# Helping online communities to semantically enrich folksonomies

Freddy Limpens
Edelweiss, INRIA
F-06902 Sophia Antipolis
freddy.limpens@inria.fr

Fabien Gandon
Edelweiss, INRIA
F-06902 Sophia Antipolis
fabien.gandon@inria.fr

Michel Buffa
KEWI, I3S - UNSA CNRS
F-06903 Sophia Antipolis
buffa@unice.fr

## ABSTRACT

This paper presents our approach to collaborative and semi-automated semantic structuring of folksonomies. Tags freely provided by users of online communities are not semantically linked, and this hinders significantly the potentials for browsing and exploring these data. We propose a socio-technical system combining automatic handlings of tags, using state of the art algorithm, and user friendly interfaces designed after a careful analysis of the usage of our target communities. Much like folksonomies, our socio-technical system lets each user maintain his own view while still benefiting from others contributions. As a complement to similar approaches, our approach supports conflicting point of views all along the life-cycle of semantically enriched folksonomies.

## 1. INTRODUCTION

Our approach aims at leveraging social tagging practices with socio-technical systems including semantic tools carefully designed after an analysis of the knowledge exchange practices of online communities. Social tagging is a successful yet still promising means to involve users in the life-cycle of the content they exchange, read or publish online. However, folksonomies resulting from this practice have some limitations, in particular, the spelling variations of similar tags and the lack of semantic relationships between tags hinder significantly the navigation within tagged corpora.

One way of tackling these limitations is to semantically structure folksonomies. This can help navigate within tagged corpora by (1) enriching tag-based search results with spelling variants and hyponyms, or (2) suggesting related tags to extend the search, or (3) hierarchically organizing tags (using SKOS[1] e.g) to guide novice users in a given domain more efficiently than with flat list of tags or occurrence-based tag clouds.

In this paper, we present our approach to design a tagging-based system which integrates collaborative and assisted semantic enrichment of the community's folksonomy. Our contribution consists in proposing a formal model to support diverging points of view and to combine automatic handlings and user input. This combination takes up the form of a socio-technical system whose design is grounded on a scenario-based analysis.

A typical scenario of application of our approach can be found within the Ademe agency[2] which seeks to broaden the audience of its scientific production in the field of sustainable development and environmental issues. In this scenario, we can distinguish between three types of agents: (1) the expert-engineers working at Ademe and who are specialists of a given domain, (2) the archivists who take care of the indexing of the documents from Ademe and have transversal yet not very deep knowledge of the thematics covered at the agency, and (3) the public audience who has access to the documents of Ademe from its website. The difficulty in the structuring of the folksonomy at Ademe comes from the different points of view that may arise from the community of expert engineers, and possibly also from the public. These points of view have to be turned by the archivists into a coherent indexing. This indexing will then be used by all the members of Ademe and the public when browsing the Ademe corpus of resources.

This paper is organised as follows. In section two we first present current works in folksonomy semantic enrichment, and position our contribution. In section three we give a general presentation of our approach. In section four we give more details of each module of our socio-technical system before concluding in section five.

## 2. RELATED WORK

### 2.1 Research in bridging folksonomies and ontologies

Folksonomy enrichment has been adressed by numerous research works which cover a broad variety of approaches.

A first category of works are aimed towards extracting the emergent tag semantics from folksonomies by measuring the semantic similarity of tags. The studies from [13] and [5] propose an analysis of the different types of similarity measures and the semantic relations they each tend to convey. The most simple approach consists in counting the coocurrence of tags in different contexts (users or resources). [5] showed that this type of measures provided subsumption relations but was not sufficiently accurate. More elaborate methods exploit the network structure of folksonomies making use of the distributional hypothesis that states that words used in similar contexts tend to be semantically related. To apply this hypothesis on tags, [5] computed the

[1] http://www.w3.org/TR/skos-reference/

[2] ADEME is the French for Environment and Energy Management Agency, see http://www.ademe.fr


*Web Science Conf. 2010*, April 26-27, 2010, Raleigh, NC, USA.

cosine similarity measure in the vector spaces obtained by folding the tripartite structure of folksonomy onto distributional aggregations spaning the associations of tags with : the other tags (tag-tag context), or the users (tag-user context), or the resources (tag-resources). Their study shows that the tag-tag context performed best at a reasonable cost. They also computed the distance and relative placement in wordnet hierarchy of the pairs of tags retrieved by this method, and showed that the semantic relation conveyed by this measure was of type "related" in thesauruses terms. Mika [14] also applied and evaluated different foldings of the tripartite structure of folksonomies. Interestingly, he showed after a qualitative evaluation that exploiting user-based associations of tags yielded more representative taxonomic relations. The principle of this association is that if the community of users using tag "biological agriculture" is included in the community of users of the tag "agriculture", then the tag "agriculture" is broader than the tag "agriculture". [8] proposed an algorithm which constructs a taxonomy from tags by crawling the similarity graph computed from the cosine distance based on the Tag-Resource context. The hierarchy of tags is built starting from the tag with the highest centrality, and each tag, taken in order of centrality, is added either as a child of one of the node or the root node depending on a threshold value.

Another type of approach consists in letting users semantically structure tags or link tags to unambiguous meanings. We can mention in this category the work of [17] who proposed to tag the tags, or the work of [9] who proposed a simple syntax to specify subsumption (with ">" or "<") or synonymy (with "=") relations between tags. Some tools available online also feature semantic structuring capacities such as Gnizr[3] and Semanlink[4], and even Flickr with machine tags[5]. In the same trend, the Linked Data community seeks to weave together the content of social web sites thanks to a set of formal ontologies not aimed at describing the knowledge of the communities but rather the structure of their knowledge exchange platforms. For instance SCOT[6] describes tags as parts of shareable tag clouds, and SIOC[7] describes online communities content. MOAT[15] is an ontology aimed at linking each tagging action with a URI representing the meaning of this tag action. These URIs can link to formal ontologies concepts or any web page containing a description of a notion. Once tag actions are formaly linked to concepts, it is possible to disambiguate tags when searching, but also to exploit inference mechanisms via the formal concepts and get a richer browsing experience. NiceTag is a model that seeks to account for the usages of tags through a finer modelization of the relations between tags and the tagged resources [12]. Its flexibility and the use of named graphs mechanism allow this model to serve as a pivot model for all other tag models, adding a level of semantic pragmatics.

Another group of works seek to integrate one or several of the preceding approaches. For instance[1] and [16] make use of similarity metrics to find related tags, and then map these tags to concepts from available online ontologies in order to semantically structure tags with formal properties. [18] proposed an integrated approach to folksonomy enrichment including as many resources as possible, using each in a tailored way, and also the validation from users.

Finally, our approach can be related to ontology construction and ontology maturing. Indeed, our approach clearly echoes with older attempts to build formal ontologies from texts [2] or databases maintained by communities of users [10]. More recently, [3] addressed the problem of collaborative ontology editing and pointed out the limitations of current ontology engineering tools in that respect. They proposed integrating ontology maturing in common tasks such as information seeking, and they developped a bookmarking service with the possibility for all users to add or edit new "semantic" tags formally structured with SKOS, which is based on thesauruses formalisms.

## 2.2 Limits of current approaches

However, full automatization of semantically enriching folksonomies is difficult. First the similarity measures used by [5, 13, 16] or other methods for retrieving taxonomical structures from folksonomies ([14], [8]) are usefull to boostrap the process, but their accuracy in reflecting the communities knowledge is limited. The semantic grounding of these measures proposed by [5] can also help evaluate their accuracy. However, as this evaluation require that tags be present in Wordnet synsets or in other ontological resources, the validity of these measures can only be evaluated for common knowledge and not really for specific terms that consist in one of the most valuable benefit of folksonomies. The same argument can be used towards other approaches [1] that make use of ontological resources to formaly structure folksonomies.

On the other hand, approaches that rely on user input (to tag the tags, or to link a tag to an unambiguous concept) may induce, without user-friendly interfaces tailored to usages, a cognitive overload that regular users of tagging are not ready to bear. Integrated approaches try to overcome this limit by mixing automatic handlings with user validation. However, none of these two types of approach formaly take into account the multiplicity of points of view within a community.

# 3. COMBINING MACHINE AND HUMAN PARTICIPATION IN A COHERENT SOCIO-TECHNICAL TAGGING APPLICATION

A generic methods to semantically enrich all types of folksonomies in a fully automatic manner seems out of reach today. We believe that significant progress can be achieved by carefully analyzing the usages of the target communities of a system. Indeed, one may take advantage of the tasks already achieved by users to capture knowledge as a side effect of their daily activity. We conducted such an analysis in one of our target community, the Ademe agency. The indexing of the documents at Ademe is made with a sort of folksonomy in which tags are carefully chosen by the archivists, hence the name "controlled" folksonomy. This controlled folksonomy is flat for the moment, but the archivists seek to structure it and enrich it with new terms so as to be able to offer richer search results as well as thematic navigation capabilities within their corpus. To do so, they need contributions in both new tags and semantic structuring from all

[3]http://code.google.com/p/gnizr/
[4]http://www.semanlink.net
[5]http://www.flickr.com/groups/mtags/
[6]http://scot-project.org/
[7]http://sioc-project.org/

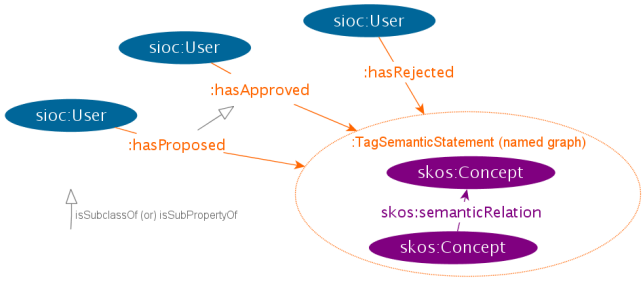


**Figure 1: SRTag RDF schema**

members of Ademe and the public.

Our approach to semantically enriching folksonomies consists in creating a synergistic combination of automatic handling, to bootstrap the process, and of users contributions at the lowest possible cost through user friendly interfaces. We propose a socio-technical system which supports conflicting points of view regarding the semantic organization of tags, but also helps online communities to build up a consensual point of view emerging from individual contributions.

## 3.1 SRTag : using named graphs to keep track of diverging points of view

In order to model the semantic structuring of folksonomies while supporting conflicting views, we propose a RDF schema, SRTag[8], which makes use of named graphs mechanisms[4, 7]. Named graphs allow for reifying the semantic relationship between two tags without the burden of classical RDF reification[9] (see figure 1). Indeed, the principle of our model is to encapsulate statements about tags within a named graph. Then these named graphs are typed with our class `:TagSemanticStatement` or more precise subclasses.

The relationships between tags can be taken from any model, but we chose to limit the number of possible relations to thesaurus-like relations as modeled in SKOS. Then we modeled a limited series of semantic actions which can be performed on a `:TagSemanticStatement` by users (represented using `sioc:User` class), namely `:hasApproved`, `has:Proposed`, and `has:Rejected`. We are then able to capture and track back users opinions (reject or approve) on the asserted relations, which allows us to collect diverging points of view.

We distinguish different types of automatic and human agents according to their role in the life-cycle of the folksonomy. We modeled different subclasses of the class `sioc:User` in order to filter tag relations according to the users who approve it. This includes `:SingleUser` which correspond to regular users of the system, `:ReferentUser (e.g. an archivist)` who is in charge of building a consensual point of view, `:TagStructureComputer` which corresponds to the software agents performing automatic handlings on tags, and `:ConflictSolver` corresponding to software agents which propose temporary conflict resolutions for diverging points of view before referent users choose one consensual point of view.

## 3.2 Folksonomy enrichment life-cycle

As a result, our model allows for the factorization of individual contributions as well as the maintenance of a coherent view for each user and a consensual view linked to a referent user. Furthermore, by modeling different types of agents who propose, approve or reject tag relations, we are able to set up a life cycle of enriched folksonomies. Figure 2 illustrates this life-cycle which can be decomposed as follows:

1. We start from a "flat" folksonomy, ie. with no semantic relationships between tag. `:TagStructureComputer` agents (automaton) performs calculation on tags using methods based on an analysis of the labels of tags and on the network structure of the folksonomy. `:TagStructureComputer` agents then add assertions to the triple store (`TagSemanticStatement`) stating semantic relations between tags . These computations are done during the night due to their algorithmic complexity.

2. Human agents, modeled as `SingleUser,` contribute through user friendly interfaces integrated in tools they use daily by suggesting, correcting or validating tag relations. Each user maintains her own point of view regarding tag relations, while benefitting also from the points of view from other users.

3. As logical inconsitencies may arise between all users' points of view, another type of automatic agent, named `ConflictSolver` detects these conflicts and proposes conflict resolutions. The statements proposed by the ConflictSolver are used firstly to avoid the noise that may hinder the use of our system when, for instance, several different relations are stated about the same pair of tags.

4. The statements from the `ConflictSolver` agent are also used to help the `ReferentUser` in her task of maintaining a global and consensual view with no conflicts. This view can then be used to filter the suggestions of related tags by giving priority to referent-validated tags over other tags suggested by computers.

5. At this point of the life-cycle we have a semantically structured folksonomy in which each user's point of view co-exists with the consensual point of view. Then a set of strategies are applied to exploit these points of view to offer a coherent navigation to all users.

6. Then, another cycle restarts with automatic handlings in order to take into accound the new tags that are added to the folksonomy.

We give more details on each step of the folksonomy enrichment in the next section

# 4. MODULES USED IN OUR APPROACH

## 4.1 Automatically handling folksonomies to extract emergent semantics

We have implemented automatic handling methods by integrating state of the art algorithms [13, 14] which are applied on the folksonomy in order to retrieve semantic relationships between tags. We first present the experiment we conducted to evaluate the performance of string-based methods to retrieving semantic relationships between tags. Then we present our implementation of the methods analyzing the structure of folksonomies.

[8]: http://ns.inria.fr/srtag/2009/01/09/srtag.html
[9]http://www.w3.org/TR/rdf-mt/#Reif

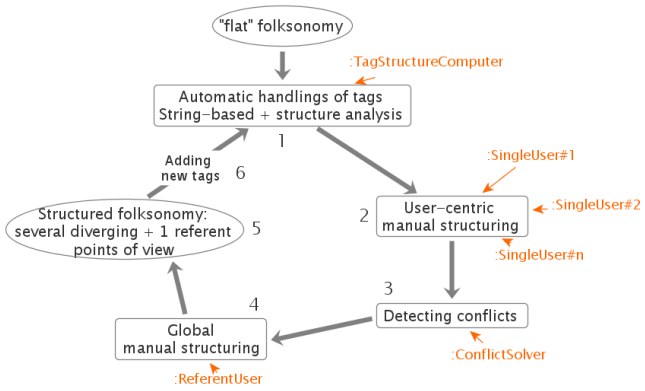


**Figure 2: Folksonomy enrichment lifecycle**

### 4.1.1 *Analyzing tag labels*

String-based similarity metrics are usually applied to tag labels to find spelling variants of tags.

*String-based metrics.* String based distance measures consider the characters string of the label of the entities, here tags, to be compared. [16] used the Levenhstein [11] distance metric to group spelling variant tags such as "new_york" and "newyork"). To go further in the use of this rather simple method,We conducted a benchmark to evaluate the ability of such metrics to retrieve other types of semantic relations such as related relation, or narrower or broader relation, also called hyponym relation. Hyponym relations reflect the relative degree of generality between two notions, such as *e.g* "pollution" is broader than "soil pollution". Two notions are merely related in the other cases, as for instance "energy" and "electricity".

We have compared the similarity metrics implemented in the package SimMetrics which give, for a pair of strings $(s_1, s_2)$, a normalized value between 0 and 1, with a value of 1 meaning that both compared strings are most similar. The similarity metrics we compared can be decomposed into several categories[10]: (a) edit distance based methods, which consider the set of operations needed to turn string $s_1$ into string $s_2$, such as *e.g.* Levenshtein, or Smith-Waterman; (b) token-based methods, which decompose strings into sets of substrings, *i.e* in our case, tokens separated by white space, such as overlap coefficient ; (c) token-based methods using vector representations of strings such as the cosine similarity; and finally (d) other types of metrics such as QGram or Soundex metrics.

*Experiment.* We have extracted a sample from the tags used at Ademe to index their documents and resources. This sample, which mixes freely chosen tags and tags chosen by the archivists, was divided into 4 sets of 22 pairs of tags $(t_1, t_2)$, each set containing tag pairs which correspond to a semantic relation, namely: spelling variant, subsumption, related, and unrelated. These relations have been validated by one member of the Ademe's archivists team so that it reflects on the knowlege of our user's domain.

The Monge-Elkan metric is hybrid metric based on edit distances which also decomposes strings into tokens, and uses a second metric to compare each token with all the others. For our experiment we used a series of 15 metrics and the combination of theses 15 metrics with the Monge--Elkan method (these are nammed for instance Monge-Elkan-_Levenhstein), which makes a total of 30 different metrics.

For all the pairs of tags, we measured their similarity value according to the 30 metrics. To evaluate the performance of each metric in retrieving the pairs that were related, we have computed for each type of semantic relation the recall, precision and the weighted harmonic mean $F_{1,25}$. These values were computed varying (between 0 and 1) the value of the threshold above which a given tag pair is retrieved or not. Then to count the false positives and true positives pair matched we applied the following rules: (a) for the related case the true positves are counted from all sets except the unrelated set, since spelling variant and broader/narrower pairs can be considered also "related"; (b) for the spelling variant and broader/narrower case, the true positives were only those from their corresponding set, and the pairs from all the other sets were counted as false positive. We chose the $F_{1,25}$ measures because we wanted to give slightly more importance to recall than precision in order to capture the fact that in a second stage users can remove false positives but they cannot guess what were the false negatives (the relation we missed) and thus recall is important.

In figure 3 we show, for each type of relation, the mean value and the statistical deviance of $F_{1,25}$ for the top 10 metrics in each category. The result is that the Monge-Elkan_Soundex method outperformed other metrics in the related case. The best for the spelling variant case is the Jaro-Winkler metric, and the best for the broader/narrower case (hyponym) is the MongeElkan_NeedlemanWunch metric. We should also notice the greater deviance in the related case than in the two other cases, and this result was expected (the fact for two notions to be related rarely translates to some terminological similarities e.g. "car" and "wheel" are related but don't even share a single letter).

Now we are interested in finding a way, using these metrics, to differentiate between the 3 types of semantic relations. First, we use the MongeElkan_Soundex metric to retrieve all related tag pairs using a threshold for which the recall is above 0.5. Then, we combine other metrics to distinguish spelling variant and broader or narrower pairs from related pairs.

To distinguish spelling variant from related pairs, we look at the mean value and deviance of the best metric in the spelling variant case. In figure 4 we show the mean value of the Jaro-Winkler metric for the four types of semantic relations. We see that if we choose a threshold above 0.9, we are more likely to retrieve spelling variant pairs. This result is confirmed when we look at the threshold value for which F1.25 is maximum for the JaroWinkler measure in the spelling variant case.

Next, we want to find a way to tell broader or narrower pairs from related pairs. The MongeElkan metrics are not symmetric, and we have calculated, for each tag pair $(t_1, t_2)$, the difference $\delta = s(t_1, t_2) - s(t_2, t_1)$, with $s$ being one of the 15 combination of MongeElkan with another metric. In figure 5 we give the mean value and deviance of $\delta$ for each set of tag pairs and for the MongeElkan_QGram metric which performed best in this respect. We only included in this computation the related tag pairs that were retrieved thanks

[10]For details on each metric and on SimMetrics package: http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

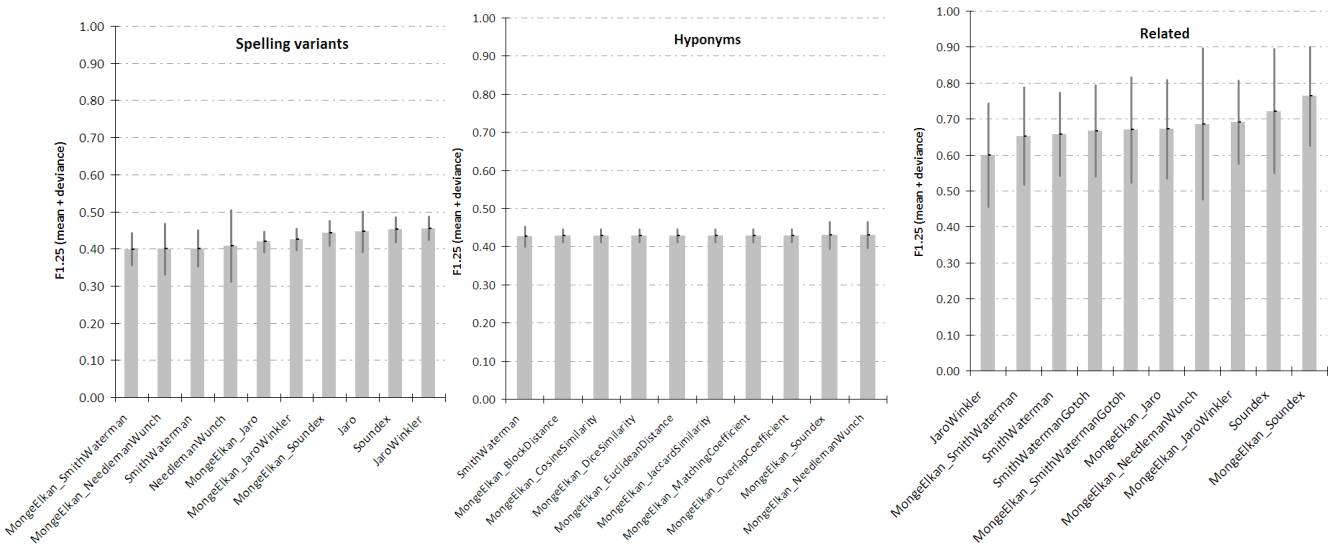


**Figure 3: Performance of metrics for retrieving (from left to right) spelling variants, subsumption, and related semantic relation. Each figure gives the mean F1.25 measure plus the statistical deviance.**

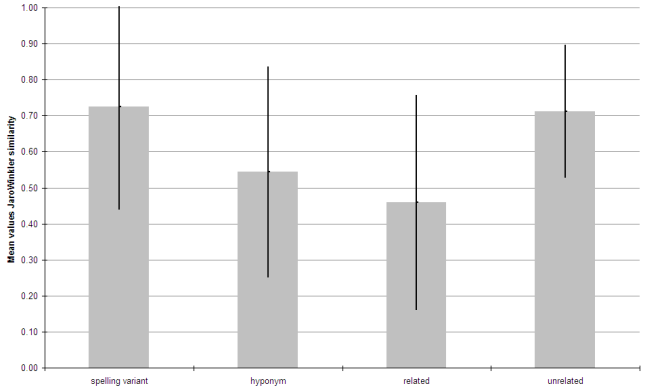


**Figure 4: Comparison of the mean value of the JaroWinkler metric for each type of semantic relation**

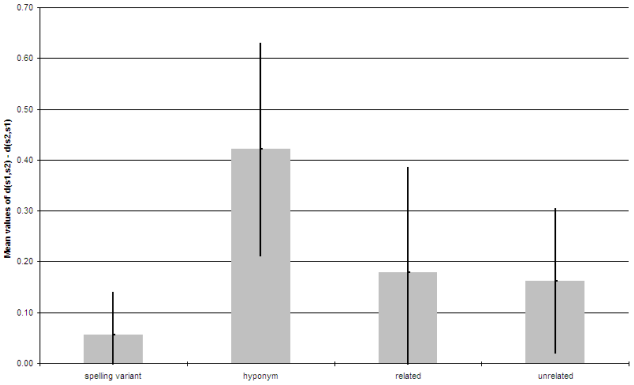


**Figure 5: Mean value of the difference $\delta = s(t_1, t_2) - s(t_2, t_1)$ with $s$ being the Monge-Elkan_QGram metric for each set of tag pairs.**

to the MongeElkan_Soundex metric. We can see that if we set up a threshold above 0.39 (the highest value for $\delta$ when including the deviance), we are able to retrieve tags sharing a subsumtion relation. When taking into account the sign of the difference, we are able to tell the direction of the subsumption relation, meaning that if we have $\delta$ negative and above a certain threshold, then $t_1$ can be considered narrower than $t_2$.

As a result we are able to propose a heuristic to combine the metrics we compared. We first look for pairs of related tags $(t_1, t_2)$ using the Monge-Elkan_Soundex with a first threshold $\tau_a$ so that we have $s(t_1, t_2) \geq \tau_a$. This first threshold is chosen so that the recall is above 0.5, ie $\tau_a = 0.8$ in our case. Then, we compare the JaroWinkler similarity with a second threshold $\tau_b$ to see wether the tags are spelling variants, such that $s(t_1, t_2) \geq \tau_b$. The threshold in this case corresponds to the best precision achieved for the spelling variant case, *i.e.* in our case, 0.94. If it's not the case, we use a third threshold $\tau_c$ and we compute the difference $\delta$ of the MongeElkan_QGram metric $\delta = s(t_1, t_2) - s(t_2, t_1)$, and if $\delta$ is such that $\delta \leq -\tau_c$, then we can infer that $t_1$ is narrower than $t_2$, or if $\delta \geq \tau_c$ then $t_1$ is considered broader than $t_2$. The third threshlod is chosen after the results shown in figure 5 by picking a value above 0.39.

We have applied our heuristic method to the same sample test. However, this heuristic is not directly comparable to the other metric as it combines different methods and retrieve 3 types of semantic relations at a time, while in the global comparison experiment each metric was dealing with one type of semantic relation at a time. In order to evaluate qualitatively the global performance of this heuristical string-based metric, we show in figure 6 the values of the precision and recall for the 3 types of relations. We can clearly see in this figure that string based metrics perform best in the spelling variant case, which confirms a natural intuition since string-based methods were originally designed to match similar strings. Nonetheless, the performance in the hyponym case is noticeable and is explained with the

ability of string-based metrics to easily detect common tokens such as in "pollution" and "soil pollution" and this cases often corresponds to a hyponym relation. The related case is more difficult to retrieve (hence the low recall) as the relation is the most fuzzy and probably the least noticeable in the actual spelling of the tags ("sun" and "energy" *e.g*). Finally, this indicates the need to use other methods to be able to cover other cases where semantically related tags are not morphologically similar.

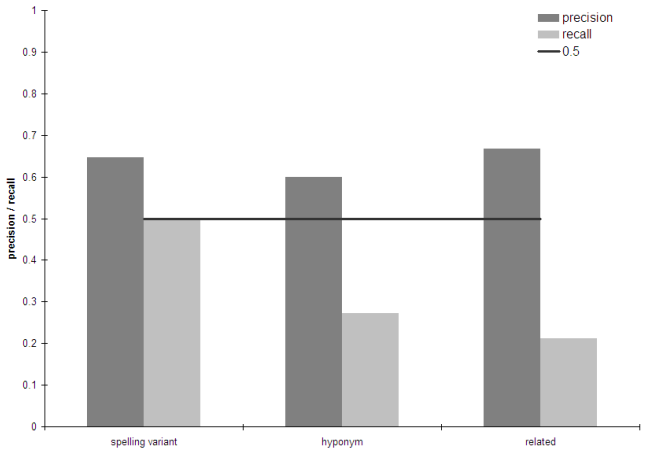


**Figure 6: Performance of the heuristic string-based metric.**

### *4.1.2 Analyzing the structure of folksonomies*

In this section we detail our implementation of the methods to extract emergent semantics which analyze the tripartite structure of the folksonomy.

In order to extract *related* relationships between tags, we use the similarity measure based on distributional aggregation in the tag-tag context[5], This means that we look at the coocurrence patterns of tags. To compute this similarity, we first construct the coocurrence matrix whose rows and columns correspond to the list of all tags, and each cell $(i, j)$ has for value the frequency of coocurrence for the tag pair $(t_i, t_j)$. Then the vector representation $v_i$ of each tag $t_i$ in this context corresponds to each row of the coocurrence matrix. The similarity value for a pair of tag $(t_i, t_j)$ in the tag-tag context is then given by the cosine distance between the vectors $v_i$ and $v_j$: $\cos(v_i, v_j) = \frac{v_i.v_j}{\|v_i\|_2.\|v_j\|_2}$. When this value is above a given threshold, we create an annotation which says that tag $t_i$ is related to tag $t_j$.

So as to extract subsumption relations, we made use of the method described by [14] which consists in looking at the inclusions of the sets of users associated to a tag. Let $S_i$ be the set of users using tag $t_i$, and $S_j$ be the set of users using tag $t_j$. We introduce an overlap coefficient defined as $overlap(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i|}$. Then if these two sets are not empty and if we have $\tau_{sub} \leq overlap(S_i, S_j) \leq 1$ with $\tau_{sub}$ a given threshlod, then we can infer that the tag $t_i$ is broader than the tag $t_j$, and conversely, if we have $\tau_{sub} \leq overlap(S_j, S_i) \leq 1$, then we can infer that the tag $t_j$ is broader than the tag $t_i$.

In terms of algorithmic complexity, these two types of computation are relatively costly and, overall, not incremental since we have to analyze the whole foksonomy to compute the similarity of newly added tags.

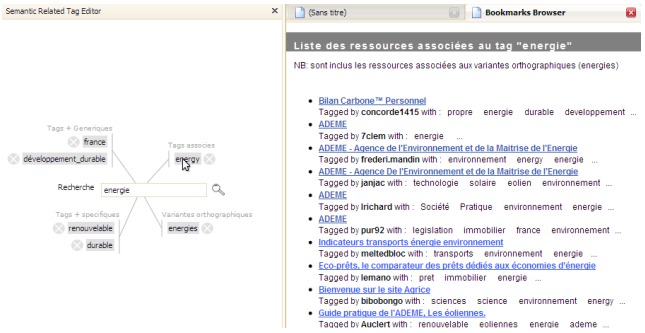


**Figure 7: Firefox extension seemlessly integrating tag structuring capabilities which can be sees in the left part. The user was about to drag the tag "energy" towards the "spelling variant" area to state that the tag "energie" (the tag currently searched for) is spelling variant of "energy". On the right side are displayed the resources associated to the current tag.**

## 4.2 Capturing users contributions

Once we are able to support diverging points of view, we want to allow users to contribute to the semantic structuring of the folksonomy while keeping as low as possible the cognitive overload that this task may involve. To achieve this goal we propose integrating simple and non-obtrusive structuring functionnalities within everyday tasks of users. For instance, in our target comunity at Ademe, we want to be able to capture the expertise of the enginneers when they browse the corpus of Ademe resources.

Our proposal consists in an interface for navigating the folksonomy in which tags are suggested and ordered according to their semantic relations with the current tag searched (see figure 7). Related and spelling variant tags are positioned on the right side (repectively top and bottom corner) and broader and narrower tags are positioned on the left side (respectively top and bottom corner). Optionaly, users can either merely reject a relation by clicking on the cross besides each tag, or they can correct a relation by dragging and dropping a tag from one category to another.

## 4.3 Maintaining a coherent view for all users

Up to this point we presented the different ways to compute tag relations, to capture them and to keep track of the diverging points of view from all users. Now we are going to see how these points of view are sorted out and arranged together in a coherent system.

### *4.3.1 Collecting contributions from computers and humans*

Automatic agents detect semantic relationships that are linked to the corresponding type of agent `TagStructure-Computer`. These automatic handlings are performed during low activity periods of time due to their algorithmic complexity (for instance, tag data from Ademe contains around 8000 distinct tags, and our heuristic string-based metric uses an average of 0.863 ms to handle a pair of tag, so that it takes approximately 7.67 hours to handle all tags from Ademe).

Then users of the system, modeled with the `SingleUser`

class, can contribute with their own point of view regarding semantic relations between tags. Each user benefits from other users contributions while her own point of view is maintained by the system. However, from a global point of view, some logical inconsitencies may arise, due to conflicts between some user's points of view.

#### 4.3.2 Solving conflicts and creating a consensual point of view

A third type of agent is introduced, named `ConflictSolver` and wich looks for conflicts emerging between all user's point of view. A conflict in the structured folksonomy emerges when two tags are linked with different relations as, for instance, when "pollution" is narrower than "co2" for a number $n_1$ of users, but for a number $n_2$of users "pollution" is broader than "co2". In this case the conflict solver will compare the ratios $r_1 = \frac{n_1}{(n_1+n_2)}$ and $r_2 = \frac{n_2}{(n_1+n_2)}$ with a given threshlod $\tau_{cs}$. If one of these ratios is above $\tau_{cs}$, then the conflict solver will approve the corresponding statement. For example if $r_2 > \tau_{cs}$, the conflict solver will approve the statement "pollution" is broader than "co2". Else, if both $r_1$ and $r_2$are below $\tau_{cs}$, this means that no strong consensus has already been reached, and the conflict solver will merely says that "pollution" and "co2" are related since this relation is the loosest and represents a soft compromise between each diverging point of view.

The users who have already rejected the narrower or broader relations will not be impacted as the relations they have rejected are more specific, and the corresponding RDF properties (`skos:narrower` and `skos:broader`) are declared in our system as subproperties of `skos:related`. Thus, thanks to inference mechanisms, their rejection of either the narrower or broader statements will be propagated to the related statements proposed by the conflict solver.

The fourth type of agent we introduced is the `ReferentUser`. The referent user will be able to approve, reject or correct all the relations already existing in the structured folksonomy in order to maintain its own and consensual point of view. The conflict solver mechanism will assist the referent user in her task by pointing out the conflicts already existing in the collaboratively structured folksonomy. Then, all the statements that the referent user has already treated will be ignored in further passes of the `ConflictSolver`.

#### 4.3.3 Exploiting and filtering points of view

At this stage of the process, we obtain a folksonomy semantically structured via several points of view, among which a global and consensual point of view emerges. We present in this section the strategies we propose to exploit these points of view in order to offer a coherent experience to all users of the system.

The consensual point of view can be used to generate a hierarchical tag cloud from the folksonomy where broader tags are printed in bigger fonts than narrower tags. This type of tag cloud may be useful to guide the users in giving him a panoramic view of the content of the folksonomy and can be presented at a starting point of the navigation, giving the broadest tags as bigger, and then, along the search, giving the semantic surrounding of the current tag by showing broader and narrower tags.

As a result, the folksonomy is structured through several semantic statements made about the tags by different types of agents. These types of agent are used to filter out the tags suggested while searching the folksonomy using our interface. By keeping track of the type of agents associated to each statement, we are able to give a priority to the suggested tags corresponding to these statements when a user u searches a tag t. The following priority order is given:

(1) all statements $S_u$ approved by the user $u$.

(2) all statements $S_{ru}$approved by the `ReferentUser`, except if they conflict with one from $S_u$.

(3) all statements $S_{cs}$ approved by the `ConflictSolver`, except if they conflict with one from $S_u$ or $S_{ru}$.

(4) all statements $S_{ou}$ approved by other users, except if they conflict with one from $S_u$, $S_{ru}$, or $S_{cs}$

(5) all statements $S_{tc}$ approved by the `TagStructureComputer`, except if they conflict with one from $S_u$, $S_{ru}$,$S_{cs}$, or $S_{ou}$.

This set of rules allows, when suggesting tags to a user during a search, to filter out the conflicting or more general points of view from the other contributions, coming from humans or machines. For example, if the user is searching the tag "energy", the system will first suggest tags coming from assertions she has approve, e.g. the user had approved that "energies" was a spelling variant of "energy". Then, the system will suggest tags coming from assertions that the `ReferentUser` has approved and that do not conflict with the ones approved by the user. For instance if the `ReferentUser` had approved that "energies" is related to "energy", this assertion will not be included, and so forth, following the priority order described above. As a consequence, it allows each user to benefit from the other users contributions while preserving a coherent experience using a referent point of view or, when this one is absent, using the conflict solver.

## 5. CONCLUSION AND DISCUSSION

In this paper, we have presented our approach to the semantic enrichment of foksonomies. As a complement to similar works, our approach supports conflicting points of views all along the life-cycle of the semantically enriched folksonomies. We propose a socio-technical system in which automatic agents help users in maintaining their personnal points of view while still benefiting from others contributions, and also help referent users in their task of building a consensual point of view. Our approach is grounded on a carefull usage analysis of our target communities which allows us to take the benefit of their daily activity to contribute to the process.

In order to bootstrap the process, we make use of automatic handlings of folksonomies which extract the emergent semantics. Automatic handling consists in analysing the labels of tags using string-based metrics, or the structure of folksonomies. To this regard, we proposed in this paper an evaluation of the main string-based methods in order to: (a) motivate the choice of the metrics performing best in our context; and (b) evaluate the ability of such metrics to differentiate the semantic relations typically used in thesaurus, ie. to be able to tell when two tags are merely related, or when one tag is broader or narrower than another tag, or when two tags are spelling variants of the same notion. As a result we proposed a heuristic metric which perfoms this task. This heuristic metric performs best for detecting spelling variants, as expected, but also gives interesting results for subsumption relations in cases such as "pollution" which is broader than "soil pollution". We have also quantitively shown that the approaches analysing the

structure of folksonomies are necessary to retrieve semantic relations when tags sharing semantic relations are not morphologically similar, even if they are more costly and not incremental, unlike string based methods. Among this second type of approaches, we have used the similarity measure based on distributional aggregations in the tag-tag context to compute related relations, and the approaches proposed by [14] to compute subsumption relations.

In order to capture diverging points of view in the semantic structuring of folksonomies, we proposed a formal ontology which makes use of named graphs to describe semantic relations between tags. The points of view of users are then attached to these asserted relations. By describing the different classes of agents who propose or reject asserted relations, we are able to model a complete life-cycle for a collaborative and automatically assisted enrichment of folksonomies. (1) This cycle starts with a flat folksonomy which is first analysed by automatic agents which propose semantic relations. (2) The users can contribute and maintain their own point of view by validating, rejecting, or proposing semantic relations thanks to a user friendly interface integrated in a navigation tool. (3) The conflicts emerging from these points of view are detected and (4) utilized to help a referent user to maintain a global and consensual point of view. (5) The result of this process is a folksonomy augmented with semantic assertions each linked to different points of view coexisting with a consensual one, and (6) the cycle restarts when new tags are added or when relations are suggested or changed. The semantic assertions are used to suggest tags when navigating the folksonomy, and a set of rules allows to filter the semantic assertions in order to offer a coherent experience to the users.

Our approach is currently tested at the Ademe agency to enhance the browsing of its corpus available online to members of the agency and to the public. In this context the expert engineers of Ademe maintain their points of view so as to reflect on their expertise on a given domain. In a second time, the archivists (our referent users) are assisted in the task of enriching with new tags and semantically structure their global point of view from the collaborative enrichment of the folksonomy.

Our future work includes a testing campaign among the users of Ademe of our approach. We also plan on exploiting the semantic relations between tags at tagging time to guide and help users provide for more precise tags, but also to provide for additional input material for semantic social network analysis [6]. We envision in this respect to propose a novel approach to indexing where users and professional indexers, such as the Ademe's archivists, are engaged in a fruitful collaborations leveraged by a tailored automated assistance.

## Acknowledgments

This work is funded by ANR-08-CORD-011-05.

## 6. REFERENCES

[1] S. Angeletou, M. Sabou, and E. Motta. Semantically enriching folksonomies with flor. In *CISWeb Worksh. at Europ. Semantic Web Conf.*, 2008.

[2] N. Aussenac-Gilles, B. Biébow, and S. Szulman. Corpus analysis for conceptual modelling. In *EKAW - Worksh. Ontologies and Texts*, 2000.

[3] S. Braun, A. Schmidt, A. Walter, G. Nagypál, and V. Zacharias. Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In *CKC*, volume 273 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[4] J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *Int. Conf. World Wide Web*, pages 613–622, New York, NY, USA, 2005. ACM.

[5] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. *Int. Semantic Web Conf.*, 2008.

[6] G. Ereteo, M. Buffa, F. Gandon, and O. Corby. Analysis of a real online social network using semantic web frameworks. In *Proc. Int Sem Web Conf, Washington, USA*, 2009.

[7] F. Gandon, V. Bottolier, O. Corby, and P. Durville. Rdf/xml source declaration, w3c member submission. http://www.w3.org/Submission/rdfsource/, 2007.

[8] P. Heymann and H. Garcia-Molina. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical report, Stanford InfoLab, 2006.

[9] Benjamin Huynh-Kim Bang, Eric Dané, and Monique Grandbastien. Merging semantic and participative approaches for organising teachers' documents. In *Proc Conf. Educational Multimedia, Hypermedia & Telecommunications*, pages p. 4959–4966, Vienna France, 07 2008.

[10] Golebiowska J. *Exploitation des ontologies pour la memoire d'un projet-vehicule - Methode et outil SAMOVAR.* PhD thesis, Univ. Nice-Sophia Antipolis, 2002.

[11] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, 10(8):707–710, February 1966.

[12] Freddy Limpens, Alexandre Monnin, David Laniado, and Fabien Gandon. Nicetag ontology: tags as named graphs. In *International Workshop in Social Networks Interoperability, ASWC09*, 2009.

[13] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *World Wide Web Conf.*, pages 641–641, April 2009.

[14] P. Mika. Ontologies are Us: a Unified Model of Social Networks and Semantics. In *Int. Sem. Web Conf.*, volume 3729 of *LNCS*, pages 522–536. Springer, 2005.

[15] A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proc. Worksh. Linked Data on the Web, Beijing, China*, Apr 2008.

[16] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. *Europ.Sem. Web Conf.*, 2007.

[17] V. Tanasescu and O. Streibel. ExtremeTagging: Emergent Semantics through the Tagging of Tags. In *ESOE at ISWC*, November 2007.

[18] C. Van Damme, M. Hepp, and K. Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007)*, pages 57–70, 2007.