sometimes makes this information unusable. At best, when the resources (i.e. articles or papers) are categorised, metadata used reside in thousands of incompatible formats and cannot be systematically managed, integrated, unified or cleaned. To make matters worse, this incompatibility is not limited to the use of different data technologies or to the multiple different “flavours” of each technology (e.g. different relational databases in existence), but also extends to an incompatibility in terms of semantics.

For example, one database or metadata source could have a term called “Tylenol”, intending to model a particular drug and classifying its function, categorizing it and relating it with some other similar drugs. Another source could simply refer to the same concept as “DCI”, “acetaminophen” or “NO2 BE01”. These concepts (despite being the same) will never be related or correlated as synonyms, except when specified by a user. If a particular researcher wants to know all the information published about “Tylenol”, he will not be able to obtain a detailed overview of the information, since the concepts used to categorize the resources are absolutely unrelated. In a larger context, this problem may be multiplied by thousands of data structures located in hundreds of incompatible sources and formats. In fact, in (Hoffmann and Valencia 2005) it is stated that there are 19 synonyms on average for every gene (with a total of 3.2 millions of synonyms) in the Gene Synonym Dictionary.

Therefore, as the information-retrieval tools currently available to researchers in biomedicine lie far behind the possibilities that a suitable management tool should offer, the need for an online service that provides a way for efficiently accessing millions of linked scientific literature documents, becomes obvious.

3 Problems

The increasing volume of textual information being stored without a clear structure makes applications in the worst cases useless. Actually, applications such as digital libraries can not be used as fully-fledged tools to create and search knowledge in an efficient way, because the information collected within these systems lies unused by computers, mainly due to the human language in which the resources are written and classified. As further processing is needed, new formal approaches are needed to make computers “understand” and interpret the Web content.

In the following list, the problems which emerge when constructing a semantically-enhanced biomedical literature environment are described, including technical and social factors:

- *Metadata representation format*. Metadata support for the actual information in any biomedical application must be explicitly declared. Some of the current social tools such as the Web 2.0 applications like *Flickr* (http://www.flickr.com) or *del.icio.us* (http://del.icio.us/) apply so-called “folksonomies” to add meta-information by means of tags chosen by the user (O’Reilly 2005). In this case, even though usability may be high, tags are different among different users; since tags are chosen freely, they cannot be fully exploited in a community in terms of knowledge discovery, as presented by statistical reports in (Chi and Mytkowicz 2007).
- *Navigation*. Hypermedia authoring applications base the relation between resources in explicit hyperlinks. These links relate one page to another basically according to user considerations (if the user authoring considers they are related). According to a study by (Désilets et al. 2005), the lack of adequate support for link creation and management is a key usability problem amongst authoring applications. If the relation between pages or pieces of data were represented by means of semantics, the application would be able to provide mechanisms to semantically navigate between related resources with real meaning.
- *Search*. Given a set of resources, the basic type of querying in current digital libraries is the keyword-based search. Structured requests for more advanced information retrieval are needed in order to convert a biomedical library to a useful knowledge repository. In addition to simple full-text searches, users must be able to search biomedical literature by querying the semantic knowledge, avoiding syntactic-related problems such as articles with related synonymous words, which are related but not returned when a search query is performed.
- *Integration*. Biomedical applications may need the integration and analysis of diverse types of data that

[1] Medline: http://www.nlm.nih.gov/medlineplus/spanish/encyclopedia.html

[2] PubMed: http://www.ncbi.nlm.nih.gov/pubmed/

can be distributed across many heterogeneous databases. Nevertheless, their structural schemas and their contents are not exposed in a standardized machine-processable way to allow seamless integration.

- *Usability*. Communities need both a critical mass of users and their participation. Without this mass, users will abandon the systems underlying these communities (Ignacimuthu 2004). For that purpose, applications enhanced with semantic functionalities have to be designed with maximum usability and minimum cognitive load for every user, independently of their expertise level.

The Semantic Web paradigm is based on the traditional Web, enhanced with formal knowledge placed below the current information. This is possible thanks to the extensibility of the Web with semantic metadata and semantic metadata processing, which allows computational reasoning and intelligent capabilities (Berners-Lee et al. 2001). The fundamental aim of the Semantic Web is to provide a response to the ever-growing need for knowledge integration on the Web. One of the main benefits of adding semantics is bridging nomenclature and terminological inconsistencies to comprehend the underlying meaning in a unified manner.

The next section describes some previous approaches which have applied Semantic Web techniques in order to develop semantically interoperable environments.

4 Related work

Applying Semantic Web techniques to digital libraries and content management applications is an advancing field which has already been the subject of extensive research.

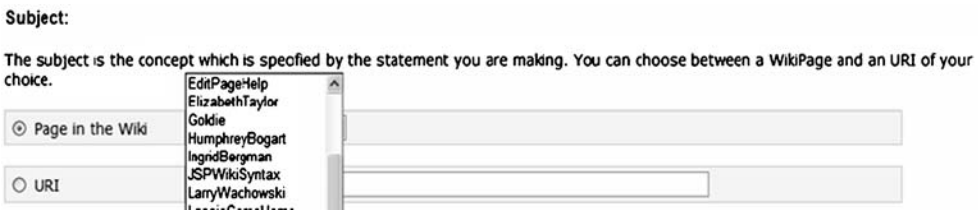
Concerning digital libraries, *JeromeDL* (http://www.jeromedl.org/) intends to apply semantic knowledge to traditional digital libraries. However, semantic data makes assumptions about the articles in general, not focusing on semantic information within their contents. Apart from that, the inclusion of semantic information still has to be carried out programmatically; that is, authoring functionalities are not available.

Approaches for authoring semantic desktop applications have already been initiated, such as the ones in Decker et al. (2006). *iAnnotate Tab* is an OWL-based *Protégé* plug-in for manually annotating text or Web documents with ontological concepts. A user can select a fragment of text and create instances for a given concept. Besides the drawback of having to install *Protégé* and the plug-in itself, there are no navigability or search functionalities and so the text annotations remain useless.

Apart from these desktop applications, Web approaches have also been implemented. Personal semantic wikis, such as *SemperWiki* (Oren et al. 2006) are created for personal use, where users can embed RDF statements within their normal text. Even though included in the text, these statements refer to information which relates to the whole page and not particular fragments in the content. *Makna and MultiMakna* (Nixon and Paslaru 2006) are also wiki-based applications extended with principles from multimedia and semantics research. However, the annotations are not particularly intuitive for a user without any specific knowledge the about Semantic Web. An example of the annotation view is presented in Fig. 1, where the user must understand what a subject is in an RDF triple statement. Furthermore, the annotation view appears separately from the content-edition view.

Semantic MediaWiki (SMW, http://ontoworld.org/wiki/) is an extension of *MediaWiki*—the wiki-system powering *Wikipedia*—with semantic technology (Krötzsch et al. 2006). However, this application is focused on the Semantic Web community, and still lacks user-friendliness for a user with no semantic knowledge. *IkeWiki* also allows users to annotate pages and links between pages with semantic annotations (Schaffert 2006). All annotations in both *Semantic MediaWiki* and *IkeWiki* are always with respect to the current document (annotations are merely properties about a document). These applications, along with *Makna*, seem to be rather more advanced than *BioSem*; however, the intended users in these applications must also have an advanced level. Moreover, the annotation view is separated from the normal editing environment (you have to use a separate annotation mode).

Regarding design principles, some of the preconditions in Möller and Decker (2005) have been taken into account for the design of the current system, as will be shown in the next sections. Besides, multi-ontologies and knowledge-layered frameworks such as ours have been provided in Xiao and Cruz (2006), however, with a different focus.



Subject:

The subject is the concept which is specfied by the statement you are making. You can choose between a WikiPage and an URI of your choice.

◉ Page in the Wiki

○ URI

EditPageHelp
ElizabethTaylor
Goldie
HumphreyBogart
IngridBergman
JSPWikiSyntax
LarryWachowski

Fig. 1 An example of the annotation view in Makna (Nixon and Paslaru 2006)

In what follows, it is explained how to solve the shortcomings existing in the above and other Semantic Web applications, both in theory and in practice, in order to achieve the development of an efficient biomedical digital library.

5 Reaching semantics in biomedical systems: BioSem

Given the aforementioned problems characteristic of those faced by traditional semantic applications, the proposed approach is explained here, based on several design principles to solve these drawbacks, and built as a socially-oriented environment for literature management, *BioSem*.

5.1 Semantic representation

Design Bearing in mind that metadata processing requires a controlled and well-defined vocabulary, the Semantic Web developed ontologies as the best mechanism to represent, share, reuse and integrate the knowledge hidden within text. One of the most well-known definitions of ontology is the one stated by Borst (extending Gruber's one (Gruber 1993)), who defines it as "a formal specification of a shared conceptualization" (Borst 1997). Progress has already been made to organize biological or medical knowledge in a conceptual way by developing ontologies and domain-specific vocabularies, such as GALEN (http://www.opengalen.org/), UMLS (http://www.nlm.nih.gov/research/umls/) or ON9 (http://saussure.irmkant.rm.cnr.it / ON9/index.html). However, we have found that ontologies designed from medical terminology are not well modeled as ontologies. Most of these 'translations' model every term as classes, when this is not exactly correct (e.g. a specific vitamin, such as vitamin B9, is not a subclass of vitamin, but an individual of vitamin class), giving as a result another taxonomy, but not a real ontology. Therefore, in *BioSem* ontologies have been selected as the representation mechanism rather than folksonomies.

Implementation *BioSem* ontologies can be represented using the various standards proposed by the World Wide Web Consortium (W3C), such as the *Resource Description Framework* (RDF) (World Wide Web Consortium 2006a) or the *RDF Schema* (World Wide Web Consortium 2004). *Web Ontology Language* may also be suitable (World Wide Web Consortium 2006b), since it also provides inference mechanisms to allow reasoning. Other research is being carried out in this area, where some of the more standardized medical ontologies are being developed using OWL, such as SNOMED (Wroe 2006). Also, some other research is developing database schemas definitions to conform to W3C standards, allowing *BioSem* to use them as domain ontologies (Lam et al. 2006, 2007).

5.2 Semantic scope

Design As every piece of content in *BioSem* is a resource, they must be first described as such. Once identified, the resources must be described with regard to their content.

Implementation Dublin Core initiative (DC, http://dublincore.org/) is appropriate as the main ontology for describing the minimal piece of data in *BioSem*; every paper is semantically enhanced with DC metadata. Then, one or more biomedical ontologies are used for formalizing the real domain of the content. The basic domain ontology used for the first version of the prototype is an OWL ontology for biological pathways called *BioPax* (http://www.biopax.org). It should be noted that more than one different domain ontology could be used, and that these potentially different domains could be integrated with ontological engineering processes such as merging or alignment. In this application stage, we have opted for this ontology because is more general and rather reduced, and it is more suitable for the interaction of the users who will participate in the future evaluation study.

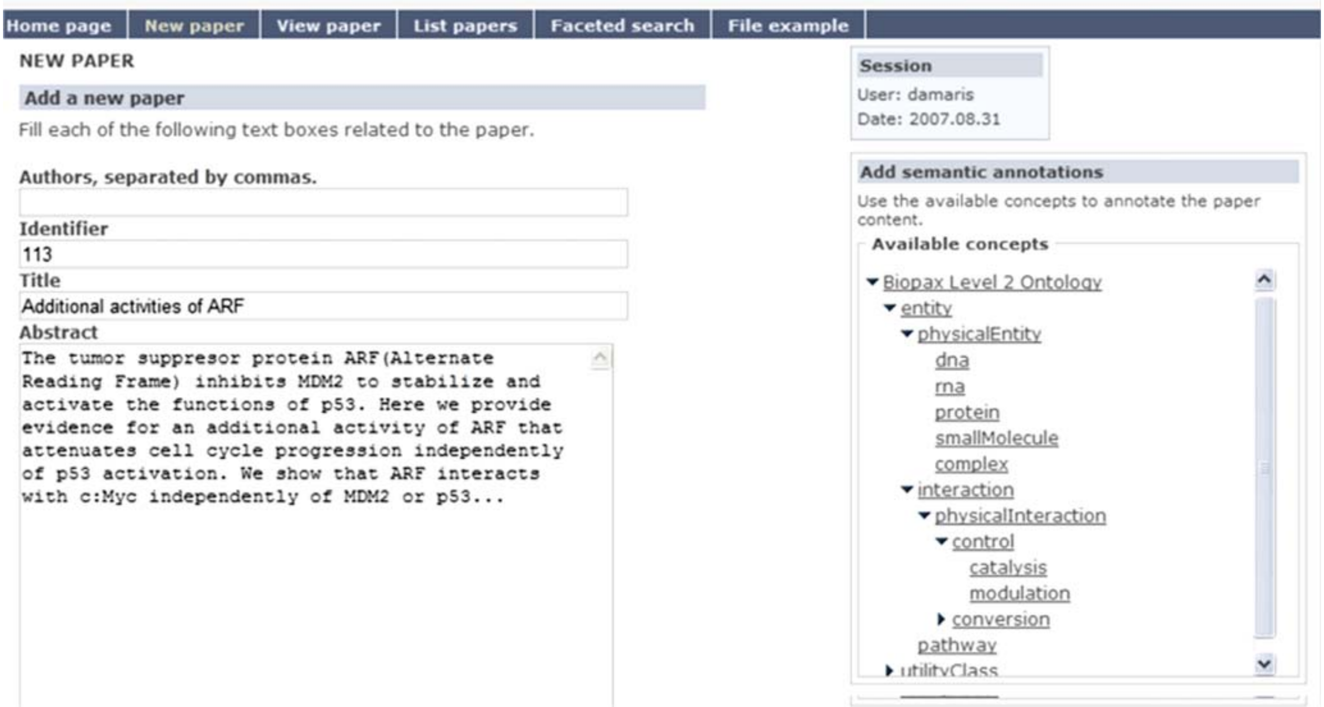
In Fig. 2 an example of how the ontology is presented to the user in *BioSem* is illustrated.

Users write the title and authors of a paper, separate from the content[3]. These metadata properties are stored as *Dublin Core* terms of this new resource. The only requirement requested from the user is to fill in the text boxes in a form; it is not necessary to hold knowledge about DC vocabulary. This is in contrast to what is expected of users in other semantic authoring applications such as *SemperWiki* (Oren et al. 2006), where users themselves have to type DC properties and their intended values.

5.3 Creating semantics

Design Authoring the resources in a semantic-based biomedical application must be made just as easy as authoring them for a traditional one. For this purpose, editing the resources must be done at the same time and in the same view panel as editing the semantic content. The way the user adds the DC metadata to the global paper has been presented in Fig. 2. To add semantic information to the content itself, *BioSem* makes use of the *semantic annotations* to fill the gap, where annotating a document means adding semantic data to it (McEnery and Wilson 2001).

[3] For this version, the content is not a full paper, but an abstract. *BioSem* may include more DC elements and terms, but these are enough for the purpose of the example.



NEW PAPER

Add a new paper

Fill each of the following text boxes related to the paper.

Authors, separated by commas.

Identifier: 113

Title: Additional activities of ARF

Abstract:

```
The tumor suppresor protein ARF(Alternate
Reading Frame) inhibits MDM2 to stabilize and
activate the functions of p53. Here we provide
evidence for an additional activity of ARF that
attenuates cell cycle progression independently
of p53 activation. We show that ARF interacts
with c:Myc independently of MDM2 or p53...
```

Session

User: damaris

Date: 2007.08.31

Add semantic annotations

Use the available concepts to annotate the paper content.

Fig. 2 BioSem: Creating a new resource

Implementation As Fig. 3 illustrates, users are provided with tree-like structured semantic information to add; therefore, while editing a paper, they can annotate a word or a set of words with semantic data, just as easy as marking the selected words on the text and associating them with a vocabulary concept from the ontology domain. The text changes automatically to include the related concept within brackets.

In *BioSem*, there are two ways of adding semantic data to the information:

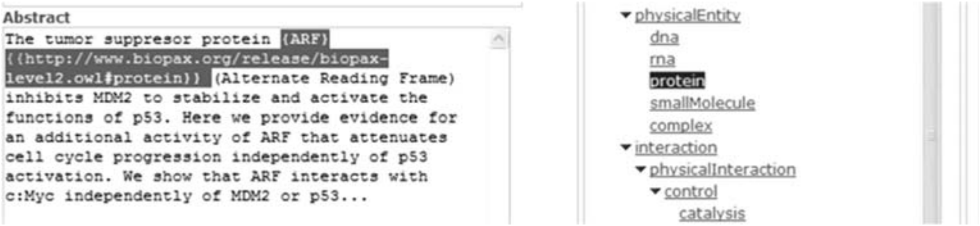
- *Relating the content with an existing instance*. In this case, the tree does not display only the concept, but also the concrete individual. To illustrate this by means of an example, suppose the concept "Gene", and one of its associated instances handling below, "Apoe E4". This is usual with ontologies in thesaurus form such as the NCI thesaurus, available at the *OBO Foundry* (http://obofoundry.org/cgi-bin/detail.cgi?ncithesaurus), also in OWL format.
- *Relating a resource with a new instance*. In this case, the concept is modelled in the ontology, but there are not yet any individuals related to the concept. The annotation has both to create the instance, and to relate it to the paper content, as Fig. 3 shows.

5.4 Navigation

Design The user interface must enable navigation between semantically-related items (Teevan et al. 2004) and ordinary hyperlinks are not sufficient to display this kind of information. For this reason, semantic links, or *semalinks* are proposed, which are ordinary hyperlinks in appearance but built upon the semantic information. This semantic link, which consists of both the ontological concept to which a certain part of the content is referring and its value (that is, the instance created while authoring), leads the user to resources with a semantically similar content, as this *semalink* information indicates.

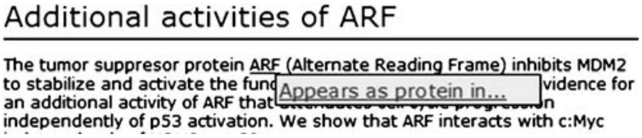
Implementation Figure 4 shows the visual information a *semalink* can provide. As depicted in Fig. 3, the word "ARF" was annotated as "protein".

Fig. 3 BioSem: Semantically-annotated content



Abstract

```
The tumor suppresor protein (ARF)
((http://www.biopax.org/release/biopax-
level2.owl#protein)) (Alternate Reading Frame)
inhibits MDM2 to stabilize and activate the
functions of p53. Here we provide evidence for
an additional activity of ARF that attenuates
cell cycle progression independently of p53
activation. We show that ARF interacts with
c:Myc independently of MDM2 or p53...
```



Additional activities of ARF

The tumor suppresor protein ARF (Alternate Reading Frame) inhibits MDM2 to stabilize and activate the func vidence for an additional activity of ARF that n independently of p53 activation. We show that ARF interacts with c:Myc

Fig. 4 BioSem: Rendered paper with graphical nodes

When this paper is provided to the user, *BioSem* transforms this annotation into a *semalink*. If the user places the mouse over this link, graphical nodes appear upon the text (see Fig. 4). These graphical nodes make reference to:

- Resources that were annotated with both the concept and value the *semalink* indicates. In the example, the user can navigate to papers annotated with a “protein” called “ARF”. Notice that if a paper was annotated with a “virus” concept called also “ARF”, this paper would not appear in the *semalink* references. This is because “ARF”-virus annotation has the same value as “ARF”-protein annotation syntactically, but not semantically.
- Resources that were annotated with any pair of semantic concept and value that fits with any of the values of any of the properties of the *semalink* instance. Imagine the “protein” class has a property called “is involved”, and the range of this property is an “illness”. Suppose then that “ARF” instance “is involved” in “cancer” (e.g., another instance). Then, the *semalink* will show graphical nodes making reference to other papers that were annotated with “cancer” as “illness”.

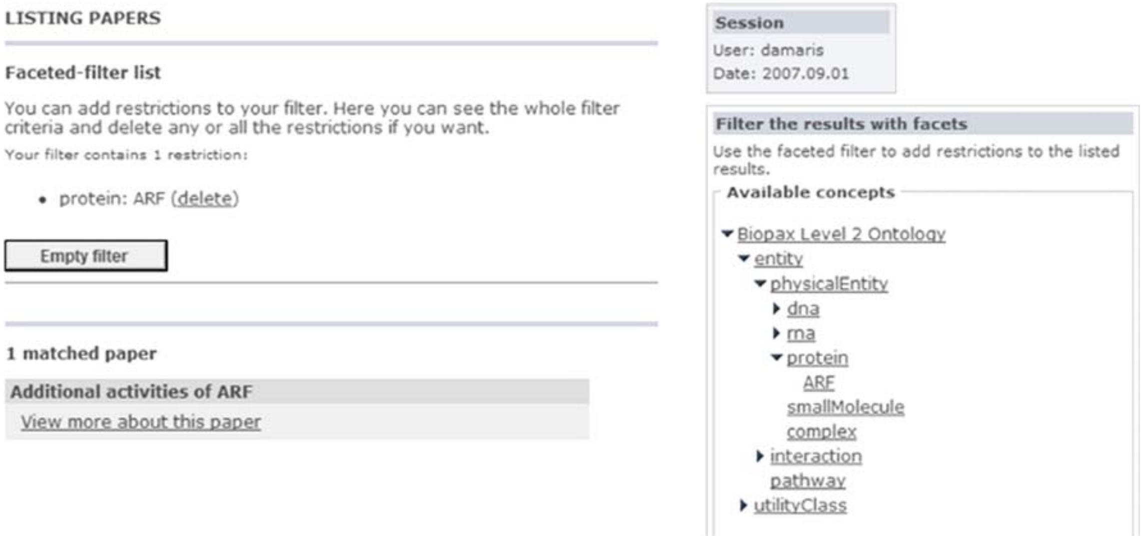
The information related could have been presented to the user keeping a metaphorical RDF network, such as *WikSar* (http://wiki.navigable.info/) or the *Fenfire* project (http://fenfire.org/) does. However, as this is not scalable for large data sets, *BioSem* shows the directly related information (that is, the one related with the direct properties of a concept).

5.5 Faceted search

Design Since keyword-based searches or other different syntactical queries are not an efficient retrieval mechanism, and provided that semantic information is the basis of the current system, a more advanced search is required. A facet-based search is the solution. With faceted metadata (Ranganathan 1962), the information space is partitioned using orthogonal conceptual dimensions of the data. These dimensions are called *facets*, and represent the characteristics of the information elements. These facets are then used to select or filter the relevant elements in a certain information space, leading users to the exact information required.

Implementation In *BioSem*, the information elements are the papers of the library; a facet is a pair formed by a concept and its value; that is, a facet is a semantic annotation.

If the user wants to search a certain resource, a faceted filter is provided (see Fig. 5). For this example, the tree view shows the possible instances in the repository, originating from their parent classes. As can be seen, a new individual appears under “protein” concept, “ARF”. When the instance is selected, a new restriction is added to the faceted filter. Figure 5 shows also the results of applying the filter with the “protein” concept and the “ARF” value as one of its restriction facets; the information space is reduced to the exact resource the user is interested in.



LISTING PAPERS

Faceted-filter list

You can add restrictions to your filter. Here you can see the whole filter criteria and delete any or all the restrictions if you want.

Your filter contains 1 restriction:

- protein: ARF (delete)

1 matched paper

Additional activities of ARF

Session

User: damaris

Date: 2007.09.01

Filter the results with facets

Use the faceted filter to add restrictions to the listed results.

Available concepts

Fig. 5 BioSem: Faceted filter for searching

5.6 Architectural and programming features

Architecture Taking into account the different levels of knowledge (ontologies, resources, etc.), these levels are divided into three layers:

- *Resource layer*. This layer manages the textual resources of the biomedical application; that is, the papers.
- *Domain layer*. This layer deals with the ontologies used to formalize the semantic information for both the resources (DC vocabulary) and their content.
- *Application layer*. This layer, supported on top of the previous one, is built with the domain ontologies the application requires and it is applied to the resources in the first layer.

Keeping these knowledge layers conceptually separated, the *BioSem* implementation guarantees the flexibility and reusability of the biomedical application for every kind of bioinformatics domain.

Figure 6 shows the framework of this approach, along with named examples for better understanding. The *Domain layer* holds the different domain ontologies that can be used. The *Application layer* uses one or more domain ontologies depending on the type of topics the application is going to deal with, and creates the possible semantic instances (assertional data). The Dublin Core Ontology is used to represent the basic metadata concepts of every resource, such as title, authors, etc.

Programming BioSem is implemented by means of *Ruby on Rails* (RoR, http://www.rubyonrails.org) (Thomas et al. 2005), a Model View Controller-based framework which eases the task of building this architectural pattern, built upon the scripting language *Ruby*. The researchers considered that this framework provides two basic characteristics which facilitates the development of Semantic Web applications:

- *Rails* eases the maintainability of the Model-View-Controller pattern (Reenskaug 1979) while developing the application.
- A dynamic scripting language, *Ruby*, as opposed to static languages, to build the semantic digital library. As stated in Oren et al. (2007a, b), there may be mismatches among RDF and compiled (static) object-oriented languages. One of them is the inheritance; in common OO type systems, classes can only inherit from at most one superclass, while RDF(S) classes can inherit from multiple superclasses. The other important issue is the flexibility. OO compiled systems usually do not allow class definitions to evolve during runtime. However, RDF is designed for integration of heterogeneous data with varying structure, from varying sources, and this is especially critical in biomedical domain.

Pages are presented to the user in XHTML 1.0 syntax, combined with Cascade Style Sheets (CSS) and visual graphics for navigation, made with JavaScript libraries such as *CoolTip* (http://www.acooltip.com). Emerging technologies such as AJAX (Garrett 2005) have been used to implement the ontology tree, which can load data asynchronously as the user needs it. Desktop applications are characterized by rich and responsive features which are frequently unavailable on the Web. This leads to better usability, in the sense of providing a quick response to user requests. With the classic Web applications, the user is obliged to wait when some process is being executed on the server side, and this is especially crucial when it is required to display large sets of data, such as those characteristic of

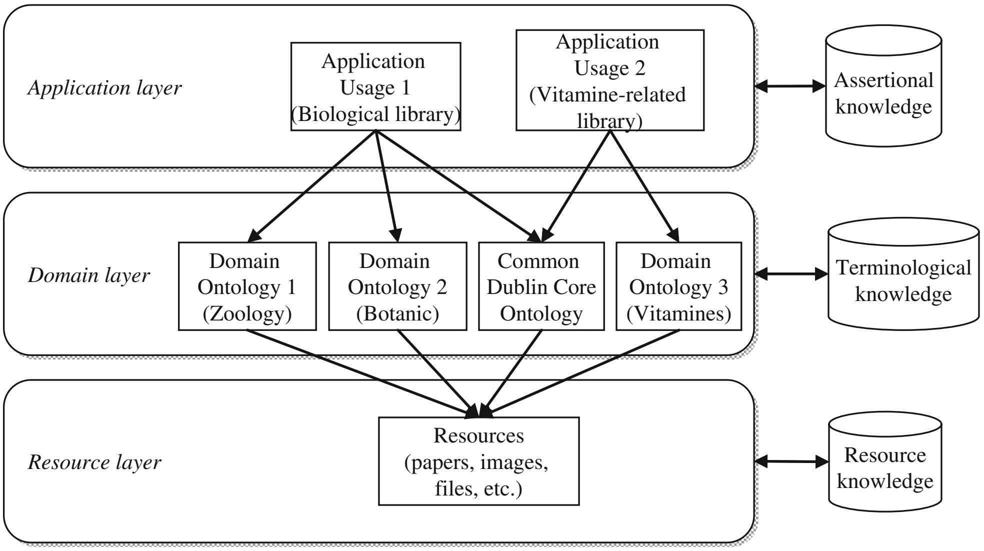

Application layer
Application Usage 1 (Biological library)
Application Usage 2 (Vitamine-related library)
Assertional knowledge
Domain layer
Domain Ontology 1 (Zoology)
Domain Ontology 2 (Botanic)
Common Dublin Core Ontology
Domain Ontology 3 (Vitamines)
Terminological knowledge
Resource layer
Resources (papers, images, files, etc.)
Resource knowledge

Fig. 6 BioSem knowledge layers

some biomedical domains. With the technologies involved in AJAX, just the appropriate fragment of a web page is updated on the server side, so the client does not have to wait any longer.

Persistence repositories are *MySQL* server for resources information and *SQLite*- based *RDFLite*, for semantic information. Finally, *BioSem* uses *ActiveRDF* (Oren et al. 2007a, b), a library for abstracting the queries for *RDFLite* within the implementation in *RoR*.

6 Comparison

The following table shows a brief comparison of *BioSem* functionality against the functionality in similar digital library applications or annotation tools (Table 1).

Notice that even though most of the applications keep the metadata models in a structured way, some of them have lightweight vocabularies such as JeromeDL[4].

As far as the approach is concerned, notice that most of the web-based applications need a prior installation of the wiki software to get running, a time-consuming task that is not needed with *BioSem*[5].

7 Conclusions

This paper has presented a new vision for a semantically-enhanced digital library in the biomedical domain, where authoring and adding semantic data are easily performed thanks to guided semantic annotations. Faceted searches and graphical visualizations help the user to find more accurate information and semantically-related data when needed. Components in the architecture are loosely coupled in order to provide flexibility in the implementation phase and to provide extensibility for applying the architecture in any domain of bioinformatics.

In spite of the fact that most of the features presented have been developed separately in other related work, the work presented here gathers all the state-of-art advantages stated there, avoiding the pitfalls of their current implementations (see Section 4 to see the most characteristic shortcomings). *BioSem* development is based on design principles which take into account the two most important ends of the line of biomedical science and the Semantic Web paradigm: research and knowledge.

Table 1 Comparison among applications

	Approach	Metadata			Hypermedia	Search	Integration	Usability		Annot. mode	Other	
	Desktop (D)/Web (W)	Structured (S)/nat. language (N)	Document (D)/content (C)	Authoring functionalities: no (N)/yes (Y)	Automat. link: no (N)/yes (Y)	Structured (S)/nat.-language (N)	Export in RDF,OWL: no (N)/yes (Y)	Tool knowledge (expertise level)	Level of usability	Content and edition in the same screen: no (N)/yes (Y)	Wiki based: no (N)/yes (Y)	Ont. browser: no (N)/yes (Y)
JeromeDL	W	S	D	N	Y	S, N	N	Medium	Medium	NA	N	Y
iAnnotate	D	S	C	Y	N	No	Y	Medium	High	Y	N	Y
KIM	D, W	S	C	Y	N	S	Y	Medium	High	Y	N	Y
SHOE	D	S	C	Y	N	S	N	High	Low	N	N	Y
SMORE	D	S	C	N	N	No	Y	High	Low	NA	N	Y
Diigo	W	N	D, C	Y	Y	No	N	Low	High	Y	N	N
SemperWiki	D	S	D	Y	Y	S	N	High	Medium	Y	Y	N
Makna	W	S	C	Y	Y	N, S	N	High	Poor	N	Y	N
SemanticMedia Wiki	W	S	C	Y	Y	N, S	Y	High	Low	N	Y	Y
BioSem	W	S	D, C	Y	Y	S	Y	Low	High	Y	Y	Y

[4] www.jeromedl.org

[5] *BioSem* is located at http://mendelson.gast.it.uc3m.es:3000/biosem. It is necessary to register at http://mendelson.gast.it.uc3m.es:3000/accounts/signup

8 Future work

The kernel of this application looks promising to be the focus of multiple approaches in E-Science. Some of them are stated below:

- *Navigating through health data records*. The *BioSem* digital library could focus on health-records annotation. This would serve as a base for better health care in medicine, with the possibility for health researchers to find similarities in illnesses and improve the current diagnoses.
- *Application to manage biomedical ontologies*. In this paper, the reader has seen *BioSem* can create the assertional knowledge, that is, the individuals (instances) of the biomedical concepts (e.g. "ARF" protein). In the future, *BioSem* could also be applied to manage the terminological knowledge: the concepts themselves (e.g. protein).
- *An application to manage special groups of studies, such as gene and proteomics analysis*. Similar work has been developed in Bresell et al. (2006) or Camon et al. (2004) to study gene lists.
- *Automatic tagging*. The annotation process could be studied to be automatic or semiautomatic, to facilitate the biomedical knowledge generation and transmission.

All these future enhancements to *BioSem* would influence the progress of information retrieval biomedical research.

References

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*, 34–44.

Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse*. Enschede: Twente University Press.

Bresell, A., Servenius, B., & Persson, B. (2006). Ontology annotation treebrowser: an interactive tool where the complementarity of medical subject headings and gene ontology improves the interpretation of gene lists. *Applied Bioinformatics*, *5*(4), 225–236.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., et al. (2004). The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Research*, *32* (Database issue), D262–266, England.

Chi, E. H., & Mytkowicz, T. (2007). Understanding navigability of social tagging systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*, San Jose, USA.

Cohen, J. (2004). Bioinformatics—an introduction for computer scientists. *ACM Computing Surveys*, *36*(2), 122–158.

Decker, S., Park, J., Sauermann, L., Auer, S., & Handschuh, S. (Eds.). (2006). *Proceedings of the semantic desktop and social semantic collaboration workshop (SemDesk 2006) located at the 5th international semantic web conference ISWC 2006*. Athens, GA, USA.

Désilets, A., Paquet, S., & Vinson, N. G. (2005). Are wikis usable? *WikiSym '05: Proceedings of the 2005 International Symposium on Wikis*, San Diego, California. 3–15.

Garrett, J. J. (2005). Ajax: a new approach to web applications. *Adaptive Path.*

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Hoffmann, R., & Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, *21*(2), 252–258.

Ignacimuthu, S. (2004). *Basic bioinformatics*. Alpha Science International.

Jurisica, I., & Glasgow, J. (Eds.). (2006). Knowledge Discovery in High-Throughput Biological Domains. *Information systems frontier, 8* (1). Hingham: Kluwer.

Krötzsch, M., Vrandecic, D., & Völkel, M. (2006). Semantic MediaWiki. *Proceedings of the 5th International Semantic Web Conference (ISWC06)*, Athens, GA, USA.

Lam, H. Y., Marenco, L., Shepherd, G. M., Miller, P. L., & Cheung, K. H. (2006). Using web ontology language to integrate heterogeneous databases in the neurosciences. *AMIA Annual Symposium Proceedings*, 464–468.

Lam, H. Y., Marenco, L., Clark, T., Gao, Y., Kinoshita, J., Shepherd, G., et al. (2007). AlzPharm: integration of neurodegeneration data using RDF. *BMC Bioinformatics*, *8*(Suppl 3), S4 England.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

Möller, K., & Decker, S. (2005). Harvesting desktop data for semantic blogging. *Proceedings of the 1st Workshop on the Semantic Desktop at the ISWC 2005 Conference*.

Nixon, L. J. B., & Paslaru, E. (2006). Makna and MultiMakna: Towards semantic and multimedia capability in wikis for the emerging web. *Proceedings of the Semantics 2006: From Visions to Applications*, Austria.

O'Reilly, T. (2005). What is web 2.0? *O'Reilly NetWork*, retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Oren, E., Volkel, M., Breslin, J. G., & Decker, S. (2006). Semantic wikis for personal knowledge management. *Database and Expert Systems Applications*, *4080*, 509–518.

Oren, E., Delbru, R., Gerke, S., Haller, A., & Decker, S. (2007a). ActiveRDF: Object-oriented semantic web programming. *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada.

Oren, E., Mesnage, C., Heitmann, B., Haller, A., Hauswirth, M., & Decker, S. (2007b). A flexible integration framework for semantic web 2.0 applications. *IEEE Software*, *24*(5).

Ranganathan, S. R. (1962). *Elements of library classification: based on lectures delivered at the University of Bombay in December 1944 and in the schools of librarianship in Great Britain in December 1956* (3rd ed.). Bombay: Asia Publishing House.

Reenskaug, T. (1979). *The original MVC reports*. Oslo: T. Reenskaug.

Schaffert, S. (2006). IkeWiki: a semantic wiki for collaborative knowledge management. *Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE 2006)*, Manchester, United Kingdom.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria.

Thomas, D., Heinemeier Hansson, D., & Breedt, L. (2005). *Agile web development with rails: a pragmatic guide*. Raleigh: The Pragmatic Bookshelf.

World Wide Web Consortium. (2004). *RDF vocabulary description language 1.0: RDF schema*. Retrieved March 30, 2007, from http://www.w3.org/TR/rdf-schema/.

World Wide Web Consortium. (2006a). *Resource description framework (RDF)*. Retrieved March 30, 2007, from http://www.w3.org/RDF/

World Wide Web Consortium. (2006b). *Web ontology language (OWL)*. Retrieved March, 30, 2007, from http://www.w3.org/2004/OWL/.

Wroe, C. (2006). Is semantic web technology ready for healthcare? Paper presented at the *3rd European Semantic Web Conference (ESWC'06)*, Budva, Montenegro.

Xiao, H., & Cruz, I. (2006). Application design and interoperability for managing personal information in the semantic desktop. *Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop (SemDesk 2006), at the 5th International Semantic Web Conference (ISWC 2006)*, Athens, Georgia, USA.

Damaris Fuentes Lorenzo holds a B.S in Technical Engineering in Computer Managements from the Universidad Carlos III, Madrid (Spain) and a M.Sc in Computer Engineering, Development of the Enterprise Information System Specialization. She also holds a M.Sc in Computer Science and Technology, specialized in Software Engineering. She has took up several scholarships in both the Departments of Telematics and Computer Science of the Universidad Carlos III, involved in the latter as a research technician. Currently, she is working as a research assistant at IMDEA Networks, while doing her Ph.D. in Telematics as a member of WebTLab group in the same university.

Jorge Morato He is currently a professor of Information Science in the Department of Computer Science at the Carlos III University of Madrid (Spain). He obtained his PhD in Library Science from the Carlos III University in 1999 on the subject of Knowledge Information Systems and its relationships with linguistics. Professor Morato has taught courses on Information Retrieval, Search Engine Optimization, Software Engineering, and Knowledge Modelling Techniques and Management Systems. From 1991–1999, he had grants or contracts from the Spanish National Research Council. His current research activity is focused on text mining, information extraction and pattern recognition, NLP, information retrieval, Web positioning, and Knowledge Organization Systems. He has published mainly on semi-automatic construction of thesauri and ontologies, topic maps, and conceptual and contextualized retrieval of semantic documents.

Juan Miguel Gomez is an Associate Professor at the Computer Science Department in the Universidad Carlos III, Madrid, Spain. He holds a PhD 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 Politecnica 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.