SOCIAL TAGGING: SEMANTICS ARE ACTUALLY USED

Biswanath Dutta and Fausto Giunchiglia

October 2008

Technical Report # DISI-08-061

Also: Published in 2nd International Conference on the Semantic Web and Digital Libraries (ICSD), 2009.

Social tagging: Semantics are actually used

Fausto Giunchiglia
Dipartimento Ingegneria e Scienze
dell'informazione (DISI)
Povo, 38050 Trento, Italia
phone: +39 0461 883938
fausto@disi.unitn.it

Biswanath Dutta
Documentation Research and
Training Centre (DRTC)
Indian Statistical Institute (ISI)
8th Mile Mysore Road
Bangalore- 560059 (India)
bisu@drtc.isibang.ac.in

ABSTRACT

This paper describes the results of a study whose goal is to analyze, evaluate and understand the use of semantics in social networks. As a paradigmatic example, the study concentrates on a fairly large portion of the tags used in del.icio.us (more than 5,000 tags). The results show that semantics are pervasively used. In particular, the large majority (in our experiment, 75%) of the tags are subject related and directly codify the semantics of the resource. Furthermore, the number of tags which are effectively used is only a small proportion of the overall tag set (in our experiment around 30% of the tags covers around 95% of the resources), and these tags tend to remain stable in time, despite the continuous growth of the number of tags and resources being indexed. Finally, it is possible to identify an implicit use of hierarchical relationships (i.e., lightweight ontologies) among the concepts denoted by the terms used.

Categories and Subject Descriptors

H.3.1 [**Content analysis and indexing**]: Dictionaries, linguistic processing, thesauruses.

General Terms

Measurement, Documentation, Experimentation.

Keywords

Web 2.0, social networks, semantics, tagging, folksonomies, lightweight ontologies, insightful analysis

1. INTRODUCTION

There are many reasons why, at least in principle, semantics (e.g., thesauri, subject heading lists, ontologies) can be very useful in retrieving meaningful information. They can serve as background knowledge both in the input and in the output stages of any information retrieval system. At the input stage, they help in the creation of subject related descriptions of a document using standard terms. At the output stage, they aid users in constructing structured queries and support precise information retrieval. However, it is a fact that, while social tagging, social networks, and the Web 2.0 in general, have found their way and marked a significant progress in the evolution of the Web, the Semantic Web, and semantics in general, have yet to have the widespread success that was originally expected.

Many explanations have been provided, mainly inside the Web community but also within various other related communities, e.g., digital libraries and information science. Two lines of thought can be identified. The first tends to highlight the intrinsic weaknesses of semantic methods. Mai [1] summarizes the difficulties of knowledge representation using established bibliographic classification schemes. Thus, for instance, as it has been noticed by many, there is a possibility that an article is not tagged by an indexer because indexing exhaustivity is low. Say, an article might mention "religion" as a secondary focus, and the indexer might decide not to tag it with "religion" because it is not important enough compared to the main focus. But for a searcher it might be relevant and hence recall fails. Another problem is that semantic information tends to become out-dated rather quickly, and it is also possible that indexers mis-interpret authors. Nicholson [2] points out that the lack of or, conversely, the excessive specificity of some controlled vocabularies as being an impediment to the adequate description of online collections within specific contexts. The need for some services which implement in-house modifications, their general dependency on significant investments of time, money, training, expertise and professional intervention further discourage their wider adoption within particular communities of practice [16]. Finally, Duval et al [3] point out that the fundamental obstacle preventing the wider deployment of controlled vocabularies is that the proliferation of digital libraries and the Web precedes the ability of any one authority to use traditional methods of metadata creation and indexing. While metadata creation is valuable and indispensable within particular communities of practice, it can be costly to implement and can present significant scaling difficulties.

The second line of thought tends to highlight the advantages of social tagging. Thus, as argued in [4,5], the emergence of social tagging becomes a useful way to supersede the subject indexing role of the information professional and to facilitate resource discovery and knowledge organization over the Web. Kipp and Campbell note that the tagging patterns follow both emerging consensus among users and emerging trends within the user's own tagging system, however idiosyncratic [6]. However, they also notice that, since users are untrained in indexing methods, users cannot create tags which arrange into useful patterns. At least, not to the extent that would justify dispensing with controlled vocabularies and faceted browsing schemes. They go even further and argue that collaborative tagging systems flaunt too many standard principles of conventional indexing to be a viable replacement, no matter how far we lower our standards [6].

The goal of this paper is to describe the result of an extensive study whose goal was to answer the following question: how much *useful semantics* are actually used, *in practice*, in social tagging systems? Here by "useful semantics" we mean semantics that are actually used to achieve their intended main goal, namely to *share meaning (understanding) among different users* who, in general, have different backgrounds, goals, use different languages or words, and so on. This study has been conducted by analyzing a fairly large portion of the tags used in del.icio.us (http://delicious.com/popular/), one of the most popular and successful bookmarking sites, a site which has been mentioned many times as the winning story over ontology based approaches.

The results of this study are quite interesting and also somewhat surprising, at least compared to the original expectations of the authors. In a nutshell, the outcome of the evaluation is that semantics are pervasively used: the large majority (in our experiment, 75%) of the tags are subject related and directly codify the semantics of the resource. Even more strikingly:

1. The number of tags which are effectively used is only a small proportion of the overall tag set (in our experiment around 30% of the tags covers around 95% of the resources), and these tags tend to remain stable over time, despite the continuous growth of the number of tags and resources being indexed.
2. It is possible to identify an implicit use of hierarchical relationships among the concepts denoted by the terms used (i.e., lightweight ontologies, as defined in [8, 20]).

The paper is organized as follows. After a description of the previous work (Section 2) and of the methodology used (Section 3), the following three sections describe the first phase of our study. They are: the analysis of the language used in tags (Section 4), the analysis of what tags are used for (Section 5), and the analysis of the source of the terms used in the tags (Section 6). The first phase provides us with evidence of the use of semantics and motivated the second phase, described in Sections 7, 8. In particular, the first analyzes the kind of semantics encoded in tags and reports the results mentioned in item 1 above, while the second studies the emergence of hierarchical organizations of terms, as mentioned in item 2 above. Section 9 concludes the paper with a brief summary of the lessons learned from this study.

2. STATE OF THE ART

This is the last of many evaluation studies which have been carried out in the field of social tagging systems. We list below some of the most relevant.

Guy and Tonkin [7] conducted a small-scale study to assess the ‘tag literacy’ of users and suggested that such literacy might impact upon the utility of the tagging approach. They consequently proposed various system specific strategies for improving the quality of tags (e.g., spelling error checking, suggestion of synonyms, etc.) and encouraged users to observe certain collaborative tagging conventions. Heymann [9] pointed out that due to the size issues, it is a bit too early for social bookmarking to have a big impact on web search, though the study showed that there are some aspects of social bookmarking which could be really useful for web search, for instance how recent pages are and the overlap of tags and bookmarks with queries and search results. The study also pointed that the tags chosen by users seem to have considerably redundancy when compared to the text and domains of the annotated pages. Kipp and Campbell [6] analysed the tagging patterns exhibited by users of del.icio.us to assess how collaborative tagging supports and enhances traditional ways of classifying and indexing documents. Their study found that the tagging practices work, to some extent, in ways that are similar to conventional indexing, using frequency data and co-word analysis matrices. They also point out that tags related to time and task suggest the presence of an extra dimension in classification and organisation, a dimension which conventional systems are unable to facilitate. Heckner, Muhlbacher and Wolff [11] performed an empirical study of the tagging behaviour in Connotea, a web-based bibliographic annotation system. The study was carried out by examining the 1191 tags from 500 ICT-related scientific articles. They observed a great overlap between tag material and document text and rather few non-content-related tags.

All the studies above are different from ours in that they all concentrate on the analysis of tag text relationships and the problems which arise. Our study, as far as we know, is the first which provides an in-depth analysis of the semantics codified in the users’ created descriptors and provides evidence that these semantics not only exist but that they actually play a crucial role in the classification of resources. This work provides evidence of the well-foundedness of the work, e.g., that by Good et al. [15], which tries to bridge the gap between social tagging and semantic annotation and also, with the analysis in Section 8, of the attempts of automatically constructing hierarchical tag relationships out of folksonomies (see, e.g., [18, 19]).

Complementary to our work, and coherently with the results described in Section 7, Mika [17] run a large experiment on del.icio.us that showed that users tend to converge on a relatively small number of tags. Mika’s work concentrates on the relation between people and tags and it provides evidence of the emergence of semantics from the users’ behaviour.

3. METHODOLOGY

Our study is articulated in three main steps. During the first step our goal is to build a meaningful data set, both in size and in content. To this extent we have collected a large amount of del.icio.us data, first in the period August 8-11, 2008 and, later, in the days before and up to October 6, 2008. We have manually selected the URLs from the bookmark list tagged “*popular*” and considered only the URLs bookmarked by a minimum of 100 users. Then we have collected the MD5-hashed (partially insecure, cryptographic hash function with a 128-bit hash value) URL values for those resources, which allowed us to download the del.icio.us tag pages for each resource (URL) in HTML format. Finally, we have automatically extracted the tags and the other related data necessary for this study. Table 1 below describes the characteristics of the resulting data set. We have a total of 13114 bookmarks from 51 popular URLs and 5181 tags out of which 3208 are single occurrences.

Table 1: Tag Database characteristics

# URLs	# Bookmarks	# Tags	# unique Tags
51	13114	5181	3208

Tags are multi-dimensional. People tag resources for various purposes, with the goal to use them in different times and places. For example, some users describe resources by directly encoding their semantics via some subject-driven keywords (which are useful for the entire tag community), other users tag resources defining their intended purpose or the intended place of use (e.g., business, office), others just tag a URL for easy identification in the future, possibly using words which have no connection with its established meaning (e.g., “A”, “%s”), and so on. The second step, therefore, is to develop a tag analysis model with the goal of establishing the role of (semantic) information in tags. To this end we have first analyzed the data of our corpora in order to get an idea of how, *in practice*, tags are used. This preliminary study was fundamental as it allowed us to get a first, superficial but meaningful feeling of the problem. Then, we have done an extensive literature survey of the evaluation studies performed in the past, e.g., the work described in Golder and Huberman [12], Gay and Tonkin [7], Kipp and Campbell [6], Heckner [11] and Heyman [9]. Then, finally, on this basis, we have defined our tag analysis model, whose top level facets are described in Figure 1 below. It is worth to declare here that we have done the study manually presented here to keep the quality of the study high. Because we believe the automatic tag analysis system cannot perform the extent a subject expert can do.

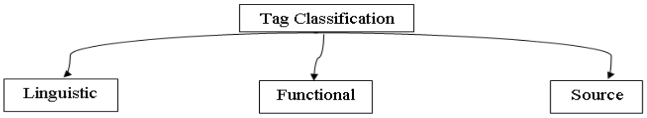


Figure 1: Tag Analysis Model

Our model consists of three top-most facets: *linguistic, functional* and *source*. These first order facets are further sub-divided into several categories discussed in detail in the following sections. Let us see here their intended overall purpose.

The *linguistic facet* concentrates on the language used in tags. The goal is to reveal user practices in creating tags. For example, with this study we have been able to identify the users preferred language, to identify the dominating parts-of-speech, to reveal user's conventions in structuring tags, and so on. The *functional facet* allows us to examine the tags functionality, namely what they are used for, independently of how they are structured. It helps us to study the tags ability to serve a particular function or use. The third and last category, the *source facet* allows us to identify the source of the tags, namely where the text codified in the tags comes from. It helps us to answer questions such as: were the tags created using the resource text? Or were they chosen from outside?

The above analysis has given evidence of a clear and strong presence of semantic information in the tags. This motivated the third phase whose goal was to analyze *which semantics* is actually used in del.icio.us. This study has concentrated on two issues: the *stability of semantics* and the *complexity of semantics*. The goal of the study on stability is to answer two simple questions, namely, first question: is semantics stable on language in the sense that users tend to converge on a small subset of all the possible and meaningful tags? Second question: does the (small) set of tags used by taggers remain stable in time despite the continuous and large growth of tags and resources? The study on the complexity of semantics has allowed us to analyze the implicit semantic structure codified in the tags' implicitly encoded semantics. In particular, the question was whether it is possible to identify some ontological (hierarchical) structure which could be used to suitably organize tags.

4. LINGUISTIC ANALYSIS

Our linguistic analysis was organized along five dimensions. Figure 2 describes the resulting five facets. They are: *language, structure*, *neologism*, *"spelling variation"* and *"starts with-special-symbols".*

The goal of the *Language* facet is to understand the natural language used in writing the tags. As it turned out, the large majority of tags are in English but other languages have also been found. As a consequence, we have organized this category only into three major sub-facets, namely: *English, Non-English* and *Unknown*. The *unknown* class accounts for the tags with undefined languages.

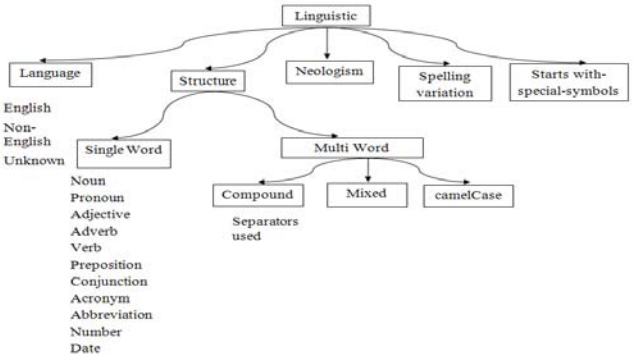


Figure 2: Linguistic Model

The goal of the *structure* facet is to understand how the natural language used in the tag labels is structured. To this extent, this class is divided into two facets, namely: "*single word*" and "*multi word*". The goal of the *single word* class is to account for the grammatical patterns used in the tags. The tags under this class are further grouped into the following categories: *part-of-speech* (noun, pronoun, verb, adjective, etc.), *acronym* (e.g., AI, A&F, etc.), *abbreviation* (e.g., tech, app, etc.), *number* (e.g., 1, 5, 100, etc.) and *date* (e.g., "08", "18092006", "20080811", "06.08.2008"). The *multi-word* class is further sub-divided into three facets, namely: *compound*, *mixed* and *camelCase*. The goal of this subdivision is to account for tag conventions and tag complexity, e.g., how many words per tag, the various patterns used in distinguishing the multi words tags. The *Compound* class includes all the labels consisting of multi-words separated by some punctuation or symbol (e.g., ajax_dhtml_archive, alive_or_dead), separated by capitalizing the first letter of each word in a multi-word tag (e.g., AttentionManagement, BasicPCKnowledge) and also the tags which look like a single word tag, but which are made by multi-words without any defined distinctions among the words (e.g., toread, datavisualization, desktopenvironments). The *Mixed* class contains the labels consisting of a mixture of alphabet elements and numbers. For example, 3twenty9, 2read, 4u, 4_y, etc. This kind of practice is quite common among taggers and is probably the reflection of SMS talk, social networking scrap languages, Internet slang and IM chat (Instant Messenger). By *camelCase* we mean a typographical way of combining words whereby the first letter of each word except the first is capitalized, e.g., bestOfBreed, databaseModel, etc. It is worth noting that for each category under *structure* we count the *number of words* per tag. For example, *blog* counts as 1, *toread* as 2, *bestOfBreed* as 3, *ajax_dhtml_archive* as 3, and so on.

The *neologism* class identifies new words. *Spelling variation* identifies the tags in English with deviant spellings against the standard dictionary spelling. We have computed the spelling variations by comparing tags with the terms in WordNet (http://wordnet.princeton.edu/) and WordWeb (http://wordweb.info/). Finally, we have found certain tags which either start with special symbols or consist only of symbols. These tags are categorised under the class *starts with-special-symbols*, for example, "::beauty::", "@Article", "|","++++","%s". Most likely, this type of tags is used to force them at the top of alphabetical lists [7].

The results of our analysis are reported in Figure 3 (for details see also Table 2). Out of a total of 3208 unique tags, the majority (86.5%) is in English and only 9.8% tags are tagged with other languages. The unknown category contains 43 tags (1.34%). Number and "date and special character" tags are 23 (0.71%) and 52 (1.62%) respectively.

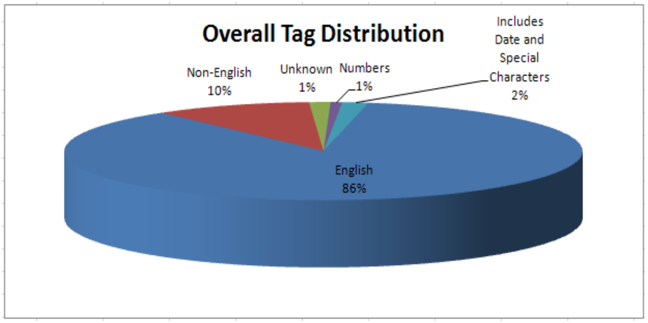


Figure 3: Overall Tag Distribution

Table 2: Overall Tag Distribution

Tags	Total	in %
English	2774	86.47132
Non-English	316	9.850374
Unknown	43	1.340399
Numbers	23	0.716958
Includes Date and Special Characters	52	1.620948
Total	**3208**	**100**

As shown in Figure 4, among the Non-English tags, tags in Spanish (2.93%) are predominant, followed by German (1.34%), Portuguese (1.28%), Italian (1.06%), etc. See also Table 3.

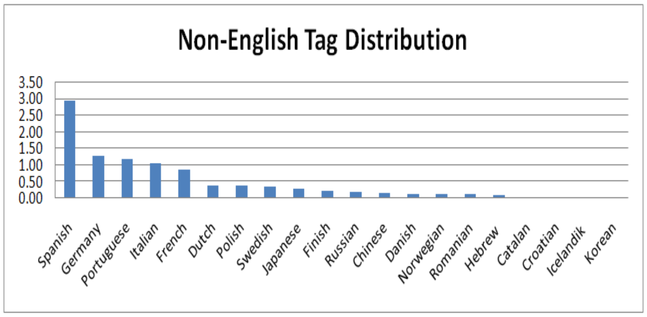


Figure 4: Non-English Tag Distribution

Table 3: Non-English Tag Distribution

Language	Total	in %	Language	Total	in %
Spanish	94	2.93	Russian	6	0.19
German	41	1.28	Chinese	5	0.16
Portuguese	38	1.18	Danish	4	0.12
Italian	34	1.06	Norwegian	4	0.12
French	28	0.87	Romanian	4	0.12
Dutch	12	0.37	Hebrew	3	0.09
Polish	12	0.37	Catalan	1	0.03
Swedish	11	0.34	Croatian	1	0.03
Japanese	9	0.28	Icelandic	1	0.03
Finish	7	0.22	Korean	1	0.03

It is interesting to notice that, even though people create *multi-word* tags, *single-word* tags are predominant. We found 2020 (63%) *single-word* tags followed by tags with *double word*, 867 (27%), *triple word* 149 (4.64%) and so on, as shown in Table 4. We found a single tag, "*StumbleUpon_-_(just_save_and_import_this_file_into_Firefox_bookmark...*" consisting of 11 words.

Table 4: # of Words per Tag

No. of words/ tag	Total	in %
1	2020	62.9676
2	867	27.0262
3	149	4.64464
4	29	0.90399
5	8	0.24938
6	2	0.06234
7	1	0.03117
9	2	0.06234
11	1	0.03117

Among the *single word* tags, the majority are *noun* tags 41.46% (1330), followed by 5.8% (186) *abbreviations*, *acronyms* 4.71% (151), *verbs* 4.61% (148) and *adjectives* 4.30% (138). Adverb, preposition, pronoun, conjunction, preposition, numbers and date tag are very rare. See Figure 5.

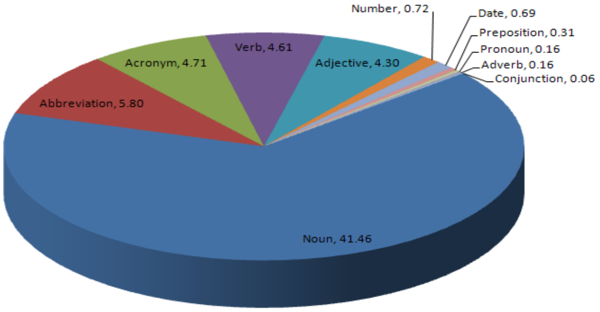


Figure 5: Single Word Tag Distribution

Out of 1330 noun tags, the *singular* and *plural* forms are 698 (52.48%) and 301 (22.63%), respectively. The rest of the nouns are proper and regular nouns. It is important to notice that word (lexical) class ambiguity is quite common in English. During our study we had very often to decide whether to label a tag as noun or verb. To resolve this ambiguity, we consulted the respective resource against the particular tag. With tags used in multiple resources, we gave preference to the resource where the tag was used most times. Finally, notice that *date* tags were found in various formats (e.g., "08", "18092006", "20080811", "06.08.2008"). The *ISO 8601* basic format (*yyyymmdd*) is the most popular.

There are 948 (29.55%) *compound* tags, 84 (2.93%) *mixed* tags and 17 (0.53%) *camelCase* tags. Figure 6 describes the various types of symbols used as separators.

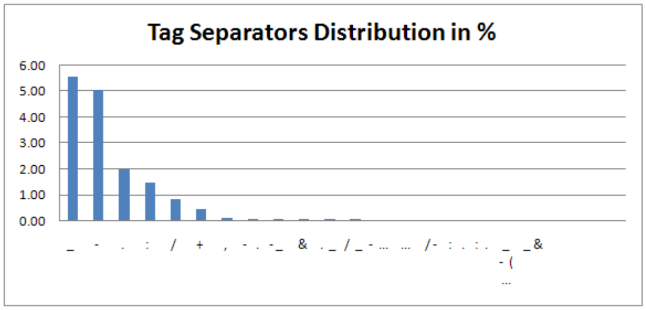


Figure 6: Tag Separators Distribution

As from Table 5, “_” is the most popular (5.58%) followed by “-“ (5.05%), and so on.

Table 5: Tag Separators

Separators Used	Total	in %
_	179	5.58
-	162	5.05
.	64	2.00
:	48	1.50
/	27	0.84
+	14	0.44
,	4	0.12
- .	2	0.06
- _	2	0.06
&	2	0.06
. _	2	0.06
/ _	2	0.06
- ...	1	0.03
...	1	0.03
/ -	1	0.03
: .	1	0.03
: .	1	0.03
_ - (...	1	0.03
_ &	1	0.03

Among the other linguistic facets, we have no *neologism* (0.00%), 67 tags in *starts with-special-symbols* (2.08%) and 86 in *spelling variation* (2.68%). We observed that most tags in the spelling variation category are spelt incorrectly. The probable causes of these tags are a typing error (e.g., applications, bittorent, boagpost, casette), lack of awareness of the correct spelling (e.g., brashes) or intentional variation (e.g., bo0oks) to distinguish from correctly spelled tags.

5. FUNCTIONAL ANALYSIS

Golder and Huberman identified seven functions of tags in bookmarks [12], such as: identifying what (or who) it is about, identifying what it is, identifying who owns it, refining categories, identifying qualities or characteristics, self-reference, and task organization. Kipp [13] attempted to relate tags with time, task and emotion. Keeping in mind this work and based upon our preliminary study of our tag corpora, we developed the tag functional model. We have split the tag functionality in two different groups. They are, *subject-related* and *non-subject related*, as shown in Figure 7. Notice that our functional model, and mainly the *subject-related* class, was highly influenced by two metadata standards, namely, Dublin Core (http://dublincore.org/) and MARC21 (http://www.loc.gov/marc/).

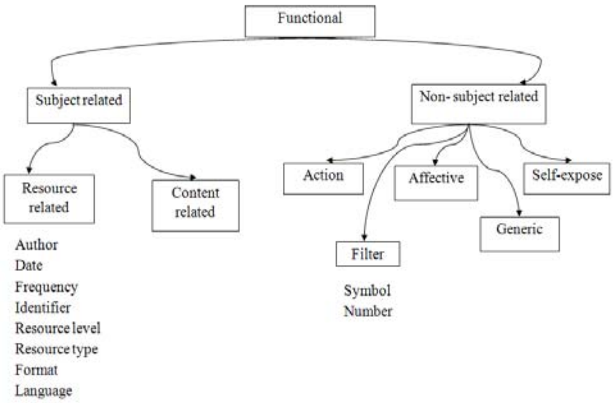


Figure 7: Functional Model

The *subject related* class collects the domain knowledge which describes the resource and its contents. It is divided into two facets, namely *resource related* and *content related*. The *resource related* class accounts for the tags containing resource related information. This category generally deals with the bibliographical information. To analyse the *resource related* tags in a more structured manner, we have identified 8 values, namely: *author*, *date*, *frequency*, *identifier*, *resource level*, *resource type*, *format* and *language*. Besides the tags whose meaning is obvious, *frequency* accounts for the tags related to the frequency of publication of the resource, e.g., *daily*, *weekly; identifier* deals with the tags related to the unambiguous reference to the resource, such as, *URL*, *ISBN*; *resource level* deals with the tags related to the use level of the resource, for example, *basic*, *primary*; *resource type* accounts for the tags defining the nature or genre of the resource, for example, *book*, *article*, *blog*, *wiki*, *poems*, *maps*, *image*, *software*, etc. The *content related* class accounts for the tags describing the content of the document. It includes descriptors, such as, *topical terms*, *personal name*, *corporate name* or *meeting name* (when acting as subject descriptors), *spatial* and *temporal* descriptors, *content categorization* (e.g., tutorial, example, etc.), and so on.

The second dimension, *non-subject related* tags, analyses the operational functionalities of the tags whose meaning is not connected to the subject of the resource. This category is divided into five categories, namely: *action*, *affective*, *self-expose*, *filter* and *generic*. *Action* contains the tags conveying information about the user’s past, present or future task or action; some examples are: *toread*, ****must_do_this* and *downloaded*. The *Affective* facet considers the tags related to the user’s opinions, and emotions towards the resource: some examples are: *boredom*, *classic*, “+++”, “****”. This class allows to learn about the quality and/or reputation of the resources. The tags under *self-expose* reflect a personal statement and involve words such as, *my*, “*I am*”, *Me* (e.g., *My_ StumbleUpon_Favorites*, “*IAm Sparticus*”, *CheckMeOut*). There are certain tags which contain only numbers (e.g., *1*, *4. 5, 101*) or only letters (e.g., *A*, *s*) or special characters (e.g., “,” “|”). They are classified under the category *filter*. Finally, we found tags which do not convey any meaning nor do they provide any link with the resource. They are classified under *generic*; some examples are: *mlf*, *BBDD*, *_abcdef*, *000s*. It would be possible that the tagger did not have the time or patience to read the document properly or to assign meaningful tags, but still (s)he was keen to revisit the resource.

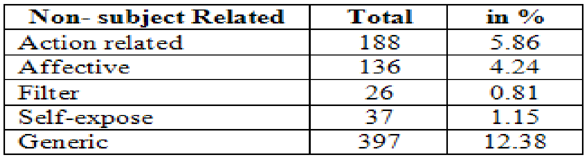
Table 6 provides a detailed description of the *subject-related* tags. *Content-related* tags are the majority (63.84%) while *resource-related* tags are only 7.70%. Within *resource-related*, *resource type* tags are the most frequent (3.99%), followed by *author* (1.18%), *identifier* (0.78%) and so on.

Table 6: Subject Related Tag Distribution

Subject Related		Total	in %
Resource Specific		247	7.70
	Author	38	1.18
	Date	22	0.69
	Frequency	4	0.12
	Identifier	25	0.78
	Resource Level	8	0.25
	Resource Type	128	3.99
	Format	15	0.47
	Language	7	0.22
Content Description		2048	63.84

Within *non-subject-related*, *generic* counts the maximum number of tags, 397 (12.38%) followed by *action*, 188 (5.86%), *affective*, 136 (4.34%) and so on. The details are in Table 7. The reason for the maximum number of *generic* tags could be due to the minimal effort needed to generate them.

Table 7: Non-Subject Related Tag Distribution



Non- subject Related	Total	in %
Action related	188	5.86
Affective	136	4.24
Filter	26	0.81
Self-expose	37	1.15
Generic	397	12.38

The overall result of this analysis is in Figure 8. *Subject related* tags (75%) largely overtake *non-subject related* tags (25%).

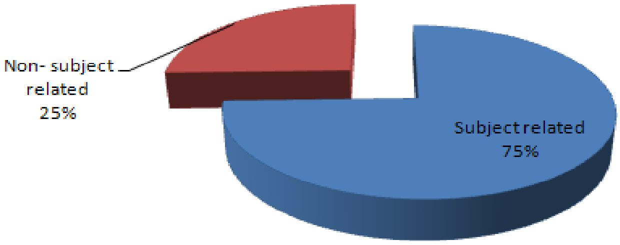


Figure 8: Functional Tag Distribution

6. TAG SOURCE ANALYSIS

A crucial issue is the extent to which tags use words from the resource they describe. Answering this question is the goal of the *tag source* analysis. The model consists of two top-level elements, *taken from resource* and *not taken from resource*, as shown in Figure 9. Element *taken from resource* identifies the tags which appear in the resource whereas *not taken from resource* element identifies the tags not appeared in the resource. Elements *taken from* resource is further categorised into two, *resource identical* and *variation from resource*. Similarly, element *not taken from resource* is categorised into two, namely, *user name* and *others*.

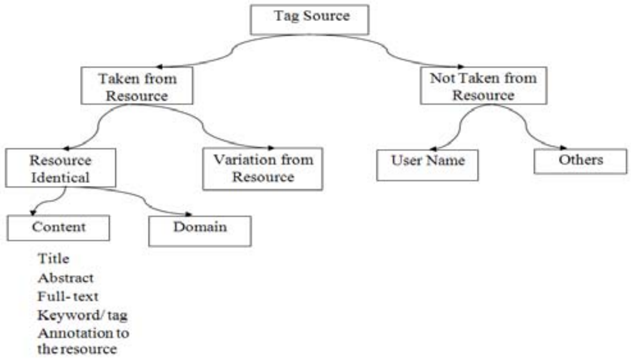


Figure 9: Tag Source Model

Resource identical identifies the tags which appear in the resource. It is further categorised into two, namely *content-identical* and *domain-identical*. For *content-identical*, we have further distinguished tags found in the *title*, *abstract*, *keyword* or *tag list* (author provided) and *annotation* (e.g., reader comments, testimonials, etc.) as the possible sources used for creating tags. *Domain-identical* distinguishes among tags found in material surrounding the documents on the screen, including, *domain name*, *external links*, *domain category list*, *tags cloud*, as available in the resource page and the *popular site list*, as possible source of tags.

We found many tags similar to the text of the resource but not exactly the same. We categorised them as *variation from resource*. We observed several types of deviation. For example, from singular to plural (practice -> practices) or vice-versa, replacement of British text with American text (colour -> color) or vice-versa, addition of some special characters either between the compound tags or at the end of the tags (Barack Obama -> barack-obama, broken -> broken,), and so on.

The user's own name or some other personal name, not related to the document were kept under the category of *user name*. For example, we found a tag "*alexvernon*" created by a user called, "*alexvernon*", also a tag "*broox*" created by a user called "*cell6*".

The fourth element of *tag source* is called "*others*". This category contains the tags not occurring in the document text. Some tags under this category can be interpreted using semantic relations, such as, synonym, homonym, hypernym, hyponym. Other tags have no relations as it happens, for instance, when tags are letters ("*A*", "*s*") or some special characters ("*%s*", "*!*"). We also included the non-English tags under this category.

As from Figure 10, *others* contains over 53% tags.

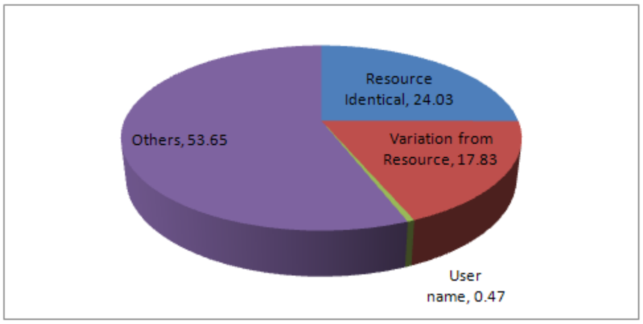


Figure 10: Tag Source Distribution

The study carried out by Heymann [9] on del.icio.us noted that 80% of the tags are from the page and surrounding text. He pointed out that tags are on the whole accurate. He concluded that a substantial proportion of tags are obvious in context, and many tagged pages would be discovered by a search engine. He further pointed out that a tagging system in general works well for media sharing sites like Flick and YouTube, while tagging may be less informative for systems which already have full text. The study carried out by Heckner on *Connotea* (http://www.connotea.org/) [11], noted that there were 54% tags identical to the full text, 16% somewhat different, while 30% showed no relation with the resource. In our study, if we exclude the non-English tags (9.8%) from "*Others*", we still have over 43% tags which show no relation with the document text (see Figure 11). On the other hand, over 17% tags are variations from the document text and only 0.47% tags are encoded from user names, whereas, over 24% tags were found matching with the document text without any variation. This sums up to 21.92% *content* identical and only 2.12% *domain* identical tags.

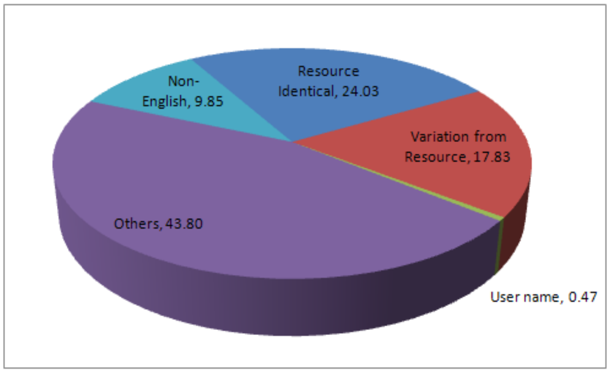


Figure 11: Tag Source Distribution (Non-English tags are treated as separate category)

If we narrow down the tag occurrence to *content-identical* tags (see Figure 12), we see 25% of the tags word appear in the title, an additional 52% of the tags word occur in the full-text (not counting words that appear in the title), tags from annotation (9%), keywords/ tags (4%) and only 1% tags match words in the abstract. The reason for this last figure is that most of the documents in our sample database have no abstract.

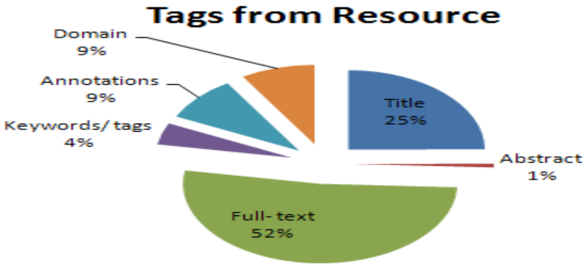


Figure 12: Tags from Resource Text (Identical)

7. WHICH SEMANTICS?

So, what evidence of the use of semantics do we have at the end of the second step of our study? And even more importantly, which semantics do we find in tags? The result of this analysis can provide us with hints about which semantics are considered *useful in practice* by taggers. Remember that, as discussed in the introduction, by *useful semantics* we mean those semantics that are used to *share meaning among different users*, thus allowing for maximal benefit in the reuse of tags. The results described in the previous sections can be summarized as follows.

Linguistic Analysis - only good news:

1. *Language*: tags in English and, more in general in any natural language, are the majority. In particular English tags are 2774 (86.47%) out of total 3208. Notice that English is still the most common and effective language for sharing knowledge on the Internet.
2. *Word type*: *nouns* are the most frequent with 1330 instances (41.46%). Nouns are words that are used to refer to a person, place, thing, quality, or idea; they are one of the basic parts of speech. Among other word classes, *verb* 148 (4.61%), *adjective* 138 (4.30) and others were negligible in quantity.
3. *# of words per tag*: single word tags are the maximum, 2020 (62.96%), followed by double word tags 867 (27%) and triple word tags 149 (4.64%). The other tags are negligible.

To sum up: people tend to largely use natural languages and in particular English. They use mostly single words and the tags with single, double and triple words cover 94.6% of all the tags, while nouns are the most commonly used words.

Functional Analysis - only good news again in that:

1. 75% (2048) tags are subject related and as such directly codify the semantics of the resource.

Tag Source Analysis - good and bad news:

2. Around 42% are resource related tags. It's good news for automated full-text indexing.
3. Around 43% (restricting to English tags) are tags not related to the document text.

At least from what we gather from our study, roughly speaking, it is possible to automate the generation of only half of the tags.

Overall observations:

1. The non-subject related tags are mostly personal tags. However there are certain tags, e.g., the *affective* tags (136), which could be useful to the community, by providing quality related information.
2. For creating compound tags, "_" is the most popular (179) separator. However "/" was found to be the most popular symbol (27) used in constructing relatively complex and multi-faceted tag structures. We will get back to this point later.
3. It appears that most of the *spelling variation* tags (86) are so simply because miss-spelled (typing mistakes).

Overall, as from above, the first part of our study provides evidence of a majority role of useful semantics in tags. To provide even more evidence of the strong role of semantics, The second step in our study concentrated in what we call the *stability of semantics*. The key intuition is that tags which are generally meaningful are most likely used by many users and they remain stable in the tag set, while tags which are personal or specialised in nature are used by very few users and get progressively pushed downwards in the tag list. Here, by *stable tagging*, we mean that tagging in time settles to a group of generally meaningful tags. The idea is not that people stop tagging but, rather, that new users mostly reuse the already available tags. Two factors are important here:

1. *The stability of language*, namely the absolute number and percentage of tags which capture most resources and the absolute and relative number of resources captured;
2. The *stability in time*, despite the increasing number of tags and resources, of the words used for tagging.

To study these factors we concentrated on a specific URL:

http://www.peoplejam.com/blogs/workouts-working-people-how-get-better-results-less-time-gym. Table 8 shows the top 25 tags with their frequency (f) and rank (r). Frequency indicates the number of occurrences of a concept in a tag set, whereas, rank indicates the respective position of a concept based upon its frequency. We collected the tags on 11th August 2008 and on 06th October 2008. The total number of bookmarks and total number of unique tags in the 1st phase of data collection were 141 and 52 respectively, whereas, in the 2nd phase, they were 420 and 82 respectively.

Table 8: Frequency and Rank of the top 25 tags as collected on 11th July 2008 and 06th October 2008.

	11th August 2008		06th October 2008	
Tag	Frequency	Rank	Rank	Frequency
health	62	1	2	150
workout	58	2	1	160
fitness	52	3	3	134
exercise	49	4	4	109
howto	33	5	9	57
gym	29	6	6	73
training	22	7	5	85
tutorial	22	8	10	54
sports	11	9	14	27
tips	9	10	8	58
weightlifting	7	11	7	63
time	6	12	15	20
fitness,	5	13	22	5
muscle	5	14	17	10
weights	5	15	12	39
workouts	5	16	11	48
less	4	17	23	4
article	3	18	18	7
blog	3	19	19	7
get	2	**20**	**31**	2
080807	1	**21**	**35**	1
At	1	**22**	**38**	1
Better	1	**23**	**26**	2
body	1	**24**	**27**	2
efficient	1	**25**	**43**	1

It is easy to notice that the generally meaningful tags, e.g, *health, workout, fitness, gym, training, tips*, are more or less stable in time in their frequency and rank. Their frequency increases at a similar rate as that of bookmarks. These generally meaningful tags make the "head" of the distribution (Figure 13). It is also interesting to see that certain tags, semantically more relevant, move up in their rank independently of their initial position. For example, *weightlifting*, 11th in the 1st phase of data collection, moved up to position 7 after the 2nd phase (Figure 14). It is worth remembering that del.icio.us show the user only the top 10 tags.

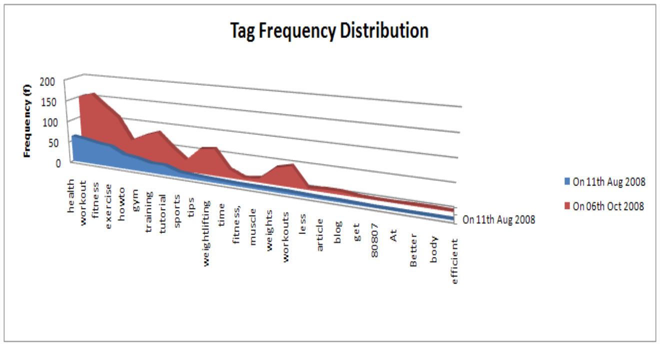


Figure 13: Top 25 Tag Frequency Distribution

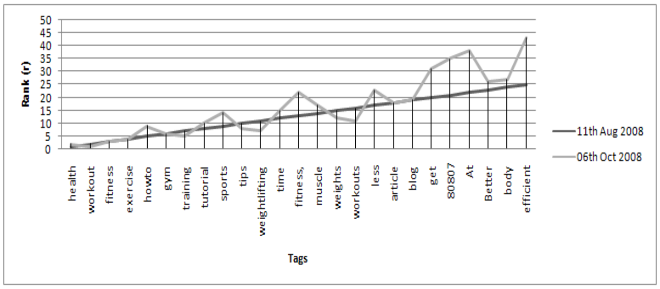


Figure 14: Top 25 Tags Rank

On the other hand the tags, like, *howto, time, less, get, at, better, body, efficient, etc.* were within the top 25 tag set during our first phase of data collection. But in the 2nd phase their frequency remained almost the same even though the total bookmarks for resource increased almost 3 times and, as a result, their rank decreased (bold ranks in Table 8). These relatively meaningless tags form the "*long tail*" of the distribution (Figure 13) along with other personal tags such as, *read, towrite, links, Miscellaneous, personal*, etc. not shown in the figure. But, and this is a crucial consideration, *the long tail does NOT apply to the generally meaningful terms, which are most useful in order to share semantics.* The meaning of a resource can be captured by very few tags and these tags largely capture the shared meaning which allows multiple taggers to categorize the resources itself. To this extent, notice that the top 25 tags, which, as of October 6, consist of only the 30% of the total number of tags, capture the 95.89% of the frequency in July and 94.71% in October.

The semantics we found in tags were not supported in any way by del.icio.us and were provided by the user entirely on a voluntarily and self-interest basis. An obvious consideration which comes next is the following: how much would these numbers improve if the system provided some support for semantics? Of course, an answer to this question depends on the type of support provided by the system. As a simple exercise we tried to identify the semantic relationship among the tags. And we did so by simulating the Natural Language Processing and disambiguation techniques described in [14]. These techniques and system are already used in an existing social network and it would be relatively straightforward to plug them inside a system like del.icio.us. We concentrated on the October 6th tag set and applied the techniques from [14]. As a result, we found many tags semantically related, but scattered because of their syntactical variations and we identified the common factors responsible for this scattering, namely: *synonym, "singular and plural form", "tags with variation", "no semantic relations found"* and "*spelling error*". Here, "*no semantic relations found*" indicates the independent tag set. Based upon the above five factors, we obtained the results in Table 9.

Table 9: Semantically related tags in collapsible form

<table>
<tr><th>Synonym</th><th>Singular and Plural form</th><th>Tags with variation</th><th>No Semantic relations found</th><th>Spelling Error</th></tr>
<tr><td>{workout | exercise}
{health | wellness}
{muscle | strength}</td><td>{workout | workouts}
{exercise | exercises}
{sport | sports}
{article | articles}</td><td>{workout | workout, | workouts46} {health | health,} {fitness | fitness,} {training | training84} {gym | gym, | gym72} {weightlifting | weightlifting61} {howto | howto57} {weight | weight, | weights37} {sports | sports27} {life | life,} {toread | to.read}</td><td>tips | tutorial | lifehacks | time | work | blog | less | out | better | body | Ejercicio | get | lifting | results | 080807 | advice | At | checknew2 | checksoon | conseil | educational | efficient | freizeit | health/exercise | Health/Sexuality | improvement | In | links | lose | Miscellaneous | personal | personaltoolbar | read | reading_is_fundamental | self-improvement | shortcut:How | szacowne.zdrowie | The | To | to_be_read | towrite | utilities | videos | weighttraining | WorkoutTips</td><td>{exercise}
{Excersize}
{shaemus,}</td></tr>
</table>

From Table 9 we can notice that the sets of synonymous terms, namely {health | wellness} and {workout | exercise}, dominate the tag set in terms of both frequency and rank. Similarly, tag sets like, e.g., {exercise | exercises}, {workout | workout, | workouts46}, {sport | sports} were scattered because of their syntactic variations. At the same time the majority of tags have no relations with other tags. But if we look at them closely, we find that most of them are either personal or special tags, and are positioned in the long tail. To provide quantitative evidence of this qualitative analysis we replicated the results of Figure 14 by collapsing terms according to the rules in Table 9. The results are reported in Figure 15 and, as it can be noticed, they show that the long tail phenomenon observed in Figure 14, becomes much more evident and sharp. The holes in the following figure (backend graph) are due to the multiple terms collapsed into a single term.

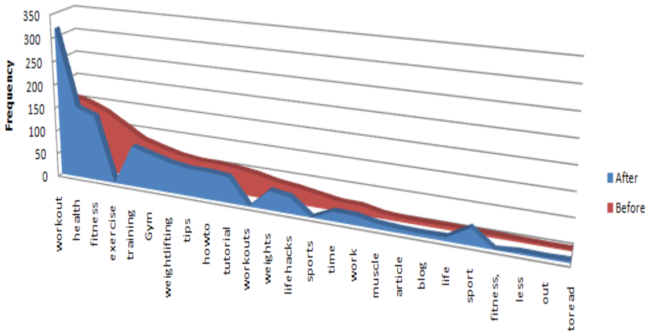
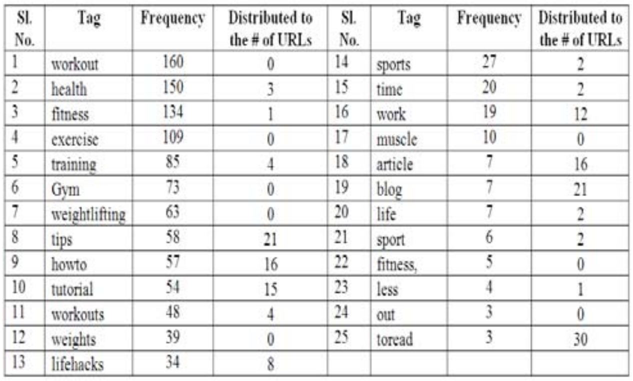


Figure 15: Top 25 Tag Distribution (before and after applying the disambiguation techniques)

What we studied so far was the role of tags associated with only one URL. What happens if we extend this study to all the 51 URLs in our corpora? In this regard we tried to quantify the distribution for those top 25 tags across the rest 50 URLs in our corpora. As we found, out of 25 tags, 17 tags (68%) are distributed across the URLs. in our corpora. See Table 10 for details. The first column represents the ranking and the theird column represents the frequency of the tags.

Table 10: Top 25 tag distribution across the rest 50 URLs in our corpora



Sl. No.	Tag	Frequency	Distributed to the # of URLs	Sl. No.	Tag	Frequency	Distributed to the # of URLs
1	workout	160	0	14	sports	27	2
2	health	150	3	15	time	20	2
3	fitness	134	1	16	work	19	12
4	exercise	109	0	17	muscle	10	0
5	training	85	4	18	article	7	16
6	Gym	73	0	19	blog	7	21
7	weightlifting	63	0	20	life	7	2
8	tips	58	21	21	sport	6	2
9	howto	57	16	22	fitness,	5	0
10	tutorial	54	15	23	less	4	1
11	workouts	48	4	24	out	3	0
12	weights	39	0	25	toread	3	30
13	lifehacks	34	8				

As can be seen, the percentage of reuse for all the tags which are subject-specific is very low and substantially higher for those few tags which have a general meaning (e.g., *howto*, *tutorial*, *article*, *blog*, *toread*, etc.). The more specific a tag is the less reused is. This observation, which in a sense, a posteriori confirms what we should have expected but, at the same time, confirms the fractal nature of the Semantic Web as hyopthesized by Tim-Berners Lee in [10].

8. HIERARCHICAL SEMANTICS

As from the previous sections we have evidence of the use of *simple* semantics. By simple semantics we mean the use of an *atomic* concept as the way to codify the meaning of one or more resources. In most cases in del.icio.us (see Section 4, Table 4) these concepts are denoted by a single term, but some of them need multiple words. It is in fact very well known that it is often the case that a Natural Language (e.g., English) does not provide the (single) word for a concept [12]. This is the main motivation for the use of separators inside tags (see Table 5 in Section 4). But this is not the only use for separators, and in particular for "/". Consider, for instance, the following tags extracted from our corpora: *Computers/Apps*, *computers-Linux*, *computer/windows*, *programming/database*, *programming/css*, *programming/objc/cocoa*, *programming /webdesign*, *web2.0 /client/bookmarks*, *web2.0_tools*, *web2.0apps*, *web2.0tools*, *linux.news*, *linux.port*, *linux_app*, *linux_configure*, *linux_daily*, *linux_gaming*, *oss*, *OSS/GNU/Linux*, *science.computers. language.understanding*, *shopping.brands*, *shopping_stores*, *software*, *support.microsoft.xbox360*. All the above examples are attempts, most likely by knowledgeable users, NOT to define a single concept but, rather, to build a hierarchy of concepts, what, technically, are paths in lightweight ontologies. Even more clear evidence of these attempts can be provided by assembling the tags above into a "standard" tree like representation of lightweight ontologies, as we have done in Figure 16 below.

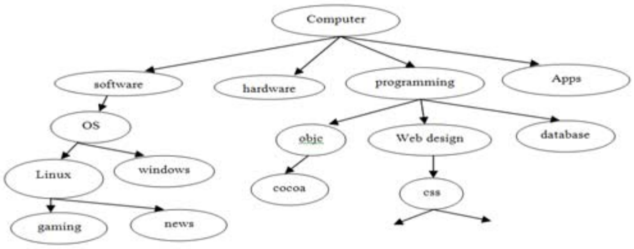


Figure 16: Tags Hierarchy

We found total 105 such complex tags out of total 3208 unique tags. Furthermore these tags tagged by only 123 (1.32%) users out of total 9257 users. Note that, 9257 counts the total number of users who tagged the resources, not only bookmarked. Due to the API support of del.icio.us, it is easy to bookmark in bulk without tagging a resource.

The numbers above are very small, still, they show that the need for more complex hierarchical semantics is there. More evidence of this need can be gathered by studying the semantic (hierarchical) relationships existing among tags. We organized this last study as follows: we took the most frequent terms, as identified in Section 7, and we categories them into three, namely, *synonyms terms (synsets)*, *broader/ narrower terms and associative terms*. *Synonymous* terms were used to establish equivalent relationships among the tag terms, *broader/ narrower* terms to establish hierarchical relationships among the terms, whereas, *associative* terms were used to establish related terms. It is worth to mention that we considered tags modulo syntactic variations and completely avoided misspelt, non-English and personal tags. A part of the resulting lightweight ontology is shown in the following Figure 17.

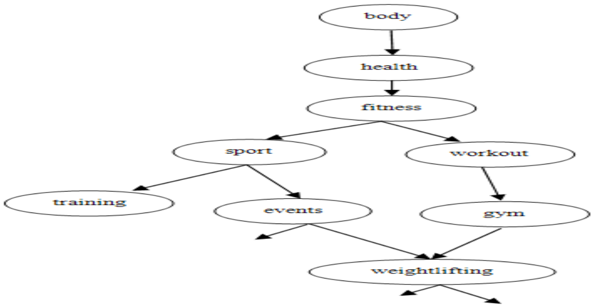


Figure 17: Tag-to-Tag relationships. The figure only includes the broader/ narrower terms.

As we see from the above figure (Figure 17) the simple single word tag corpus leads us to a relatively complex hierarchical tag structure. So *even if at the moment a small minority of users seem able or interested in building hierarchical structures, lightweight ontologies emerge and are used as the result of a collective effort.* Remember that the tags listed above are the ones which are most used.

9. LESSONS LEARNED

So, what have we learned from this study? We report below a few findings, some more grounded in the results, others a bit more speculative.

1. Semantics are used and useful. Not only in the form of simple tags but also in the form of hierarchical structures, i.e., lightweight ontologies. They help users in concentrating on the few meaningful tags thus not considering the long tail of all the possible terms. The open issue of course is how to properly integrate semantics support into systems.
2. There seems to be a huge potential in building semantics as a collective effort, more than the effort of single users.
3. The fractal nature of the Semantic Web pushes even more towards the generation of semantics as a collective distributed effort. Users will join and contribute to the communities which better fit their interests and within these communities they will contribute to the construction of the "right" minimal subset of relevant tags.
4. Natural language plays a fundamental role and it seems the "right" interface for expressing the users' interests. Semantics, in the form of concepts, not necessarily shown to users, can help in handling the many syntactic variations of a concept (including synonyms and the use of different natural languages)
5. Only half of the subject-related tags are generated from the text in the resource. This percentage can surely be increased (e.g., by some clever image analysis techniques). Still, users have a fundamental role in the generation of metadata. The interplay of users, interaction, and knowledge extractors is of course the big open issue to be resolved.

10. REFERENCES

[1] Mai, J.-E. (2004). Classification in Context: Relativity, Reality, and Representation. Knowledge Organization. 31(1), pp.39-48.

[2] Nicholson, D., Neill, S., Currier, S., Will, L. Gilchrist, A., Russell, R. and Day, M. (2001). HILT: High Level Thesaurus Project – Final Report to RSLP & JISC. Centre for Digital Library Research, Glasgow, UK. http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc

[3] Duval, E., Hodgins, W., Sutton, S. & Weibel, S, L. (2002). Metadata Principles and Practicalities. DLib Magazine, 8(4). http://www.dlib.org/dlib/april02/weibel/04weibel.html

[4] Quintarelli, E. (2005). Folksonomies: power to the people. Proceedings of the 1st International Society for Knowledge Organization (Italy) (ISKOI), UniMIB Meeting, June 24, Milan, Italy. http://www.iskoi.org/doc/folksonomies.htm

[5] Shirky, C. (2005a). Ontology is Overrated: Categories, Links and Tags. Shirky.com, New York, USA. http://shirky.com/writings/ontology_overrated.html

[6] Kipp, M. E.I. and Campbell, D. G. Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. http://eprints.rclis.org/archive/00008315/01/KippCampbellASIST.pdf

[7] Guy, M. and Tonkin, E. (2006). Folksonomies: Tidying up Tags? D-Lib Magazine, 12(1). http://www.dlib.org/dlib/january06/guy/01guy/html

[8] Giunchiglia, F. and Zaihrayeu, I. (2008). Lightweight Ontologies. in Encyclopedia of Database Systems, Springer/Verlag.

[9] Heymann, P. (2008). Can social bookmarking improve web search? http://heymann.stanford.edu/improvewebsearch. html

[10] Berners-Lee, T. and Kagal, L. (2008). The fractal nature of the Semantic Web. AI Magazine. Fall 2008.

[11] Heckner, M., Mühlbacher, S. and Wolff, C. Tagging Tagging. Analysing User Keywords in Scientific Bibliography Management Systems. http://journals.tdl.org/jodi/article/viewFile/246/208

[12] Golder, S. A. and Huberman, B. A. The Structure of Collaborative Tagging Systems. http://arxiv.org/ftp/cs/papers/0508/0508082.pdf

[13] Kipp, M. (2007). @toread and cool: Tagging for time, task and emotion. http://eprints.rclis.org/archive/00011414/

[14] I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang. From web directories to ontologies: Natural language processing challenges. In 6th International Semantic Web Conference (ISWC 2007). Springer, 2007. http://www.dit.unitn.it/~ilya/Download/Publications/classification_nlp_challenges_may2007.pdf

[15] Good, B. M., Kawas, E. A. and Wilkinson, M. D. (2007). Bridging the gap between social tagging and semantic annotation: E.D. the Entity Describer. precedings.nature.com/documents/945/version/1/files/npre2007945-1.pdf

[16] Macgregor, G. and McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, 55 (5), 291-300.

[17] Mika, P. Ontologies are us: a unified model of social networks and semantics. Journal of Web Semantics, 5(1).

[18] Zhou, M., Bao, S., Wu, X., and Yu, Y. (2007). An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations. ISWC 2007, Busan, Korea. http://iswc2007.semanticweb.org/papers/673.pdf

[19] Paul, H.; Hector, Garcia-Molina (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report, Stanford InfoLab. http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=2006-10&format=pdf&compression=&name=2006-10.pdf

[20] Giunchiglia, F., Marchese, M. and Zaihrayeu, I. (2007). Encoding Classifications into Lightweight Ontologies. Journal of Data Semantics VIII, LNCS 4380, pp. 57-81.