Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report

Sofia Angeletou[1], Marta Sabou[1], Lucia Specia[2], and Enrico Motta[1]

[1] Knowledge Media Institute (KMi)
The Open University, Milton Keynes, United Kingdom
{S.Angeletou, R.M.Sabou, E.Motta}@open.ac.uk
[2] Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, São Carlos, Brazil
lspecia@icmc.usp.br

Abstract. While folksonomies allow tagging of similar resources with a variety of tags, their content retrieval mechanisms are severely hampered by being agnostic to the relations that exist between these tags. To overcome this limitation, several methods have been proposed to find groups of *implicitly* inter-related tags. We believe that content retrieval can be further improved by making the relations between tags *explicit*. In this paper we propose the semantic enrichment of folksonomy tags with explicit relations by *harvesting the Semantic Web*, i.e., dynamically selecting and combining relevant bits of knowledge from online ontologies. Our experimental results show that, while semantic enrichment needs to be aware of the particular characteristics of folksonomies and the Semantic Web, it is beneficial for both.

1 Introduction

Folksonomies [13] are typical Web2.0 systems that allow users to upload, tag and share content such as pictures, bookmarks etc. One of their distinctive features is that they are open, uncontrolled systems where users can annotate resources with different tags depending on their social or cultural backgrounds, expertise and perception of the world [2, 3, 9, 14]. For example, a zoologist can tag a photograph of a lion with {`felidae, pantherinae, mammal`}, while a non-zoology expert can use {`lion, king, animal, jungle`} for the same purpose. This freedom of tagging largely contributed to the success of folksonomies: users need neither to have prior knowledge or specific skills to use the system [5, 15], nor need to rely on a priori agreed structure or shared vocabulary.

Unfortunately, the simplistic tag-based search used by folksonomies is agnostic to the way tags relate to each other although they annotate the same or similar resources. For example, a search for {`mammal`} ignores all resources that have not been tagged with this specific word, even if they are tagged with related concepts such as {`lion, cow, cat`}. As a result, content retrieval activities such as searching, subscription and exploration are limited [2], they provide low-recall and hardly lend themselves to query-refinement [11]. Therefore, to obtain satisfactory results, a searcher needs to build multiple complex queries to cover all the possible tags that could have been used by taggers [3, 9, 14]. As searchers rely on their own view about what inter-related tags best describe the resource they are looking for, it follows that content retrieval could be enhanced if folksonomies were aware of the relations between their tags.

Following this intuition, a variety of approaches have been proposed to identify inter-related tags. The existing work considers tag co-occurrence for the organisation of related tags into clusters. For example, [11] uses a subsumption-based model, derived from the co-occurrence of tags, to find groups or related tags. [2] organises the tag space as an undirected graph, representing co-occurring tags as vertices, weighting the edges between them according to their co-occurrence frequency, and applying a spectral clustering algorithm to refine the resulting groups. [15] uses a probabilistic model to generate groups of semantically related tags based on the co-occurrence of tags, resources, and users. These are represented as a multi-dimensional vector, where each dimension refers to a category of knowledge. Both the number of dimensions and the relation values of entities to each dimension are determined using log-likelihood estimates. [7] uses co-occurrence information to build graphs relating tags with users and tags with resources, and applies techniques of network analysis to discover sets of clusters of semantically related tags. [12] groups tags according to their co-occurrence using a clustering algorithm similar to clustering by committee [8]. Finally, most of the folksonomies provide funtionalities to derive "clusters" and "related tags", which apparently also rely on co-occurrence information and clustering techniques.

All the approaches, except from [12], focus on finding groups of related tags rather than identifying the semantics of those relations. In this work the authors envisaged tag space enrichment with semantic relations by exploring online ontologies. Their preliminary experiments on Flickr and Del.icio.us data confirmed that this is a promising strategy. Indeed, the recent growth of the Semantic Web has resulted in an increased amount of online available semantic data and has led to the first search engine to exploit this data, Swoogle [6]. These facts made it possible to build applications that *harvest the Semantic Web* (i.e., dynamically select, combine and exploit online knowledge) to successfully solve a variety of tasks, such as query disambiguation [4] and ontology matching [10].

Applying this novel paradigm to folksonomies would make them explicitly aware of the inherent semantic relations between their tags. For example, subsumption relations such as the ones depicted in Fig. 1 could be derived between the tags of the cluster {`lion, animal, mammal, feline, tiger`} by combining information from different online ontologies. The knowledge that *Lions* and *Tigers* are *kind of Mammals* would expand the potential of folksonomies. Users could make generic queries such as *"Return all mammals"* and obtain all the resources tagged with `lion` or `tiger` even if they are not explicitly tagged with `mammal` .

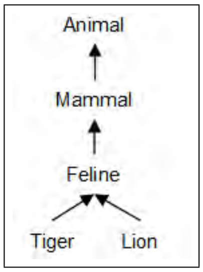

Fig. 1: Related Tags.

While previous work has experimentally shown that harvesting online knowledge yields good results when applied to ontologies [10], the folksonomy tag enrichment algorithm proposed in [12] was not fully automated. Therefore, an important research question is: *Can we enrich folksonomies by automatically harvesting the Semantic Web?* In particular, we are interested in finding out: *What are the major characteristics of the Semantic Web and folksonomies that need to be taken into account to perform such enrichment?* And if this enrichment is possible: *What are its benefits?* To answer these questions, we propose a method to enrich the tag space of folksonomies which assumes the existence of previously defined groups of potentially related tags (these can be obtained by any of the above mentioned techniques) and which is entirely focused on the exploitation of the Semantic Web (Section 2). This approach is automated by using the algorithm described in [10]. We present and discuss our experimental results which give an insight in the major characteristics of the Semantic Web and folksonomies that need to be considered when performing such enrichment (Section 3). We conclude and point out future work in Section 4.

2 Semantic Enrichment of Folksonomy Tag Space

In this section we describe our approach for semantically enriching the folksonomic tag spaces. Our method is based on [12], which describes a hybrid approach that combines harvesting the Semantic Web with using other Web resources such as Wikipedia and Google. As the goal of our work is to understand the potential and limitations of the Semantic Web when used to semantically enrich folksonomies, we have modified their algorithm so that it only relies on online ontologies. Our algorithm, presented next, takes as input a cluster of implicitly related tags and returns 1) a knowledge structure obtained by making explicit the semantic relations among them and 2) a set of tags which could not be semantically related to any other tag in their cluster or were not covered by the Semantic Web.

2.1 Semantic Enrichment Method

The semantic enrichment of each cluster is depicted in Fig. 2 and consists of two phases: Phase 1, concept definition for each tag (i.e., linking tags to ontology concepts) and Phase 2, relation discovery between all the possible pairs of tags.

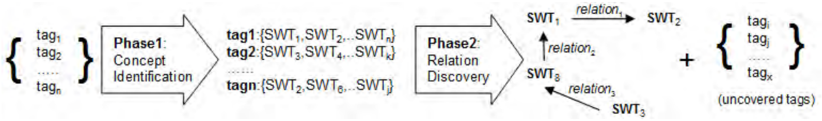
{ tag_1 tag_2 ...... tag_n }
Phase1: Concept Identification
tag1:{SWT_1,SWT_2,..SWT_n}
tag2:{SWT_3,SWT_4,..SWT_k}
....
tagn:{SWT_2,SWT_6,..SWT_j}
Phase2: Relation Discovery
SWT_1
$relation_1$
SWT_2
$relation_2$
SWT_8
$relation_3$
SWT_3
+
{ tag_i tag_j ...... tag_x }
(uncovered tags)

Fig. 2. Semantic Enrichment Method

Phase 1. Concept Identification: The first step *explicitly defines the meaning* of each tag by extracting all Semantic Web Terms (SWT) whose label or localname matches the tag. The matching between the tag and the SWT can be achieved using anchoring techniques ranging from strict to flexible string matching as described in [10].

Using the Semantic Web for extracting concepts is proposed in the work of [4] as a first step to query disambiguation. The authors search for candidate senses in online ontologies and then perform disambiguation based on the semantic similarity of the retrieved senses (e.g., `bass` can refer to either a fish or musical notes depending on the context in which it is used). While we use the same technique for SWT identification we do not explicitly disambiguate between them. In our case, disambiguation is a side effect of relation discovery (Phase 2).

The disambiguation of the tag sense (i.e., finding the right concept for a tag given its context) is approached differently in [12]. The authors rely on the heuristic that if pairs of tags from a cluster appear in the same ontology, then this leads to an implicit disambiguation (i.e., searching for `apple` and `fruit` leads to ontologies about fruits, while when searching for `apple` and `computer` they identify ontologies about computers). While this intuition holds in the case of domain-specific ontologies, it is problematic when the tags appear in broad, cross-domain ontologies such as WordNet[3] or TAP[4]. Also, by considering only ontologies that contain both tags, this approach potentially misses important information that might be declared in ontologies defining only one of the tags. This information can prove to be useful when combined with information from other ontologies. For example, an ontology containing *Apple* and *Mac*, can be combined with information from another ontology containing information about *Mac* and *Computer*. For these reasons, we retrieve all the potential SWTs for each tag and discover relations between them in Phase 2.

Phase 2. Relation Discovery: This step identifies *explicit semantic relations* among all the pairs of SWTs (`T1` and `T2`) discovered in the previous phase:

- **Subsumption Relations:** when one of the two SWTs is a subclass of the other, `T1 subClassOf T2`. This relation can be either declared in an ontology or derived by different levels of inference (no inference, basic transitivity, Description Logics reasoning). An example of inferred relation is: if `T1 subClassOf T2` and `T2 subClassOf T3` then `T1 subClassOf T3`.
- **Disjointness Relations:** when `T1` and `T2` are disjoint, `T1 disjointWith T2`. Again this relation can be declared or inferred. We use the algorithm described in Section 2.2 to discover disjointness and subsumption relations.
- **Generic Relations:** when a generic relation holds between the two SWTs, e.g., `Property1 hasDomain T1` and `Property1 hasRange T2` or inversely.
- **Sibling Relations:** when the two SWTs share a common ancestor, which can be either a direct or an indirect parent. Note that our definition covers the three sibling definitions described in [12].

[3] `http://wordnet.princeton.edu/`

[4] `http://tap.stanford.edu/data/`

- **Instance Of Relations:** such as `T1 instanceOf T2` or inversely. Unlike the previous relations, this relation is not considered by [12].

The identification of these relations can be made in two ways. First, a relation between SWT's might be declared **within a single ontology**. Second, if no single ontology mentions both SWT's, then a **cross-ontology relation discovery** can be performed by combining knowledge from several ontologies.

Cross-ontology relation discovery has been successfully implemented in the case of ontology matching [10]. An important issue to be considered is how to deal with potential contradictory relations, e.g., `T1 subClassOf T2` and `T1 disjointWith T2`. This remains a future work topic.

The semantically connected tags form the knowledge structures mentioned in the beginning of Section 2.1 and the tags not linked to SWTs or not related to other tags compose the set of uncovered tags. The study of the latter is expected to provide hints about how to evolve the Semantic Web, as described in Section 3. Next we describe the current implementation of our approach which identifies only subsumption and disjointness relations found in single ontologies.

2.2 Subsumption/Disjointness Discovery Based on One Ontology

The discovery of subsumption and disjointness relations between two terms within one ontology has been described and implemented on Swoogle'05 in [10]. Given two candidate concept names (`A` and `B`) as input, corresponding concepts are selected in online ontologies (`A'` and `B'`) by using strict string based anchoring. The possible semantic relations occurring between concepts in an ontology are shown using description logic syntax, e.g., $A' \sqsubseteq B'$ means that `A'` is a subconcept of `B'`. The returned relations are expressed with arrows, e.g., $A \xrightarrow{\sqsubseteq} B$. The steps of this strategy in detail are:

1. Select ontologies containing concepts `A'` and `B'` corresponding to `A` and `B`;
2. If no such ontology is found, then `A` and `B` do not relate;
3. If there are returned ontologies, for each:
   - if $A' \equiv B'$ then derive $A \xrightarrow{\equiv} B$;
   - if $A' \sqsubseteq B'$ then derive $A \xrightarrow{\sqsubseteq} B$;
   - if $A' \sqsupseteq B'$ then derive $A \xrightarrow{\sqsupseteq} B$;
   - if $A' \perp B'$ then derive $A \xrightarrow{\perp} B$;

In a simple implementation we can rely on *direct* and *declared* relations between `A'` and `B'` in the selected ontology. But for better results *indirect* and *inferred* relations should also be exploited. For our experiments, we used an implementation relying on basic transitivity reasoning (i.e., taking into account all parents of `A'` and `B'`) and stopping as soon as a relation is found.

3 Experimental Results

The goal of our experiments is twofold. On the one hand, we wish to reveal how much of the semantic enrichment of folksonomy tags can already be automated by using the software developed in [10] which partially implements the current version of our envisioned algorithm (the part described in Section 2.2). On the other hand, we wish to understand any problematic issues so that they can be addressed in the design of the final, complete algorithm. At a higher level, these issues give an insight in how folksonomies and the Semantic Web relate. In a first experiment (Section 3.1) we applied the software developed in [10] to Flickr and Del.icio.us clusters generated by [12]. This experiment lead to valuable insights into issues that hamper the enrichment and prompted us to repeat the experiments with another set of clusters selected directly from Flickr. We discuss the second set of experiments in Section 3.2.

3.1 Experiment 1

The number of results obtained by running our algorithm with the clusters generated in [12] were surprisingly low. Two major reasons explain this. First, our implementation only searches for `subClassOf` and `disjointWith` relations. Unfortunately, the majority of tags in the clusters we work with are not related by these relations but by generic relations. The second major reason is that few of the tags in the analysed clusters could be identified in ontologies in the Semantic Web. Taking a closer look to the tags that were not found we individuated the following cases:

Novel terminology. Folksonomies are social artifacts, built by large masses of people and dynamically change to reflect the latest terminology in several domains. As such, they greatly differ from ontologies which are generaly developed by small groups of people and evolve much slower. Therefore, it is not surprising that many of the tags used in folksonomies, e.g., {`ajax, css`}, have not yet been integrated into ontologies. Identifying frequent folksonomy tags that are missing from ontologies has a great potential for the Semantic Web as it can provide the first step towards enriching existing ontologies with these novel terms.

Instances. When people tag resources, especially pictures, they more often tend to tag them with specific names rather than more abstract concepts. In particular, we frequently found names of people {`monica, luke, stephanie`}, names of places {`japan, california, italy`} and particular dates {`august2005, aug292005`}. Unfortunately, the current version of our system only works at terminological level (it deals only with concepts and not with ontology instances), so we did not identify any of these instances in the experiments. Apart from that limitation it is unlikely that instances related to people and specific dates can be reliably identified in ontologies anyway.

Photographic jargon. Given the scope of Flickr as a photo annotation and sharing site, many of the tags that are used reflect terms used in photography, such as {nikon, canon, d50, cameraphone, closeup, macro}. Unfortunately, this domain is weakly covered in the Semantic Web.

Multilingual tags. Both Flickr and Del.icio.us (but especially Flickr) contain tags from a variety of languages and not only English. These tags are usually hard to find on the Semantic Web because the language coverage of the existing ontologies is rather low. Indeed, statistics[5] performed on a large collection of online ontologies (1177) in the context of the OntoSelect library indicate that 63% of these ontolgies contain English labels, while a much smaller percentage contains labels in other languages (German 13.25%, French 6.02%, Portuguese 3.61%, Spanish 3.01%).

Concatenated tags such as {christmasornament, xmlhttprequest, librariesandlibrarians} appear frequently but obviously it is hard to identify concepts with the same spelling.

Given the very low coverage of the Semantic Web for the above mentioned categories of tags, we decided to repeat the experiments for clusters of tags that are well-covered in the Semantic Web. Also, since at this stage our system only discovers subsumption and disjoint relations, we decided that the experiments should consider significantly larger clusters than those provided by [12].

3.2 Experiment 2

In the second set of experiments we relied on the lessons learnt from the first experiment to identify clusters of tags that would be more appropriate for our goal. To address the first conclusion (i.e., that clusters should be potentially well covered in the Semantic Web), we relied on the results of previous work in the context of ontology matching [10]. Follow up experiments revealed that domains related to food and animal species are well covered in the Semantic Web. Therefore, we selected a couple of tags from these domains, based on the concepts for which the most mappings were found during the matching experiments. We selected the tags: mushroom, fruit, beverage and mammal.

The next step was to identify clusters of tags related to each of these tags. As we said, we were looking for large clusters that would be more likely to accommodate subsumption relations and not just generic relations between tags. We chose the cluster generator provided by Flickr[6], since it returns much larger clusters of related tags than Del.icio.us and Technorati (moreover, since Del.icio.us and Technorati are mostly oriented towards news, business and web technologies, the clusters they provide for our tags in the food and animal domains are quite small).

The same algorithm as in Experiment 1 was then applied to these clusters. As expected, we found several relations among tags as depicted in the figures below (directed arrows represent subClassOf relations, dotted lines depict disjointWith relations). 23% of the investigated tags was discovered in

[5] http://olp.dfki.de/OntoSelect/w/index.php?mode=stats

[6] http://www.flickr.com/services/api/flickr.tags.getRelated.html

ontologies. Besides the tags between which we found relations, there were also sets of tags that could not be linked with any other tag in their cluster. We analyze these tag sets and describe possible causes that led to this failure.

The case of Mushroom. The semantic relations identified among the 21% of the tags related to `mushroom` by using online ontologies are depicted in Fig. 3. *Mushroom* was identified as a kind of *Fungi* and a kind of *Plant*. Also, we have learnt that it is disjunct with *Pizza*, *Pepper*, *Cheese* and *Tomato* and so are these with each other. *Mushroom* also co-occurs with *Soup*, *Rice* and *Onion*. As expected, there is no subsumption relation between these concepts and *Mushroom*. However, they are all subclasses of *Food*, as are *Tomato* and *Cheese* as well.

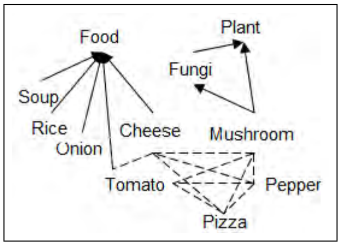

Fig. 3: Mushroom in the Semantic Web.

Type	**Tags**
Not covered by the SW	{`amanitamuscaria, toadstool, flyagaric`}
Generic relation (location)	{`nature, forest, garden, grass, moss`}
Generic relation (seasons)	{`autumn, fall, herfst`}
Generic relation (usage)	{`cooking, dinner, pasta, lunch`}
Colors	{`green, white, yellow`}
Photo jargon	{`macro, nikon, closeup`}

Table 1. `mushroom` related tags that could not be connected semantically

Table 1 shows some of the tags in the cluster of `mushroom` that could not be related semantically to any other tag, grouped according to the reason why they could not be linked. These are:

Tags that are not covered by the Semantic Web. These tags refer to kinds of mushrooms or scientific names that are not described in the Semantic Web. Generally, our experience is that currently very few online ontologies cover scientific labels.

Tags generically related to mushroom. The next three sets of tags are related to mushroom through other generic relations than subsumption or disjunction and describe locations, time and potential ways to use mushrooms.

Tags about colors. This set of tags is not surprising reflecting the fact that we retrieved the tag clusters from a photo-sharing system where users add color names to describe the image content of their photos. Note, however, that these colors might be meant to describe the rest of the tags associated to a resource, e.g., {`green pepper, white mushroom, yellow cheese`}. Unfortunately, because the creation of compound tags such as these is not well handled by folksonomies, users have to add each tag separately, thus loosing the relationship between them.

Photo jargon. The remaining group of tags are Flickr related tags, as we discussed in Experiment 1, and are not covered in the Semantic Web. Also, given the fact that they describe the photographs rather than their content, even if they were covered it is quite unlikely that they could be related to mushrooms or any other tag describing image content.

The case of Fruit We obtained interesting results for the cluster of `fruit` (Fig. 4) and the highest percentage of related tags, 29%. As fruits are well-covered by the Semantic Web, the generated semantic structure contains much more information than a single relation between the tags of the cluster. For example the multiple relations that exist between *Fruit* and *Vegetable*, and how this affects their common subclass, *Tomato*. In a biological context, a tomato is indeed the fruit of a tomato plant, however, normally one would classify tomatoes as types of vegetables. While such different views can co-exist, the fact that *Fruit* and *Vegetable* are disjoint makes this bit of knowledge inconsistent. Therefore, once such structures are derived from multiple ontologies, their consistency should be verified.

Also, according to online ontologies, *Fruit* is disjoint with *Dessert*. The validity of this statement depends on the point of view we adopt: some would argue that fruits are desserts, while others might consider desserts generally inappropriate catogorisation for fruits. Finally *Strawberry* and *Watermelon* were also found as subclasses of *Fruit*, but declaring them as subclasses of *Berry* and *Melon*, respectively, automatically infers they are also subclasses of *Fruit*.

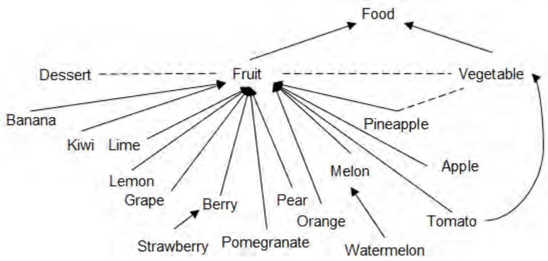

Fig. 4. *Fruit* in the Semantic Web

The tags that could not be connected to *Fruit* fall into five categories (see Table 2), two of which are related to colors and photo jargons, as discussed before. A new set of interesting tags describes attributes generally related to fruits: {`juicy, yummy, delicious, fresh, sweet`}. Unfortunately, most concepts in ontologies model nouns. Attributes are often modeled as properties (geneneric relations). Finally, the other two sets of interesting tags refer to fruit cultivation methods and possibly best seasons for consumption of specific fruits, which again share generic relations with fruits, currently not in the scope of our software.

Type	Tags
Attributes	{juicy, yummy, delicious, fresh, sweet}
Generic relation (cultivation)	{tree, nature, plant, seeds, leaves}
Generic relation (seasons)	{summer, autumn, fall, red, pink}
Colors	{brown, green, white, red, pink}
Photo jargon	{closeup, macro, canon}

Table 2. `fruit` related tags that could not be connected semantically

The case of Beverage. Beverage is the least covered tag with 18% of its related tags found to be connected in the Semantic Web. The knowledge structure that emerged from the semantic enrichment of the cluster related to `beverage` is shown in Fig. 5. As in the case of `fruit`, the cluster for `beverage` contains many concepts that were more specific than *Beverage*. Accordingly, these were identified to be in a subsumption relation with *Beverage* by our system. The two most interesting cases are of *White* being a subclass of *Beer* (white beer as a type of beer) and *Water* not being connected to *Liquid*. *Water*, though, was found to be related with *Fluid* which doesn't belong to the related tags of `beverage`. The tags that could not be related fall under the types of categories that we have already discussed in the previous cases and are presented in Table 3.

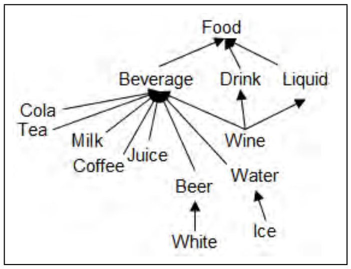

Fig. 5: Beverage in the Semantic Web.

Type	Tags
Not covered by the SW	{energy_drink, soda, martini, latte}
Generic relation (container)	{straw, mug, can, bottle, glass, cup}
Generic relation (event/place)	{breakfast, restaurant, party, starbucks}
Generic relation(ingredient)	{lemon, fruit, cream, orange}
Attributes	{hot, delicious, refreshing}
Colors	{brown, black, orange, green, red, pink}
Photo jargon	{closeup, macro, canon}

Table 3. `beverage` related tags that could not be connected semantically

Some types of beverages are not covered by the Semantic Web. It is interesting to note here that `latte` is not just an English word for a type of coffee, but also Italian for milk. The fact that it is not covered can be a side-effect of the low level of multilinguality in online ontologies, as we discussed in Experiment 1. Additionally, certain tags could be related to *Beverage* by generic relations, but these are not discovered by the current version of our system. These tags express types of containers, events and locations where beverages are served, as well as the ingredients of drinks. It is worth noticing that `orange` could belong both to the categories representing colors and ingredients. The final set of tags that could not be related refer to attributes which, as discussed before, have generally a weak coverage on the Semantic Web.

The case of Mammal The last tag that was investigated is `mammal`. Relations for the 25% of its tags were found in the Semantic Web. Fig. 6 shows the structure derived from its cluster. It is interesting to observe that the subclasses of *Mammal* do not represent the same level of abstraction. We note many common names of animals like *Horse*, *Monkey*, *Rabbit*, but also two subclasses of higher abstraction, *Rodent* and *Feline*. This is another evidence that users annotate their content with a variable level of generality: although *Squirrel* and *Rabbit* appear in the graph as subclasses of *Mammal*, their superclass, *Rodent*, appears as well. This confirms the hypothesis put forward by [3] according to which different users will settle at different "basic levels" depending on their level of expertise.

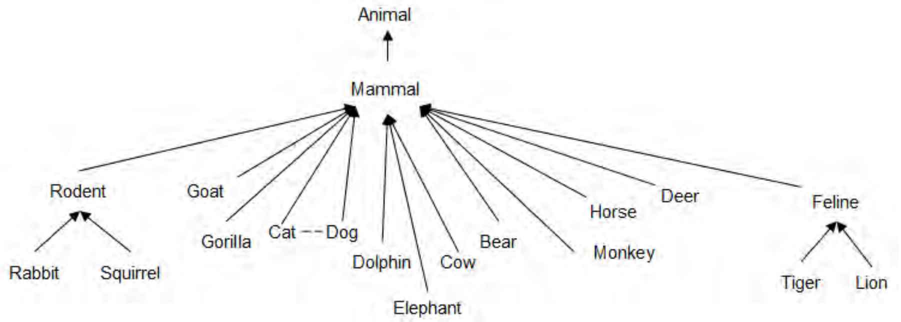

Fig. 6. *Mammal* in the Semantic Web

The tags that could not be related are displayed in Table 4. Most of these categories have been discussed previously, along with a set of tags that could have been related by generic relations indicating the location or habitat of mammals. Two tags were found to describe the state of the mammal when it was shot {`eating, sleeping`}. Finally, an interesting set of tags depicts body parts which should be related to mammals through a part-of relation.

Finally, it is worth pointing out that in all of the above here cases we identified certain tags, which were also found in Experiment 1, describing the places shown in the images, such as `barcelona`, `japan`, or the interests of the users, such as `ilovenature`, `stilllife` (we found 84.077 pictures annotated with `ilovenature` and 39.320 with `stilllife`).

Type	**Tags**
Not covered by the SW	{giraffe, seal, zebra}
Generic relation (location)	{zoo, nature, water, ocean, wild, farm, outdoors}
Generic relation (action)	{eating, sleeping}
Part-of	{fur, whiskers, eyes, face, nose}
Attributes	{cute, pet, funny, bunny}
Photo jargon	{portrait, closeup, macro, canon}

Table 4. `mammal` related tags that could not be connected semantically

4 Conclusions and Future Work

As an answer to our main research question, which is to explore whether folksonomies can be automatically enriched by harvesting the Semantic Web, based on the results of the preliminary experiments presented above, we can already conclude that it is indeed possible to automate the semantic enrichment of folksonomy tag spaces by harvesting online ontologies. By using these ontologies, we were able to automatically obtain semantic relations between the tags of several clusters of related tags. An immediate goal of our future work is to apply our approach on folksonomies and evaluate it in terms of Information Retrieval performance values (recall and precision). As an answer to our second research question, which is to identify the inherent characteristics of folksonomies and the Semantic Web and how they should be approached, the experiments also yielded relevant observations about these characteristics which have an impact on folksonomy enrichment process:

1. Folksonomy Characteristics. Our experiments show that many folksonomy tags fall in specific categories that require special attention. First, by being dynamically updated by large masses of people, folksonomies reflect the newest terminology within several domains (**novel terminology**). Second, many folksonomy tags refer to specific **instances** (names of people, places, dates). Third, folksonomies contain tags representing words in a variety of languages (**multilinguality**). Fourth, some of the tags that are frequently used depend on the purpose of the folksonomy and usually describe the resource itself rather than its content (**folksonomy jargon**). Fifth, folksonomy tags often describe **attributes** of the content, for example, colors (especially in Flickr). Sixth, there are many **concatenated tags** which describe a large number of photographs and need to be exploited. Finally, a **broad range of semantic relations** can exist between tags, including subsumption, disjointness, meronymy and many generic relations (e.g., location).

2. Semantic Web Characteristics. The most important observation regarding the Semantic Web is that even if it is growing fast it still suffers from *knowledge sparseness* (i.e., it presents good coverage for certain topics, but very low coverage for others). Due to this limitation, we needed to restrict our experiments to domains that are well-covered (related to animals and food). Also, some of the categories of tags that appear frequently in folksonomies are difficult to find in online ontologies. First, **novel terminology** that emerges from folksonomies is often missing from ontologies. Second, the majority of **specific instances** that appear in folksonomies cannot be found (e.g., `aug2004`) or are difficult to reliably map to ontology instances (e.g., `monica`). Place names are an exception to this. Third, few of the online ontologies contain **multilingual labels**, therefore tags in languages other than English are unlikely to be found in ontologies. Fourth, **specific jargons**, such as those related to photography are weakly covered as well. Fifth, online ontologies are rather **poor in describing generic attributes** such as color. One of the reason for this is that attributes are most often modeled as part of properties rather than concepts.

We are confident, however, that surpassing some of the current limitations is a matter of time as many of them will be solved as more ontologies will appear online. For example, the AGROVOC[7] ontology contains roughly 16000 concepts and their labels in 12 different languages. Making this single ontology available online will positively impact on the issue of anchoring multilingual tags. Nevertheless the appearance of more online ontologies can also be seen as a potential risk for this work as different ontologies reflect different views which often lead to contradictory bits of knowledge. Combining these bits may result in inconsistencies in the derived semantic structures. However, existing reasoning techniques can be used to filter out and eliminate possible inconsistencies.

Being aware of these characteristics help us to identify the **current limitations of our software**. Our software only implements a subset of the functionality envisioned for the enrichment algorithm. First, it is currently implemented on Swoogle'05 which lags behind in ontological content. Our final algorithm will be built on top of up-to-date semantic search engines [1]. Second, the anchoring mechanism is based on strict string matching and therefore needs to be extended to more flexible anchoring. Third, from the broad range of semantic relations that can exist between tags, our software only identifies subsumption and disjointness. Obviously, extensions are needed that can discover the other types of relations as well. Finally, note that we have only experimented with finding relations within a single ontology and excluded cases when knowledge can be derived by combining facts from multiple ontologies. Another important future work will be to implement this cross-ontology relation derivation.

The experimental work reported in this paper indicates that the proposed enrichment process has the potential to benefit both folksonomies and the Semantic Web, thus answering our third research question. On the one hand, even using a software with limited functionality we were able to derive *explicit* semantic relations between tags, thus going beyond existing methods that identify *implicitly* inter-related tags. We believe this could considerably enhance content retrieval in folksonomies. On the other hand, the differences between folksonomies and ontologies (such as novel terminologies emerging in several languages) can be used to evolve the Semantic Web. This valuable knowledge available in folksonomies could allow keeping online ontologies up to date, extending them with

[7] `http://www.fao.org/agrovoc`

multi-lingual information and evolving them towards being truly *shared* conceptualisations of a much broader range of domains.

Acknowledgements

We thank M. d'Aquin for his useful comments on earlier versions of this paper. This work was funded by the Open Knowledge and NeOn projects sponsored under EC grant numbers IST-FF6-027253 and IST-FF6-027595.

References

1. M. dAquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. Watson: A gateway for the semantic web. In *Poster Session at ESWC'07*, 2007.
2. G.Begelman, P. Keller, and F.Smadja. Automated Tag Clustering: Improving search and exploration in the tag space. In *Proc. of the Collaborative Web Tagging Workshop at WWW'06*, 2006.
3. S. Golder and B.A. Huberman. The Structure of Collaborative Tagging Systems. HPL Technical Report, 2005.
4. J. Gracia, R. Trillo, M. Espinoza, and E. Mena. Querying the Web: A Multiontology Disambiguation Method. In *Proc. of ICWE'06*, 2006.
5. A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information Retrieval in Folksonomies: Search and Ranking. In *Proc. of ESWC'06*, 2006.
6. L.Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V.C. Doshi, and J. Sachs. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proc. of the 13th ACM Conf. on Information and Knowledge Management*, 2004.
7. P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. of ISWC'05*, 2005.
8. P.A. Pantel. *Clustering by Committee*. PhD thesis, 2003.
9. E. Peterson. Beneath the Metadata: Some Philosophical Problems with Folksonomy. *D-Lib Magazine*, 12(11), November 2006.
10. M. Sabou, M. d'Aquin, and E. Motta. Using the Semantic Web as Background Knowledge for Ontology Mapping. In *Proc. of the Int. Workshop on Ontology Matching (OM-2006)*, 2006.
11. P. Schmitz. Inducing Ontology from Flickr Tags. In *Proc. of the Collaborative Web Tagging Workshop at WWW'06*, 2006.
12. L. Specia and E. Motta. Integrating Folksonomies with the Semantic Web. In *Proc. of ESWC'07*, 2007.
13. Thomas Vander Wal. Folksonomy coinage and definition. 2007.
14. H. Wu, M. Zubair, and K. Maly. Harvesting Social Knowledge from Folksonomies. In *In Proc. of HYPERTEXT '06*, 2006.
15. X. Wu, L. Zhang, and Y. Yu. Exploring Social Annotations for the Semantic Web. In *Proc. of WWW'06*, 2006.