A Unifying Framework for Semantic Annotation

Davide Eynard, David Laniado, and Marco Colombetti

Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5, 20133 Milano, Italy
{eynard,david.laniado,colombet}@elet.polimi.it

Abstract. In the modern Web, users have the technical possibility to write anyhing about anything else; however this information is unstructured and not always easy to find and access. Semantic annotation systems merge user participation with the advantages that derive from structured knowledge, but they still haven't gained much success among the average internet users. We believe that a unifying model, which takes inspiration from existing annotation systems and extends them by allowing users to define their own vocabularies, might help the diffusion of this kind of tools and the creation of structured metadata. In this paper, after an analysis of requirements and best practices for collaborative semantic annotation, we develop a model which should be general enough to be compatible with many available social applications and that allows to semantically express both annotation metadata and structured knowledge about resources being annotated. We test the model by applying it for the development of an annotation system based on Semantic MediaWiki, with an architecture which relies on semantic standards and technologies such as RDF and SPARQL endpoints. Finally, we show how metadata could be employed inside annotation vocabularies to aggregate and elaborate annotations, providing more interesting results as an incentive for user participation and augmenting user navigation experience in many different ways.

1 Introduction

With Web2.0, the dream of a Read/Write Web seems realized: today Internet users have the technical possibility to write anyhing about anything else on the Web. However most of this information is published by humans for humans, it is unstructured (often just in the form of plain text), thus difficult to search and hard to reuse in other applications.

Semantic annotation systems face this problem by allowing users to specify metadata about a resource in a semantic format: this way, for instance, it is much easier to query for annotations related to a given resource or written by a specific author. However, semantic annotations still haven't gained enough success among the average Internet users: the reasons might be various, such as a technological barrier at the entrance, the lack of tools with a friendly user interface, the difficulty of setting up annotation servers[1], and most important

[1] For instance, `http://www.w3.org/1999/02/26-modules/User/Annotations-HOWTO`.

to us the fact that users are often constrained by common, not very expressive vocabularies for annotations.

We believe the times are ready for a richer and more expressive model for semantic annotation of Web resources, allowing for the specification of structured knowledge both concerning annotation metadata and the resources being annotated, where users can not only insert new content according to existing ontologies, but also collaboratively edit the semantic model, reusing existing ontologies and vocabularies and extending them. Of great importance for the success of this kind of systems is the capability, on one hand, to provide easy interfaces for contribution requiring the minimum user effort, and on the other hand to exploit the annotations in a useful way and provide instant gratification to users, enriching the possibilities and the efficiency of their navigation experience.

In this paper, after an overview on the state of art (Section 2), in Section 3 we describe the motivations for this work and our vision of a collaborative approach for semantic annotation of Web resources, and we define a coherent list of requirements; in Section 4 we propose a general model for semantic annotation; in Section 5 we show how a system complying with our approach and model can be implemented using existing technologies, describing our first prototype based on the Semantic MediaWiki platform; finally, in Section 6, we summarize conclusions and directions for future work.

2 Related work

One of the first systems to semantically annotate Web pages was SHOE, a platform that allowed to mark-up HTML documents on the basis of existing ontologies [7]; another framework supporting ontology-based annotation of Web pages is CREAM[5]. [11] introduce a Semantic Markup Tool, based on templates to hide ontological complexity from end users and allow them to easily specify new instances in the knowledge base. [16] propose the Mangrove tool with the intention to "entice ordinary people onto the semantic Web", by providing them an easy graphical interface to annotate HTML documents with semantic metadata, and on the other hand by making these metadata immediately available to a series of semantic services, such as semantic search and calendar. Saha is an annotation editor supporting the usage of different metadata schemas and domain ontologies[22].

A milestone is for sure the W3C project Annotea, aimed at providing a semantic annotation framework, to enhance collaboration via shared metadata based Web annotations, bookmarks, and their combinations[8]. It uses an RDF based annotation schema for describing annotations as metadata and XPointer[2] for locating the annotations in the annotated document. Different client softwares for Annotea have been built: among these Annozilla[3], created as an ex-

[2] `http://www.w3.org/XML/Linking`
[3] `http://annozilla.mozdev.org/`

tension of the Mozilla browser, and the Web editor Amaya[4]. Whereas the (extensible) vocabulary allows a certain richness of expressivity for describing annotation metadata and also the type of annotation (e.g. *Comment*, *Example* and *Change*), the content of annotations is just limited to unstructured data.

Another tool to share annotations about any Web page (or part of a page) is CritLink[25]; of particular interest for our work is the idea of a *mediator* in the user navigation experience, providing additional information related to the page they are visiting, and in particular showing *extrinsic* links, defined collaboratively, in addition to the *intrinsic* ones (i.e., the link embedded by the author in the source web page).

The scarce success that semantic annotation systems have encountered so far is counterbalanced in the recent years by the rapid diffusion and growth of *folksonomies*, or collaborative tagging systems. A tag can be considered as a very simple kind of annotation, where users just assign a keyword to a resource; the semantics provided by each user is shallow, but the strength of applications like Flickr[5] or del.icio.us[6] resides in the high number of active users, achieved also thanks to the extremely low effort required. [26] propose a model for semantic annotation generation, exploiting emergent semantics from folksonomies by mining tag co-occurrences.There have been several proposals of vocabularies for tagging systems[18, 13, 3]; a broadly accepted starting point for a formalization of tagging is a tripartite model, where a tagging action is seen as a triple (User, Resource, Tag)[4, 17]. A comprehensive revision and comparison of tagging ontologies is provided in [12]. The MOAT project proposes to associate a *meaning* (i.e. an external URI) to each occurrence of a tag, to integrate tags in the Semantic Web[19]. [14] shows the use of Annotea for (semantic) social bookmarking. Another quite rich ontology for tags and annotations is NAO[7], defined under the NEPOMUK Social Semantic Desktop project.

Revyu[6] is not a generic annotation tool, but rather a reviewing and rating Web site, built with great attention towards Linked Data principles and best practices. The architecture is based on a centralized server and the vocabulary is fixed and not extendible. Only a few projects, to our knowledge, provide a richer possibility of editing the semantic model of annotations and their domain. SMORE is a semantic editor to enhance the creation of RDF metadata marking-up documents, using and extending existing domain ontologies, as well as creating new ones[10]; related to this project is the SWOOP web ontology editing browser[9].

However, though not explicitly and specifically aimed at semantic annotation, during the last years we have seen the birth and the growth of many collaborative systems for the creation and the management of structured knowledge. To mention some notable examples, Freebase[8] is a collaborative knowledge base which,

[4] `http://www.w3.org/Amaya/`
[5] `http://www.flickr.com/`
[6] `http://del.icio.us/`
[7] `http://www.semanticdesktop.org/ontologies/nao/`
[8] `http://www.freebase.com/`

beyond containing a huge amount of structured data automatically imported from all over the Web, allows users to contribute both to the population of the KB and to the definition of the data model. Semantic MediaWiki[9] (SMW) is an extension to the popular wiki platform Mediawiki, that allows users to insert additional structured data inside the wiki pages, using already defined concepts and properties as well as specifying new ones[23]. Other examples of semantic wiki platforms are KIWI[20] and OntoWiki [1], more focused on the collaborative creation and maintenance of OWL ontologies. With Collaborative Protégé[10] users can simultaneously edit the same ontology, annotating its components and its changes, and discussing and voting new modification proposals.

3 Our Approach

One of the main motivations for our work is the lack of a common, widely accepted annotation system that employs semantics both to describe annotation metadata and the domain of knowledge related to the annotated resource.

Many tools do not provide information about annotation metadata, making it impossible to filter them; other tools, like Annotea, use semantics just to model annotations and not to describe the domain of knowledge. As a result users produce structured information for annotations, rather than for annotated resources. In a system which is based on user contributions, we think that being able to easily search, access, and elaborate structured metadata about resources as well as annotations would be a great incentive for participation.

From a theoretical point of view, Annotea has been built in a way that allows it to be extended with new annotation types, however it does not provide a simple way to do so (Annotea Bookmarks, while having ideas in common with Annotations, are not exactly their extension). In systems like Revyu, users can provide structured information about resources; however they have to stick to a general, limited vocabulary they cannot extend. Moreover, many of these systems have a rather high technological barrier at the entrance, requiring users to get accustomed to new applications or to install and configure the server-side software. This contributes to the "bootstrap problem" of the annotation system: not only it does not provide contents, but also it is not able to reach the critical mass of users it needs to start.

Nevertheless, existing semantic annotation systems suggest many fundamental features: for instance, providing annotation metadata such as server, authors, and dates to let users filter annotations; the decentralized architecture of many of these systems; the use of standards both to describe knowledge and to query it.

Keeping the basic principles of current annotation systems in mind and trying to address, at the same time, some of their main limitations, we have created a new list of requirements and best practices. Following these requirements, we designed a model for semantic annotation which should be general enough to

[9] `http://semantic-mediawiki.org/`

[10] `http://protegewiki.stanford.edu/index.php/Collaborative_Protege`

be compatible with many available social applications, and we applied it to the specific case of annotating Web contents using a semantic wiki.

3.1 Requirements for a Semantic Annotation Framework

Our list of requirements is an extension and reorganization of the ones defined in the Annotea project[8] and in [21]. We have divided the requirements in four main categories:

Standards and technologies

- **Use open standards and (semantic) technologies**. Knowledge is expressed in triples and saved into an RDF store which is then made available through a SPARQL endpoint.
- **Annotations are RDF resources** with their own URI. Their types and properties should also be formally defined (i.e. in RDFS).
- **Look like formal**. Whenever an existing tool is not capable to describe an annotation with the structure and the expressiveness of our model, it should be at least possible to obtain the desired result by applying an automatic transformation to its data.

Decentralization and participation

- **Allow for collaboration and collective contributions**. It should be possible for users both to write annotations collaboratively and to contribute them individually. As described in [24], through collaboration a group builds one understanding by capturing many perspectives (such as in a Wikipedia article); by collecting user contributions, instead, it is possible to keep the different user perspectives and draw conclusions by aggregating them (such as in tag-based system).
- **Local and remote annotations**. As it is possible to refer to any local URI, it would be meaningful to keep at least these kind of annotations local. Moreover, it should be possible to easily move annotations from a remote server to a local machine for backup or replicate them to another server.
- **Multiple annotation servers**. The system should not be completely centralized, not just for redundancy purposes but also to reflect the point of views of different communities on the same resources. Users should not need to install a server to read and write remote annotations, but they should be able to just use an available one; it should be possible to subscribe to one or more annotation servers. Each server may have its own policy and eventually require users to be registered to access or to contribute to its content. This can be necessary to better express the point of view of a community and to allow for "spam free" annotation sources.
- **Collaborative vocabulary editing**. Following the recent advancements in participative tools for ontology authoring, we think that users should

also be able to collaboratively create their own vocabularies for annotation domains. Once created, these vocabularies could be shared with other users and servers so that they can customize or directly use them.

User-centered design

- **Allow for automation**. Automatic import of both domain ontologies and annotations from other tools and sources should be allowed and encouraged to ease the system bootstrap.
- **Allow for different levels of participation**. As described in [15, 2] users contribute to a participative system in different ways, accordingly to their level of expertise: the more they become expert, the more complex their contributions are (and the higher is the impact of their actions on the system). A social system like ours should allow users to participate differently according to their expertise, from viewing annotations in the simplest case up to editing a domain ontology in the most complex one.
- **Provide an easy and immediate user interface**. In particular exploit the browser user interface and allow most of the actions to be easily performed just inside the normal interface while browsing the Web.

Filtering

- **Annotation metadata** Annotations should bring some information to allow users to filter them (e.g. by author, group or community, annotation server, and date of creation).
- **Annotation history**. Provide access to the history of each annotation. This might be useful also for annotating annotations: users might refer to the URI of a specific revision of the annotation rather than the URI of the annotation itself.

4 A Model for Semantic Annotations

One of the most important features of annotations is the possibility to filter them according to some of their properties. Annotea[11], for instance, allows filtering by author, annotation type and server.

When extending annotations with domain semantics, we need to describe two different types of metadata: those that are asserted about the annotated resource and the ones about the annotation itself. As we consider information of the first kind as valid only within the context of the annotation, we have decided to keep the annotation as the main subject, characterized by properties related to it and containing one or more reified statements about the resource: the resulting model[12] is the one shown in Figure 1.

[11] See Annotea annotation schema at `http://www.w3.org/2000/10/annotation-ns#`.
[12] See `http://davide.eynard.it/rdf/annotations#`.

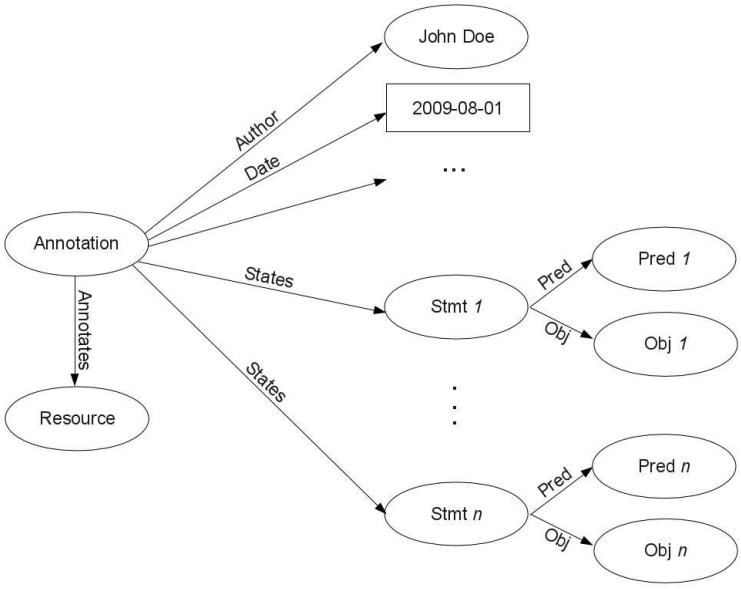

Fig. 1. A model for semantic annotations.

The model is based on a couple of assumptions. First, any number of statements can be grouped inside an annotation. The reason for this choice is that users do not think in triples and it would be too complicated for them to split what they consider a single annotation into a sequence of separate statements; moreover, it does not represent a real limitation as it is still possible to create annotations containing a single statement. The second assumption is that, being an annotation about a specific resource, all the reified statements have the resource as a default subject: this avoids inconsistencies and reflects our view that a statement is meaningless if separated from the annotation that originated it.

5 A Wiki-like approach for Semantic Annotation

As already mentioned, a considerable and growing attention is being addressed in recent years for a "collaborative way" towards the semantic Web and many tools for collaborative creation and management of semantic metadata are becoming available. Several of the systems mentioned in the end of Section 2 already offer features that can satisfy most of the requirements illustrated in Section 3, without the need to create a new system from scratch. In particular, among other features they offer:

- possibility of importing existing ontologies and vocabularies, as well as defining new properties and classes;
- possibility of exposing or exporting data in RDF format;

– user management, with the possibility of setting different levels of privileges;
– availability of metadata like authorship and creation date for all contents;
– versioning;
– community management tools, such as MediaWiki discussion pages;
– easy interfaces for novel user (like semantic templates in SMW);
– no entrance barrier for an already existing wiki-centered community;
– possibility of importing data from external sources, to solve the bootstrap issue.

As a further advantage, some of these platforms are already available online and can easily be adapted for our purpose; for example, Wikia[13] provides a SMW farm, and other websites offer hosting with SMW[14]. It should also be taken into account that many communities are already growing around wikis relative to specific topics, and some of these are already using semantics.

5.1 An experiment based on the Semantic MediaWiki platform

The main purpose of our experiment is to show that the model we developed is general enough to describe an existing system, and at the same time that this adaptation can be made automatic through the use of already known standards and technologies. For this reason we decided to develop a prototype that, given a URI, shows annotations gathered from a wiki system based on Semantic MediaWiki[15].

The architecture of our prototype is shown in Figure 2. The client application is a Firefox extension, which is able to access different annotation servers using the SPARQL protocol and query language. The application is totally parametric (that is, it does not depend on a specific domain ontology) and works with any SPARQL endpoint which exposes data described using our ontology.

Annotation servers are seen by external applications in a consistent way, that is as SPARQL endpoints, independently from how information is stored inside them. In our case, due to the way Semantic MediaWiki represents its metadata, it was necessary to apply a transformation between the data stored by the wiki and our final format.

In fact, Semantic MediaWiki allows users to specify triples in a page as predicate-object pairs, keeping the page URI as the default subject of every triple. We decided to use wiki pages for annotations, allowing users to state in the same place both the properties related to the page and the ones referring to the annotated resource; the only difference is that these latter are defined as "external properties". Basically, all of the internal properties were imported from our main ontology, while the external ones were defined inside the wiki. As an application example we created page-related properties such as tag, rating,

[13] `http://www.wikia.com/`
[14] A list can be found at `http://semantic-mediawiki.org/wiki/Help:Introduction_to_Semantic_MediaWiki`.
[15] The annotation wiki used for the prototype can be found at `http://davide.eynard.it/elc`.

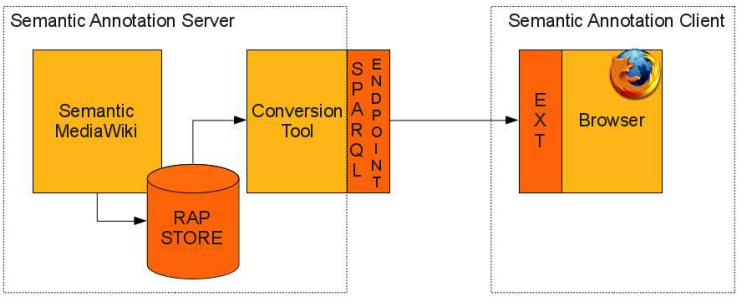

Fig. 2. The architecture of our prototype.

seeFirst (for pages that help to understand the annotated one), and seeNext (for pages that provide additional information). Advanced users can customize their annotations by manually adding new properties, while beginners can use a ready-made template which makes annotating much easier.

To expose wiki metadata with our representation format we used a conversion tool which applies the transformation on the fly, sending a CONSTRUCT query to the wiki's RDF store. The query gets all the triples for an annotation, replicates the internal properties and translates the external ones in reified statement pairs: the final result is another view of the knowledge base which complies to our model.

Figure 3 shows how annotations look like inside the browser, when the annotated page is loaded. To obtain more interesting results, we have classified properties inside the wiki and customized the tool so it sends an additional SPARQL query directly to the wiki endpoint and asks for property types. Using this approach, properties can be visualized differently depending on their category, so tags are shown in a ranked list (or a tag cloud), ratings as an average, and so on. We think this solution could yield very interesting results and we have planned to extend the approach to the general model so that it will be possible to apply it for different systems.

6 Conclusions and future work

In this paper we have spotted some limitations of current annotation systems, and we have illustrated our approach, based on the idea that both metadata regarding the annotations and knowledge about the annotated resources should be expressed in a semantic format, with the possibility for users to collaboratively extend the underlying ontologies and vocabularies. Starting from this assumption we have defined a list of requirements and best practices for a semantic annotation framework, based on a distributed server architecture and one common semantic model as an interface with the services consuming data.

We have then shown a prototype that relies on the Semantic MediaWiki platform to allow for collaborative creation and maintenance of the annotation

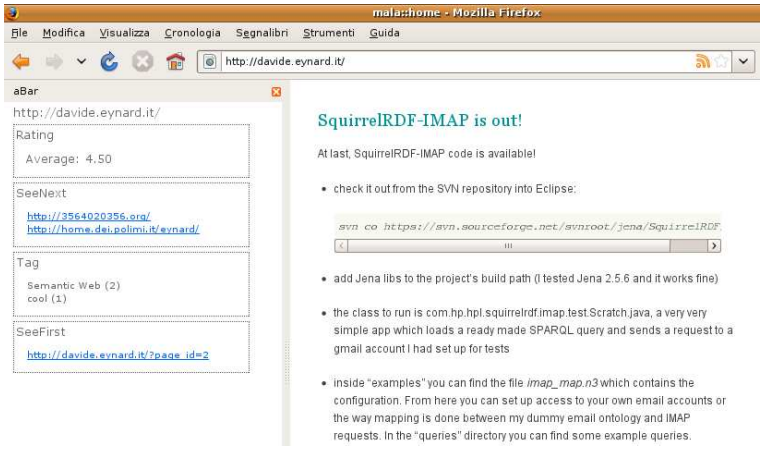

Fig. 3. Semantic annotations as shown by our browser extension.

knowledge base, with the possibility for expert users to also define new classes and properties. On the client side, a Firefox extension shows how semantic annotations can be used to improve the user navigation experience, especially by adding information and possible actions related to the URL they are visiting, filtered according to their personal choices and visualized in a variety of manners according to the different kinds of properties shown.

We believe that the added value of a general and extensible semantic model to aggregate information for a variety of sources under a unifying framework can sensibly increase the user navigation experience, allowing for possibilities that also go much further the examples we were able to imagine and to show in this paper. Just to give a further example, any set of properties of the kind like "seeFirst" and "seeNext" might be used to create navigational paths inside the Web, whereas a collection of resources corresponding to places with geocoordinates might be used to suggest itineraries, which might be visualized in a map.

It shall then be noted that the generality of the model we have proposed allows to access annotations from a variety of existing systems. For instance, as tagging is just a particular kind of annotation, data from social bookmarking applications like del.icio.us and bibsonomy can be easily translated to be accessible within the semantic framework we have proposed. In particular we are working at making del.icio.us annotations about any URI accessible through a SPARQL endpoint. The same thing might be done for review and rating systems, like Revyu or TripAdvisor, though in this case the fact that the system uses internal URIs to identify resources may be a limitation.

As another direction for future work, we would like to enrich our model integrating it with existing ontologies, like SIOC, SKOS and MOAT. By adopting the MOAT vocabulary, for instance, tags might be associated to external URIs, and each URI could be of course subject to annotations.

References

1. S. Auer, S. Dietzold, and T. Riechert. Ontowiki - a tool for social, semantic collaboration. In I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, editors, *The Semantic Web - ISWC 2006, 5th International Semantic*, volume 4273 of *Lecture Notes in Computer Science*, pages 736–749. Springer, 2006.
2. S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10, New York, NY, USA, 2005. ACM Press.
3. F. Echarte, J. J. Astrain, A. Crdoba, and J. E. Villadangos. Ontology of folksonomy: A new modelling method. In S. Handschuh, N. Collier, T. Groza, R. Dieng, M. Sintek, and A. de Waard, editors, *SAAKM*, volume 289 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.
4. First on-line conference on Metadata and Semantics Research (MTSR'05). *Ontology of Folksonomy: A Mash-up of Apples and Oranges*, 2005.
5. S. Handschuh and S. Staab. Authoring and annotation of web pages in cream. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 462–473, New York, NY, USA, 2002. ACM.
6. T. Heath and E. Motta. Revyu.com: A reviewing and rating site for the web of data. pages 895–902. 2008.
7. J. Heflin, J. Hendler, and S. Luke. Shoe: A knowledge representation language for internet applications. Technical report, University of Maryland, 1999.
8. J. Kahan and M.-R. Koivunen. Annotea: an open rdf infrastructure for shared web annotations. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 623–632, New York, NY, USA, 2001. ACM Press.
9. A. Kalyanpur, B. Parsia, E. Sirin, B. C. Grau, and J. Hendler. Swoop: A web ontology editing browser. *Journal of Web Semantics*, 4(2):144–153, 2006.
10. B. P. A. Kalyanpur and J. Golbeck. Smore semantic markup, ontology, and rdf editor, 2005. http://www.mindswap.org/papers/SMORE.pdf.
11. B. P. Kettler, J. Starz, W. Miller, and P. Haglich. A template-based markup tool for semantic web content. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 446–460. Springer, 2005.
12. H.-L. Kim, S. Scerri, J. Breslin, S. Decker, and H.-G. Kim. The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies. In *International Conference on Dublin Core and Metadata Applications*, Berlin, Germany, 2008.
13. T. Knerr. Tagging ontology - towards a common ontology for folksonomies, 2006. http://tagont.googlecode.com/files/TagOntPaper.pdf.
14. M.-R. Koivunen. Semantic authoring by tagging with annotea social bookmarks and topics. In *Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006)*, 2006.
15. J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, New York, 1991.
16. L. Mcdowell, O. Etzioni, S. D. Gribble, A. Halevy, H. Levy, W. Pentney, D. Verma, and S. Vlasseva. Mangrove: Enticing ordinary people onto the semantic web via instant gratification. In *2nd International Semantic Web Conference*, volume 2870, pages 754–770, 2003.

17. P. Mika. Ontologies are us: A unified model of social networks and semantics. In *The Semantic Web - ISWC 2005, Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer, 2005.
18. R. Newman, D. Ayers, and S. Russell. Tag ontology, December 2005. http://www.holygoat.co.uk/owl/redwood/0.1/tags/.
19. A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, Apr*, 2008.
20. S. Schaffert, J. Eder, S. Grnwald, T. Kurz, and M. Radulescu. Kiwi - a platform for semantic social software (demonstration). In *ESWC'09: The Semantic Web: Research and Applications, Proceedings of the 6th European Semantic Web Conference*, pages 888–892, Heraklion, Greece, June 2009.
21. V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14 – 28, 2006.
22. O. Valkeap, O. Alm, and E. Hyvnen. Efficient content creation on the semantic web using metadata schemas with domain ontology services (system description). In E. Franconi, M. Kifer, and W. May, editors, *ESWC*, volume 4519 of *Lecture Notes in Computer Science*, pages 819–828. Springer, 2007.
23. M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland, 2006.
24. T. V. Wal. Getting to know collective and collaborative, March 2008. `http://www.personalinfocloud.com/2008/03/getting-to-know.html`.
25. K. P. Yee. Critlink: Advanced hyperlinks enable public annotation on the web, 2002. http://zesty.ca/pubs/cscw-2002-crit.pdf.
26. L. Zhang, X. Wu, and Y. Yu. Emergent semantics from folksonomies: A quantitative study. *Journal on Data Semantics VI*, pages 168–186, 2006.