

# Folksonomies and Ontologies in Authoring of Adaptive Hypermedia

Fawaz Ghali<sup>1</sup>, Mike Sharp<sup>2</sup>, and Alexandra I. Cristea<sup>1</sup>

<sup>1</sup> Department of Computer Science, University of Warwick,  
Coventry, CV4 7AL, United Kingdom  
{F.Ghali, A.I.Cristea}@warwick.ac.uk

<sup>2</sup> School of Engineering, University of Greenwich,  
Chatham Maritime, Kent, ME4 4TB, United Kingdom  
M.W.Sharp@gre.ac.uk

**Abstract.** *Semantic Web* and *Social Web* are two rapidly growing areas, evolving independently but complementing each other. On one hand, ontological aspects in the Semantic Web represent the top-down model, which lacks flexibility and scalability, in addition to facing the technical challenges of deployment on the current web. On the other hand, folksonomies on the Social Web represent the bottom-up model, which consists of unstructured data and carries no semantics, leading to questions of accuracy and reusability. Therefore, we worked on merging folksonomies from the Social Web with ontologies from the Semantic Web. This merge has the following advantages: 1) Creating semantic relations between tags of folksonomy; 2) Enabling reasoning on the Social Web. 3) Augmenting the authoring process of adaptive hypermedia, by providing rich, free, but also hierarchically structured data from the combined Social and Semantic Web.

**Keywords:** Semantic Web, Social Web, ontology, folksonomy, semantic enrichment, authoring, adaptive hypermedia.

## 1 Introduction

Adaptive and adaptable hypermedia authoring is challenging, especially with respect to the move to the Social Web and/or the Semantic Web. The *Semantic Web* is a web of data, in which the machines are able to understand, leading to an effective way of finding and sharing information [23]. Moreover, in [6] we find that “the Semantic Web is all about authoring”, in the sense of rigorous, rich creation of annotated data. The *Social Web* is also based on annotations, however, free ones. The term refers to the activities of users on the web, such as chatting, discussion groups, and online communities; these activities are supported by social network services, which collaborate to make the web an open social network [22]. The Semantic Web organizes and categorizes data into *ontologies*, a more rigorous, formal way of creating machine-processable data structures. On the other hand, the Social Web uses folksonomies, which represent a method of collaborative categorization using freely-

chosen keywords called *tags* [19]. In this paper, we present a mechanism of mapping unstructured data from Web 2.0 to structured ontologies. This further allows augmenting the authoring of adaptive hypermedia by providing rich, free, but also hierarchical structured data from the social and semantic web. This can be used by authors in different ways. Either to add adaptivity to linear social web data, or as a pool of keywords to draw from, as will be explained. Such automatic authoring steps can be a help in the “*difficult problem*” [7] of adaptation authoring.

## 2 Merging Methodology

Our social and semantic web merging methodology consists of three main phases: 1) *Filtering misspelled* tags from the Social Web. 2) *Grouping* unstructured tags based on co-occurrence values. 3) *Mapping* grouped tags onto matching elements of ontologies (using Swoogle [21] and Jena [11]). Fig. 1 illustrates these main phases. The input is a set of unstructured tags from Flickr. These are filtered using the Google API<sup>1</sup>. Next, they are grouped, using the relations between these tags (i.e., their mutual *co-occurrence values*, as explained in section 2.2). Thirdly, we used Semantic Web tools (i.e., Swoogle and the ontologies available on the web) for the mapping process between grouped tags and elements of selected ontologies. Finally, we create a structured hierarchy from the structured tags. Fig.1. also illustrates that our work is divided into two sections; the first section is done on the Social Web, whereas the other section is done on the Semantic Web.

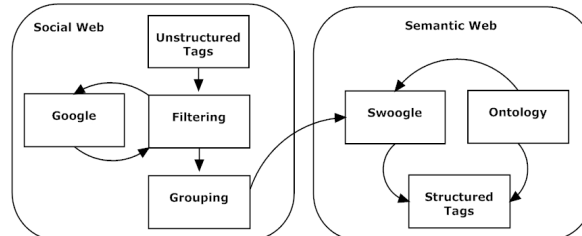


Fig. 1. Merging Social Web with Semantic Web.

### 2.1 Filtering Misspelled Tags

Tags on Flickr are generated in abundance by end users. However, the quality often suffers. E.g., some of those tags are misspelled. We used the Google API to correct misspelled keywords, where the Google spell checker software analyses the query, examining if the tag represents the most used version of a word. If it finds more related search results with an alternative spelling, based on occurrences of all keywords on the Internet, then the misspelled tag is replaced with the suggested one.

<sup>1</sup> API: Application Programming Interface; Google API: <http://code.google.com/>

## 2.2 Grouping Filtered Tags

In this phase, we group similar tags based on statistical information retrieved from their *co-occurrence values*, by using the Flickr API. The *co-occurrence value* between two tags represents how many times these two tags are used together in tagging multiple resources. A formalization of a classical formula for computing the co-occurrence is shown below:

$$\text{co-occurrence}(tag1, tag2) = \sum_{res \in resources} istag(tag1, res) * istag(tag2, res) \quad (1)$$
$$istag(tag, res) = \begin{cases} 1, & tag \in tags(res) \\ 0, & rest \end{cases}$$

$tags(res)$  – all tags of resource  $res$ ;  $resources$  – the set of all resources;

For example, consider that the algorithm found two images on Flickr tagged with “love” and “harmony”. Then, the co-occurrence value between the tag “love” and the tag “harmony” is 2. If these two tags are used together again, the co-occurrence value will be incremented accordingly. This means that the relatedness between the two tags is determined by the number of the times these two tags appeared together in the whole dataset.

## 2.3 Mapping between Grouped Tags and Elements of Ontologies

Grouping tags is a first important step. However, the inclusion of a number of tags in the same group does not provide any information about the type and structure of the relation between these tags. Therefore, in order to enrich the grouped tags with semantic relations, the next phase requires adding semantic content (i.e., from ontologies). Thus, matching between the grouped tags and ontologies is performed.

We used Swoogle [21] to retrieve the required ontologies. Swoogle is a semantic search engine for semantic contents on the web. It returns the ontologies as RDF files; however, it does not provide reasoning and extraction of fully automated semantic relations between the grouped tags. This can be performed with another semantic API, such as Jena [11]. Jena is an open source semantic web framework for Java. It provides an API to extract data from RDF and write to RDF graphs. The purpose of using Jena is to parse and serialize RDF files retrieved from Swoogle to determine the semantic relations between the tags within a group.

## 3 Experiment

We used data from Flickr, a social website, allowing users to upload their photos and describe them via “tags” (a form of metadata). The initial subset is called *tag cloud* of Flickr<sup>2</sup>, which is the set of the most popular tags. This amounts to about **145 tags** ( $size(tag\_cloud) = 145$ ). However, this set of tags does not define relations between tags, or contain quantifiers of these relations, which are necessary to analyze the co-occurrence among the tags. We extract relations by extending the initial tag cloud to a

---

<sup>2</sup> <http://www.flickr.com/photos/tags/>

larger set of *related tags* (**11,138 tags**). Flickr has an open API<sup>3</sup> for data retrieval (e.g., photos, tags, profiles or groups).

### 3.1 Conditions of the Experiment

In the experiment, we combined a set of APIs from different sources to accomplish our goal of enriching tags on Flickr with semantics relations. In order to limit the results, we also set up a set of conditions as follows:

- 1) Each tag has to occur at least ten times to be a part of a group (this value is determined experimentally, and is based on the fact that spam tags and/or unrelated tags present in our database have a lower average of co-occurrence value, i.e., *co-occurrence* (*spam\_tag*, *any\_tag*) < 10).
- 2) Every two groups that share more than five tags must be combined into one group, to avoid redundancy (this result in an average of number of tags in per group of ten tags).

### 3.2 Steps of the Experiment

We developed a specific application that can work with the APIs mentioned in this paper. Firstly, we retrieved tags from the *tag cloud* and its related tags, and then we calculated the initial co-occurrences between those tags, using the appropriate Flickr API functions. Secondly, we filtered the misspelled tags using Google's spell checker API, as below. If the word is correctly spelled, the check spelling function will return NULL. Following, we calculate the final grouping. One group is created for each *tag cloud* tag. Matching the rest of the tags to these groups is based on a *co-occurrence value* threshold of 10 (based on an average experimental value). Next, the initial number of groups are reduced, by searching for common tags between groups. In the final phase of the experiment, we interpret relations between grouped tags by mapping them onto ontology elements.

### 3.3 Results

As previously mentioned, the initial input of the experiment was a set of **145 tags** from the *tag cloud*. Each of these tags is used to retrieve related tags. The extended resulting input was a set of **11,283 tags**. This set of tags was manipulated in the following three phases: 1) Filtering. 2) Grouping. 3) Mapping. During the filtering process, all misspelled keywords (e.g., "aple" instead of "apple") were corrected using Google API spell checker. **2,124 tags** were misspelled, **1,854 tags** of the misspelled tags were corrected, whilst the rest of **2,302 tags** were removed from the total set, because the co-occurrence value was below the threshold. Some tags appeared in different groups with different meanings (e.g., "food" appearing in both food and hobby group). This is solved in the mapping process, by mapping semantic

---

<sup>3</sup> <http://www.flickr.com/services/api/>

relations to different ontologies. After the filtering process, the remaining **8,710 tags** were categorized into groups, based on the *co-occurrence values* between tags.

To concretely show some results<sup>4</sup>, we illustrate the processing of the tag “Food” from the tag cloud. For it, we retrieved a set of **71 related tags**. We filtered the previous set using the Google API spell checker, and after applying the conditions of our experiment, we obtained a set of **14 tags**, as follows: {**food, fruit, dessert, pasta, cake, red, seafood, fish, meat, grape, spaghetti, vegetable, bread, green**}. Finally, we mapped those 14 tags onto the elements (concepts, properties, or instances) of ontologies, using Swoogle (to retrieve ontologies that contain the selected tag, ‘food’) and Jena (to define relations among tags in this group by calling the appropriate function). Due the lack of semantic contents on the web, not all groups are mapped onto ontologies; however this group on this well-known concept, “food”, mapped almost completely on the Food Ontology<sup>5</sup> (Please note that even the food group doesn’t completely map onto the Food ontology. For instance, there is no mapping of the tag ‘cake’). The list of the mapping results are too numerous to list in this paper, but here is a sample: (items in **Bold** are elements of the processed group ‘Food’, shown below as they map onto the Food ontology.)

```
{Food → Fruit}
{food:EdibleThing → Desert → food:SweetDessert}
{Desert → food:CheeseNutsDessert}
{food:EdibleThing → Pasta → food:PastaWithWhiteSauce}
{Pasta → food:PastaWithRedSauce}
{vin: property colour = Red}
{Meat → RedMeat → NonSpicyRedMeat}
{food:EdibleThing → seafood → food:Fish}
{RedMeat → NonSpicyRedMeat}
{seafood → food:Shellfish}
```

Hierarchical cluster analysis is based on a similarity matrix amongst tags. This matrix presents a table in which both rows and columns are the tags in same group and the cell entries are a measure of similarity (co-occurrence values) for any pair of tags. There are many measurement techniques for similarity between pairs of tags in the same group, such as (intervals, counts, binary), each with its own functions. For example, for the interval measurement, we can use Euclidean distance, Squared Euclidean distance, City block distance, Pearson correlation, and Cosine.

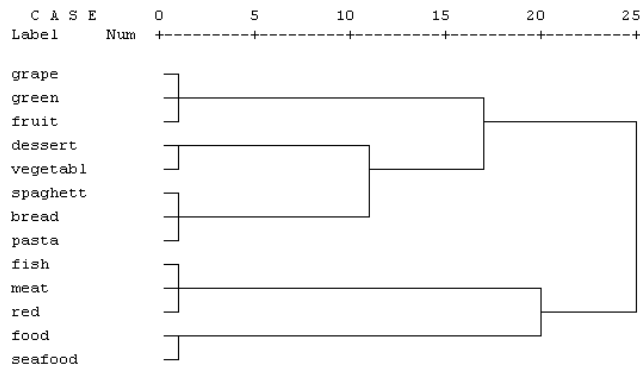
For within-cluster analysis and validation, the Count measurement using the Chi-square measure, the default for counting data, is most appropriate, as we compare the co-occurrence of the various tags within the cluster with the initial cloud tag that started the cluster. The Count Chi-square test in HCA compares observed counts of particular cases to expected counts (in our experiment, co-occurrence values between two tags). The result of the analysis can be depicted as *dendrograms*, also called hierarchical tree diagrams, which show the relative size of the proximity coefficients at which cases (tags) were combined (Figure 2). Each row represents a case (tag) on

---

<sup>4</sup> The result shown in this experiment is only a sample of the complete result, as there is no place to mention **8,710 tags**.

<sup>5</sup> <http://www.w3.org/2001/sw/WebOnt/guide-src/food>

the Y axis, while the X axis is a rescaled version of the proximity coefficients. Tags with higher similarity are close together, with a line linking them a short distance from the left of the dendrogram. When, the linking line is to the right of the dendrogram, the linkage occurs with a lower similarity coefficient, indicating that tags were agglomerated, even though much less alike (e.g. pasta and spaghetti are closely linked, but not as close as fruit and bread).



**Fig. 2.** Hierarchical cluster analysis for the food group

The previous dendrogram presents a tree of visual classification of similarity (based on co-occurrence values), which illustrates another valid hierarchy. This hierarchy represents the bottom-up ontologies from folksonomies. In Figure 2, “green”, “grape”, and “fruit” are clustered first; similar to “fish”, “meat” and “red”; and “pasta”, “bread” and “spaghetti”; and so on. Next, these generated clusters (those are close to each other) are grouped together. The final result is a hierarchy which can be used to determine semantic relations among tags in same group, especially when the group does not have a corresponding ontology from the Semantic Web. It is thus possible, although not the main objective of this paper, to compare Web 2.0 bottom-up generated ontologies with expert-made, top-down ontologies from the semantic web. (e.g., the ‘Food’ ontology presented at the beginning of this section, with the resulting ontology in Figure 2). For here it suffices to note that our clustering method can thus be justified and analyzed via a statistical method independent to our initial clustering choice. However, our method is (arguably) computationally simpler to statistical methods, and thus preferable, as it is expected to scale better.

#### 4 Discussions and Critical Analysis

We divided findings into three separate sections, reflecting our three processing phases, accordingly.

#### 4.1 The Major Findings of Filtering Tags

The procedure outlined here has the effect of *decreasing the cost of creating and authoring semantic content*. We have obtained our set of tags ‘for free’, as the Social Web offers a great resource of data at a low cost.

*Tags are context-specific*, which means different users use the same tags with different meanings for different resources (e.g., user *A* uses the tag "mouse" for a photo of a type of rodent, whilst user *B* uses it to describe a computer input device).

*Tags can be system-specific*; we obtained our dataset from Flickr by using its API. However, other social websites exist, and it is reasonable to expect that the mapping results are different, as tag clouds can be different. For instance, extracting semantics from wiki content [1] requires using templates of Wikipedia only.

#### 4.2 The Major Findings of Grouping Tags

*All tags are connected* to each other on the social web, thus, in principle, all tags belong to one set. However, this option is not practical for most application cases. However, the number of sets should be kept low. So, due to scalability issues, we initialized our set of groups with the elements in the tag cloud.

However, the number of sets should not be too low. Thus, as *large sets can lead to complex, time-consuming computations* and can contain too many false entries, we based the group inclusion condition on the *co-occurrence* value. This solution however could lead to deleting important tags from the group. Thus, there is a delicate balance between manageability and preciseness.

*Not all social web tag-groups are covered by ontologies*: some folksonomy tags have *no* matched ontologies. This finding is due to the lack of semantic content available on the web. A solution would be the building of ontologies from social data, e.g., via the users’ behaviour mapped on the time dimension, as proposed previously.

*Several groups share the same subset of tags*, which leads to a high degree of similarity amongst those groups. The similarity leads us to merge those groups into larger groups, in order to avoid redundancy. Such similarity between groups may come, for example, from describing the same content in different ways. E.g., in the tag cloud there are two tags “Trip” and “Travel” whose groups have 7 shared tags: *trip* {travel, **vacation**, **sky**, **sea**, **water**, **beach**, holiday, nature, sun, ocean, **landscape**, sunset, road, clouds, light, street, boat, **Europe**, people, tree, family, red, night, friends, fun, California, island, car, parispark, roadtrip}; *travel* {**vacation**, nature, **sky**, **Europe**, **beach**, trip, **landscape**, **sea**, **water**, italy, architecture, blue, people, summer, mountains, paris, clouds, ocean, mountain, asia}.

#### 4.3 The Major Findings of Mapping Tags

Since the experiment used tags from Flickr, we found that the most used ontology element was “instance” (not “concept” or “property”). The reason is that, when users tag a specific resource (here, pictures) they often tend to describe it with specific tags

rather than more abstract concepts (e.g., names of people, models of cars, names of places, etc.).

Merging knowledge from multiple ontologies could provide a much richer perspective on the underlying semantics of tagging systems (as supported by result in [20]). Mapping to multiple ontologies can also help decrease the ambiguity level. For example, "mouse" (as a type of rodent or a computer input device) and can be mapped on two different elements of two different ontologies, clarifying the meaning. This also solves one of the main obstacles noted by [14], which focuses on polysemy (same word with multiple meaning) in a semantic collaborative tagging system.

It is possible to have tags in a group which are not mapped onto same ontology. This highlights differences between *usage* (folksonomy) and *semantics* (ontology).

The mapping process, compared to the related work [17], has the *advantage of automating* the *retrieving* ontologies from the semantic web, as well as *enriching* folksonomy tags with semantic relations.

## 5 Scenarios of Using the Authored Hierarchical Structure

Different scenarios can be applied to augment authoring environments with the produced hierarchy. The most straightforward is to import data from the social web into authoring environments, after it has been structured by mapping to ontologies, as shown in this paper. The imported data can then be used with different adaptation strategies, in order to personalize the display of the information. For instance, a simple rule would be ‘show the current user all figures labelled with his current tag (from the user model (UM))’:

```
IF (concept.tag == UM.tag) then concept.show = TRUE
```

Another scenario is using a structured keyword pool for *suggestions* or *corrections*, so that authors will start using more structured data in their keyword choice instead of free-chosen tags. This can be achieved as follows: 1) **Auto-replace**: feature found also in text editors or word processors. It involves automatic replacement of a particular string with another one, usually one that is longer and harder to type, such as "NYC" – unstructured tag from social web - with "New York City" – structured tag taken from the city ontology. 2) **Auto complete**: feature provided by many text editors and/or word processors where the system suggests a word or phrase that the user wants to type. This would require a similar processing of the keywords annotating concepts in authoring environments to the tags in the social web.

## 6 Related Work

Merging ontologies and folksonomies can benefit from the strengths of each of them [9]. The ontology provides the benefit of enabling semantic search queries, as well as applying reasoning on structured data. On the other hand, the folksonomy provides the benefit of flexibly generating tags. This merging utilises direct relations between tags based on *explicit* co-occurrence values rather than using complex mathematical

relations between tags (as in [20]). Our work uses simpler co-occurrence computation methods that should allow for swifter processing, higher performance and scalability. Related work can be categorized in two main directions [18]: 1). *Enrich the social web with semantic content*, as it is possible to apply the semantic enrichment between tags in folksonomy by using ontologies from the semantic web [17]. An example of semantic enrichment of tags are the weblogs in [3]; as well as the method of extending the (object, tag) pair in the folksonomy towards the triple (object, ontology, tag) in [12]. [15] aimed at extracting *structured* information from folksonomies, where the co-occurrence value among tags in folksonomies is the main factor to determine to which group the tag belongs to. 2). *Add information on users' activities or other social aspects to the semantic web*. Examples include the method of applying ontologies in a folksonomy model [2,4], or defining multiple labels for ontology nodes [13,8].

Our research is useful for both research directions, as it could both be used to enrich social web sites (such as Flickr) with semantic relations, or be used to correct and extend given ontologies with missing terms from the social web. Unlike the above presented related work, in our work we use explicit co-occurrence values among tags, which impacts on the performance, as the reasoning process is simplified. Moreover, we believe that co-occurrence values between tags are more important in the tag grouping process, rather than in the process of mapping onto elements in ontologies, especially when adding semantics to social web applications.

## 7 Conclusion and Future Work

The social web is driven by the power of its users, who collaborate and produce a massive amount of data. However, this data is unstructured and has no semantics, which leads to problems of retrieval accuracy and processing effectiveness. The semantic web is driven by the power of machine computing, where the computers are able to understand the data. Nevertheless, semantic concepts lack simplicity and scalability, and they are difficult to extend over the already existing large scale of information available on the current web. Our work thus aims at merging the Social Web with the Semantic Web by enriching tags of folksonomies with semantic relations. For this purpose, we used multiple APIs from different sources.

Concluding, the added value of our work is that it combines, in a simple, reproducible way, the strengths of the Social Web's bottom-up approach (i.e., the flexibility and simplicity of folksonomies) with the strengths of the Semantic Web's top-down approach (i.e., the accuracy and the hierarchy of ontologies). In addition, the work avoids the weaknesses in both Social and Semantic web, as the generated structures are extracted from folksonomies, rather than by applying predefined ontologies. For future work, we plan to investigate how to concretely integrate our work with authoring systems such as MOT (My Online Teacher) [5].

**Acknowledgments.** The work accomplished in this paper is supported by the Socrates Minerva ALS Project (Adaptive Learning Spaces) - 229714-CP-1-2006-1-NL-MPP and the GRAPPLE FP7 STREP (215434).

## References

1. Auer, S. and Lehmann, J. Extracting Semantics from Wiki Content, The 4th European Semantic Web Conference 2007, Tyrol region of Innsbruck, Austria (2007)
2. Bateman, S., Brooks, C. and McCalla G. Collaborative tagging approaches for ontological metadata in adaptive e-learning systems. The 4th International Workshop on Applications of Semantic Web Technologies for E-Learning (2006)
3. Cayzer, S.: What next for semantic blogging? In Proceedings of the SEMANTICS 2006 conference, pp. 71–81, Vienna, Austria (2006)
4. Christiaens, S Metadata Mechanisms From Ontology to Folksonomy and Back, On the Move to Meaningful Internet Systems 2006 OTM 2006 Workshops, pp. 199-207, Montpellier, France (2006)
5. Cristea, A. and De Mooij, A. Adaptive Course Authoring: My Online Teacher, ICT'03, Papeete, French Polynesia, IEEE, IEE, ISBN: 0-7803-7662-5, pp. 1762-1769 (2003)
6. Cristea, A. I. What can the Semantic Web do for Adaptive Educational Hypermedia? Educational Technology & Society, 7 (4), 40-58. (2004).
7. Cristea, A., Smits, D., and De Bra, P.: Towards a generic adaptive hypermedia platform: a conversion case study. Journal of Digital Information (JoDI) Special Issue on Personalisation of Computing & Services, Vol 8, No 3. (2007)
8. Gruber, T. Folksonomy of Ontology A Mash-up of Apples and Oranges. First on-Line conference on Metadata and Semantics Research (MTSR'05) (2005)
9. Herzog, C., Luger, M. and Herzog, M. Combining Social and Semantic Metadata for Search in a Document Repository, 4<sup>th</sup> ESWC 2007, Innsbruck, Austria (2007)
10. J. B. MacQueen (1967): Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297
11. Jena Semantic Web Framework, <http://jena.sourceforge.net/> (accessed on 14/07/2007)
12. Lawrence, K. and Schraefel, M. Freedom and restraint: Tags, vocabularies and ontologies. 2nd IEEE Int. Conference on Information & Communication Technologies, Syria (2006)
13. Maedche, A. Emergent semantics for ontologies – support by an explicit lexical layer and ontology learning. IEEE Intelligent Systems - Trends & Controversies, pp. 78–86 (2002)
14. Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M. and Minutoli, S. SemKey. A Semantic Collaborative Tagging System, 16<sup>th</sup> Int. WWW Conf., Banff, Canada (2007)
15. Mika, P. Ontologies are us: A unified model of social networks and semantics. The International Semantic Web Conference 2005, pages 522–536, Galway, Ireland (2005)
16. Monteroa, Y. and Solana, V. Improving Tag-Clouds as Visual Information Retrieval Interfaces, International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006. Mérida, Spain. October 25-28, 2006.
17. Sabou, M., Angeletou, S. and Specia, L. and Motta E. Bridging the Gap Between Folksonomies and the Semantic Web An Experience Report, 4th European Semantic Web Conference 2007, Tyrol region of Innsbruck, Austria (2007)
18. Schaffert, S. Semantic social software: Semantically enabled social software or socially enabled semantic web? Semantics 2006 conference, pp. 99–112, Vienna, Austria (2006)
19. Shelly, P. Television Disrupted: The Transition from Network to Networked TV. Focal Press. ISBN 0240808649. pp. 200 (2006)
20. Specia, L. and Motta, E. Integrating Folksonomies with the Semantic Web, 4th European Semantic Web Conference 2007, Tyrol region of Innsbruck, Austria (2007)
21. Swoogle Semantic Web Search Engine, <http://swoogle.umbc.edu/> (accessed on 15/07/2007)
22. The Planetwork Journal, <http://journal.planetwork.net/article.php?lab=reed0704> (accessed on 23/06/2007)
23. The World Wide Web Consortium, <http://www.w3.org/2001/sw/> (accessed on 06/07/2007)