Tagpedia: a semantic reference to describe and search for Web resources

Francesco Ronzano
Institute for Informatics and Telematics (IIT) - CNR
Via Moruzzi, 1
Pisa, Italy
francesco.ronzano@iit.cnr.it

Andrea Marchetti
Institute for Informatics and Telematics (IIT) - CNR
Via Moruzzi, 1
Pisa, Italy
andrea.marchetti@cnr.it

Maurizio Tesconi
Institute for Informatics and Telematics (IIT) - CNR
Via Moruzzi, 1
Pisa, Italy
maurizio.tesconi@iit.cnr.it

ABSTRACT

Nowadays the Web represents a growing collection of an enormous amount of contents where the need for better ways to find and organize the available data is becoming a fundamental issue, in order to deal with information overload. Keyword based Web searches are actually the preferred mean to seek for contents related to a specific topic. Search engines and collaborative tagging systems make possible the search for information thanks to the association of descriptive keywords to Web resources. All of them show problems of inconsistency and consequent reduction of recall and precision of searches, due to polysemy, synonymy and in general all the different lexical forms that can be used to refer to a particular meaning. A possible way to face or at least reduce these problems is represented by the introduction of semantics to characterize the contents of Web resources: each resource is described by one or more concepts instead of simple and often ambiguous keywords. To support these task the availability of a global semantic resource of reference is fundamental. On the basis of our past experience with the semantic tagging of Web resources and the SemKey Project, we are developing Tagpedia, a general-domain "encyclopedia" of tags, semantically structured for generating semantic descriptions of contents over the Web, created by mining Wikipedia. In this paper, starting from an analysis of the weak points of non-semantic keyword based Web searches, we introduce our idea of semantic characterization of Web resources describing the structure and organization of Tagpedia. We introduce our first realization of Tagpedia, suggesting all the possible improvements that can be carried out in order to exploit its full potential.

Categories and Subject Descriptors

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

General Terms

semantic resource, knowledge organization, semantic web

Keywords

semantics, web, social, wikipedia, data mining

SWKM'2008: Workshop on Social Web and Knowledge Management @ WWW 2008, April 22, 2008, Beijing, China.

1. INTRODUCTION: KEYWORD BASED SEARCHES

Currently, keyword based Web searches are the preferred way to seek for resources of interest over the Web. Each resource, usually identified by its URL, can be accessed by one or more keywords describing its content. The most widespread methods to explore links between Web resources and keywords are **the exploitation of a search engine** or **the access to a collaborative tagging service** (see Figure 1).

Search engines like *Google*, *Yahoo*, *Ask* and so on are examples of automated information extraction systems: they analyze the data and the structure of Web contents as well as the search behaviour of users and the frequency of usage of different search strings to collect the most appropriate keywords that can be used to access a Web resource (see the lower portion of Figure 1).

On the other side, collaborative tagging systems like *delicious*, *Flickr*, *YouTube* and *Technocrati* rely upon user contribution. They are examples of social classification systems: each person who belongs to the community of users of a collaborative tagging system describes Web resources of interest by means of one or more freely chosen keywords, called tags. All the tags associated to Web resources are collected and exploitable by every user in order to find many resources of interest. A popularity value is usually associated to each tag describing a Web resource to point out the number of times it has been chosen to characterize that resource and consequently the importance of the tag itself among those related to the specific resource (see the upper portion of Figure 1).

Even if they are very popular, **keyword based Web search approaches show many weak points in managing language expressivity**. Many keywords can identify distinct concepts (*polysemy*): as a consequence the precision of search results decreases. Moreover if we don't search for a common sense of that keyword, it is often very difficult to explore the search results space so as to find Web resources of interest among those retrieved. For example, let us suppose that we want to find all the resources dealing with 'ajax' intended as the Greek hero: choosing 'ajax' as search text string, there are no links related to mythology among the first 30 search results of Google. If we better specify the search string in order to solve the problem, we partition the space of relevant search results depending on the particular word added to 'ajax' to disambiguate its meaning. For instance, depending on the addition of the word 'hero' or the word 'mythology' to 'ajax' in the search string, considering

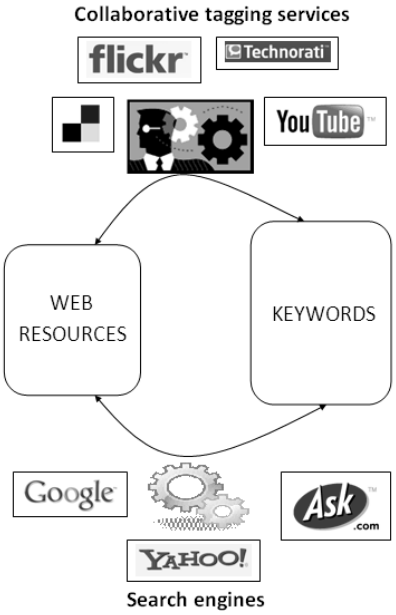

Figure 1: Two ways to associate keywords to resources

the first 10 search results shown by Google, only two of them are present in both cases. Besides polysemy, also *synonymy* affects precision and recall of keyword based Web searches. In fact, when a specific meaning can be accessed through two or more keywords, the set of search results is different depending on the particular keyword chosen. Moreover, the different *level of precision* and the *many possible users points of view* that can be considered describing a particular resource, often cause a considerable loss of quality of Web searches. For a deeper analysis of all the factors that affect efficiency and effectiveness of keyword based Web search systems see [4] [10] [12] [25].

In order to face the different drawbacks of the systems just analyzed, many distinct methods have been applied. The **aggregation of search results from different search engines and their post elaboration** is experimenting a growing diffusion. Systems like **Vivisimo** [15], **Grokker** [18] and **Kartoo** [19] are meta search engines. They collect search results from other search engines and group them exploiting, for example, the category hierarchy of Yahoo and Wikipedia (Grokker) or creating clusters of similar search results and characterizing each of them by one or more additional keywords (Vivisimo). They also display search results cartographically through very expressive maps that connect the most relevant resources to the most used keywords (Kartoo).

Also considering tagging systems, we can find many proposals to better organize search results to improve their quality and the effectiveness of the search. **FolkRank** [5] is an algorithm created to rank search results in a tagging system, calculating a ranking value for each of them and thus evaluating their relevance. Also user profile is exploited in order to adapt ranking calculation to the information needs of every single user.

The rest of this paper is organized as follows. In Section 2 we describe our idea of semantic characterization of Web resources, underlining the need for a general-domain semantic resource of reference in order to support this task, taking into account also our past experience with the semantic tagging and SemKey. In Section 3 we introduce Tagpedia, the semantic resource of reference we have created by mining Wikipedia, explaining its organization and structure (Subsection 3.1). In Section 4 we describe how Tagpedia can be utilized, describing the Tagpedia Web API and showing all the possible improvements to Tagpedia to exploit its full potential. Conclusions are described in Section 5.

2. FROM KEYWORDS TO CONCEPTS: SEMANTIC CHARACTERIZATION

We can solve, or at least substantially reduce, Web resources organization and classification problems by adding a further level of completeness in their characterization: **the semantics**. Instead of relying on post processing of search results, we can directly semantically describe resources thanks to their association with one or more properly chosen concepts. In this way we extend the characterization of resources introducing the semantic level: each resource (R) is described by one or more concepts (C) and in turn each concept can be accessed through one or more keywords (K) (see Figure 2). When we search for some informaton of interest, we can better specify our informative needs and we can easily and effectively access relevant results thanks to the support and the exploitation of the collection of concepts used to describe Web resources, referred to as semantic resource in what follows.

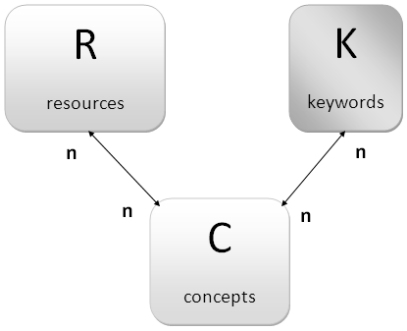

Figure 2: Relations between resources, keywords and concepts

This way of improving Web contents organization represents an attempt to *realize the semantic description of information that stands at the basis of the Semantic Web vision.*

At present there are many proposal of semantic classification methods for Web contents. **FolksAnnotation** [13], for instance, tries to extract the tags that describe a Web resource from a collaborative tagging system, automatically mapping them to the corresponding concepts of a predefined domain ontology. Such kind of systems usually require a strongly and well organized ontological frame of reference that is difficult to realize; they have not provided significant improvements in comparison with the classical keyword based methodologies. A different approach is those exploited by systems like **Semantic Halo** [3]: it improves tag based search systems adding semantic information without relying on ontologies. Analyzing co-occurrences and frequencies of

tags, Semantic Halo algorithm extracts groups of tags useful to better specify and drive user search, like more general or more specific ones or group of keywords defining a particular naming of the selected tag. Not enough experimental data on the effectiveness and usefulness of this method to improve tag based searches is currently available. Summarizing, *a strong and widespread infrastructure that organizes and provides access to Web resources on the basis of semantic classificatory information is still absent.*

During the first half of 2007, we have tried to realize the possibility to semantically describe Web resources developing SemKey [4], a semantic collaborative tagging system. It extends current tagging systems allowing to characterize resources by referring to concepts. Each user can point out and describe Web resources of interest: starting from a freely chosen tag, he can disambiguate it thanks to the support of *Wikipedia* [14] and *WordNet* [17] in order to identify one or more defined concepts. In this way he produces a semantic assertion that is the description of a specific feature of Web resources through one or more chosen concepts. Thus we can potentially overcome the limits in the description of Web resources related to the complexity of language, exploiting their semantic characterization as well as the semantic relations between concepts present in WordNet and Wikipedia.

We have implemented a working prototype of SemKey; by analyzing the usage patterns and the semantic classification support provided by our system, we have identified **two key factors that need to be improved in order to really make possible semantic characterization of Web resources**, as described in the previous part of this Section.

Both Wikipedia and WordNet, even if they show important features to support the semantic description of Web resources, are weakened by relevant lacks. WordNet presents a rich set of parts of speech and a strongly structured set 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. Thus Wikipedia has strong proper names coverage and it has been proposed as a named entity disambiguation resource in [7] and [8]; 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 and loosely classified through categories. As a consequence the semantic resources considered are in some way complementary, but they have been built and structured for purposes different from the semantic characterization over the Web. In order to better support this task **we need a semantic resource built and structured ad hoc**, which is still absent: it must feature all the advantages of those just analyzed, removing pointless informative contents.

Moreover, a great limit to the usability of SemKey and to an easy definition of new semantic metadata is represented by the different steps users must carry out to compose a semantic assertion. This often discourages them from creating semantic metadata. **Some sort of automation is necessary in order to speed up the tag disambiguation process or to execute it through automated procedures**.

3. TAGPEDIA: A GENERAL DOMAIN SEMANTIC RESOURCE OF REFERENCE

Starting from the need for a global semantic resource exploitable as a reference to describe Web contents and therefore comprehensive and updated, we have proposed a possible solution to this demand, designing and building Tagpedia. It is **a semantic organization and classification of tags, intended as words or in general brief textual expressions, that people may use to describe Web Resources**. Tagpedia is **based on the model of term-concept networks** [11], **structured ad hoc to support the semantic characterization of Web contents and initially populated exploiting Wikipedia data**. In particular we have tuned a new way of mining Wikipedia to extract the information needed to build Tagpedia so as to support concept based descriptions of Web resources also through tag disambiguation.

We have chosen **Wikipedia** as the starting point because **it represents the most rich and constantly updated encyclopedic reference over the Web with a huge set of semantic contents included, even if not explicitly exposed and easily accessible**. During the last few years many studies have been carried out finding new ways to extract useful semantic data exploiting the great amount of information contained in Wikipedia. Information organizational patterns like infoboxes, internal and external links, redirect and disambiguation pages have been analyzed in order to extract valuable data. The DBPedia Project [16], for instance, is a relevant attempt to extract semantic data from Wikipedia, making them available over the Web complying with Semantic Web standards [6]. DBPedia is a global knowledge base derived from Wikipedia, not specifically intended for Web resources description as Tagpedia is. In [24] there is a description of KLYN, a system that autonomously semantifies Wikipedia, automatically suggesting data inconsistencies, lacks or incompletenesses. Wikipedia has been also successfully exploited to compute semantic relatedness between words [21] and natural language texts [9], but also to tune new named entities disambiguation methodologies [7] [8]. Semantic relationships between Wikipedia categories have been studied in order to make the search of information easier and to give articles editors relevant suggestions [20]. Moreover some research has been done to understand and measure the way Wikipedia articles are created and their contents become mature [22] or to analyze statistical information about the growth of the data that constitute Wikipedia, the types of articles, the editors, the link and category structure and so on [23].

3.1 The structure of Tagpedia

The main aim of Tagpedia is the semantic characterization of data over the Web. In particular it must allow to describe a Web resource through the association with one or more univocally referenced concepts. Thus, **the main constitutional unit of TagPedia is the concept**. Each concept must be unequivocally identified but also easily accessed. The main way to point out a concept is through the words that refer to it. Such words will also be called tags in the following. As a consequence, each concept is identified by *the set of all the words or, more generally, all the alfanumeric expressions of any kind that can be adopted by a community of users to refer to it*, thus constituting a

set of synonymous tags or **syntag set**. Syntag sets are the molecules which form Tagpedia.

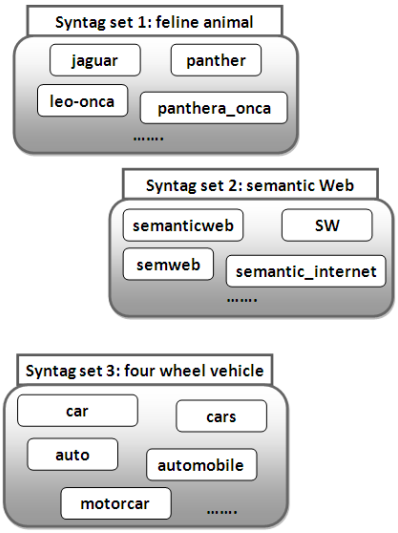

Figure 3: Three syntag sets

The creation of an initial rich collection of syntag sets is the first necessary step that must be carried out to build our semantic resource. Wikipedia shows many features exploitable to create such a collection of syntag sets. In particular, in Wikipedia an article usually defines a specific concept. As a consequence in order to bootstrap Tagpedia, we create syntag sets from the articles of Wikipedia. In Figure 3 we show three examples of syntag sets made up by tags collected mining Wikipedia.

To be more precise, Wikipedia pages can be substantially divided into three groups:

- **article pages**: each describes a particular concept, identified by the title of the same page;
- **redirect pages**: each links an alternate literal expression, that constitutes the title of the redirect page, to the corresponding concept, usually identified by the title of an article page;
- **disambiguation pages**: each lists all the possible concepts, usually identified through the titles of article or redirect pages, that can be referred by the literal expression constituting the title of the disambiguation page.

The redirect and the disambiguation page mechanisms are two important Wikipedia organizational solutions that can be exploited to build and enrich syntag sets.

Once identified a concept referring to a particular article page, we create an initial version of a syntag set, pointed out by a unique identifier, including only the tag corresponding to the title of the page (in Figure 3, considering the syntag set 1, the tag 'jaguar' is the title of an article page). Then we collect all the words and expressions that may be used to refer to that concept.

As previuosly mentioned, in Wikipedia the **redirect** mechanism is used to link alternate literal expressions to the original encyclopedic article that describes a specific entity. It is usually used to manage synonyms, abbreviations, acronyms, misspellings, other spellings, different punctuations, particular capitalization rules and so on. In TagPedia we mine Wikipedia content and extract all the redirect information analyzing redirect pages; for each of them we enrich the syntag set related to the referred concept by adding the title of the page as a new tag (in Figure 3, considering the syntag set 1, the tag 'leo onca' is extracted from a redirect page).

Moreover, Wikipedia usually manages polysemy through the **disambiguation pages**. As said, each disambiguation page represents a collection of links to all the different article pages that identify the distinct meanings pointed out by the page title (textual string). For example, the word '*ajax*' is highly polysemous and has 49 different meanings in Wikipedia: its disambiguation page contains links to 49 distinct article pages; each one identifies a particular concept. We analyze Wikipedia disambiguation pages as a futher source of information to enrich the syntag sets of Tagpedia through the addition of new words that refer to a defined meaning. In particular, for every disambiguation page, we point out each syntag set related to the concepts referenced inside its Wikipedia text and we add the title of the same disambiguation page as a new tag exploitable to access to the selected syntag sets (in Figure 3, considering the syntag set 1, the tag 'panther' is extracted from a disambiguation page).

Summarizing, let us define C_i a concept derived from a specific Wikipedia article page P_i. To populate with tags the syntag set for C_i we extract:

- the title of P_i;
- the title of every redirect page to P_i;
- the title of every disambiguation page containing a link to P_i.

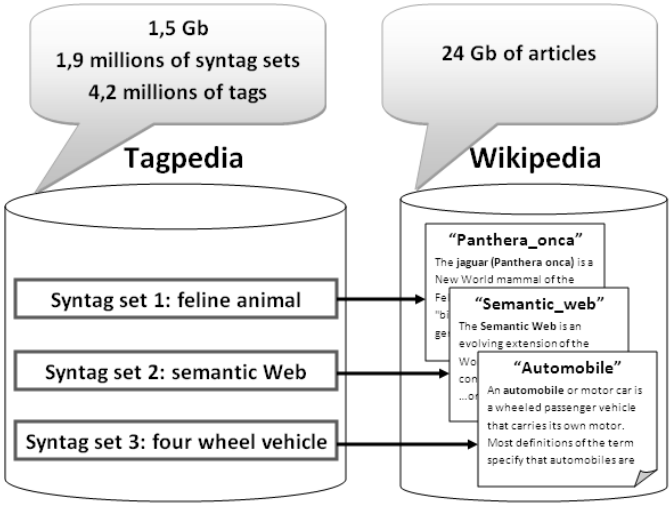

Figure 4: The structure of Tagpedia

Starting from a dump of the English version of Wikipedia, we have developed a set of C++ routines, that automatically analyze the text of Wikipedia articles. By mining structural

elements of Wikipedia syntax as well as by considering texts punctuation and by exploiting pattern matching techniques mainly based on regular expressions and string analysis, our routines gather all the concepts as well as all the possible tags used to refer to each single meaning, thus defining a huge collection of syntag sets. The meaning of each concept, identified by a syntag set, is also better specified by pointing to the corresponding article in Wikipedia.

All these data are collected in a relational database properly designed and optimized for a fast access. It is constituted by two basic collections: *the concept table* and *the tag table*. The first one gathers all the concepts of Tagpedia assigning to each of them a unique identifier, the Concept ID and a brief definition, extracted from the English version of Wikipedia. For every concept we also collect the URL of the corresponding Wikipedia article. On the other side, the tag table contains links between each concept, referenced through its identifier, and all the tags used to access to it.

By mining September 2007 dump of the English version of Wikipedia, we have obtained more than **1,9 millions of syntag sets** and more than **4 millions of tags** used to point out the intended concepts, each one referencing a specific Wikipedia article (see Figure 4).

Considering Figure 5, we can visualize the weight of the different sources of the 4.230.740 tags of Tagpedia. The number of tags extracted from article pages (P) is equal to the number of syntag sets, that is 1.927.378. Among the 2.303.362 remaining tags, 481.250 have been generated by mining disambiguation pages and 1.822.112 by analyzing redirect ones.

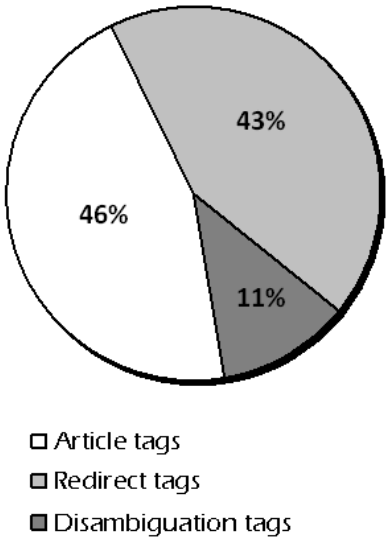

Figure 5: Sources of the tags in Tapgedia

This group of syntag sets constitutes the basis of Tagpedia providing a way to unequivocally access and refer to concepts when users must semantically describe or search for Web resources.

Number of del.icio.us URLs:	100	
Number of distinct tags:	1087	
Percentage of successfull disambiguations:	84	%

Table 1: Tagpedia tag disambiguation support: preliminary evaluation results

4. EXPLOITING AND IMPROVING TAGPEDIA

In order to support the generation of semantic descriptions of Web resources or to semantically search for Web contents, the information contained in Tagpedia should be easily accessed, querying the whole collection of syntag sets. For this purpose we have developed the **Tagpedia Web API** that is a simple set of procedures that may be invoked via Web to exploit the semantic support offered by Tagpedia. These procedures carry out few fundamental tasks and may be composed to realize more complex functions; their execution can be easily requested by other external Web applications so as to integrate semantic features.

The main tasks that Tagpedia Web API supports are:

- the definition of *all the possible meanings for a given tag*, i.e. all the syntag sets that contain the tag;
- the collection of *all the tags belonging to a specific syntag set*, i.e. all the words or expressions exploitable to access that particular meaning;
- the retrieval of *the short textual description of a specific syntag set*.

Exploiting Tagpedia Web API, we have integrated this semantic resource into SemKey, our semantic collaborative tagging system, substituting WordNet and Wikipedia so as to support the disambiguation of the meaning of tags. Once chosen one or more tags, the user specifies the right meaning for each of them, choosing a particular syntag set among those including the intended tag. An early prototypal Web-based interface useful to explore and interact with Tapgedia is accessible at the URL *www.tagpedia.org*.

In order to evaluate the coverage of Tagpedia and also to obtain suggestions to improve this semantic resource, we have tried to manually **point out the right meaning of the tags associated to the 100 most popular Web resources over del.icio.us**, tagged by more than 25000 users. Relying upon Tagpedia Web API, we have developed a Web based procedure that, starting from the URL of a Web resource retrieves all the related tags in del.icio.us. All the possible meanings of each tag are retrieved from Tagpedia along with their short descriptions and the user manually verify if the right concept is present. In this way, collecting all the results of our user based tests, we have obtained a first evaluation of the disambiguation effectiveness of our semantic resource. The results are shown in Table 1.

Tagpedia provides a valid support to the process of disambiguation for 84% of the total number of tags considered.

Anyway we have identified several different ways to improve its contents and, as a consequence, its semantic coverage and its usefulness. In the following part of this section we will describe these proposals for future works.

Despite its good disambiguation coverage, there are different particular tags like 'sem_web', 'inplaceedit', 'web_dev'

and similar ones that are not managed by Tagpedia, because they are non conventional words, often created by a user to describe a particular concept and then accepted and exploited by many others. One possible solution to this problem is **the introduction of collaborative Web editing techniques for Tagpedia contents**. Giving users the possibility to create new syntag sets or to merge or extend existing ones through new tags is fundamental for such a kind of resource. Indeed the effectiveness of Tagpedia in the description of Web resources is proportional to the possibility to adapt and enrich this semantic resource in respect to the variability of user descriptive needs. In this context, the introduction of the possibility to collaboratively collect and manage data, following a Wiki-like paradigm, represents a key factor of current Web and is a crucial issue considering Tagpedia.

Another aspect of Tagpedia that can be substantially improved is **the enrichment of its semantic contents with the addition of semantic relations between syntag sets**; they are useful to better identify concepts or to easily search for them. Each syntag set, representing a meaning, may be connected to other ones through relationships like specialization, generalization, relatedTo and similar ones. Possible ways to mine relevan relations between syntag sets are the analysis of the internal links between Wikipedia article pages as well as the exploitation of the hierarchy of Wikipedia categories. For instance, relying upon relations, when we specify the concept to search for or when we must choose a specific concept to semantically characterize a resource, the system can show the most general or the most specific ones to simplify this task. Similarly, during a semantic search, starting from a specific syntag set, if we can browse all the related ones, we can better specify our search needs and thus easily retrieve the desired information.

A third way to improve and enrich Tagpedia is **the definition of semi-automated procedures to extend its data, exploiting other resources and importing their contents into Tagpedia**. Other relevant free Web thesauri or dictionaries or other language tools can be valid sources of information. For instance the *Dictionary of Automotive Terms* [1] or the *Free Online Medical Dictionary* [2] are two domain specific resources that can be integrated in Tagpedia. Moreover, mapping rules between Tagpedia syntag sets and other Web semantic resources can be defined to integrate different sources of information thanks to the common gronud represented by Tagpedia itself.

Another aspect that must be further addressed in Tagpedia, is **the support for multilinguism**. In Tapgedia, each syntag set is language independent. The tags constituting that particular syntag set are specific to the particular language. Managing the possibility to collect different tags belonging to different languages into a syntag set, we can deal with different languages and once identified one or more particular concepts we can make language indipendent semantic searches. We think that this possibility should be better explored and defined, trying to determine specific semantic search patterns.

As already mentioned in the concluding part of Section 2, **the definition and tuning of automated or semi-automated procedures to create semantic descriptions** is a further important issue to be faced. Users should be allowed to semantically describe Web resources in an easy way; they must be supported in the task of turning simple keywords into concepts or browsing the collection of syntag sets constituting Tagpedia without complicating their usual interaction patterns or compromising the usability of the systems they interact with. Moreover, automated methodologies to derive semantic descriptions of Web resources from simple keyword based ones can also be tuned, so as to create an initial solid collection of semantic metadata and bootstrap this new way to characterize resources over the Web.

5. CONCLUSIONS

In this paper we have presented Tagpedia, a collection of tags semantically structured, built ad hoc to describe Web contents.

Starting from a brief analysis of the weak points of keyword based methodologies for information organization and searching and considering also the current approaches to face these issues, we have introduced the possibility to semantically describe Web resources through concepts. To make it possible, we have developed an initial version of Tagpedia a general domain semantic resource of reference, created by mining Wikipedia. After a description of its structure and organization and an overview of the Tagpedia Web API, useful to easily access and exploit the information collected in Tagpedia, we have focused our attention on the possible improvements to this semantic resource. Collaborative wiki authoring, syntag set relations enrichment, automated procedures for content extraction from external sources, support for multilinguism and automated generation of semantic descriptions of Web resources are some of the many improvements considered that can be carried out, underlining its broad enhancement possibilities.

On the base of all these considerations, we believe that Tagpedia, despite its initial stage of development, represents an important attempt to support the introduction of semantics over the Web, trying to put in practice the principles of the Semantic Web on a global scale and to better structure and manage the huge amount of data constituting the actual Web.

6. REFERENCES

[1] Dictionary of automotive terms. http://www.motorera.com/dictionary/.

[2] Free online medical dictionary. http://cancerweb.ncl.ac.uk/omd/.

[3] Alessio Malizia Alan Dix, Stefano Levialdi. Semantic halo for collaboration tagging systems. *In the Social Navigation and Community-Based Adaptation Technologies Workshop - June 20th, 2006, Dublin, Ireland.*

[4] Francesco Ronzano Marco Rosella Salvatore Minutoli Andrea Marchetti, Maurizio Tesconi. Semkey: A semantic collaborative tagging system. *In the Tagging and Metadata for Social Information Organization Workshop at the World Wide Web Conference 2007 - May 8, 2007, Banff, Alberta, Canada.*

[5] Christoph Schmitz Gerd Stumme Andreas Hotho, Robert Jĺaschke. Folkrank: A ranking algorithm for folksonomies http://www.kde.cs.uni-kassel.de. *In the Lernen - Wissensentdeckung - Adaptivität Workshop - October 9-11, 2006, Hildesheim, Germany.*

[6] Soren Auer and Jens Lehmann. What have innsbruck and leipzig in common? extracting semantics from

wiki content. *In the 4th European Semantic Web Conference - June 5th, 2007, Innsbruck, Austria.*

[7] Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. *In the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics - April 9-16, 2006, Trento, Italy.*

[8] Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. *In the Empirical Methods in Natural Language Processing Conference - June 28-30, 2007, Prague, Czech Republic.*

[9] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *In the Proceedings of the 20th International Joint Conference on Artificial Intelligence - January 6-12, 2007, Hyderabad, India.*

[10] Scott A. Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *In the Journal of Information Sciences, vol. 32, April, pag. 198-208, 2006.*

[11] Andrew Gregorowicz and Mark A. Kramer. Mining a large-scale term-concept network from wikipedia. *Mitre Technical Report, October 2006.*

[12] Marieke Guy and Emma Tonkin. Tidying up tags? *D-Lib Magazine*, 12, January 2006.

[13] Hugh C. Davis Hend S. Al-Khalifa and Lester Gilbert. Creating structure from disorder: using folksonomies to create semantic metadata. *In 3rd International Conference on Web Information Systems and Technologies - 3-6 March, 2007, Barcelona, Spain.*

[14] http://en.wikipedia.org/wiki/. The english version of wikipedia.

[15] http://vivisimo.com/. Vivisimo, search done right!

[16] http://wiki.dbpedia.org. Dbpedia.

[17] http://wordnet.princeton.edu/. Princeton wordnet.

[18] http://www.grokker.com/. Grokker enterprise search management.

[19] http://www.kartoo.com/. Kartoo meta-search engine.

[20] Wolfgang Nejdl Sergey Chernov, Tereza Iofciu and Xuan Zhou. Extracting semantic relationships between wikipedia categories. *In the 1st Workshop on Semantic Wikis at the 3rd European Semantic Web Conference - June 11-14, 2006, Budva, Montenegro.*

[21] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. *In the Proceedings of the 45th Annual Southeast Regional Conference, pag. 106 - 110 - March 23-24, 2007, Winston-Salem, North Carolina, USA.*

[22] Cristopher Thomas and Amit P.Sheth. Semantic convergence of wikipedia articles. *In the Proceedings of Web Intelligence Conference, pag. 600-606 - Silicon Valley, November 2-5, 2007.*

[23] Jakob VoSS. Measuring wikipedia. *In the Proceedings of the 10 th International Conference of the International Society for Scientometrics and Informetrics - July 24-28, Stockholm, Sweden.*

[24] Fei Wu and Daniel S. Weld. Autonomously semantifying wikipedia. *In the Proceedings of the 16th ACM conference on Conference on information and knowledge management, pag. 41-50 - November 6-9, 2007, Lisboa, Portugal.*

[25] Jianchang Mao Zhichen Xu, Yun Fu and Difu Su. Towards the semantic web: Collaborative tag suggestions. *In the Proceedings of the Collaborative Web Tagging Workshop at the World Wide Web Conference 2006 - May 23-26, 2006, Edinburgh, Scotland.*