SemKey: A Semantic Collaborative Tagging System *

Andrea Marchetti
Andrea.Marchetti@iit.cnr.it

Maurizio Tesconi
Maurizio.Tesconi@iit.cnr.it

Francesco Ronzano
Fr.Ronzano@libero.it

Marco Rosella
marco.rosella@iit.cnr.it

Salvatore Minutoli
salvatore.minutoli@iit.cnr.it

Institute of Informatics and Telematica (IIT) - CNR
Via Moruzzi, 1
Pisa, Italy

ABSTRACT

By analysing the current structure and the usage patterns of collaborative tagging systems, we can find out many important aspects which still need to be improved. Problems related to synonymy, polysemy, different lexical forms, mispelling errors or alternate spellings, different levels of precision and different kinds of tag-to-resource association cause inconsistencies and reduce the efficiency of content search and the effectiveness of the tag space structuring and organization. They are mainly caused by the lack of semantic information inclusion in the tagging process. We propose a new way to describe resources: the semantic tagging. It allows user to state semantic assertions: each of them expresses a defined characteristic of a resource associating it with a concept. We present SemKey, a semantic collaborative tagging system, describing its global architecture and functioning along with the most relevant organizational issues faced. We explore the adequacy of the support offered by the entries of Wikipedia and WordNet in order to access to and reference concepts.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; D.2 [**Software**]: Software Engineering

General Terms

tagging systems, semantic web, knowledge organization

Keywords

collaborative tagging, semantics

1. INTRODUCTION

In the last few years, many kinds of collaborative tagging systems have experienced a great diffusion and the related communities of taggers have considerably increased [13] [24]; taggers are actively involved in the process of pointing out and cataloguing resources of interest assigning them one or more descriptive keywords or tags; they exploit the growing amount of information collected to improve their searches and content discovery process. Some of the most used and representative collaborative tagging services are [23]:

- **Del.icio.us** (http://del.icio.us): it allows user to assign a free set of tags to a Web resource identified by its URL; this kind of tagging schema is also known as 'social bookmarking', because users can create and share resources annotations in a way similar to local bookmarking systems integrated in existing browsers;
- **Flickr** (http://www.flickr.com): this is a photo sharing system; each user can share and tag his personal photos. He can also access and tag photos of other users;
- **Technorati** (http://www.technorati.com): it allows authors to tag their blog's posts, aggregating information contained in weblogs and facilitating their search.

All the tagging systems listed above are usually adopted by particular communities of users; del.icio.us by Computer Science experts, Flickr mainly by amateur photographers and Technocrati by bloggers.

As we can infer, also by reading this short description of significant examples, tagging represents a collaborative social effort of a community of users constituted around a tagging service; with his tagging action, every user, mainly on the basis of its interests, directly contributes to the creation of a shared metadata collection, progressively increasing its relevance and its richness of useful and shared data. The three main components of collaborative tagging systems are: users, resources and tags [19]. Users may be connected in groups with common interests; resources may be related by the different kinds of links which constitute the basis of current Web; tags provide the connection between a single user and a particular resource.

When every user can assign a freely defined set of tags to a resource, the tag collection reflects the social attitudes of the community of users and a shared social organization and structuring of the tag-space will emerge: this phenomenon is referred to as emergent semantics. The result of this process of adaptive social structuring of the tag-space has recently been defined as folksonomy [22]. A folksonomy is the outcome of the fusion of two words: 'folk' and 'taxonomy'. When we speak about folksonomy, we refer to the collaborative (folk) and progressive definition of a relaxed categorization and organization of content (taxonomy), not

*A full version of this paper is available at `CNR TECHNICAL REPORT`

WWW2007, May 8–12, 2007, Banff, Canada.
.

based on a rigid hierarchical structure, and the related semantic specification of concepts, or better of the meaning of tags.

Many formal (research articles references [12]) and informal (blogs references [21] [18]) analyses of collaborative tagging systems have also identified the low user learning curve and the relatively low bootstrapping cost of this kind of services as two relevant factors influencing their spread and rapid diffusion.

When we analyse in more depth the current structure and usage patterns of collaborative tagging systems, we can discover many important aspects which still need to be improved to really exploit their real potential.

2. SEMKEY: SEMANTICS GETS INTO COLLABORATIVE TAGGING

Starting from the weak features of existing collaborative tagging systems and introducing the semantic tagging, a new way to describe resources, we have developed SemKey, a semantic collaborative tagging system. We describe its architecture and its main organizational features, giving also some practical example of functioning.

2.1 Semantic tagging

By analysing existing collaborative tagging systems, we can point out some relevant weak features. In particular, we can identify the following main causes of weakness related to different ways of using keywords:

- **Polysemy** : the same word can refer to different concepts (the word 'field' can refer to a piece of land cleared of trees and usually enclosed, but also to a branch of knowledge);
- **Synonymy** : the same concept can be pointed out using different words ('auto', 'car', 'machine' are three different words that refer to the same concept: a four wheels vehicle);
- **Different lexical forms** : the same concept can be referred to by different noun forms, for instance plural nouns ('car'/'cars'), different verb conjugation ('buy'/'buying'), name-adjective couples ('energy'/'energetic'), multiple words ('pc'/'personal computer') and so on;
- **Misspelling errors or alternate spellings** : typing errors that occours when we write a word ('staton' in place of 'station') or different possible spelling of the same word ('color'/'colour');
- **Different levels of precision** : the specifity of the word chosen to tag a resource ('jazz' is more specific than 'music');
- **Different kinds of tag-to-resource association** : implicit kinds of relations that links a tag to a specific resource ('interesting' expresses an opinon on the resource, 'car' expresses the topic of the resource and so on) .

Many of the problems described can be related to the absence of any semantic information in the process of assigning descriptive keywords to a resource ([16], [25], [13], [17], [20]). They can be traced back to the existence of $n : m$ relations between concepts and tags/keywords used to identify an intended concept.

When a single tag is used to express different concepts ($Tag(1) \longrightarrow Concept(n)$), polysemy issue occurs. When we adopt that tag to find all resources related to a specific intended concept, *the percentage of all retrieved resources that are actually relevant to the query (called precision) decreases* because of the noise generated by the other retrieved resources dealing with different concepts but tagged using the same tag.

On the other side, multiple tags can be used to refer to the same concept ($Tag(n) \longrightarrow Concept(1)$); this can be caused by synonymy, different lexical forms, misspelling errors or alternate spellings. In this case, using one of the different tags that refer to a concept to find all related resources, *the percentage of all relevant resources present in the system that are returned by the search (called recall) decreases* because of the presence of other relevant resources that are not retrieved since they are tagged using different words that point to the same concept.

To solve or at least reduce these problems we suggest to introduce the support of semantics in collaborative tagging activity. In particular we propose to substitute semantic assertions for current keywords or tags. They don't consist of simple strings releted to a particular resource like existing tags; each semantic assertion describes a specific property of a resource. It associates a concept with a resource and specifies the semantics of their relation. One or more different strings, called lexical forms, can be used to identify a particular concept; the set of strings related to a concept includes synonyms, alternate spellings or misspelling errors and all other possible lexical forms used to pick out a particular meaning. The activity of describing resources formulating semantic assertions is referred to as **semantic tagging**. It considerably improves search efficiency and effectiveness and makes exploitable new important information access and oraganization patterns.

2.2 Concept identification support: WordNet and Wikipedia

Every semantic assertion provides descriptive intormation about a resource referring to a particular concept. In order to identify a specific concept, we must disambiguate the meaning of a lexical form. As a consequence we need to exploit some resource that should support the following tasks:

- given a particular lexical form it should identify all its possible meanings (or concepts), providing for example a short textual description in order to express each one;
- it should allow for a univocal reference to every single concept.

Considering these fundamental requirements, we have identified two different and may be complementary kinds of resource currently available over the Web:

- **WordNet**: a lexical database which is based on the concept of set of synonym words, called *synset*, which defines a particular concept; it is sufficiently structured and includes a lot of lexical and semantic relations between words and synsets. Wordnet is updated by a group of lexicon experts and presents quite a complex net of internal relations, in fact it has been developed

in order to support text mining and information extraction. WordNet has a broad coverage of all common parts of speech (names, verbs, adverbs and adjectives). At present, WordNet version 3.0 is available; it includes 117597 concepts (or distinct synsets). To obtain additional information see [9].

- **Wikipedia**: the famous collaboratively-edited free encyclopedia, which is the result of the efforts of many editors worldwide, directly involved in this project; it is rich of extensively described and easily referenced definitions of concepts and it is continuously increasing its dimension and completeness. Wikipidia does not cover all parts of speech like WordNet, but it is extremly rich and constantly updated. It provides descriptions of many specific proper-named concepts that are not present in WordNet. Wikipedia is obviously less strongly structured than WordNet, but thanks to the possibility to collaboratively edit its data, it is constantly enriched with new updated contents. It supports the disambiguation of polysemous words through the introduction of *disambiguation pages* which allow users to choose a specific meaning among those available. For words with synonyms, in Wikipedia, it is possible to use the *redirect mechanism* in order to redirect all of them to the same document and so to the same concept. Moreover since May 2004 Wikipedia includes also a sort of relaxed classification system of its documents: the *Wikipedia categories*. Every description included in the encyclopedia can be assigned to one or more categories in order to provide a new way of accessing and cataloguing it. Users can create new categories arranging them in a hierarchical-like structure. Every document is also related to many other documents through simple links usually used to point to extended descriptions of terms. In Table 1 we mention some important numerical data [10] regarding the English version of Wikipedia in order to better quantify the great amount of information collected. To obtain additional information see [3].

Number of articles included	1,4 Millions
Number of active editors (who edited at least 10 times since they arrived)	150.000
Number of links between Wikipedia articles	32,1 Millions
Number of redirects	1,4 Millions
Number of categories	176.000
Percentage of categorized articles	86%

Table 1: Wikipedia statistics - English - October, 2006

WordNet and Wikipedia are the resources exploited by SemKey to support semantic tagging activity.

2.3 Requirements and global architecture

2.3.1 General requirements

The main idea that supports our system, which is also its main requirement, is the following: *giving the user the possibility to express semantic assertion about a Web resource.*

Usability. One of the most important features that have supported the wide diffusion of current collaborative tagging systems is the directness of the process of tagging; every user can immediately tag a Web resource using one or more keywords. We have considered the importance of this aspect trying *not to excessively increase the cognitive weight of the semantic tagging process.* It is obvious that we need a greater amount of information to be provided by a user in order to specify the intended meaning of a tag, but we have paid attention to organize and graphically arrange the interactions in order to make the semantic tagging process as fast as possible.

Motivation. Moreover we must consider *user's motivation to produce semantic assertions.* To achieve a wide diffusion of our system we have to provide actual advantages in information organization and accessibility, when adopting this new way of tagging. For this reason we have placed great emphasis also on the completeness and the availability of added information organization and search features.

2.3.2 Main user interaction patterns

To fulfill the system requirements just analyzed, we can define the structure of the principal interactions between our semantic tagging system and its typical users.

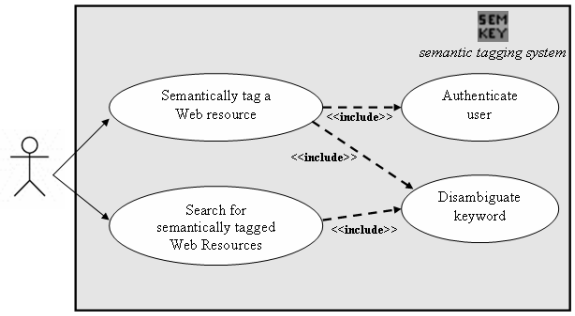

Figure 1: System UML use case diagram.

First of all a generic user can access our system from two fundamental different perspectives. He may use the system only *as a search engine performing a semantic assertion search* in order to find relevant Web resources references; this is a passive exploitation of the semantic information collected by the system, meaning that the user accesses to the contents already present in the system without giving any contribution to its enrichment. In this way the user takes advantage of only a part of all the possibilities offered by our system, not exploiting one of its fundamental aspects: the social component. On the other end, the user, after having completed a registration phase, may *authenticate himself and access his personal area.* Thus he can exploit all the possibilities of SemKey. He can semantically tag Web resources of interest producing semantic assertions, manage his collection of resources and semantic assertions and organize them. Moreover, through his tagging activity, every user gives his contribution to the enrichment of the informative data collected by the system, thus making searches possibly more effective. We can graphically schematize the

described user-system typical interactions through the UML use case diagram in Figure 1.

2.3.3 Global architecture of the system

Our system architecture is based on *three main modules*: two server side components and a client side one. The main functionalities provided by each module are:

- **Semantic tagging manager** (client side): this module is intended to be strictly integrated in user browsers so as to allow a fast process of semantic tagging in order not to alter the usual Web browsing activity of a common user. In this way while using our semantic tagging system, we aim not to introduce any change in the diffused browsing interaction patterns;
- **Sense disambiguation module** (server side): this module provides access to all information and services needed during the lexical form disambiguation process; it mainly supports the client in the choice of the intended concept described by a lexical form collecting the different meanings and thus allowing the definition of a semantic assertion;
- **Metadata store and access module** (server side): this is the principal module of our system. It mainly stores and provides Web access to all collected semantic tagging information. It is also responsible of the users' management.

In Figure 2 we represent SemKey high-level modules just described.

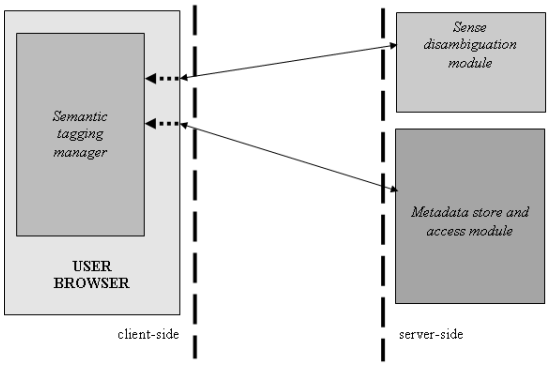

Figure 2: SemKey high-level modules.

When a user browses the Web and visits a resource of interest, he can decide to semantically tag it. He activates the 'Semantic tagging manager' that retrieves the URL of the resource and, if selected by the user in the Web page, a tag (or lexical form) (1). If the user isn't still logged in SemKey, logging credentials are requested in order to identify him; they are validated interacting with the 'Metadata store and access module' (2). After the authentication phase is successfully completed, the user will be driven in the choice of the intended meaning of the selected tag. Interacting with the tagging Web APIs of del.icio.us [1] and Yahoo My Web 2.0 [11] the 'Semantic tagging manager' retrieves and shows the user the most popular tags concerning the selected resource, in order to provide possible suggestions (3). Once the user has chosen a tag, it will be sent to the 'Sense disambiguation module' in order to receive a list of all possible concepts that can be referred to using that tag (4). The user selects the intended meaning of his tag and the specific property of the Web resources to describe: thus he formulates a semantic assertion. It is sent to the 'Metadata store and access module' to be stored (5). Then the 'Semantic tagging manager' ends its execution and the user can resume his browsing (6).

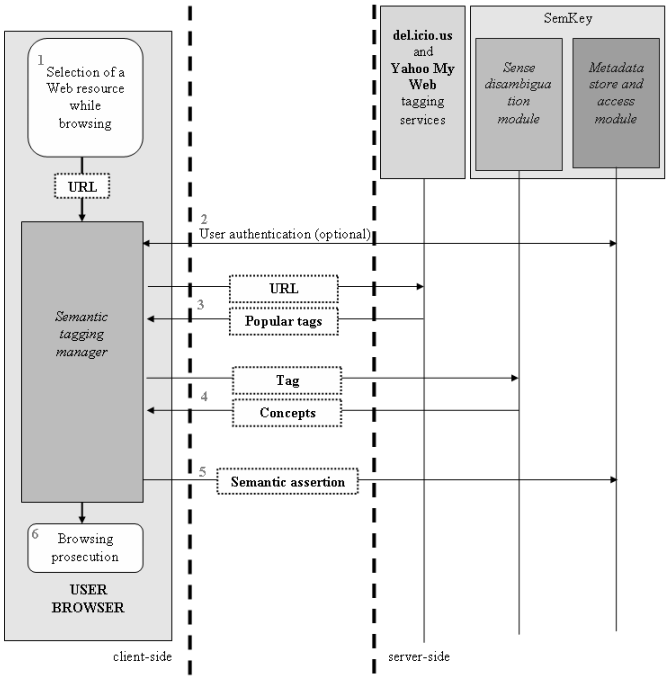

Figure 3: Modules interaction to produce a semantic assetion.

2.4 Main organizational issues

In this section we analyze the basic organizational issues faced when structuring the modules which constitute our system. We discuss and motivate every adopted solution, but we also describe possible improvements and future scenarios.

2.4.1 The structure of Sense disambiguation module: WordNet and Wikipedia exploitation

The 'Sense disambiguation module' is the core of our system; its purpose is to support the client-side 'Semantic tagging manger' during the process of semantic tagging and it is responsible for the collection of available meanings of user tags. As stated before, in this initial version of SemKey we have decided to explore the semantic contents of WordNet [9], and Wikipedia [3]. During the disambiguation process, the 'Sense disambiguation module' accesses the Web interfaces of WordNet and Wikipedia to collect the meanings of the typed tag. In particular, given a tag we consider:

- **in WordNet**, all the synsets the tag belongs to;
- **in Wikipedia**, the description of the meaning of the tag or the different meanings associated to a polysemyc tag through its disambiguation page.

The 'Sense disambiguation module' selects for every concept two information:

- an **URI** which identifies the concept;
- a short textual **description** or gloss of the concept.

Compared to WordNet, Wikipedia contents are often more difficult to manage for the disambiguation of a tag because of its relaxed organizational structure that doesn't provide many facilities to support this task.

2.4.2 Semantic assertion model

Another relevant issue is represented by the organization of the set of data stored during the semantic tagging of a Web resource. In a generic collaborative tagging system as del.icio.us a user can associate a tag to a resource without specifying the relation type. Normally the tag represents the topic of the resource but this is not always true and this semantic information remains in the head of the user. The solution could be to force the user to explicitly define the kind of relation for each tag, but this can make the semantic tagging activity boring, so we adopted an intermediary solution.

By analyzing the different kinds of tags managed by existing collaborative tagging systems, we have decided to manage only *three different relations*:

1. **hasAsTopic** : this relation will be used to describe the topic of the resource such as book, Web design, sport, politics, cars, animal, medicine, etc.;
2. **hasAsKind** : this relation will be used to characterize the kind of informative content of the resource such as blog, application, mashup, podcast, official Web site, streaming, video, e-commerce, Web API, etc.;
3. **myOpinionIs** : this relation concerns all subjective opinions such as cool, funny, interesting, boring, amazing, expensive, etc..

The choice of the right relation to connect a concept to a particular resource is left to the user.

In this way the model of a semantic assertion is a particular type of the RDF triple composed by the following parts:

- **Subject** : the URL of the Web resource that has been semantically tagged;
- **Predicate** (property) : the particular property of the resource that has been described (chosen among the three relations previously described);
- **Object** : the URI that identifies a concept.

Other data need to be added to those just mentioned in order to fully describe the association of a concept to a particular Web resource:

- the *lexical form* (string) employed by the user to identify the particular concept referred to at the time of the generation of the semantic keyword;
- the *username* adopted in our system in order to uniquely identify the user, author of the semantic keyword;
- the *date* and the *time* of the generation of the semantic keyword.

These data are all descriptive information that is added to the core RDF-triple previously mentioned. To represent them we need to exploit another RDF expressive conventionalism: the reification [15]. It is used to make RDF statements that describe an entire RDF triple referring to it through a unique identifier (ID). Once determined this ID, we can define other properties referred to the entire RDF-triple, in particular:

- **semkey:word** : the lexical form used to refer to the concept during the semantic tagging process;
- **dc:date** : date and time of the generation of the semantic keyword;
- **dc:creator** : username of the user that generated the tagging data.

In the properties just described, the namespace 'semkey' refers to the local RDF Schema namespace of our system and the namespace 'dc' refers to the Dublin Core Metadata RDF Schema namespace [2]. As a consequence the information contained into a semantic assertion on a Web resource made by a particular user in a precise moment is represented as shown in Figure 4.

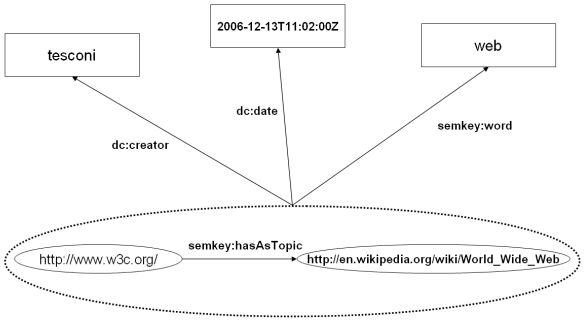

Figure 4: RDF graph of the information related to a disambiguated tag.

The main RDF-triple is represented by the information contained in the dotted circle; through the RDF reification conventionalism three other descriptive data are added to the main RDF-triple. If we want to represent those data using the XML/RDF serialization, we obtain the following xml fragment:

```
<?xml version='1.0'?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:semkey='http://www.semkey.org/schema/'
xmlns:dc='http://purl.org/dc/elements/1.1/'>

<!-- Triple 1 -->
<rdf:Description rdf:about='http://www.w3.org/'>
<semkey:hasAsTopic rdf:nodeID='id00001'
rdf:resource='http://it.wikipedia.org/wiki/World_Wide_Web'/>
</rdf:Description>

<!-- Descriptive information added to triple 1 exploiting its
reification -->
<rdf:Description rdf:nodeID='id00001'>
```

```
<semkey:word>web</semkey:word>
<dc:date>2006-12-13T11:02:00Z</dc:date>
<dc:creator rdf:resource='http://www.semkey.org/users/tesconi'/>
</rdf:Description>

</rdf:RDF>
```

This RDF data structure is exploited to export semantic assertions, making them externally available.

2.4.3 Semantic search patterns

When a user searches for relevant resources, he must specify the structure of one or more *generic semantic assertions*; they are semantic assertions defined without referring to a particular resource. Each of them specifies a concept that describes a particular characteristic (or property) of the resource to find. All the resources that are described by the set of generic semantic assertions specified by the user are considered to form search results.

For instance, the user could ask the system to find all 'blogs' (property: kind of resource) which deal with 'Web design' (property: topic of resource) and are reputed to be 'interesting' (property: personal opinion); 'blog', 'Web design' and 'interesting' are disambiguated lexical forms, referring to specific concepts. The search parameters just described are composed by three generic semantic assertions. The user could specify none, one or more semantic assertions.

2.4.4 Exploitation of WordNet and Wikipedia net of relations.

Those just described represent only the basic search capabilities and content structuring possibilities that such systems offer. A possible relevant improvement could be obtained considering all the nets of relations that could connect the concepts used to support the disambiguation of lexical forms or could relate two or more different tag lexical forms. This further part of informative content is usually strongly present in lexical resources.

In this first development phase of our system, we have decided to exploit the sense disambiguation information provided by the lexical resource *WordNet* [9]. In WordNet, the meanings and the lexical forms used to refer to a particular concept are connected by a set of 18 different kinds of relations. Some of these are very specific and have been introduced in order to exploit the semantic information available with an original distinct purpose: text mining and information extraction. However, other relations could be exploited to further enrich search possibilities and structured exploration of contents. For example, *the hyponymy/hypernymy relations* that represent the hierarchical specialization / generalization of concepts may be used to suggest, during the disambiguation of lexical forms, all their hyponyms or hypernyms in order to better define the level of precision adopted by the user; in this way we can solve or at least reduce the basic level of precision problem, mentioned before. Moreover we can allow users to extend the coverage of their search including all the hyponym concepts of those related to particular chosen concept; in a similar way we can suggest users to choose one of the hyponyms of a disambiguated tag to better specify the search parameters. He can also substitute a disambiguated tag with one of its hypernyms in order to eventually increase search coverage.

For instance, if a user wants to find all the resources tagged with 'automobile', after the choice of the intended meaning for this word, he could then examine all the concepts hyponyms of this concept: 'jeep', 'coupe', 'station wagon', etc. so as to extend the search coverage including all resources tagged with a least one of the hyponyms or to further refine his search replacing, for example, 'car' with 'jeep' and increasing the level of precision adopted. Part of the considered WordNet subsumption hierarchy of concepts is schematized in Figure 5.

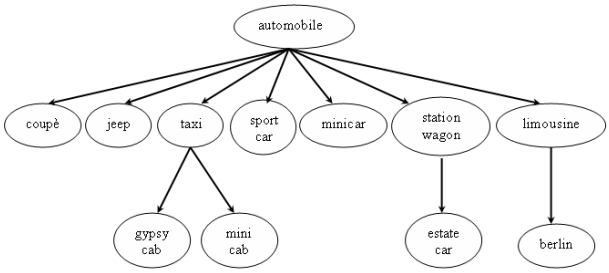

Figure 5: Part of WordNet hierarchy of concepts referred to the automobile world.

Another exploitable WordNet's relation is *the meronym or 'part of' relation.* It connects a concept with other concepts which constitute its parts. For example, the disambiguated tag 'automobile' has the following parts or meronyms: 'accelerator', 'air bag', 'auto engine', etc. It could be useful to show all meronyms of a concept in order to help users to better structure and organize their search tag set.

When we analyze *Wikipedia* [3] and the semantic concept references extracted from this other resource, we should consider that it is not a coherent lexical resource, but a collaboratively edited encyclopedia. Wikipedia also presents a sort of content categorization system: *the Wikipedia categories.* They are collaboratively edited and managed and don't constitute a hierarchical structure; they form a direct acyclic graph. Every category could be included in one or more general ones, and sometimes there are also cyclic inclusions, even if editors are explicitly advised to avoid them. All these categories constitute a sort of specialization / generalization structure similar to that previously described speaking about WordNet, with more relaxed constraints. We can consequently exploit this added informative content in a way similar to that described above, in order to improve search completeness. Besides the category structure, in Wikipedia there is a highly dense net of simple inter-document references and every concept description or encyclopedia entry presents a collection of related Web resources which could be exploited to provide the user with useful links suggestions in order to deeply examine a concept.

2.5 Detailed system modules' architecture

Considering the main high-level modules of our system, their interactions and the fundamental organizational issues faced when specifying their architectural structure, in Figure 6 we detail the internal organization of each of them.

The 'Semantic tagging manager', implemented as a browser extension, can directly interact with del.icio.us [1] and Yahoo My Web 2.0 [11] Web APIs to retrieve poular tags suggestions.

The 'Sense disambiguation module' can be accessed di-

rectly from the Web browser when the user must single out a specific concept considering a particular tag, in order to support SemKey search functionalities; this module can be also queried by the 'Semantic tagging manager' during the formulation of a semantic assertion in order to express a particular concept.

The 'Metadata store and access module' can be accessed by the 'Semantic tagging manager' in order to save one or more semantic assertions, by a request to SemKey Web APIs or by the user browser in order to execute semantic searches or to manage the personal data of every user of our system.

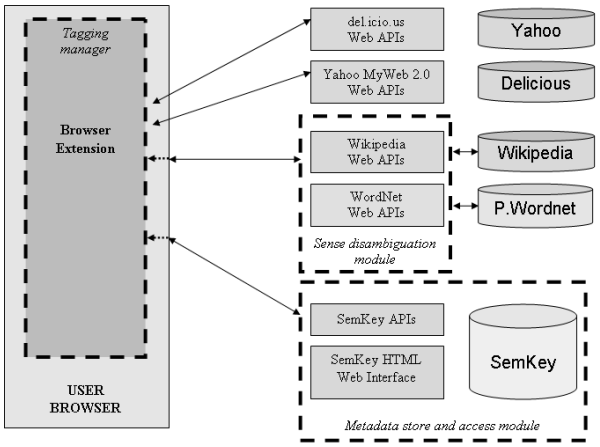

Figure 6: Detailed system modules' architecture.

2.6 Implementation and functioning

We describe some implementation details with some examples of SemKey in action, considering each one of its three main modules. To test a first version of this tool see the site http://www.semkey.org.

2.6.1 The Semantic Tagging Manager (STM).

STM is the client-side module of our system: it must provide support to make the semantic tagging process (choice of the concept starting from a lexical form and its relation) easy for the user and to *maintain high the global usability of our system.* We have decided to implement it as a Mozilla Firefox extension [7].

When the plug-in is installed, a multi-coloured button is added to the user interface of the browser. It is used to add one or more semantic assertion to the current resource displayed on the browser by activating a dialog window as shown in Figure 7.

If the user is not still logged in the system, it requests logging credentials (username and password) in order to identify him (2), interacting with the 'Metadata store and access module' to validate them. After the log-in phase is successfully completed, the user will be driven in the composition of the semantic assertion.

In fact the STM proposes initially some tags corresponding to the most popular ones used by delicious users to annotate the current resource. The user can select one of these tags or insert a new one and the relative relation (by default it is selected the 'hasAsTopic' relation).

Immediately the STM answers with a list of available meanings. Once selected the intended meaning of the considered tag, the semantic assertion is completed and STM will save it sending all data to the 'Metadata store and access module'.

Figure 7: Semantic tagging manager dialog window.

2.6.2 The Sense Disambiguation Module (SDM)

This module supports the process of disambiguation of the tag chosen by the user. SDM gathers the different meanings of a particular lexical form by exploiting the available concepts of WordNet and Wikipedia. To carry out this goal, the SDM filters the web pages of these two lexical resources producing a list of the collected meanings associated to the lexical form. This list is serialized in order to compose a JSON array [5] with all collected couples of concept URLs and respective short concept descriptions; this array is sent back as the reply to the STM.

If Wikipedia and Wordnet provides some suitable Web APIs to access their contents, we could bypass the SDM.

2.6.3 The Metadata Store and Access Module (MSAM)

This module provides storage functionalities to save and retrieve all semantic annotations. It is also responsible for the users management. All these features are available through a Web based HTML interface.

User Management. In our system each user must be registered in order to be identifiable; in this way we can manage his personal data and his tagging metadata and we can support him with additional system functionalities.

Semantic Annotation. The main goal of our system is the collection of the semantic assertions produced by the semantic tagging activity. Every semantic assertion has been generated by a particular user in a precise *moment.* All these data are stored by the 'Metadata store and access module' as the outcome of semantic tagging activity.

User oriented views. When we talk about user oriented views, we mean all the available ways that a registered user, after his authentication, can exploit to interact with the system and visualize his personal profile data and his tagging metadata. Here is a list of all those implemented in our system:

- **Visualization of user semantic tagging metadata** :

  - *my Web resources* view : all the Web resources semantically tagged by the user ordered by date, each one with all the associated semantic assertions;
  - *my semantic assertions* view : all the semantic assertions created by the user ordered by referred concept and property (every semantic keyword is a link to the next view);
  - *my Web resources tagged with* view : all the Web resources semantically tagged with one particular concept (is a list of Web resources links ordered by date).

- **Deletion of a semantic assertion from a web resource** (*delete a tag* view).
- **Visualization / partial modification of user personal profile data** (*my profile* view);

Generic search-oriented view. This section includes all available ways that a generic user, authenticated in the system or not, can exploit to execute a semantic search among the metadata collected:

- *basic search* view : search of all the Web resources semantically tagged with one or more generic semantic assertions:
  - the user chooses one or more lexical forms;
  - every lexical form is disambiguated interacting with 'The sense disambiguation module' previously described and retrieving all the meanings used to refer to it; in this way the user can define the intended concept, choosing between the multiple meanings presented;
  - once a concept has been chosen, the user can select a particular property that links the Web resource he wants to find to the concept, thus defining a generic semantic assertion;
  - SemKey, interacting with the 'Metadata store and access module', will retrieve all resources matching the set of generic semantic assertions previously defined.

Figure 8 illustrates an example of the *basic search* view system interface; the user has chosen the word 'ajax' and, among the list of concepts retrieved disambiguating it, has pointed out the concept of 'AJAX (programming) (Asynchronous JavaScript and XML), a technique used in Web applicatins...'. Then, selecting the property 'The topic of the resource is', has formed a generic semantic assertion, requesting to find all Web resources that talk about the Asynchronous JavaScript and XML Web programming technique. SemKey shows a list of all the resources semantically tagged that match the parameters previously specified.

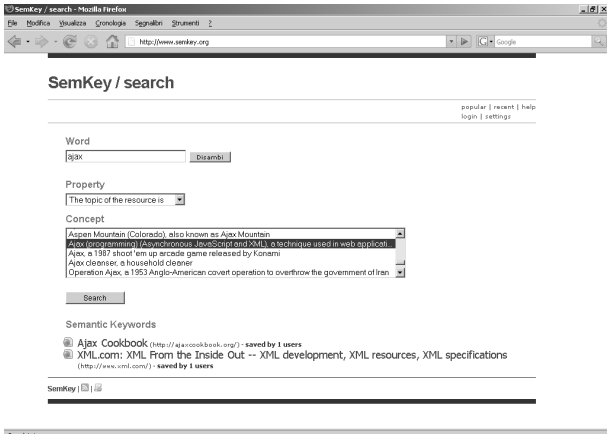

Figure 8: Example of semantic search.

3. CONCLUSIONS

In this work we have described the architecture and the main implementation choices of SemKey: a semantic collaborative tagging system.

By considering the fundamental weak points of existing social tagging systems, we have deduced that most of them are referable to the absence of any semantic support in tagging activity. We have proposed the semantic keywords: they are new semantic-aware metadata that substitute existing tags allowing tagging in a semantic context. Each semantic keyword refers to a particular concept that can be identified using one or more different strings, called lexical forms. In SemKey, we have chosen to exploit WordNet, a widely adopted lexical resource and Wikipedia, the biggest collaboratively edited Web encyclopedia in order to disambiguate lexical forms to refer to a concept, thus identifying a semantic keyword; we have pointed out the main advantages and disadvantages experimented in their adoption. We have described in detail the architecture of SemKey along with all the organizational issues faced. Then we have illustrated the fundamental implementation choices, providing some practical examples of its functioning.

At the basis of our idea of semantic tagging stands the availability and completeness of a global collection of concepts and lexical forms in order to express and univocally reference semantic keywords; both WordNet and Wikipedia have been used in order to test their possible support to this tasks. We have explored their main organizational features: WordNet presents a rich set of parts of speech and a strongly structured net of relations between them, but it lacks many data useful to support proper names disambiguation and it is not collaboratively edited; Wikipedia is an encyclopedia so its content is composed mainly by a very rich set of names along with their extended descriptions. Wikipedia has strong proper names coverage; it is also continuously updated, but lacks a structured set of relations between the concepts described, even if its documents are interconnected by a huge number of links: at present, only the system of Wikipedia categories is available as an attempt to provide some sort of relaxed structure to its informative content. Besides Wikipedia and WordNet we must mention an early project OmegaWiki [8]; it is attempting to build a free socially-edited multilingual thesaurus; it organizes concepts and terms adopting a structure that seems suitable to support lexical forms disambiguation and concept referenceability as needed by semantic tagging. In parallel with the growing of OmegaWiki informative contents, future works should be oriented to better explore its possibilities of cooperation with semantic tagging systems.

Since now we have analysed semantic tagging activity concerning a global domain. Recently, the advantages provided

by tagging activity has been introduced also in enterprise networks; IBM has announced its version of an internal social bookmarking system: Dogear [14]. The exploitation of SemKey in specific knowledge domains represents another important potential field of application. Indeed, our semantic tagging system can be used referring to defined collections of concepts in order to describe a particular domain of interest. We think that future works could concern the possibility to exploit our semantic tagging tool as a corporate knowledge management and organizational support; it can support the organization and improvement of the accessibility to shared information like internal collections of documents or, in general, any huge amount of data which needs to be collaboratively organized. Many domain specific concepts collections are currently available: for instance MeSH [6], the National Library of Medicine's controlled vocabulary thesaurus is a terminological medical reference currently widely used and that could be adopted as a specific tagging reference. The analysis of this possibilities constitutes an interesting new semantic tagging application scenario.

Summarizing, in this work we have suggested a new semantic tagging system in order to combine semantic technologies with the collaborative tagging paradigm in a way that can be highly beneficial to both areas.

4. REFERENCES

[1] del.icio.us tagging system web site. http://del.icio.us/.

[2] Dublin core metadata initiative web site. http://dublincore.org/.

[3] English wikipedia web site. http://en.wikipedia.org/wiki/.

[4] The friend of a friend (foaf) project. http://www.foaf-project.org/.

[5] Json (javascript object notation) - web site. http://json.org/.

[6] Medical subject headings - official web site. http://www.nlm.nih.gov/mesh/meshhome.html.

[7] Mozilla developer center - plugin. http://developer.mozilla.org/en/docs/Plugins.

[8] Omegawiki web site. http://www.omegawiki.org/Main_Page.

[9] Princeton wordnet web site. http://wordnet.princeton.edu/.

[10] Wikipedia statistics english. http://stats.wikimedia.org/EN/TablesWikipediaEN.htm.

[11] Yahoo my web 2.0 apis reference. http://developer.yahoo.com/search/myweb/.

[12] D. Beckett. Semantics through the tag. Conference presentation XTech 2006: Building Web 2.0, Amsterdam, The Netherlands, May 2006. Yahoo! Inc.

[13] D. B. M. D. Cameron Marlow, Mor Naaman. Position paper, tagging, taxonomy, flickr, article, toread. Technical report, Yahoo! Research Berkeley 1950 University Avenue, Suite 200 Berkeley, CA 94704-1024 - UC Berkeley School of Information 102 South Hall Berkeley, CA 94720-4600, 2006.

[14] J. F. David Millen and B. Kerr. Social bookmarking in the enterprise - ibm. *ACM Queue*, vol. 3, no. 9, 2005.

[15] J. Futrelle. Harvesting rdf triples. Technical report, Natioanl Center for Supercomputing Applications 1205 W. Clark St., Urbana IL 61801, US, 2006.

[16] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. Technical report, Information Dynamics Lab, HP Labs, 2005.

[17] M. Guy and E. Tonkin. Folksonomies: Tidying up tags? *D-Lib Magazine*, Volume 12 Number 1, January 2006.

[18] E. Kroski. The hive mind: Folksonomies and user-based tagging. Blogsite.

[19] R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network. Technical report, SUPRATECS, Universit de Lige,B5 Sart-Tilman, B-4000 Li'ege, Belgium, 2005.

[20] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Technical report, Computer Mediated Communication - Graduate School of Library and Information Science - University of Illinois Urbana - Champaign, 2004.

[21] R. Shena. A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular). Blogsite, September 2005.

[22] G. Smith. Folksonomy: social classification. Blog article, August 2004.

[23] B. L. Tony Hammond, Timo Hannay and J. Scott. Social bookmarking tools (i) - a general review. *D-Lib Magazine*, Volume 11 Number 4, 2005.

[24] D. Weinberger. Tagging and why it matters. Technical report, Harvard Berkman Center for the Internet and Society, 2005.

[25] J. M. Zhichen Xu, Yun Fu and D. Su. Towards the semantic web: Collaborative tag suggestions. Technical report, Yahoo! Inc 2821 Mission College Blvd., Santa Clara, CA 95054, 2006.