Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

WHERE ARE THE CONCEPTS IN THE FOLKSONOMY MODEL?

Pierre Andrews, Juan Pane
and Ilya Zaihrayeu

December 2010

Technical Report # DISI-10-066

Where are the Concepts in the Folksonomy Model?*

Pierre Andrews, Juan Pane, and Ilya Zaihrayeu

Department of Information and Communication Technology
University of Trento
Italy
{andrews,pane,ilya}@disi.unitn.it

Abstract. The folksonomy annotation model has become very popular on the web and is now used for diverse research, in particular in distributional semantics. Currently, this model links tags, resources and users in a tripartite graph where the atomic units of meaning are the tags. However, tags might not be atomic concepts (for example "sunny italy") and thus the current folksonomy model might not be optimal for semantic applications. In this paper we illustrate how this formalisation would gain in being extended to discriminate between tags and concepts. We do this by studying the meanings of tags in a del.icio.us dataset and we propose an extension to the tripartite graph folksonomy model.

1 Introduction

One of the cornerstones of what we now call the "Web 2.0" is unconstrained user collaboration and creation of content. Some of the first sites to allow such features were del.icio.us and Flickr where users could share resources – bookmarks and photos respectively – and freely annotate them. Both websites allowed the creation of so called *folksonomies*: social classification of resources created by the community that have shown to be very important for organising the large amount of content online, but also for, later on, studying the collaborative creation of shared vocabularies and lightweight ontologies.

These folksonomies are now widely studied, in particular with the model of tripartite graphs of *tags-users-resources*. However, in this model, *tags* are free-form terms with no explicit semantic, therefore a number of issues arise from their use, such as:

- the loss in precision due to the ambiguity of tags – for example, the tag "java" can refer to the "Indonesian island", the "programming language", and a "beverage".
- the loss of recall due to the synonymy of terms – for instance, if you search for the tag "travel", you might be interested by the results for the tag "journey".

* This work has been partly supported by the INSEMTIVES project (FP7-231181, see http://www.insemtives.eu).

The use of different forms of the same word also exacerbate these issues as some users would, for example, use the tag "running", others would use instead "run", "runs", "torun", etc.

In the Semantic Web field, the folksonomy model has already been formalised and expressed with dedicated ontologies. In the meantime, knowledge organisation systems have been formalised to create organisations of concepts to express their linguistic meaning to the user. It is our opinion, as we demonstrate in this paper, that the current models of folksonomy could be extended to include a stricter representation of the tags folksonomy. This would then allow a better understanding of the semantics used by the creators of the tags and to provide better quality of service over system based on folksonomies.

In this paper, we first summarise the existing formalisation of folksonomies in Section 2. In Section 3 we argue for the use of a stricter representation of the concepts described by tags. To illustrate the need for such extension to the model, we propose a case study of a delicious dataset – described in Section 4 – and a detailled study of some features of the folksonomy without tags to concept mappings followed by a study of a disambiguated subset of such dataset in Section 5 and 6 respectivelly.

2 Current Modelling of Folksonomies

The term folksonomy was coined in 2004 by T. Vander Wal [1] who characterised the new social tagging web sites that were appearing at the time. He defined a folksonomy as "*the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval*". This "result" is one of the simplest form of annotation of resources with metadata that can serve to help the indexing, categorisation or sharing of such resources: a tag annotation.

Mika [2] introduced a formalisation of this results to ease its processing in multimodal graph analysis. Doing so, the author enables the formal representation of the social network resulting from the folksonomy building activity. Mika represents a folksonomy as a tripartite graph composed of three disjoint types of vertices, the *actors* A (the user creating the tag annotation), the *concepts* C (tags, keywords) used as metadata and the *objects* O or resources being annotated. A tag annotation is thus a triple combining the three vertices:

$$T = \langle u, t, r \rangle \text{ where } u \in A, t \in C \text{ and } r \in O$$

According to Mika, such tripartite graph can be used to describe an ontology representing the knowledge of the community that created this folksonomy. This model has been used since to exploit different social networking analysis tools and distributional semantic models to extract a more formal representation of the semantic knowledge encoded in these tripartite graphs.

3 The Semantics of Tags

For the sake of clarify, we define the *semantics* of a word as an explicit mapping from lexical atomic units of the word (such as its tokens) to elements of an

ontology defined in a formal language suitable for automated machine processing. This ontology can be defined at different levels of expressivity and formality and can vary from a lightweight formalization of classification hierarchies (e.g., see [3]) to rigorously formalized logical theories (see [4] for a report on the ontology kinds).

An important point in Mika's [2] description of the folksonomy model is that "tags" are considered to be mapped one-to-one to the *concepts* of the ontology and that these are the semantic units of the language used in the community that created the folksonomy. However, we believe that a more granular model has to be used to represent the conceptual part of folksonomies. This will enable a better understanding of its underlying semantic and of the overlap of vocabularies between the users of the folksonomy.

In fact, tags and keywords, while they represent a specific concept and have a known semantic for the agent that creates them, are just stored and shared in the folksonomy as purely free-form natural language text. Because of the ambiguous nature of natural language [5], a number of issues arise when sharing only the textual version of the annotations:

Base form variation This problem is related to natural language input issues where the annotation is based on different forms of the same word (e.g., plurals vs. singular forms, conjugations, misspellings) [5] or to word combinations such as "sunny italy".

Homography Annotation elements may have ambiguous interpretation. For instance, the tag "Java" may be used to describe a resource about the *Java island* or a resource about the *Java programming language*; thus, users looking for resources related to the programming language may also get some irrelevant resources related to the Island (therefore, reducing the precision);

Synonymy Syntactically different annotation elements may have the same meaning. For example, the tags "image" and "picture" may be used interchangeably by users but will be treated by the system as two different tags because of their different spelling; thus, retrieving resources using only one of these tags may yield incomplete results as the computer is not aware of the synonymy link;

Specificity gap This problem comes from a difference in the specificity of terms used in annotation and searching. For example, the user searching with the tag "cheese" will not find resources tagged with "cheddar[1]" if no link connecting these two terms exists in the system.

One suitable formalism for the representation of the semantics of tags is the one defined by Description Logics (DL) [6]. Among other things, this formalism introduces the notion of a *concept*, whose semantics (or, extension) is defined as the set of elements (or, instances). For example, the extension of the concept `Person` is the set of people existing in some model (e.g., in the model of the world). Because they are defined under a set-theoretic semantics, operators from the set theory can be applied on concepts, e.g., one could state that concept

[1] which is a kind of cheese

`Organism` *subsumes* (or, is more general than) the concept `Person` because the extension of the former concept is a superset for the extension of the latter concept. Among other things, the subsumption relation can be used for building taxonomies of concepts similar to knowledge organisation system such as the one proposed in SKOS [7]. These properties lead to a number of useful reasoning capabilities such as computing the instances of concepts through the concept subsumption, computing more specific or general concepts – these capabilities can be used for building services for the end users such as semantic search, as shown, for example, in [8]. A more complete introduction to DL is out of the scope of this article; interested readers are referred to [6] for details.

We adopt the approach reported in [3][2] in order to apply the DL formalism for the explicit codification of the semantics of tags. Namely, recognised adjectives and nouns in tags are converted into concepts whose extension is defined as the set of resources which are about objects that possess the properties denoted by adjectives or about the objects that are described by nouns. For example, the extension of the concept `Fast` is the set of resources about fast objects, and the extension of the concept `Car` is the set of resources about cars. Note that in this model named entities can be represented as concepts too and not as concept instances. For example, the extension of concept `Italy` is the set of resources about the "Italian Republic".

With the proposed model, tags consisting of several words can be converted into DL conjunctive formulas that codify the semantics of the tags. For example, the tag "fast cars" can be converted into the formula ($\texttt{Fast} \sqcap \texttt{Car}$) that results into the intersection of two sets of resources – those about fast objects and those about cars, i.e., those about fast cars. The procedure that encode a natural language phrase into a DL formula are not trivial and require part-of-speech detection, lemmatisation, word sense disambiguation, coordinating conjunction disambiguation, and other algorithms. For the sake of simplicity, in this article we are not discussing these algorithms; interested readers are referred to [9].

These tag formulas can be reasoned about to find semantically related tags using logical reasoning. For example, the formula ($\texttt{Fast} \sqcap \texttt{Car}$) results to be more specific than the formula ($\texttt{Fast} \sqcap \texttt{Vehicle}$) if the knowledge base contains the fact that the concept `Car` is more specific than the concept `Vehicle`. It enables a number of useful services, for example, a user query "fast vehicles" would return resources annotated with "fast cars" if the concepts in the query and annotation formulas were disambiguated properly by the user or by the system. Note that more than one word can be mapped to the same concept, e.g., the word "auto" can be disambiguated to concept `Car`; thus, the above mentioned query can also return resources annotated with "fast autos". This allows us to address the known problem with search related to the inconsistent use of polysemous and synonymous terms in queries and data, as discussed in [8].

We thus introduce two new formalisations in the model to create a fourpartite graph representing the user-resource-tag-concept link:

[2] In this works DL logics were used to codify the meaning of web directory labels.

– A *controlled tag* ct is a tuple $ct = \langle t, \{lc\}\rangle$, where t is a term, i.e., a non-empty finite sequence of characters normally representing natural language words or phrases such as "bird", "sunnydays" or "sea"; and $\{lc\}$ is an ordered list of linguistic concepts, defined as follows:
– A *linguistic concept* lc is a tuple $lc = \langle c, ct\rangle$, where c is a concept as defined in DL (see above); and ct is a term in a natural language that denotes the concept c.

Consider an example of a controlled tag: $ct = \langle$"$sunnydays''$, $\{lc_1, lc_2\}\rangle$, with $ls_1 = \langle$`Sunny`, "$sunny''\rangle$ and $ls_1 = \langle$`Day`, "$days''\rangle$.

Recall the syntactic folksonomy model definition (see Section 2 that we now extend to the definition of a controlled tag annotation, T^C:

$$T^C = \langle u, ct, r\rangle \text{ where } u \in A, ct \textit{ is a controlled tag}, \text{ and } r \in O$$

4 A Case Study Dataset

To study our model we analyse a subset of the widely used del.icio.us[3] folksonomy.

del.icio.us is a simple folksonomy as was defined by [1] and formalised by [2] in that it links resources to users and tags in a tripartite graph. However, these tags are totally uncontrolled and their semantic is not explicit. In the current datasets, for instance the ones provided by Tagora[4] or listed in [10], no-one has yet, to the best of our knowledge, provided a golden standard with such semantics. In that, the del.icio.us dataset is not perfectly what we are looking for, the Faviki[5] website could provide such dataset, however it does not contain so many users and annotations as del.icio.us and the quality of the disambiguations is not guaranteed. To make the del.icio.us dataset fit our problem statement, we have thus decided to extend a subset of a del.icio.us dump with disambiguated tags by manual validation. We used WordNet 2.1 [11] as the underlying ontology for finding and assigning senses for tag tokens.

4.1 del.icio.us Sample

We obtained the initial data from the authors of [12] who crawled del.icio.us between December 2007 and April 2008. After some initial cleaning the dataset contains *5 431 804* unique tags (where the uniqueness criteria is the exact string match) of *947 729* anonimized users, over *45 600 619* unique URLs on *8 213 547* different website domains. This data can be considered to follow the syntactic folksonomy model $\langle t, r, u\rangle$ where the resource r is the URL being annotated, containing a total of *401 970 328* tag annotations.

To study the semantic used in these tags, we have thus decided to extend a subset of the data with disambiguated tags; i.e., convert $t \rightarrow ct$ (See the previous

[3] `http://del.icio.us`
[4] `http://www.tagora-project.eu/data/`
[5] `http://faviki.com/`

Section). This means that for each tag t in this subset, we have explicitly split it in its component tokens and marked it with the WordNet synset (its sense) it refers to and thus get to the semantic folksonomy model described in Section 2.

The golden standard dataset we have built includes annotations from users which have less than 1 000 tags and have used at least ten different tags in five different website domains. This upper bound was decided considering that del.icio.us is also subject to spamming, and users with more than one thousand tags could potentially be spammers as the original authors of the crawled data assumed [12]. Furthermore, only $\langle r, u \rangle$ pairs that have at least three tags (to provide diversity in the golden standard), no more than ten tags (to avoid timely manual validation) and coming from users who have tags in at least five website domains (to further reduce the probability of spam tags) are selected. Only URLs that have been used by at least twenty users are considered in the golden standard in order to provide enough overlap between users. After retrieving all the $\langle r, u \rangle$ pairs that comply with the previously mentioned constraints, we randomly selected 500 pairs. We thus obtained 4 707 tag annotations with 871 unique tags on 299 URLs in 172 different web domains.

The validation application for creating this dataset is available as open source code on the sourceforge repository of INSEMTIVES[6] and the first batch discussed here has been distributed as a LOD RDF dataset with the schema presented in Section 8 at `http://disi.unitn.it/~knowdive/dataset/delicious/`. This RDF export will grow as we extend the dataset with new samples of del.icio.us.

5 Considerations on the Raw Folksonomy

del.icio.us is used in many research groups that work on folksonomies as a large dataset showing how users use tags to organise and share their resources. We have thus started by a basic analysis of how users used tags in the dataset and what we could observe from this. In the following paragraphs, we discuss the analysis that we performed on the whole dataset of 45 600 619 URLs, with all the users and tags available. The analyis and first conclusion on the manual disambiguation batch of 500 $\langle URL, u \rangle$ pairs is discussed in the next section.

While the annotation task on del.icio.us is quite simpler as it does not require the specification of semantics, we can already see that the users are not motivated to provide a large amount of annotations. Note that we cannot make any conclusions on why this might be the case as this would require a direct users study, however, as illustrated by Figure 1, we can see that in 35.5% of the cases, users use only one tag per bookmark and only in 12.1% of the cases they would add more than five tags per bookmark.

This might be because each user only uses very specific tags to classify/categorize the bookmark and thus does not require many indexing terms to find the resource in the future. This assumption would be a "dream" scenario as it would

[6] `http://www.sourceforge.net/projects/insemtives/`

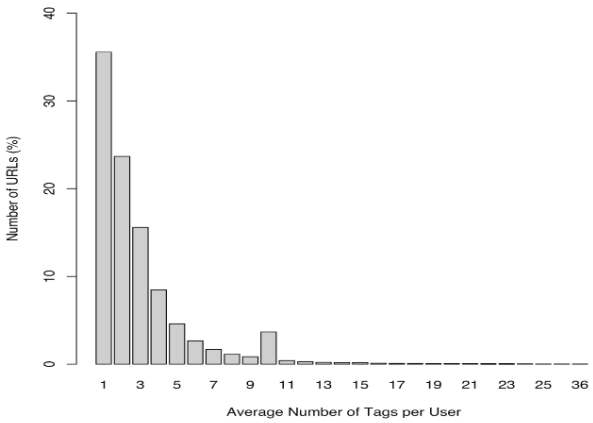


Fig. 1. Number of Tags per URL per User

mean that the users are already ready to provide very specific descriptors for their resources and if they are linked to the underlying ontology, we can retrieve them using synonymous and/or more general terms very easily. However, it might just be that the users are not interested in adding more tags as they do not see the value of adding many indexing terms for future retrieval.

An interesting point is that there is an out-of-the-norm peak at ten tags per bookmark that seems too strong to be coincidental. We have not yet studied in details why this happens but hypothesise that it might be created by spambots providing a lot of bookmarks with exactly ten tags.

In Figure 2, we consider another interesting feature of the tagging behaviour of users on del.icio.us. While an often used assumptions in folksonomy based algorithms is that we can learn a lot from tag collocations on different resources, we can see that users do not often reuse the same tag more than once.

In fact, from our analysis, in 73% of the cases, a tag is used only once on the whole set of bookmarks by a single user. This means that in most cases, a tag will not be found located on different resources, at least not by the same user. Only in 7.3% of the cases a tag is reused on more than seven resources.

This might support our previous assumption that the users use very specific tags when they annotate resources and thus they do not use them on multiple documents. However, this might create difficulties when sharing knowledge between users as they might not use the same vocabulary (as they use very specific/personal terms). It might also impair the ontology learning algorithms [13] that are based on the measure of collocation of tags.

When annotating shared goods such as web pages, if their is no agreement between the users on what the resource means, it is difficult to reuse these annotations to improve search and ranking of resources. It is also difficult to learn the meaning of the resource or of the annotations attached to it. We have

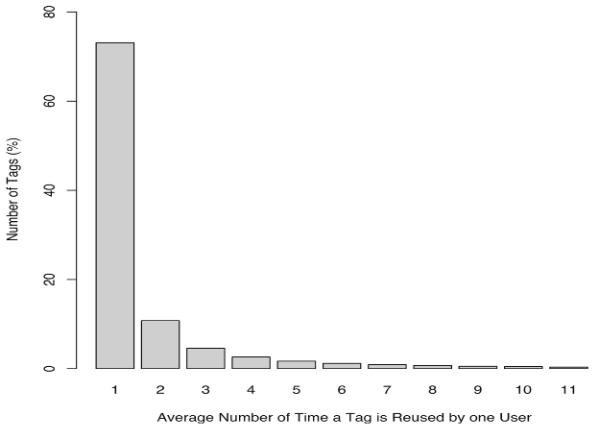


Fig. 2. Number of Time a Tag is Reused by the same User on all the Bookmarks

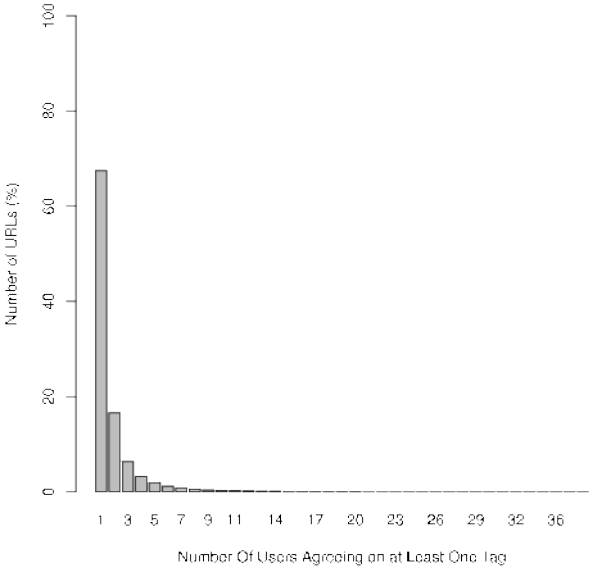


Fig. 3. Average Agreement on Tags for the same Resource

thus done a preliminary analysis of the general agreement of the users in the del.icio.us dataset when they tag a resource. Here we are interested to see how many tags are used by more than one user on the same resource.

To do this, we have adopted a non chance corrected measure of agreement where we count how many users have used the same tag on the same resource. For instance, if there is user U_1 who tagged a resource R_1 with T_1 and T_2 while user U_2 tagged this resource with T_3 and T_4, then there is only one user using any of the four tags. If U_3 tagged R_2 with T_5 and T_6, U_4 tagged it with T_6 and T_7 and U_5 with T_8 and T_9, then there are two users agreeing on at least one tag for that resource. Note that only URLs that are bookmarked by at least two users are considered.

Figure 3 shows the results of this measure. In 67.5% of the cases, there is only one user "agreeing" on at least one tag, which means that every users used different tags on the same resources. In only 9.3% of the cases more than three users agreed on at least one tag.

In a sense this is a good result in that users do provide very diverse tags for the same resource and thus we can learn more about the resource itself. However, if there is no agreement between the users, it is difficult to consider that tags are valid as they might be very personal or subjective.

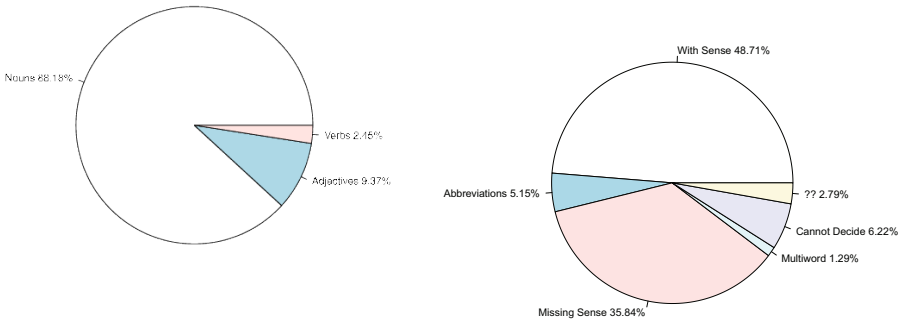
It is interesting to note that these percentages apply on millions of tags, resources and users and in this, a small percentage still represent a large mass of resources and users on which automatic semantic extraction algorithms can be applied. Also, these figures were computed without any preprocessing of the different forms of tags, or without their disambiguation. As we show in the next section, this might be an important factor for the lack of overlap of tags between resources and users that we are seeing.

6 Features of the Dataset when Linking to Concepts

While in the previous section we discussed some observations that could be made on the uncontrolled tags, we have developed a subset of these uncontrolled tags that are cleaned and disambiguated to a controlled knowledge organisation system (WordNet). It is thus interesting to analyse this subset to see the tagging behaviour when tags are disambiguated to the terms in an ontology. In the following paragraphs we present some first conclusions on the use of an ontology and how it maps to the users' vocabulary. In the following analysis, we only consider entries that were validated and agreed upon by two validators.

6.1 Use of Nouns, Verbs and Adjectives

In a previous study, [14] points out that the users of del.icio.us tend to use mainly nouns as descriptors of the urls. In the current dataset we have a validated sense (with all its metadata provided by WordNet) for each term and thus we can easily reproduce such observation.



a) Distribution of Part of Speech on the validated Tokens

b) Distribution of Ignored Tokens (part of a Tag)

Fig. 4. Properties of the Tokens in the del.icio.us dataset

Figure 4a) shows that we can come to the same conclusions as [14]. In fact, nouns are used most of the times (88.18%) while verbs and adjectives, while being used sometimes cannot be found in great numbers in the annotations.

Note that Adverbs seem to be never used, at least in the sample of del.icio.us that we are studying.

6.2 WordNet as an Underlying Ontology

While disambiguating the tags to a sense in WordNet, the manual annotators could decide that no sense provided by the ontology was adequate to express the sense meant by the user. For example, the tag "ajax" was found in the dataset and usually referred to the ajax technology used in web applications[7]. However, the only sense present in WordNet for this tag is "a Greek hero".

As shown in Figure 4b), the case of the missing sense happened in 35.8% of the cases. However, the validators were able to find a matching sense in WordNet for 48.7% of the terms used in the validated batch. For diverse reasons (the users use abbreviations, there is no sense in WordNet, etc.) less than half of the vocabulary used by the users can be mapped to WordNet.

This is an important observation as it shows the inadequacy of fully automatic folksonomy processing systems based on fixed knowledge organisation systems such as WordNet. For instance, if we consider the issue of Word Sense Disambiguation (WSD), the state-of-the-art tools cannot often achieve more that 60% accuracy. However, given the fact that only half of the terms from our dataset can be found in a vocabulary such as WordNet, from the end user

[7] `http://en.wikipedia.org/wiki/Ajax_(programming)`

Fig. 5. Decrease in the Amount of ambiguities after pre-processing and after sense disambiguation

perspective, it means that user will be suggested the right sense for a given tag token in less than 30% of the cases.

6.3 Sense Disambiguation

One of the issues presented in the raw analysis we discussed in Section 5 is that there is not a great agreement between users in the tags they use and there is not a great overlap in their personal vocabularies. One of the hypothesis for this is that there are many lexical variations of the same term that cannot be matched without preprocessing the tags (for example, "javaisland", "java_island", "java" and "island", etc.) and as we have already discussed earlier, there are different terms that can be used for the same concept (for example, "trip" and "journey").

In the validation process for the batch, we have actually cleaned all these issues by collapsing different lexical variations and linking them to their relevant concepts. We can thus evaluate the amount of ambiguity that is added by these different type of variations.

Figure 5 shows a summary of this decrease in ambiguity when going from *tags* – that can represent the same word in different forms – to *tokens* – that are preprocessed tags collapsed to the normal form of the world – and then to *synsets* – that disambiguate the meaning of the tag. The top bar represents the number of tags we started from (742), the middle bar represents the number of tokens to which they collapse (265) and the bottom bar represent the number of synsets from WordNet to which these tokens can be mapped (258).

We can thus see that by preprocessing alone (splitting and lemmatazing tags), the vocabulary size shrinks by 64.7%, thus reducing the ambiguity of the annotations significantly without the need to disambiguate them to the terms in an ontology (e.g., a user searching for "blog" will be able to find bookmarks tagged with "blogs", "coolblog", "my_blog", etc.).

The disambiguation provided by linking to the ontology does not actually provide a great amount of reduction in the vocabulary size in the current batch. In fact, only seven tokens can be mapped to a smaller set of synsets. This means that there is not a great amount of synonymy in the tags that we have studied.

We believe that this is not a general feature of the full del.icio.us folksonomy and that synonymy will happen more in different domains. We are now extending the size of our study batch to observe this hypothesis. In fact, in the current batch, the main topic was focused on computer and web technologies that use a very restricted vocabulary where words do not often have synonyms. We believe that this phenomenon might appear more often in less technical domains and we are thus extending our study to the domains of cooking, education and travel.

Another important observation that we can do from this disambiguated dataset relates to the issue of considering that one tag maps directly to one concept. In the current model, there is a one to one mapping between the string used in a tag and a concept and thus the fact that multiple sense can exist for the same linguistic representation is forgotten; this is usually dealt with by only considering the most popular sense for a given tag used in the folksonomy. However, when we look at our annotated dataset, if we look at the most frequent sense for a word as provided by WordNet, we can see that the most frequent sense in the English language is only the right disambiguation for a tag in 69.8% of the cases. That is, in 30.2% of the cases, taking the most frequent sense yields an error of disambiguation and thus will later impair the accuracy of the reasoning when using the conceptual folksonomy.

7 Related Work

As we have discussed earlier, Mika [2] has already proposed a tripartite graph representation of the folksonomy model. This model has been widely used in the study of folksonomies, their use in automatic ontology building, and their export to Linked Open Data (LOD).

There has been a number of RDF representation of folksonomies since the introduction of this term. In particular, one of the most popular has been the one proposed by Newman [15], which is very close to the proposed model of Mika as the *Tag* class is a subclass of the SKOS [7] *Concept* class. Kim et al. [16] have proposed a large review of the different ontologies available to distribute folksonomies in the LOD. From all of these models, our preoccupations are probably dealt with best with the SCOT [17] ontology that adds the special constructs *scot:spellingVariant*, *scot:delimited* and *scot:synonym* to link different linguistic variations of the same tag together. However, these construct do not take into account the fact that tags can have the same linguistic form (homographs) but not relate to the same concept (as we have discussed earlier) and how different concepts relate to each other, independently of their linguistic representation. This issue of separation of the linguistic and the semantic levels of ontologies is also strongly pointed out by Buitelaar et al. [18].

There has recently been a number of research on the disambiguation of folksonomies and their use to automatically build ontologies to represent the vocabulary of the users. Garcia-Silva et al. [13] provide a good survey of the field of semantic discovery in folksonomies and we recommend the reading of this article to understand better the field. The method used to extract the semantics from

folksonomies is based on the principles of distributional semantics and is called *tag clustering*, which is based on machine learning clustering algorithms [19]. This clustering is based on the principle that similar tags will have the same semantic and can thus be attached to the same *"concept"* in the created vocabulary. For instance, if the algorithm finds out that "opposition" and "resistance" are similar, then it can associate it to one concept for that meaning.

However, these approaches suppose that tags are all different single terms that represent a single concept and try to attach each single tag to its own distinct concept in the knowledge organisation system created. This is based on the current model of folksonomies, however, as we illustrate in Section 6, it appears that many tags are actually just different linguistic variations of the same term. By making this simplification of mapping one tag to one concept only, only the most popular sense of the tag can be detected and the other, homograph terms cannot be mapped to the right sense. As we have shown earlier, this is the wrong assumption in around 30% of the cases, thus creating a large amount of errors in the semantic disambiguation of a tag that will propagate to the semantic services implemented on the folksonomy.

To improve the general understanding of folksonomies and the distributional semantics work that is based on them, we thus propose to extend Mika's model [2] by separating tags and atomic concepts in the tripartite graph, thus creating a quadripartite graph as discussed in the Section 3.

8 RDF Model for Tags and Concepts

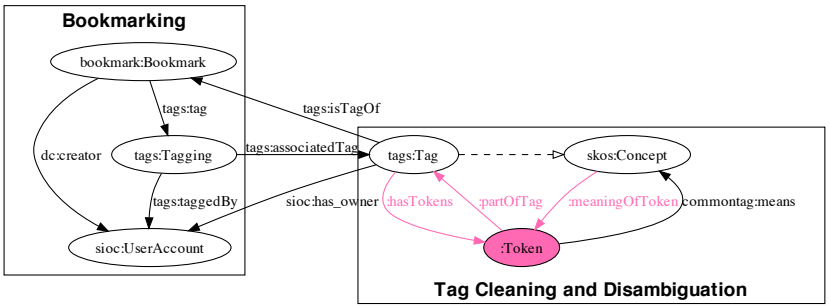


Fig. 6. Extension of the RDF Model

While the Newman's tagging ontology [15] and the SCOT [17] extension can represent the tripartite graph model of folksonomies, they do not discriminate between a tag and a concept.

In Figure 6[8], we propose an extension to the Newman's ontology where a *tags:Tag* can be split in *tags2con:Token*s that then link to the actual semantic in a knowledge organisation system (in this case a *SKOS:Concept*) that can be used in reasoning. In this proposal, for compatibility with the existing RDF models that widely use the Newman's *tags:Tag* class, we also use this one. However, as we have already pointed out, it is our belief that this creates a confusion between the linguistic layer of the folksonomy and its conceptual layer that can lower the accuracy of reasoning services based on this data. Thus, we would recommend to drop such compatibility in the future.

9 Conclusion

In this paper, we reported on a study of a del.icio.us dataset that supports the assumption that there is a need to separate the linguistic representation of a tag (i.e., the free text string) and its meaning to the user (i.e. the semantics of the tag).

First, we observed that in the current model of folksonomies, where there is a one-to-one mapping between a tag and a concept, there seems to be very little overlap of vocabularies between users and little agreement in the tagging of shared resources. However, it appears that many tags are actually different linguistic variations of the same concept, due to different writings variations and multi-word tags. By collapsing these diverse representations of the same concept, we can reduce the vocabulary size by almost 65%. We have also shown that assuming that one tag maps to only one concept ignores the issue of homography and polysemy and can thus yield up to 30% errors when ignoring the possible multiple senses of one linguistic representation.

To resolve these issues, we propose an extension of the standard tripartite graph model mapping tags, resources and users together. We add a fourth layer to the graph that links the tags to their meaning in a knowledge organisation system with a strictly defined semantics. This enables a formal reasoning on the meaning of the tags, taking into account the case of multiword tags or homograph tags.

We believe that such stricter formalisation of the semantics of tags will improve the quality of service of semantic algorithms, such as semantic search, and we are currently performing studies on the dataset described in this paper to show such an improvement.

References

1. Vander Wal, T.: Folksonomy: Coinage and definition. http://www.vanderwal.net/folksonomy.html

[8] The ontology described in Figure 6 is detailed at `http://disi.unitn.it/~knowdive/dataset/delicious/tags2con.n3`.

2. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web **5**(1) (March 2007) 5–15
3. Giunchiglia, F., Zaihrayeu, I.: Lightweight ontologies. In Liu, L., Ozsu, M.T., eds.: Encyclopedia of Database Systems. Springer (July 2009)
4. Uschold, M., Gruninger, M.: Ontologies and semantics for seamless connectivity. SIGMOD Rec. **33**(4) (2004) 58–64
5. Golder, S., Huberman, B.A.: The structure of collaborative tagging systems. Journal of Information Science **32**(2) (April 2006) 198–208
6. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press (2003)
7. Miles, A., Pérez-Agüera, J.R.: Skos: Simple knowledge organisation for the web. Cataloging & Classification Quarterly **43**(3) (2007) 69–83
8. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept search. In: ESWC. (2009) 429–444
9. Autayeu, A., Giunchiglia, F., Andrews, P.: Lightweight parsing of classifications into lightweight ontologies. In: ECDL. (2010) 327–339
10. Körner, C., Strohmaier, M.: A call for social tagging datasets. SIGWEB Newsl. (January 2010) 2:1–2:6
11. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). illustrated edition edn. The MIT Press (May 1998)
12. Wetzker, R., Zimmermann, C., Bauckhage, C.: Analyzing Social Bookmarking Systems: A del.icio.us Cookbook. In: Proceedings ECAI 2008 – Mining Social Data, IOS Press (2008) 26–30
13. Garca-Silva, A., Corcho, O., Alani, H., Gmez-Perez, A.: Review of the state of the art: Discovering and associating semantics to tags in folksonomies. The Knowledge Engineering Review (2010, (To be published))
14. Dutta, B., Giunchiglia, F.: Semantics are actually used. In: ICSD, Trento, Italy, University of Trento (September, 8-11 2009) 62–78
15. Newman, R.: Tag Ontology design (2005) `http://www.holygoat.co.uk/projects/tags/`.
16. Kim, H.L., Scerri, S., Breslin, J.G., Decker, S., Kim, H.G.: The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In: DC-2008, DCMI (2008) 128–137
17. Kim, H.L., Decker, S., Breslin, J.G.: Representing and sharing folksonomies with semantics. J. Inf. Sci. **36** (February 2010) 57–72
18. Buitelaar, P., Cimiano, P., Haase, P., Sintek, M.: Towards linguistically grounded ontologies. In: ESWC. (June 2009) 111–125
19. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on Neural Networks **16**(3) (May 2005) 645–678 PMID: 15940994.