Semantically Enriching Folksonomies with FLOR

Sofia Angeletou[1], Marta Sabou[1], and Enrico Motta[1]

Knowledge Media Institute (KMi)
The Open University, Milton Keynes, United Kingdom
{S.Angeletou, R.M.Sabou, E.Motta}@open.ac.uk

Abstract. While the increasing popularity of folksonomies has lead to a vast quantity of tagged data, resource retrieval in folksonomies is limited by being agnostic to the meaning (i.e., semantics) of tags. Our goal is to automatically enrich folksonomy tags (and implicitly the related resources) with formal semantics by associating them to relevant concepts defined in online ontologies. We introduce FLOR, a method that performs automatic folksonomy enrichment by combining knowledge from WordNet and online available ontologies. Experimentally testing FLOR, we found that it correctly enriched 72% of 250 Flickr photos.

1 Introduction

The popularity of many Web2.0 applications such as Del.icio.us[1], Flickr[2] and YouTube[3] has led to a massive amount of freely accessible, user contributed and tagged content. Despite the presence of tags, the lack of structure and explicit semantics hampers the creation of intelligent user interfaces for annotation, navigation and querying and the integration of content from diverse and heterogeneous data sources. A popular hypothesis, expressed by many web experts ([4, 8, 9, 11, 17]), is that Web2.0 data sources can be used more efficiently by structuring and semantically organising them and that the Semantic Web can provide the needed semantics to achieve that.

This hypothesis motivated two different research approaches to enrich folksonomies. First, some methods rely on the statistical analysis of tagspaces based on tag co-occurrence to identify clusters of related tags. In this cases the meaning of a tag is given by its cluster but it remains implicit, i.e., it is not explicitly stated. Second, more recent methods shift from this statistical view to a knowledge-intensive approach where a semantic definition of tags is obtained by aligning them to a knowledge source. The majority of works use WordNet to define the semantics of tags for organizing resources or enhancing their navigation.

Our work is part of the second type of approaches, with the difference that we rely on all online available ontologies as a background knowledge source to define the meaning of tags. In this paper, we present the **FLOR**, **FoLksonomy Ontology enRichment**, algorithm which takes as input a set of tags

[1] http://del.icio.us
[2] http://www.Flickr.com
[3] http://www.youtube.com

(either the tagset of a resource or clusters derived by the statistical analysis of folksonomies) and automatically relates them to relevant semantic entities (concepts, relations, individuals) defined in online ontologies. An immediate advantage of this correlation between tags and semantic entities is that the tag is automatically associated with the semantic neighborhood provided by the corresponding ontology. For example, for the tag `canine` apart from identifying that *Canine* `SubClassOf` *Carnivore* we also acquire the knowledge that *Canine* `DisjointWith` *Feline*.

In the following we describe the related work (Section 2), our methodology (Section 3) and discuss our experimental results (Section 4). We discuss future work in Section 5.

2 Related Work

Since the term *folksonomy* was coined, research has focused on comprehending the inherent characteristics of folksonomies and exploring their emergent semantics. Two of the primer works exploring and analysing their structure, the types of their tags and the user incentives in tagging are described in [7] and [14]. Additionally, there are two main lines of folksonomy related research.

The first works on folksonomies ([3, 15, 16, 20], see [18] for a detailed analysis of the specific methods) were based on the assumption that frequent co-occurrence of tags translates to tag association. They used various statistical methods to identify clusters of related tags but they did not define the exact relations among them. An exception is the work detailed in [18], where, in addition to clustering the tags, the semantic relations among them are identified.

The second research line focuses on the semantic definition of tags, primarily by using WordNet. For example, [13] try to identify the meaning of tags in order to enrich the relevant resources with RDF descriptions. The authors distinguish six conceptual categories of tags in Flickr. Using WordNet and other knowledge resources for these conceptual categories they organise the tags accordingly. Then they enrich the Flickr photos with RDF triples created for each of the tags in a photo. These triples are generated either by predefined predicates or from WordNet signatures depending on which of the above categories they belong to.

The authors of [10] describe a method that expands the related tags clusters of Del.icio.us with more related tags based on co-occurrence. The expanded clusters are presented as navigable hierarchical structures or semantic trees. These semantic trees are derived from WordNet. Using a combination of WordNet based metrics they identify the possible WordNet sense for each tag. Then they extract the path of this tag from the WordNet hierarchy and they integrate it into the semantic tree of the tag's cluster.

The TagPlus system described in [12] uses WordNet to disambiguate the senses of Flickr tags by performing a two step query. First a user looks for a tag, then the system returns all the possible WordNet senses that define the tag and the user selects (disambiguates) which sense he meant. Finally the system looks for all the Flickr photos tagged with this tag and its synonyms.

T-ORG ([1]) performs ontology based organisation of Flickr photos into a set of predefined categories according to the tags describing them. At first the user selects an ontology of interest. Then, the system extracts the concepts and tries to identify semantic relatedness between these concepts and the tags by querying the web with various linguistic patterns between them. Then each tag is categorised under a superclass of the concept to which was more related by the web search.

All the aforementioned works present methods for tag disambiguation, resource organisation and tag cluster enrichment. Our work aims to address the following additional issues. First, the existing works require some initialising from the user's side (e.g., a priori selecting ontology or knowledge resources for the relevant categories of tags) or they require the user contribution to perform the disambiguation of the tags. FLOR is aimed to run entirely *automatically* (i.e., without user contribution). Second, FLOR uses more than one resources (all the online ontologies and WordNet) aiming to achieve higher coverage of tags compared to the coverage from single resources. Finally, the proposed enrichment links each tag with a relevant semantic entity but also with its semantic neighbourhood as demonstrated in the `canine` example in Section 1.

3 FLOR components and methodology

The goal of FLOR is to transform a flat folksonomy tagspace into a rich semantic representation by assigning relevant Semantic Web Entities (SWEs) to each tag. A SWE is an ontology entity (class, relation, instance) defined in an online available ontology. While in this paper we describe the process of enriching a set of tags, the ultimate goal of our system is not just to connect to SWE's but also to bring in other knowledge related to these SWE's. An example of the inputs and expected outcomes to FLOR is demonstrated in Fig. 1. The input consists a cluster of related tags and the output is a set of semantically enriched FlorTags. Note that FLOR is agnostic to the way in which this cluster was obtained. It can either be the set of all tags associated to a resource, or a cluster of related tags obtained through co-occurrence based clustering methods. The experiments reported in this paper used sets of tags associated with a given resource.

Intuitively, FLOR performs three basic steps (see Fig. 1). First, during the **Lexical Processing** the input tagset is cleaned and all potentially meaningless tags are excluded. We rely on a set of heuristics to decide which tags are likely to be meaningless. Second, during the **Sense Definition and Semantic Expansion** we attempt to assign a WordNet sense to each tag based on its context (i.e., the other tags in its cluster) and to extract all relevant synonyms and hypernyms so that we migrate to a richer representation of the tag. Finally, during the **Semantic Enrichment** step each tag is associated to the appropriate SWE.

Note that there is a strong correlation between the steps of FLOR and the components of the final FlorTag structure. The first step results in the **Lexical Representations** which is a list of lexical forms for the tag, such as plural and singular forms for nouns, or various delimited types of compound tags (sanFran-

cisco, san.Francisco, e.t.c). The second step identifies **Synonyms** and **Hypernyms** for each tag. The last step generates the list of **Entities** containing the associated SWE's. Note that a tag can be associated to several relevant SWE's.

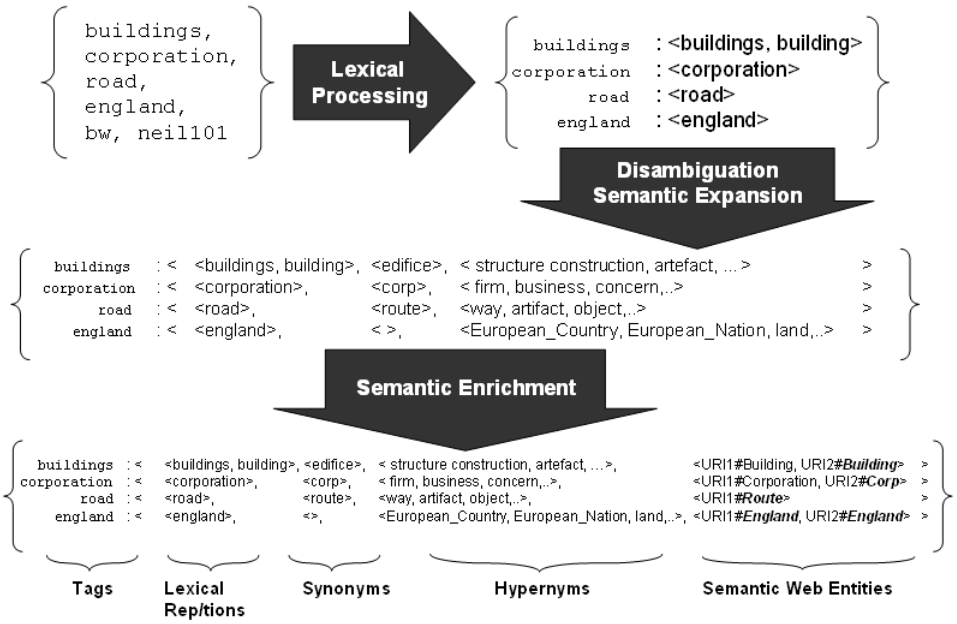

Fig. 1. FLOR Methodology

3.1 PHASE1: Lexical Processing

Due to the freedom of tagging as a basic rule of folksonomies, a wide variety of different tag types are in use. Understanding the types of tags used is the first step in deciding which of them are meaningful and should be taken into account as a basis of a semantic enrichment process. Previous work ([2, 7, 13]) has identified different conceptual categories of tags (event, location, person), as well as tag categories that can be described by syntactic characteristics. For example, there are many tags containing special characters (e.g., :P), numbers (e.g., aug07), plurals as well as singular forms of the same word (e.g., building, buildings), concatenated tags (e.g., littlegirl) or tags with spaces (e.g., little girl) and a big number of non-English tags (e.g., sillon). The role of the lexical processing step is to identify these different categories of tags and exclude those that are meaningless and should not be further processed. This is done in two sub-steps.

The Lexical Isolation phase idenfies sets of tags that should be excluded as well as those that can be further processed. Currently we isolate and exclude all tags with numbers, special characters and non English tags. The reason for excluding non-English tags is that our method explores various external knowledge sources (WordNet, Semantic Web ontologies) that are primarily in English. As future work, we will extend FLOR to isolate additional types of tags as well and deal with non-English tags.

The Lexical Normalisation phase aims to solve the incompatibility between different naming conventions used in folksonomies, ontologies and thesauri such as WordNet. This phase produces a list of possible **Lexical Representations** for each tag aiming to maximise the coverage of this tag by different resources. For example, the compound tag `santabarbara` in folksonomies appears as *Santa-Barbara* or *Santa+Barbara* in various ontologies and as ***Santa Barbara*** in WordNet. However, as the lexical anchoring to these resources is a quite complex problem we try to address it by producing all the possible lexical representations for each tag such as: {santaBarbara, santa.barbara, santa_barbara, santa barbara, santa-barbara, santa+barbara, ...}.

3.2 PHASE2: Sense Definition and Semantic Expansion

Due to polysemy, the same tag can have different meanings in different contexts. For example, the tag `jaguar` can describe either a car or an animal depending on the context in which it appears. Before connecting a tag with a relevant SWE, it is important to determine its intended sense in the given context. This task is performed in the first step of this phase.

Another issue to take into account is that, despite its significant growth, the Semantic Web is still sparse. A direct implication is that while online ontologies might not contain concepts that are syntactically equivalent to a given tag, they might contain concepts that are labeled with one of its synonyms. To overcome this limitation, we perform a semantic expansion for each tag, based on its previously identified sense, in the final step of this phase.

The Sense Definition and Disambiguation phase deals with discovering the intended sense of a tag in the context it appears. As context we consider the set of tags with which the given tag co-occurs when describing a resource. For example, in the tagset: {`panther`, `jaguar`, `jungle`, `wild`} the context of `jaguar` is {`panther`, `jungle`, `wild`}. We use WordNet as a sense repository and rely on its hierarchy of senses to compute the similarities between the senses of all tags in the tagset and thus achieve their disambiguation. WordNet also provides rich sense definitions which facilitate the semantic expansion in the next step.

To define the senses of the tags in a tagset, we identify all the lexical representations for each tag in WordNet. In the cases that a tag has more than one senses in WordNet (synsets) we exploit the contextual information of the tagset to identify the most relevant sense. For this, we calculate the similarity between

all the combinations of tags in the tagset using the Wu and Palmer similarity formula ([21]) on the WordNet graph. The similarity degree between two senses is calculated based on the number of common ancestors between them in the WordNet hierarchy and the length of their connecting path. The result for each calculation is a couple of senses and a similarity degree for these senses. We select the two senses of the tags that return the higher similarity degree provided that the similarity degree is higher than a specific threshold.

Our experiments indicate that similarity values over 0.8 almost always are assigned to tag senses that are indeed similar. For example, in the tagset: {`girl`, `eating`, `red`, `apple`} the similarity between `red` and `girl` is 0.7 for the senses:

Bolshevik, Marxist, Pinko, Red, Bolshie (emotionally charged terms used to refer to extreme radicals or revolutionaries)

Girlfriend, Girl, Lady_friend (a girl or young woman with whom a man is romantically involved)

These two senses are connected through the concept ***Person*** in the WordNet hierarchy. However we are not sure if this is the intended meaning for this resource. If the similarity returned for the couples of tags is lower than this threshold or if there is no similarity, hence no relation, between a tag and any of the rest of the tags we select the most popular sense for this tag from WordNet.

Thanks to the modular architecture of FLOR, the disambiguation and sense selection method can be replaced by other methods (e.g., such as those used in [19] and [22]). Or our current method could be modified to exploit a different similarity measure between two concepts such as the Google Similarity Distance [5]. Another possible improvement could be achieved by further expanding the resource tagset with more related tags. These can be discovered with statistical measures based on tag co-occurrence as described in [18]. For example, the expanded tagset of {`apple`, `mac`} could be {`apple`, `mac`, `computer`, `macOs`}. So instead of trying to disambiguate with two tags we increase the possibilities of finding the correct sense by disambiguating with a more specific context.

The Semantic Expansion includes the synonyms and hypernyms of a tag in the FlorTag. For the purpose of this work we used WordNet to extract the synonyms of the correct sense and the synonyms of this sense's hypernym in WordNet. For example if we decide that in the specific context the tag `jaguar` refers to the animal then the semantic expansion would include a list of synonyms: {***Panther***, ***Panthera onca***, ***Felis onca***} and a list of hypernyms: {***Big cat***, ***Feline***, ***Carnivore***}. The Semantic Expansion step produces the lists of synonyms and hypernyms for FlorTags as demonstrated in Fig. 1 and provides the input for the next phase of the algorithm.

3.3 PHASE3: Semantic Enrichment

This phase of FLOR identifies the SWEs that are relevant for each tag by leveraging the results of lexical cleaning and semantic expansion performed in the

previous two phases. The final output of FLOR is produced by this phase (see Fig. 1) and it is a set of FlorTags enriched with relevant SWEs and their semantic neighbourhood (e.g., parents, children, relations).

The relevant SWEs are selected by querying the WATSON semantic web gateway[6]. We search for all possible ontological entities (Classes, Properties and Individuals) that contain in their local name or in their label one of the lexical representations or the synonyms of a given tag.

Such queries often result in several SWEs some of which are very similar (or the same in case they appear in ontologies that are versions of each other). To reduce the number of SWEs, we perform an entity integration process similar to the one described in [19]. The goal of this process is to "collapse" entities that have a high similarity into a single semantic object, thus reducing redundancy. To compute similarity between two entities we compare their semantic neighbourhoods (superclasses, subclasses, disjoint classes for classes; domain, range, superproperties, subproperties for properties) and their localnames and labels. The similarity $simDgr$ for two SWEs e_1 and e_2 is calculated as:

$$simDgr = W_l * simLexical(e_1, e_2) + W_g * simGraph(e_1, e_2)$$

$simLexical(e_1, e_2)$ is the similarity between the lexical information of two entities, i.e., their labels and localnames, computed with Levenshtein distance metric. $simGraph(e_1, e_2)$ is the similarity of the entities' neighbourhoods, where the similarity of each neighbourhood element is computed based on string similarity. Because we consider the similarity of the semantic neighbourhoods more important than the similarity of the labels, we set the weights as $W_l = 0.3$ and $W_g = 0.7$. If the similarity between two entities is higher than a threshold we merge them in one entity by integrating their neighbourhoods into one. Then we repeat the process until all entities are sufficiently different from each other, i.e., their similarity falls under a chosen threshold.

Consider for example Fig. 2 where five SWEs $e_{1,5}$ are compared against a threshold value of 0.5. We start by performing their pair-wise comparison and observe that the pairs (e_1, e_4), (e_1, e_5), (e_2, e_3) and (e_2, e_5) have a similarity equal or above the set threshold. We proceed by merging the first two entities with the highest similarity, e_1 and e_5, to one entity e_1+e_5 and compute the similarities between the new entity and the remaining ones. This process continues until all similarities are lower than the set threshold, which directly demonstrates that the obtained entities are sufficiently different.

Once the merged entities are created we enrich the tag with the relevant entities. This is done by comparing the ontological parents of the merged entity with the hypernyms retrieved from WordNet. The ontological parents are the superclasses of classes, the superproperties of properties and the classes of individuals. For example, as we can see in Fig. 3, the tag `moon` is enriched with two entities. The superclasses of both the entities have as localname one of the hypernyms extracted from the WordNet sense of moon. Also, apart from the semantic definition of the tag with the respective entity, we further enrich the tag with the information carried by the entity, *EarthsMoon* `TypeOf` *Moon*.

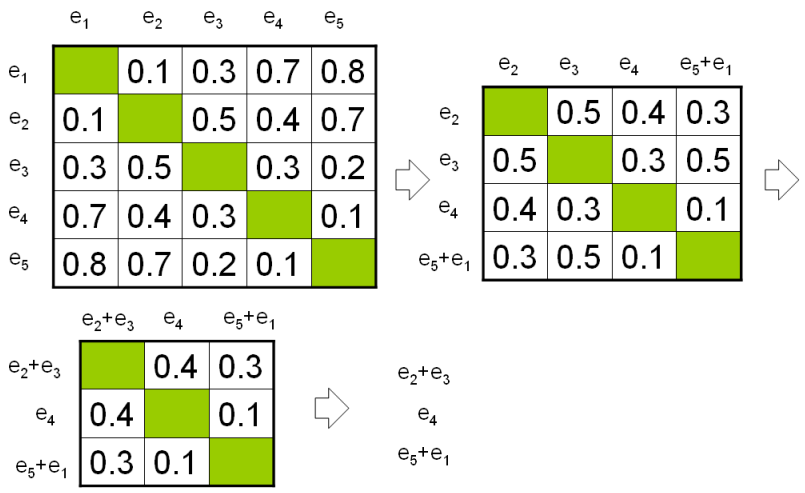

Fig. 2. Merging Strategy

moon

Lexical Representations	Synonyms	Hypernyms	Entities
moon		satellite celestial_body heavenly_body natural_object object physical_object entity	**http://www.ida.liu.se/~adrpo/modelica/rdf/inheritance.owl#moon** type (of) http://www.ida.liu.se/~adrpo/modelica/rdf/inheritance.owl#***CelestialBody*** **http://www.cyc.com/2003/04/01/cyc#moon** subClassOf http://www.cyc.com/2003/04/01/cyc#***NaturalSatellite*** type http://www.cyc.com/2003/04/01/cyc#***EarthsMoon***

Fig. 3. Enriched FlorTag `moon`

3.4 An Enrichment Example

In this section we demonstrate a full cycle of the FLOR semantic enrichment method for the tag `lake`, which was found in the following five tagsets: {`rush`, `lake`, `pakistan`, `rakaposhi`, `mountain`, `asia`, `kashmir`, `snow`, `glacier`, `green`, `white`, `sky`, `blue`, `clouds`, `water`}, {`moraine`, `alberta`, `banff`, `canada`, `lake`, `lac`, `rockies`, `scan`}, {`rising`, `sunlight`, `lake`, `quality`, `bravo`}, {`lake`, `nature`, `landscape`, `sunset`, `water`, `organisms`} and {`lake`, `finland`, `suomi`, `beach`, `bubbles`, `blue`, `sunlight`, `kids`, `natural`}. Note that these tagsets contain the tags that remain after the lexical processing performed in the first phase of FLOR. Fig. 4 shows the information contained in the automatically obtained FlorTag.

For the second phase of FLOR, the available WordNet senses for ***Lake*** are considered. These are:

WordNet 1: ***Lake→Body of water***, ***Water→Thing→Entity***
(a body of (usually fresh) water surrounded by land)
WordNet 2: ***Lake→Pigment→Coloring material→Material → Substance→Entity***
(a purplish red pigment prepared from lac or cochineal)
WordNet 3: ***Lake→Pigment→Coloring material→Material →Substance→Entity***
(any of numerous bright translucent organic pigments)

lake

Lexical Representations	Synonyms	Hypernyms	Entities
lake		lake body_of_water water thing entity	**http://lonely.org/russia#lake** subClassOf http://lonely.org/russia#***waterway*** http://lonely.org/russia#***Lake_Baikal*** – type **http://lsdis.cs.uga.edu/proj/semdis/testbed/#lake** subClassOf http://lsdis.cs.uga..../testbed/#***Water_Feature*** subClassOf http://lsdis.cs.uga.edu/proj/semdis/testbed/#***Thing***

Fig. 4. Enriched FlorTag `lake`

Applying the Wu and Palmer formula for `lake` and the rest of the tags in these tagsets we obtained variable similarities from 0 to 0.86. The zero similarities were obtained for location names such as `banf`, `pakistan`, `suomi` and for

generally unrelated with `lake` tags such as `quality`, `scan`, `sunlight`, `sunset`. Interestingly, `lake` returned zero similarity for the tags `glacier` and `mountain` while they should be related. This is due to the fact that, in WordNet, ***Glacier*** and ***Mountain*** are hyponyms of ***Geological formation*** which is a hyponym of ***Natural object*** while ***Lake*** is a hyponym of ***Body of water*** which is a direct hyponym of ***Thing***. Furthermore ***Glacier*** is a hyponym of ***Ice mass*** but there is no subsumption relation between ***Ice mass*** and ***Ice*** or ***Water***, so that would allow for a connecting path between ***Lake*** and ***Glacier***. This fact motivates further research on how to identify similarities between tags of a tagset beyond the subsumption relations provided by WordNet.

The highest similarity, 0.86, for `lake` was obtained with the tag `water`, because Sense 1 of ***Lake*** is related to ***Body of water*** (Sense 2 of ***Water***) with a direct hyponymy relation. Note that, in most of tagsets the first sense of ***Water***, ***Liquid***, is selected as this is the most common sense in which the tag is used. Therefore, this is a nice example of phase 2 identifying a non-trivial correlation.

Sense 1. ***Water***, ***H2O***: (binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid) → ***Binary Compound*** AND → ***Liquid***

Sense 2. ***Body of water***, ***Water***: (the part of the earth's surface covered with water) → ***Thing***

Once the correct sense is selected and the tag is semantically expanded with hypernyms (there are no synonyms for this sense of ***Lake*** in WordNet) then the final step of FLOR selects the Semantic Web Entities that correspond to this sense. As shown in Fig. 4 both selected entities have the term *Lake* in their localname and their superclass in the ontology contains one or more of the hypernyms returned by WordNet, *Water* and *Thing*, as a whole or as a compound. This example demonstrates that our anchoring to ontologies is strict for the tags to be defined (their lexical representations and synonyms) and the localnames and labels of the entities and flexible for the ontological parents and hypernyms. Note also that the selected SWEs carry additional information about two superclasses of *Lake* (*Waterway*, *Waterfeature*) and an instance of *Lake* (*Lake Baikal*) thus semantically enriching the tag `lake`.

4 Experiments and Results

In order to assess the working of our method, we applied it to a sample of Flickr data. In particular, we wanted to assess the correctness of SWE assignment (i.e., whether tags were linked to relevant SWEs).

The data set comprised of 250 randomly selected Flickr photos with a total of 2819 individual tags. During the Lexical Isolation we removed 59% of the initial tagset resulting in a data set of 1146 tags in total. We isolated 45 tags with two characters (e.g., `pb`, `ak`, `fc`), 333 tags with numbers (e.g., `views200`, `356days`, `tag1`), 86 tags with special characters (e.g., `:P`, `(raw → jpg)`, `??(m0)??`), and 818 non English tags (e.g., `turdus merula`, `arbol`, `tormenta`). Then we filtered

out the photos that exclusively contained the isolated tags (24 photos) and obtained a dataset of 226 photos with a total of 1146 tags. After running the FLOR enrichment algorithm for these 226 photos, one of the authors has manually checked all the assignments between tags and SWE's. The evaluation results are displayed in Table 1.

Enrichment Result	**# of Photos**	**Percentage**
All Tags Correctly Enriched	179	79.2%
All Tags Incorrectly Enriched	3	1.3%
Mixed Enrichments (some correct, some incorrect)	17	7.5%
Unclear Enrichments	4	1.8%
No Tags Enriched	23	10.2%
Total	226	100%

Table 1. Evaluating the correctness of SWE assignment.

According to our evaluation, 179 photos (about 80%) were correctly enriched, meaning that at least one of their tags was enriched and all the enriched tags were assigned to a relevant SWE. Note that these results were highly superior to the ones we have obtained in previous experiments where we did not rely on WordNet as an intermediary step. Indeed, the WordNet sense definition and expansion of the tags with synonyms and hypernyms (FLOR phase 2) increased their discovery in the Semantic Web.

Our method has failed to correctly enrich some tags, thus resulting in 3 photos where all the enrichments were incorrect and in 17 photos where at least one tag was enriched incorrectly. One example of incorrect enrichment is for the tag `square` in the context {`street`, `square`, `film`, `color`, `documentary`}. While its intended meaning is ***Geographical area***, because during the disambiguation phase `square` did not return high similarity with any of the rest of the tags, the WordNet sense assigned to it was the most popular one, ***Geometrical shape***. This lead to the assignment of non-relevant SWE's namely, *Square* `SubClassOf` *Rectangle* and *Square* `SubClassOf` *RegularPolygonShaped*. Despite this error, the rest of the tags were correctly enriched.

FLOR could not enrich 23 photos (i.e., none of their tags could be enriched). A major cause for these failures was that their WordNet derived hypernyms did not match the superclasses of these tags's concepts in the Semantic Web. For example, the definition of `love` in WordNet and the relevant entity found in the Semantic Web are:

WordNet: ***Love→Emotion→Feeling→Psychological feature***
(a strong positive emotion of regard and affection)
Semantic Web: *Love* `SubClassOf` *Affection*

Although both these definitions refer to the same sense, and additionally the superclass *Affection* belongs to the gloss of ***Love*** in WordNet, they were not

matched as *Affection* was not included in the hyponyms of *Love*. Current work investigates alternative ways of Semantic Expansion.

For four of the enriched photos the correctness of the enrichment was difficult to assess. This is because the meaning of the tag was unclear even when considering its context (the rest of the tags) and the actual photo. For example, in the photo depicted in Fig. 5 the meaning of the tag `volume` is unclear. In the second phase of FLOR the tag was expanded with the hypernyms ***Measure*** and ***Abstraction***. Then, it was related to the SWE *Volume* `SubClassOf` *Measure*. Because the meaning of the tag was not clear for the evaluator, she could not assess whether this enrichment was correct or not. More generally, there are several cases when tags only make sense to their author (and maybe to his social group) and thus will be difficult to enrich by FLOR.

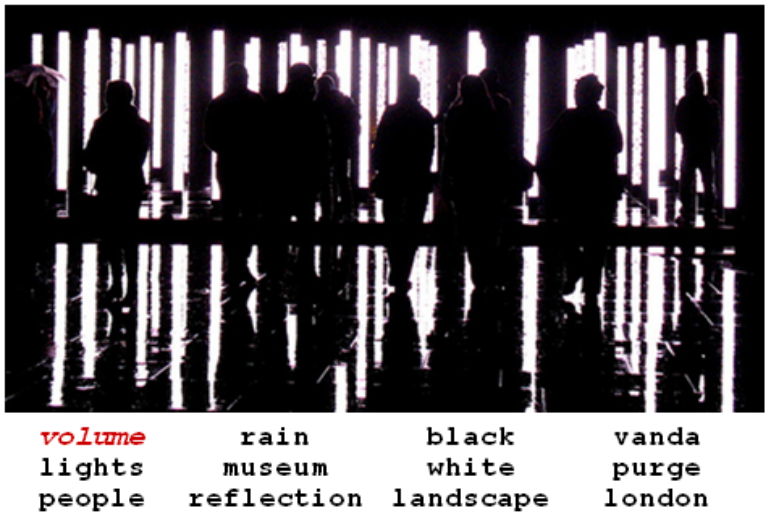

Fig. 5. Ambiguous Enrichment

5 Conclusions and Future Work

We presented the methodology and the experiments we performed to test the hypothesis that **enrichment of folksonomy tagsets with ontological entities can be performed automatically**. As demonstrated in Section 4, we selected a subset of Flickr photos and after performing lexical processing and semantic expansion we correctly enriched the 72% (179 of 250) of them with at least one Semantic Web Entity. Compared to our previous efforts to define the tags with Semantic Web Entities without previously expanding them with

synonyms and hypernyms, this is a significant improvement. Analysing the experimental results we identified a number of issues to be resolved in order to enhance the performance of FLOR. These issues also form our future work.

The **Lexical Processing** phase requires supplementary methods to identify and isolate additional special cases of tags (e.g., photography jargon, dates). Furthermore, the implementation of strategies to deal with these isolated tagsets and the integration of these strategies to the FLOR methodology are intended to be addressed by our future work, as our higher goal is to apply FLOR on the tagspaces of folksonomies as a whole without excluding cases of tags.

As demonstrated by the results in Section 4, the cases of incorrect enrichment were mainly caused due to the failure of the **Sense Definition and Semantic Expansion** phase. The following issues are currently investigated in order to correct the errors and enhance the performance of this phase. First, it is essential to extend the tag similarity measure to also identify generic relations rather than only subsumption relations. This flaw was demonstrated in the case of `lake` and `glacier` (Section 3.4) where they were not found to be related in the hierarchical structure of WordNet. Also, in the example of `square` co-occurring with `street` (Section 4) we saw that the incorrect sense definition for `square` caused further incorrect enrichment. This happened because `square` did not return a high relatedness with any of the rest of the tags. One of the possible solutions to this is the context expansion based on tag co-occurrence. For example, expanding the {`square`, `street`} tagset with their frequently co-occurring tags e.g., {`building`, `park`} can increase the semantic relatedness between the tags and potentially lead to mapping the tags to the correct sense.

The quality of the results returned from the **Semantic Enrichment** phase, which are also the results of the overall FLOR enrichment algorithm, depends on two factors. First, on the input provided to this phase by the Semantic Expansion step and second on the anchoring of the tags' lexical representations and synonyms to the online ontologies. Alternative strategies for flexible anchoring to increase the number of successful enrichments and the same time keep the number of irrelevant matches low, are investigated by our current work.

Finally, we aim to evaluate the FLOR three phase enrichment method by performing large scale experiments. This is to identify the possible implications of the overall process that are not apparent in a small scale study like the current.

To conclude, we demonstrated that the **automatic enrichment of folksonomy tagsets using a combination of WordNet and online ontologies is possible** without user intervention in any step of the methodology and by using straightforward methods for lexical isolation, disambiguation, semantic expansion and semantic enrichment. The goal is to create a semantic layer on top of the flat folksonomy tagspaces, that allows intelligent annotation, search and navigation as well as the integration of resources from distinct, heterogeneous systems.

References

1. Rabeeh Abbasi, Steffen Staab, and Philipp Cimiano. Organizing resources on tagging systems using t-org. In *4th European Semantic Web Conference*, pages 97–110, Innsbruck, Austria, 2007. International Workhsop: Bridging the Gap between Semantic Web and Web 2.0.
2. Sofia Angeletou, Marta Sabou, Lucia Specia, and Enrico Motta. Bridging the gap between folksonomies and the semantic web: An experience report. In *4th European Semantic Web Conference*, pages 30–43, Innsbruck, Austria, 2007. International Workhsop: Bridging the Gap between Semantic Web and Web 2.0.
3. Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *15th International World Wide Web Conference*, Edinburgh, Scotland, 2006. Collaborative Web Tagging Workshop.
4. Richard Benjamins, John Davies, Ricardo Baeza-Yates, Peter Mika, Hugo Zaragoza, Mark Greaves, Jose Manuel Gomez-Perez, Jesus Contreras, John Domingue, and Dieter Fensel. Near-term prospects for semantic technologies. *Intelligent Systems, IEEE*, 23:76–88, 2008.
5. Rudi Cilibrasi and Paul Vitanyi. The google similarity distance. *Transactions on Knowledge and Data Engineering, IEEE*, 19(3):370–383, 2007.
6. M. dAquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. Watson: A gateway for the semantic web. In *4th European Semantic Web Conference*, Innsbruck, Austria, 2007. Poster Session.
7. Scott Golder and Bernardo Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
8. Mark Greaves. Semantic web 2.0. *Intelligent Systems, IEEE*, 22(2):94–96, 2007.
9. Jim Hendler. The dark side of the semantic web. *Intelligent Systems, IEEE*, 22(1):2–4, 2007.
10. David Laniado, Davide Eynard, and Marco Colombetti. Using wordnet to turn a folksonomy into a hierarchy of concepts. In *Semantic Web Application and Perspectives - Fourth Italian Semantic Web Workshop*, pages 192–201, Bari, Italy, Dec 2007.
11. Ora Lassila and Jim Hendler. Embracing "Web 3.0". *Internet Computing, IEEE*, 11(3):90–93, 2007.
12. Sun-Sook Lee and Hwan-Seung Yong. Tagplus: A retrieval system using synonym tag in folksonomy. In *International Conference on Multimedia and Ubiquitous Engineering*, pages 294–298, Seoul, Korea, 2007.
13. Mohamed Zied Maala, Alexandre Delteil, and Ahmed Azough. A conversion process from flickr tags to rdf descriptions. In *10th International Conference on Business Information Systems*, Poznan, Poland, 2007. 1st Workshop on Social Aspects of the Web.
14. Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *15th International World Wide Web Conference*, Edinburgh, Scotland, 2006. Collaborative Web Tagging Workshop.
15. Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *4th International Semantic Web Conference*, pages 522–536, Galway, Ireland, 2005.
16. Patrick Schmitz. Inducing ontology from flickr tags. In *15th International World Wide Web Conference*, Edinburgh, Scotland, 2006. Collaborative Web Tagging Workshop.

17. Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *Intelligent Systems, IEEE*, 21(3):96–101, 2006.
18. Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *4th European Semantic Web Conference*, pages 624–639, Innsbruck, Austria, 2007.
19. Raquel Trillo, Jorge Gracia, Mauricio Espinoza, and Eduardo Mena. Discovering the semantics of user keywords. *Journal of Universal Computer Science*, 13(12):1908–1935, 2007.
20. Xian Wu, Lei Zhang, and Yong Yu. Exploring social annotations for the semantic web. In *15th International World Wide Web Conference*, pages 417–426, Edinburgh, Scotland, 2006. ACM.
21. Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico, USA, 1994.
22. Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In *International Semantic Web Conference*, Busan, South Korea, 2007. International Workshop on Emergent Semantics and Ontology Evolution.