LexiTags: An Interlingua for the Social Semantic Web

Csaba Veres
University in Bergen, Fosswinckelsgt. 6, 5020 Bergen,
Norway. Csaba.Veres@infomedia.uib.no

Abstract. The paper describes *lexitags*, a new approach to social semantic tagging whose goal is to allow users to easily enrich resources with semantic metadata from WordNet. This is a paradigm example of the Social Web and the Semantic Web working together: ordinary users help create the metadata so needed by the Semantic Web and in turn, Semantic Web technologies help those users get a richer experience from the Social Web. A family of simple user interfaces for lexitagging is described, as are some methods for the subsequent, automatic generation of lightweight ontologies. These ontologies are presented as an ideal *interlingua* for the Social Semantic Web.

Keywords: Social Web, Semantic Web, ontology, metadata, WordNet, linked data, rdf, folksonomy

1. Introduction

Two of the most exciting innovations for transforming the World Wide Web are "Web2.0" [1] and the "Semantic Web". Each has a separate vision for moving a relatively static Internet driven by focused content providers, to a dynamic and largely self managing entity enabled by large volumes of metadata. But while the general vision is shared, the details of the two approaches appear to be opposites. While Web2.0 is focused on free-form, user generated ad hoc metadata and opportunistic social organization, the Semantic Web is a vision containing strict and enforced data structures suitable for automated machine processing. Web2.0 has proven advantages in the ease of data creation and a correspondingly lower threshold for user adoption, but the lack of predefined structure may inhibit effective retrieval as the amount of unstructured metadata grows in volume. An obvious idea is to combine the two sets of technologies so that the users can have systems which behave as Web2.0 at the point of insertion, yet as Semantic Web at the point of retrieval. Following papers such as [2], it is now widely agreed in the community that the Semantic Web and the Social Web can benefit from each other.

In particular, the Information Architecture community has embraced *folksonomy*[1] as a way to enhance information management practices. An analogy is often made with the term *desire lines,* which comes from landscape architecture. The basic idea originates in the observation that, in spite of the careful planning undertaken by architects to lay out walking tracks in their meticulously designed spaces, one will often find emergent paths that have been forged by people who deviate off the planned tracks onto the grass or gravel of the spaces. The paths become entrenched when particular tracks are found useful by many people. It is similar in information spaces, where folskonomy describes the desire lines, representing informal tag based classification schemes that people find useful. The addition of formalized "desire lines" on the web would benefit emerging semantic platforms that rely on such metadata, especially with respect to querying and mining social semantic data.

This paper describes a set of tools and principles currently under development, which will help formalize folksonomy for the web. In the next section we describe some of the main problems with current tagging practice. Then we describe our approach to cleaning up tags and generating formal rdf based *semantic tags*. Following this, a method for automatically generating lightweight ontologies from semantic tags is described, and their use as a universal *interlingua* is

1 http://vanderwal.net/folksonomy.html

explained.

2. Folksonomy term problems

The basic problems with folksonomy terms from user tags are nicely summarized in [3]. The problems Mathes identifies are as follows.

- **Ambiguity** Since tags are mainly natural language terms, they are characterized by the inherent ambiguity of those terms. A special case of ambiguity can be seen in the proper identification of acronyms. As noted by Mathes: "Examining the front page on November 14, 2004 revealed one user tagging sites with ANT. After examining the other sites the user tagged with ANT, it was apparent this was an acronym for Actor Network Theory, in the domain of sociology. However, when examining the ANT tag across all users (Delicious apparently is not case sensitive in tags) most of the bookmarks were about Apache Ant, a project building tool in the Java programming language. Two completely separate domains and ideas are mixed together in the same tag."
- **Spaces**, multiple words Many services do not allow users to enter multiple word tags separated by spaces, so users improvise as in the example: "vertigovideostillsbbc". Perhaps more creatively, users concatenate words to express alternative names (design/css), or even hierarchical groupings (Devel/C++).
- **Synonyms**. Since there is no control on the vocabulary, one often finds multiple words or variants expressing the same concept, as in mac, macintosh, and apple (apple of course has the added problem of being ambiguous). Another manifestation of this problem is the indiscriminate use of the plural and singular of a term. The NISO guidelines for controlled vocabularies recommends the singular use.

[4] also summarize problems in tag use, and based on these observations they provide guidelines for creating "tidier tags". They conclude that the main problem with tags is *imprecision*. They flesh out this remark to include the already mentioned problems with ambiguity, synonymy and number, but add a few additional observations: "… the tags are often ambiguous, overly personalised and inexact. .. Plural and singular forms, conjugated words and compound words may be used, as well as specialized tags and 'nonsense' tags designed as unique markers that are shared between a group of friends or co-workers. The result is an uncontrolled and chaotic set of tagging terms that do not support searching as effectively as more controlled vocabularies do."

[4] performed some quantitative analyses on a set of randomly selected tags from delicious as well as the photo sharing site, Flickr. They made the following observations about the prevalence of various errors:

- **Misspellings, incorrect encodings, and compound words**: "By testing against multilingual dictionary software, we found that 40% of flickr tags and 28% of del.icio.us tags were either misspelt, from a language not available via the software used, encoded in a manner that was not understood by the dictionary software, or compound words consisting of more than two words or a mixture of languages."
- **Words that did not follow system conventions**: Almost 8% of the flickr tags and over 11% of the del.icio.us tags were plural forms of words.
- **Symbols used in tags**: "Symbols such as "# " were used at the beginning of tags, probably for an incidental effect such as forcing the del.icio.us interface to list the tags at the top of an alphabetical listing."

They also note after the quantitative evaluation that "However, we did find that single-use tags were less common than we had expected", suggesting a high degree of consensus in tagging behavior, and a correspondingly low degree of "personalised" tags. They additionally note that the high number of tags that were not words that can be found in a standard dictionary may be artificially high. In many cases the tags were misspelled or creative variants of dictionary words. Many examples of misspelling consisted of the transcription of characters across languages. For example, the Norwegian æ can be written as "ae". Sometimes the reason was the compounding of

words and letters as in “17thjuly”. Another prominent practice was the inclusion of geotagging information (latitude and longitude) in the tag. This was particularly popular in Flickr (perhaps unsurprisingly).

Based on the previous observations, it is fair to say that the predominance of tags are dictionary words, or compounds formed from dictionary words (or numbers). In support of this conclusion, [5] report that 82% of the top 100 tags on delicious.com appear in WordNet, and that this drops to a still respectable 79% for the top 1000, and 61% for the top 10000 tags. Apparently, all the mysticism surrounding tags notwithstanding, in the vast majority of cases tags are simply dictionary words or word compounds. It is in this vein that [4] recommend a number of simple guidelines to improve tagging practice. They propose a number of practices like standardized spelling and hyphenation practices, and a handful of useful heuristics for tag selection.

The problem with recommendations is that they can be difficult to enforce, or even convince people that they should try to follow them. For example, one could stipulate that tags should follow NISO recommendations [6] that count nouns appear in the plural form (e.g. dogs, toys) and mass nouns in the singular (e.g. water, furniture). But it is difficult to imagine that people will accept that they should always use the tags *movies, toys, knives,* rather than *movie, toy, knife* for example.

The solution which is suggested in this paper is at the outset a simple way to gently enforce these best practices through the tagging interface, by allowing people to simply tag with dictionary words, otherwise known as *lexical items*. It is for this reason that the tags themselves are called *lexitags*. Lexitags guarantee that every tag can be unambiguously connected with a known lexical item, while still allowing some flexibility in user behavior. For example, both *cat* and *cats* are allowed in the user interface, but both are linked to the lexical item {cat}. By keeping information about the surface form and lexical item separate, no information about user behavior is lost: the underlying semantics of the tag is captured, as well as the potentially significant choice of plural or singular. On the other hand, only acceptable spellings can be used, so the problem of misspellings, idiosyncratic spelling variations and so on, disappears.

3. Creating RDF-based knowledge using social media services

The primary lexical database in this project is WordNet. This is supplemented by DBPedia which provides terms missing in WordNet, such as names for emerging technologies and people. WordNet is perhaps the most well established electronic lexical database, whose development at Princeton University dates back to 1985. WordNet represents disambiguated word senses with synonym sets (*synsets*), which are equivalent terms enclosed in braces. For example some of the unique senses of the word *cat* are: {cat, true cat}, {guy, cat, hombre, bozo}, {cat, gossip}, {kat, khat, qat, quat, cat, Arabian tea, African tea}, {cat-o'-nine-tails, cat}, and so on. WordNet is a very large database, containing in total 206941 word-sense pairs including nouns, verbs, adjectives and adverbs. In addition, each synset contains lexical pointers to related synsets, where the relations are specific to grammatical category. For example nouns are included in (amongst other things) *hyponymy* and *meronymy* relations, but adjectives in *antonymy*. In summary, WordNet is an extensive database of English words, together with a rich set of lexical and semantic relations defined over the lexical items.

The simple idea, then, is to use WordNet as a source for disambiguating tags that are applied to resources by users, and to provide a simple interface where this can be achieved. The disambiguated tags are referred to as *lexitags* to honor their origins in the lexicon, or *semantic tags* to indicate that their interpretation is fixed relative to a semantic resource. In order to realize the interface design in a reference implementation, we designed a platform for social bookmarking, which we will refer to as LexiTags, and which was developed through a commercial startup company called LexiTags D.A. The company was jointly established by the author and his colleague Andreas Opdahl at the University in Bergen, and supported by a seed grant from Innovation Norway. It should be noted that there are at least two existing commercial ventures that are advertised as a “semantic bookmarking service”, making them similar to

LexiTags in this regard. These are Faviki and Zigtag. Zigtag, which has been running since early 2009[2] is the most similar in that it uses its own dictionary as tag definitions. Faviki uses WikiPedia concepts instead. However, there are fundamental differences in the motivation for these services and LexiTags. Zigtag and Faviki are bookmarking services, pure and simple. Their interest in "semantic tags" is to enhance findability on their site by providing equivalences between differently spelled tags (NYC, New York City, Big Apple, …), returning results for only one sense of an ambiguous tag, and so on. On the other hand LexiTags is simply a reference implementation of the lexitagging interface, with the focus being on the generation and exploitation of the metadata itself, rather than the underlying purpose of the reference implementation (which in this case happens to be bookmarking). The principles and algorithms are meant to be portable to any application, using the semantics in the generated metadata to bridge the divide in content across the services and applications. Metadata in the form of lexitags becomes an interlingua between applications.

To demonstrate the idea of a simple portable tagging interface, we will discuss here an iPhone interface which is currently in development. The iPhone interface communicates with the LexiTags service over http and can upload html bookmarks, but also photographs taken with the iPhone camera. The application can therefore serve as a tagging interface for bookmarking as well as a photo upload service.

The design principle is that lexitagging must be no more difficult than ordinary tagging otherwise people will not be inclined to use it. One key research problem is how to rank the possible senses so that the sense intended by the user is immediately available in the interface. The current iPhone tagging interface is shown in figure 1. On the left are two ambiguous tags, and on the right *cat* has been disambiguated (which is evident from the text below the tag). Notice that both URL entry field and photo choser are both available, and the user must chose which they use. URLs have to be manually entered at the moment, but ideally this will be linked to the web browser. In figure 1 we see that the user has chosen (or snapped) an image of a cat, and has assigned two tags "cat" and "cute". The tags are simply typed in the "Tags" field, and automatically marked as "undefined". This allows people to initially add tags freely. Once they have typed a few tags, users must tap each one to define it, which brings up the selection interface in figure 2.

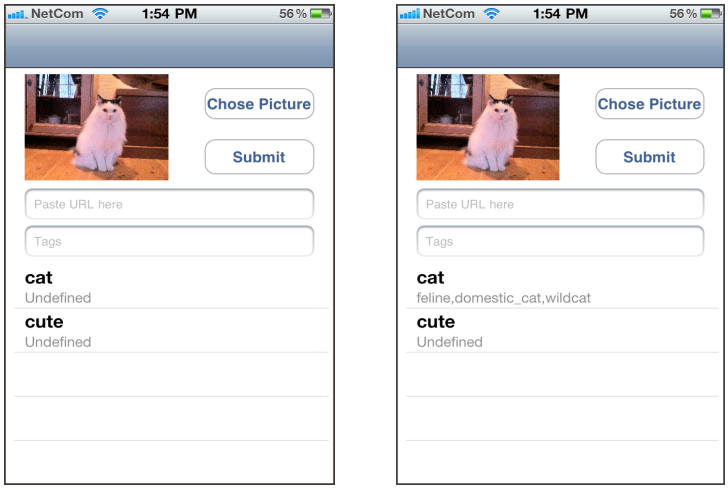

Fig. 1. Picture with two undefined tags and one defined tag on the right

On the left of figure 2 is the initial display, showing 5 possible choices. There are actually 10 senses of the word "cat", so one must scroll to see the others. It is here where the ranking becomes important, since ideally the desired sense will always appear in the top 5 choices. The iPhone choser currently uses a series of words related to each sense as a

2 http://www.readwriteweb.com/archives/semantic_tagging_service_zigtag_finally_launches.php

disambiguator, rather than the actual synset (in this example {cat, true cat}). This is because not all senses have near equivalent synonyms, so the synset in these cases is simply the word itself. For example the synset for the third sense of *cat* is simply {cat}, which would not be a useful disambiguator. We are still experimenting with the usefulness of this method for selecting the correct sense. We are also experimenting with an alternative display in the larger, browser based desktop client. The display in figure 3 shows for each sense the synsets from WordNet, as well as the full explanatory gloss. Usually one or other at least is available. In addition, each sense is preceded by a determiner in brackets. This is meant to help people with selecting the grammatical category: "(a, an, the)" are nouns, "(to)" are verbs and "(is)" are adjectives. Our feeling is that the simple iPhone interface is adequate, but we will need to experiment extensively before making any conclusive claims regarding usability, since the selection of the appropriate word sense is the most difficult and important role for the lexitags interface.

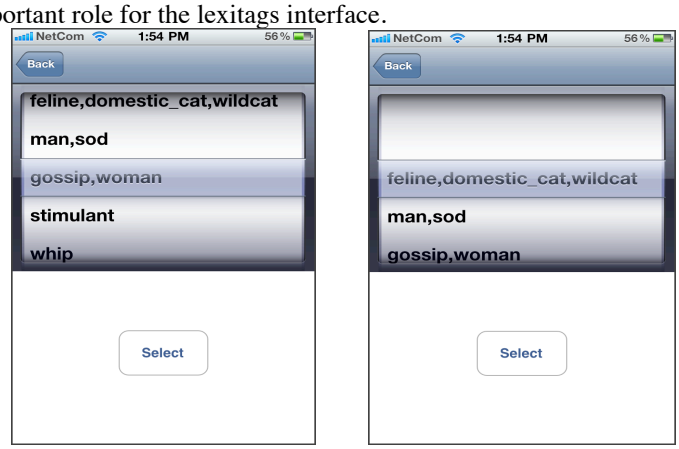

Fig. 2. Two stages of sense selection

Once the process is complete, the entry can be submitted. Any tag that has not been disambiguated by the user is simply discarded at this stage.

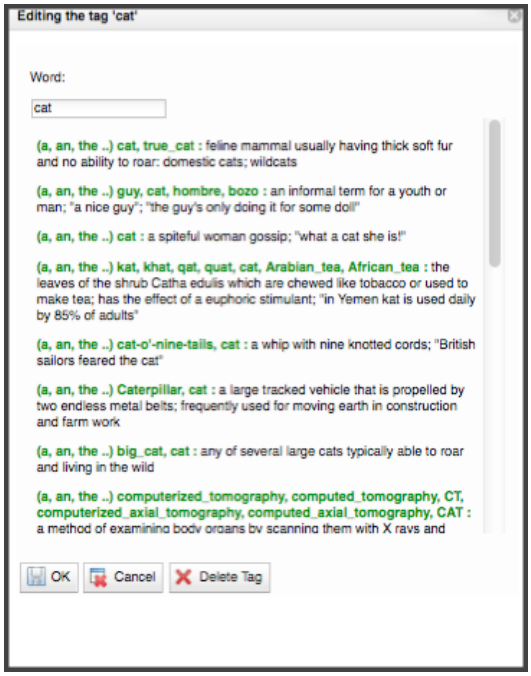

Fig. 3. An alternative tagging interface

A key problem is in ranking the possible senses such that the desired alternative is very near the top of the list for every tagging episode. Clearly this involves an estimate of the likelihood that a particular sense matches the content of the to-be-tagged item. If one can obtain

disambiguated key terms from the resource itself, then there are a number of useful algorithms for computing the similarity between those terms and each candidate sense of the tag [7]. When the resource is an html page, there are a number of obvious possibilities to obtain such contextual information. The simplest is to extract the title, or any metadata that is available, and use any of a number of open solutions which are available from the text processing community to disambiguate these terms. Of course the more sparse the retrieved text, the more difficult the disambiguation. Another option is to scrape the entire text of the html page and extract key summary terms. However, while this method could give the most accurate candidate ranking, it can become computationally expensive and may not return results sufficiently quickly for use in the tagging interface.

Currently we are using a much simpler approach, which is to use the tags themselves for disambiguating other tags. That is, once the user has sense selected the appropriate lexitag, then that can be used to rank any successive tags. The more tags that have been selected, the more accurate the algorithm can become. The biggest problem with this simplistic approach is that there is no disambiguating evidence for the first tag. However, in these cases we simply use the relative frequency of use, arguing that people are less likely to use infrequent senses of words as tags. There is no reason why these various techniques could not be used in complementary ways, combining ranking estimates based on the different sources. For example the initial disambiguating context could be a fast analysis of the title and some metadata, which would be replaced as the analysis of the text becomes available. In turn, this could be combined with the disambiguated lexitags as the user works his way through the tagging session. It is of course an empirical question to see which combination of these methods results in the best user experience.

The results of a tagging session are recorded in RDF, using a number of common standards including Dublin Core[3], FOAF[4] and Common Tag.[5] Figure 4 shows the format we have adopted from the Common Tags specification. The representation is straightforward, so we only point out the two relations ctag:label and ctag:means. The former is the word string used by the tagger, and the latter is a "dereferencable Resource that identifies the concept expressed by the Tag".[6] Of course in LexiTags this is a WordNet synset. This allows some separation between the word string and its meaning, accommodating the case where *NYC* and *NewYork* can both be used as tags, yet refer to the same concept. Because the application is based on open standards, all web sites which expose their data in the Common Tags format will automatically inter operate. Lexitags give extra value in that they can add semantically rich, disambiguated metadata to a URL that may be recorded on another site without rich metadata.

```
@prefix  ctag: <http://commontag.org/ns#> .
@prefix  wn:   <http://www.w3.org/2006/03/wn/wn20/instances/> .
@prefix  rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix  foaf: <http://xmlns.com/foaf/0.1/> .
@prefix  xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix  dc:   <http://purl.org/dc/terms/> .
[] rdfs:type ctag:TaggedContent ;
  ctag:isAbout "http://commontag.org/QuickStartGuide"^^xsd:anyURI ;
  ctag:tagged [
  ctag:means wn:dog-n-1;
  ctag:label "dog"@en;
  foaf:maker u1234;
  dc:created "20.01.2200"^^xsd:Date;
  ctag:taggingDate "22.22.2200"^^xsd:Date  ] .
```

Fig. 4. RDF representation of a tagging session

---

3 http://dublincore.org/
4 http://www.foaf-project.org/
5 http://commontag.org/Home
6 http://commontag.org/Specification#means

4. Ontologies for the Social Web

We have already mentioned the most straightforward advantages of using semantic tags for finding content on a bookmarking site. But the use of WordNet as the reference semantics provides far greater benefits. Lexitagging provides us with collections that are marked up with semantically disambiguated lexical items, which have rich associations to other lexical items in WordNet. We have taken advantage of this in developing a method for creating lightweight ontologies for social media sites. [8] reports an algorithm to extract general terms from a set of resources annotated with WordNet synsets. Basically, the algorithm infers maximally informative hypernyms (SuperTags) for user generated tags with the simple algorithm shown in figure 5. Nodes are only retained with this algorithm if they have two or more children, and are more than six nodes from the root nodes. These parameters are variable.

```
algorithm enrich Bookmark collection {
    forall Bookmarks in the collection {
        find all hypernyms and store them in chains
    }

    forall hypernym chains {
        find every hypernym that {
            either only has one unique child in the
                set of chains it appears in
            or appears as the sixth or higher element
                in any chain
        }
        mark these hypernyms as irrelevant
    }

    for all unmarked hypernyms {
        convert the hypernyms to a SuperTag of the
            Bookmark to which the tag chain belongs
    }
}
```

Fig. 5. Algorithm for maximally informative SuperTags.

We have implemented this algorithm in a web service which generates visualizations for a set of lexitagged resources. Figure 6 shows a typical visualization for a small set of tags.

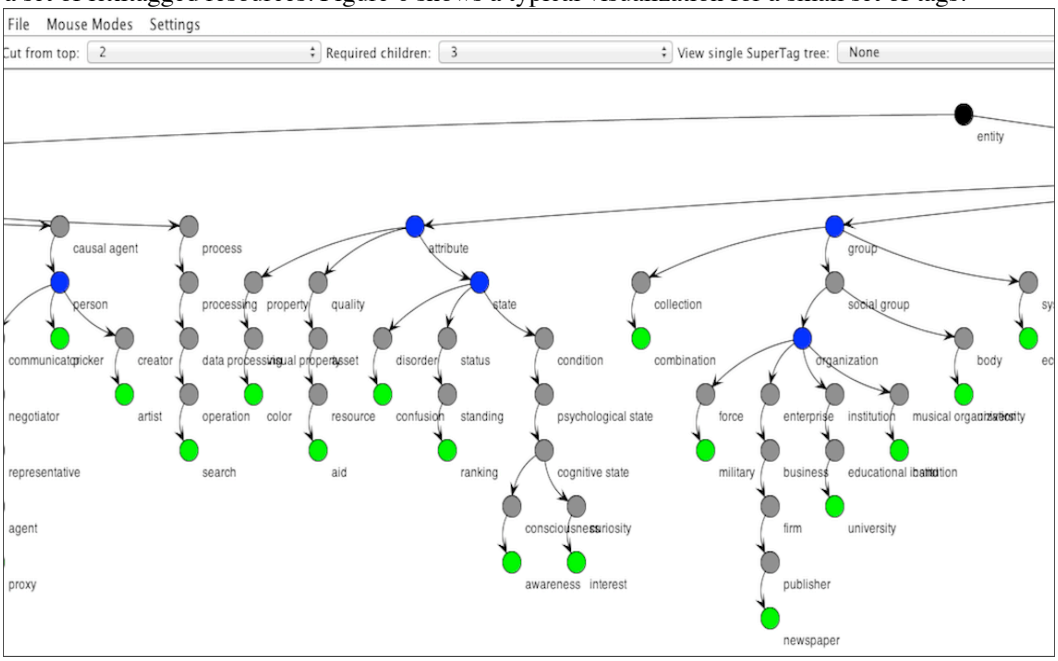

Fig. 6. Inferred SuperTags.

The figure shows the user assigned lexitags in green (light grey leaf nodes), and their

respective hypernym chains. The hypernyms eventually intersect at common nodes. The nodes colored blue (dark grey) are retained as SuperTags because they have three or more children. Light grey nodes (except the leaf nodes) are discarded because they have less than three, and black ones because they are too close to the root nodes. All of these parameters are adjustable in the interface, so it is possible to adjust the generality and inclusiveness of the nodes which are finally retained. Already in this small example we see some useful nodes emerge: *group, organization, attribute, state, person*. The emerging nodes can tell us about the nature of this particular collection. For example, in the sub tree originating from *group* we see *organization* but not *social group* emerge as an important node. The reason is that most resources in this part of the collection deal with either *university, newspaper,* or *military*. It is easy to imagine other collections where the predominant node would be *social node*.

The real power of using lexitags as a basis for lightweight ontologies becomes apparent when relations from WordNet are added to the inferred SuperNodes. For example, figure 7. shows the meronyms of *organization*. These can be added to the emerging ontology as properties. Once embellished with properties, the ontology becomes a rich representation of the key concepts in a social site, and can be used for various inference tasks. For example, if someone uses the tag *Apple inc.,* then this will be an example of an *organization*. Since the ontology tells us that organizations have a *quorum*, this could prompt an application to automatically fill in the names of the Apple board of directors and even suggest them as contacts in the social site. The result is a rich, dynamically emerging ontology which reflects the users attitudes to the underlying domain, and which can change if the concepts or the tagging behavior of users change.

organization, organisation -- (a group of people who work together)
HAS MEMBER: quorum -- (a gathering of the minimal number of members of an organization to conduct business)
HAS MEMBER: membership, rank -- (the body of members of an organization or group; "they polled their membership"; "they found dissension in their own ranks"; "he joined the ranks of the unemployed")

Fig. 7. Meronyms of *organization*

The results presented in this paper are preliminary, but the way forward is clear. The implementation of portable tagging interfaces will result in a growing number of resources tagged with lexitags. The resources could include traditional http bookmarks, geo tagged photographs, wiki and blog entries, and even local file systems. The automatically generated lightweight ontologies will add unique metadata to each site. However, because each site is marked up with the same lexitags, this will facilitate comparisons and sharing between the sites. In fact, we make the bold claim that lexitags (WordNet synsets) are an ideal *interlingua* for the social semantic web because it has the expressive power to align concepts between any arbitrary ontologies, yet is intuitive in the most basic sense of the word.

Notice that we are not advocating WordNet as a universal ontology. In fact, we are sympathetic to [9], who details a number of reasons for why the lexicon ought not be construed as an ontology at all. Ontologies attempt to model domains of interest with strict, mutually exclusive classes, while lexicons often use overlapping words to cover the semantics of the world. For example, consider the English words *error* and *mistake* and some of their hyponyms, which by definition denote kinds of mistakes or errors: *blunder* (an embarrassing mistake), *slip* (a minor inadvertent mistake), *lapse* (a mistake resulting from inattention), *faux pas* (a socially awkward or tactless act). But notice that a *slip* can also be a *blunder* and that a *faux pas*, which is itself a kind of of *blunder*, could also be just a *slip*. What licenses the use of the different words in natural language conversation is that they emphasize different dimensions of the concept being communicated: a *slip* is distinct from *mistake* because it does not (presumably) result from an *error* in judgment (i.e. it is inadvertent), whereas a *blunder* is distinguished by the fact that it causes embarrassment. But there is no reason that a *blunder* could not be inadvertent, and therefore also a *slip*. Words at a given level in the hyponym tree sometimes shift attention from

one distinguishing feature to another, rather than being non overlapping sub types of their hypernym.

WordNet may not be a universal ontology, but is powerful as an interlingua precisely for the same reasons that make language so powerful at communicating concepts. Flexibility allows one to finesse levels of detail but still communicate, and also allows one to reach arbitrary levels of precision when needed. When using lexitags as an interlingua, designers of individual ontologies can map their terms to specific interpretations in WordNet as the requirements demand. They can chose the mappings that reflect their particular world view: for example domains that require attribution of blame can map their terms to *slip* or *mistake* while everyone else can map to *error*. If it is important that people who use *cinema* are kept away from people who use *movie* [10] then this is possible, but they can still become acquainted when the distinction no longer matters.

Another interesting possibility is that the lexitags interface may help solve another problem with using WordNet as an ontology: *lexical gaps*. [9] points out the problem where an easily demonstrable *covert category* exists, but there is no word for it. For example, *things that can be worn on the body*. Since the lexitagging interface allows multi word tags, someone could use a general tag *body wear* with the two words appropriately disambiguated. This would then establish a new link between *body* and *wear*, as the lexical representation of the covert category.

Lexitags can also serve as an interlingua between formal ontologies and the social web. For example, SUMO [11] has an extensive set of links to WordNet which can be explored with the SIGMA knowledge engineering environment.[7] The links include equivalent as well as subsuming mappings. Any ontology that is mapped to SUMO is therefore automatically aligned with lexitags ontologies. Perhaps equally importantly, the EuroWordNet project oversees the creation of wordnets for many European languages,[8] and there are attempts at Chinese wordnets.[9] These projects constitute a major step towards making lexitags a universal interlingua for formal and semi formal metadata.

5. Related work

There is a large body of work whose aim is to exploit folksonomies for more effective information management. Most of the existing literature concerns the exploitation of statistical regularities in the way tags are assigned to resources by users. [12] suggests that the efforts can broadly be classified as (a) extracting semantics of folksonomies, including measuring relatedness, clustering, and inferring subsumption relations or (b) semantically enriching folksonomies, including collaborative structuring, and linking tags with professional vocabularies and ontologies.

One of the earliest demonstrations in the first vein was *clustering* on Flickr, where polysemous tags are displayed with co-occurring tags in different sets of images. For example, the tag *apple* has the following clusters: <mac, macbook, macintosh, computer, laptop, imac, keyboard, powerbook, osx, macbookpro>, <fruit, red, green, food, tree, macro, canon, orange, blossom, apples>, <ipod, iphone, music, nano, touch, shuffle, mp3, black, phone, ipodtouch>, and <nyc, newyork, manhattan, newyorkcity, ny>. The algorithm can identify photographs tagged with the different uses of *apple*: apple the fruit, apple the company, and the "Big Apple". However, this form of clustering is not simply lexical disambiguation since the company sense of *apple* is listed in two different clusters which reflect different distinguishing product lines for the company. An additional benefit is that different spelling variations of a tag are bundled into the same cluster as in *nyc, newyork, ny,* because these tags tend to co-occur with the same pictures. While the details of the Flickr algorithm are proprietary, various clustering algorithms were explored by [13].

In another interesting use of co-occurrence, [14] report a study in which their algorithm

---

7 http://sigma.ontologyportal.org:4010/sigma/WordNet.jsp

8 http://www.illc.uva.nl/EuroWordNet/

9 http://cwn.ling.sinica.edu.tw/

suggests new tags to users just in case they used an ambiguous tag for a resource. The tag is ambiguous because it also appears in a cluster of unrelated resources, as in the Flickr example. For example the spatially ambiguous tag *Cambridge* can co-occur either with *MA* or *UK*. In these situations one of these will be suggested as an additional disambiguator.

Clustering algorithms can identify different uses of tags, but they do not provide any semantics beyond this. [15] show that an analysis of the temporal and spatial distribution of tags can determine if a tag belongs to a place and/or an event. For example, they can identify that the tag *bay bridge* corresponds to a place, but *www2007* to an event.

Researchers have also investigated the possibility of inferring hierarchical relations between folksonomy terms. [16] also consider a probabilistic model of tag semantics in which ambiguity is directly observed through graphs of the distribution of concepts labeled by individual tags. For example *cooking* has a single very distinct distributional peak, whereas *XP* has several peaks corresponding to the various uses of the term. Because semantics is defined relative to the resources in the data set, the results are dynamic and depend on the current state of the concepts in the data set. But more interestingly, their probabilistic model can also infer hierarchical ordering among the tags by considering overlaps in the concepts covered. Another interesting attempt to infer hierarchies is to use conditional probabilities rather than distributional data. [17] inferred subsumption relations through conditional probabilities in tags. They say that X potentially subsumes Y if P(x|y >= t) and P(y|x < t), where t is a co-occurrence threshold. The algorithm can discover interesting subsumptions, like that between *san francisco* and *goldengatebridge, fishermanswharf, pier39*. On the other hand there are spurious probabilistic dependancies that lead to poor examples like *glass* subsuming *magnifying, blow, stained*. This highlights the problem with purely statistical procedures that are oblivious of syntactic or semantic constraints.

In terms of semantic enrichment there are several attempts to extend statistical approaches by extending folksonomies using resources such as Wikipedia, on line ontologies, and WordNet [18-20]. These resources are used in various ways, including to effectively cluster tags, for disambiguation, adding synonyms, and linking to annotated resources and ontology concepts. During this process the terms of the folksonomy are cleaned up and disambiguated, linked to formal definitions and given properties which make them more useful as ontologies. [21] also suggest a rich framework by which tags can acquire post hoc assignments to formal interpretations, including the categories of use suggested in [22].

There are also a few studies in which users are expected to contribute semantics at the time of tagging. [23] studies a corporate blogging platform which included a tagging interface. The tagging interface was linked to a domain ontology, and whenever someone typed a tag that had interpretations in the ontology the interface would present a choice of possible concepts to link the tag to. The ontology would also evolve as users typed new tags which were initially not in the ontology, but the scope of defined tags was limited by the ontology. [24] discuss a sophisticated Firefox plugin, *Semdrops*, which allows users to annotate web resources with a complex set of tags including *category, property,* and *attribute* tags. These are aggregated in a semantic wiki of the user's choosing. [25] reports on an open source bookmarking application (SemanticScuttle) that has been enhanced with *structurable tags* which are tags that users can enhance with inclusion and equivalence relations at the time of tagging. [26] describes *extreme tagging* in which users can tag other tags, to provide disambiguation and other relational information about tags.

Finally, the two previously mentioned commercial ventures Faviki and Zigtag should be mentioned as existing bookmarking services which make use of defined tags. Faviki uses Wikipedia concepts as common tags, and is able to aggregate tagged content according to Wikipedia categories. Since the defined tags are Wikipedia concepts, Faviki cannot semantically ground tags like *interesting, cool,* and *useful*. Zigtag uses dictionary entries, but also allows undefined tags, which make up a significant proportion of their tags.

This birds eye view of the literature shows that existing work is focused almost exclusively on the problem of extracting latent semantics from naive folksonomies composed of messy vocabularies rife with the problems of ambiguity and indeterminacy. In this respect the

work presented here represents a much less well explored effort in eliciting precise semantic tags at the time of tagging. The current work is distinguished from similar research along four major dimensions. First, Lexitags aims to provide a lightweight tagging tool that can be used to tag a wide range of content including html bookmarks, pictures, and local filesystem content. Second, we use WordNet as the primary semantic reference, exploiting the structure of WordNet to construct new relationships and lightweight ontologies. Third, no tags are allowed to be completely undefined, which makes for a more coherent tag collection. Fourth, Lexitag users are not expected to make any complex decisions when assigning semantic tags. They are not expected to contribute relational tags, and so on. They simply chose the sense of the word which they already had in mind when writing the tag.

6. Conclusion

The paper introduced the lexitags approach to social semantic tagging with simple lightweight tagging interfaces. Lexitags are tags whose semantics are grounded in disambiguated lexical items, and which stand in useful relations to other disambiguated lexical items. These form the basis of automatically generated lightweight ontologies which can take the role of universal interlingua between social applications in any domain, and in many non English languages.

Tags which have rich, unambiguous definitions make some aspects of previous work to make sense of tags, un necessary. There is no need to infer that spelling variations on a term have the same meaning, for example, because the distinction between word form and word meaning in lexitags already accommodates spelling variations. Similarly there is no need for disambiguation or clustering for the purpose of identifying different word senses. However, many of the current ideas can still be used in more refined ways. For example clustering is still useful but now at a more detailed level because we can focus on clusters within each sense. If we ignore the fruit sense of *apple*, for example, it may be possible to discover interesting clusters in the way the company name is used. Similarly, taxonomy inference for "tags-in-use" with any of the methods mentioned is still possible, but now it can be refined by taking into consideration the semantics of the tags. For example if subsumption can only occur between nouns, then *glass* will never subsume *magnifying*.

Semantic enrichment becomes much easier too, because lexitags are primarily WordNet synsets. As an example, WordNet already has a rich mapping to DBPedia, so embellishing the dynamically constructed ontology with Wikipedia facts is much simplified. This is the essence of the linked data movement, removing uncertainty and probability from data integration.

One of the most important claims is that WordNet is the ideal means by which to ground the semantics of common tags. This differentiates Lexitags from previous efforts such as Faviki, [21], and [24]. Faviki has chosen to use Wikipedia concepts instead, but we argue that WordNet is more useful as an interlingua because it is more flexible, has more general coverage of terms, and already has many mappings defined to resources such as DBPedia and SUMO.

In summary, this paper suggests that the tagging world be turned upside down. Rather than using clever algorithms for making sense of messy user generated tags, the clever algorithms should be used to help users generate tags that make sense in the first place.

References

1. Oreilly, T.: What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. (2007).
2. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Semantic Web and Information Systems. 3, (2007).
3. Mathes, A.: Folksonomies-cooperative classification and communication through shared metadata. Computer Mediated Communication. (2004).
4. Tonkin, E., Guy, M.: Tidying up tags. D-Lib Magazine. 12, (2006).
5. Cattuto, C., Benz, D., Hotho, A.: Semantic grounding of tag relatedness in social bookmarking

systems. The Semantic Web-ISWC 2008. (2008).
6. National Information Standards Organization: NISO_vocabularies. (2005).
7. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics. 32, 13–47 (2006).
8. Veres, C., Johansen, K., Opdahl, A.: Browsing and Visualizing Semantically Enriched Information Resources. 2010 International Conference on Complex, Intelligent and Software Intensive Systems. 968–973 (2010).
9. Hirst, G.: Ontology and the Lexicon. Handbook on ontologies. (2009).
10. Shirky, C.: Ontology is Overrated--Categories, Links, and Tags. http://www.shirky.com/writings/ontology_overrated.html. (2007).
11. Niles, I., Pease, A.: Towards a Standard Upper Ontology. Proceedings of the international conference on Formal Ontology in Information Systems - FOIS '01. pp. 2–9. ACM Press, New York, New York, USA (2001).
12. Limpens, F., Gandon, F., Buffa, M.: Linking Folksonomies and Ontologies for Supporting Knowledge Sharing,
13. Begelman, G., Keller, P., Smadja, F.: Automated Tag Clustering:
Improving search and exploration in the tag space. Proc. of the Collaborative Web Tagging Workshop at WWW'06. (2006).
14. Weinberger, K., Slaney, M., van Zwoi, R.: Resolving tag ambiguity. Proceeding of the 16th ACM international conference on Multimedia, Vancouver, Canada. (2008).
15. Rattenbury, T., Good, N., Naaman, M.: Towards extracting flickr tag semantics. Proceedings of the 16th international conference on World Wide Web (2007).
16. Zhang, L., Wu, X., Yu, Y.: Emergent semantics from folksonomies: A quantitative study. Journal on Data Semantics VI. (2006).
17. Schmitz, P.: Inducing ontology from flickr tags. Collaborative Web Tagging Workshop at WWW2006. (2006).
18. Specia, L.: Integrating folksonomies with the semantic web. The semantic web: research and applications. (2007).
19. Angeletou, S., Sabou, M., al, E.: Bridging the gap between folksonomies and the semantic web: An experience report. In ESWC workshop. Bridging the Gap between Semantic Web and Web 2.0 (2007)
20. Van Damme, C., al, E.: Folksontology: An integrated approach for turning folksonomies into ontologies. In ESWC workshop. Bridging the Gap between Semantic Web and Web 2.0 (2007)
21. Limpens, F., Monnin, A., Laniado, D., Gandon, F.: NiceTag Ontology: tags as named graphs. In Proceedings of the 6th International Conference on Semantic Systems (I-SEMANTICS '10). (2010).
22. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems, (2005).
23. Passant, A.: Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. Proceedings of International Conference on Weblogs … (2007).
24. Torres, D., Diaz, A., Skaf-Molli, H., Molli, P.: Semdrops: A Social Semantic Tagging Approach for Emerging Semantic Data. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2011).
25. Huynh-Kim Bang, B., Dané, E., Grandbastein, M.: Merging semantic and participative approaches for organising teachers' documents. In J. Luca & E. Weippl (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008 (pp. 4959-4966). Chesapeake, VA: AACE., Vienna (2008).
26. Tanasescu, V., Streibel, O.: Extreme tagging: Emergent semantics through the tagging of tags. Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution, ESOE 2007, co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 12th, 2007. (2007).