

Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art

Freddy Limpens, Fabien Gandon, Michel Buffa

▶ **To cite this version:**

Freddy Limpens, Fabien Gandon, Michel Buffa. Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art. 2009. hal-00530371

HAL Id: hal-00530371
https://hal.science/hal-00530371

Submitted on 28 Oct 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art

Emetteur	Freddy Limpens
Contributeurs	Freddy Limpens, Fabien Gandon, Michel Buffa
Relecteurs	Olivier Corby
Date de livraison prévue	TO+06 : 2009/08/01
Date de livraison	2009/07/16
Work package	T3. Social management of shared knowledge representations
Deliverable	T3.5 State of the art on ontology & folksonomy hybrid techniques
Référence	ISICIL-ANR-EA01-FolksonomiesOntologies-20090716.pdf
Version	3
Destinataires	Membres ISICIL

Projet ISICIL :
Intégration Sémantique de l'Information
par des Communautés d'Intelligence en Ligne

Appel ANR CONTINT 2008

ANR-08-CORD-011-05

Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art

Freddy Limpens, Fabien Gandon, and Michel Buffa

Abstract

Social tagging systems have recently become very popular as a means to classify large sets of resources shared among on-line communities over the social Web. However, the folksonomies resulting from the use of these systems revealed limitations: tags are ambiguous and their spelling may vary, and folksonomies are difficult to exploit in order to retrieve or exchange information. This report compares the recent attempts to overcome these limitations and to support the use of folksonomies with formal languages and ontologies from the Semantic Web.

Contents

1 Introduction **3**

1.1 Freely categorizing . . . 3
1.2 The need for a shared vocabulary: tidying up online communities . . . 4
1.3 Comparison of different types of knowledge representations used to index resources . . . 5
1.4 Different ways of considering the link between folksonomies and ontologies . . . 7
1.5 Organization of the report . . . 8

2 Nature and structure of Folksonomies **8**

2.1 Folksonomies as collaborative classification means . . . 8
2.2 Formal definition . . . 9
2.3 Structure and dynamics of social tagging . . . 9
2.4 Looking for common associations in folksonomies . . . 11
2.5 Comparison and intermediary conclusions . . . 12

3 Extracting the semantics of folksonomies **13**

3.1 Measuring the relatedness between tags . . . 14
3.2 Inferring subsumption relations . . . 22
3.3 Clustering tags . . . 24

3.4 Comparison of the approaches and intermediary conclusions . . . 25

4 Semantically enriching folksonomies 26
4.1 Collaborative semantic structuring of folksonomies . . . 27
4.2 Ontologies for modeling folksonomies and online-communities . . . 28
4.3 Infrastructure for linking tags with ontologies . . . 29
4.4 Linking tags with professional vocabularies . . . 32
4.5 Assisting semantic enrichment of tagging . . . 33
4.6 Tagging and collaborative ontology maturing processes . . . 35
4.7 Comparison and intermediary conclusions . . . 37

5 Knowledge sharing in the social and semantic Web 38
5.1 Collaborative information and experts seeking . . . 38
5.2 Sharing on the semantic Web . . . 39
5.3 Semantic Wikis . . . 39
5.4 Comparison and intermediary conclusions . . . 40

6 Discussion 41
6.1 The best of both worlds . . . 41
6.2 Adapting the models and tools to the usages . . . 42
6.3 Perspectives . . . 42

1 Introduction

This report is deliverable T3.5 of ISICIL project ANR-08-CORD-011-05 and proposes a state of the art on ontology & folksonomy hybrids.

1.1 Freely categorizing

To share and index the large number of resources available on the Web raises several issues that systems based on folksonomies Vanderwal (2004), such as del.icio.us for sharing bookmarks, have recently tried to address. On the other hand, the Semantic Web aims at supporting the exchange of information by developing the interoperability between applications available on the Web. To this end, several methods, tools and principles are proposed, among which formal ontologies play a central role. Generally speaking, ontologies are knowledge representations aiming at "specifying explicitly a conceptualization" Gruber (1993). More specifically, formal ontologies use formal semantics to specify this conceptualization and make it understandable by machines. The obstacles to a generalization of ontologies lie mainly in their cost of design and maintenance.

The problem we address here is the need for the users of social Web platforms to find an agreement about the knowledge representations that support their collaborative use of the system. To this regard, folksonomies are often seen as the bottom-up approach, while formal ontologies of the Semantic Web are considered to be necessarily a top-down approach. In this report we try to show that opposing folksonomies and ontologies in this way is counterproductive, and the works we present here show the potential of combining both approaches in order to collaboratively build up solid knowledge representations that are both representative of the communities of users, and at the same time allows for better retrieval or exchange of information.

The Web 2.0 consists essentially in a successful evolution of the Web supported by some principles and technologies. Social tagging and the resulting folksonomies can be seen as two of those principles which have emerged and met a growing success within Web 2.0 applications. The simplicity of tagging combined with the culture of exchange allows the mass of users to share their annotations on the mass of resources. However, the exploitation of folksonomies raises several issues Mathes (2004) and by Passant (2009)(section 2.2.3): (1) the ambiguity of tags, for one tag may refer to several concepts ; (2) the variability of the spelling, for several tags may refer to the same concept; (3) the lack of explicit representations of the knowledge contained in folksonomies; (4) the difficulties to deal with tags from different languages. Another challenge is the need to assist the life-cycle of the folksonomies and the ontologies which support the knowledge bases of social Web applications. Our hypothesis is that the synergy of both folksonomies and ontologies may bring great benefits. Research has been undertaken to tackle the problems posed by the annotation and the exchange of the resources on the Web. The systems or methods they propose strive to reconcile ontology-based models and folksonomy-based models.

1.2 The need for a shared vocabulary: tidying up online communities

Most of the research works we present in this report take place within the social Web which includes all types of groups of people communicating online. These communities range from groups of people who do not know each other in the real life but contribute to the same sharing platform (as in Wikipedia or delicious.com where users contribute to an encyclopedia or a social bookmarking database), to collaborators who work together and exchange knowledge online.

One of the most commonly cited notion about communities with respect to knowledge sharing issues is probably the notion of Community of Practice (CoP) proposed by Lave & Wenger (1991). The notion of CoP defines a group of people gathered by a commitment to a common activity and sharing common interests, proficiencies, and knowledge. However, other notions have emerged to describe the specificity of online communities because the criterion of sharing a common commitment is not always fulfilled in communities communicating online.

Tardini & Cantoni (2005) tried to apply the concepts of semiotics (Saussure, 1916; Hjelmslev, 1963) to describe and characterize online communities. They distinguish two main types of communities. (1)

Paradigmatic communities are groups of people simply having something in common, such as the fact of using the same website for the "Wikipedia visitors" community. It is possible to belong to several paradigmatic communities at the same time, and these communities can be embedded in each other such as the community of "eye specialist surgeons" in the surgeons community. (2) Syntagmatic communities consist in groups of persons who are characterized by their differences and complementarities, and who share a common activity. This type of community is also very close to the concept of CoP, but is less constrained concerning the commitment to a common activity.

The next step consists in finding criteria to evaluate whether a group of online users form a syntagmatic or a paradigmatic community, as this distinction has some consequences about the characterization of the type of knowledge structure which will better fit their needs. For instance, the visitors of a web site form a paradigmatic community which can evolve into a syntagmatic community as soon as the visitors start exchanging more and realize they have a lot of things in common. To this respect, Tardini *et al.* give five conditions which should be fulfilled for a group of users to form a syntagmatic community: (1) a shared environment of communication, (2) a reasonable level of wealth of exchange, which allows for the discovery of common interests, (3) the arousal of a feeling of belonging to a group, (4) the development of a common symbolic space called the "semio-sphere", and (5) the development of a group identity. The development of a semio-sphere is particularly relevant to the scope of this report in that shared ontologies should depict as closely as possible these semio-spheres, and also in that it seems irrelevant to start building collaboratively an ontology if the community is still at the paradigmatic state.

To this respect, the authors have analyzed several online communities (from users of search engines to online video-game players) and came to the conclusion that out of the five conditions mentioned above, the common interests, the feeling of belonging, and the development of a common identity are the most important to constitute a syntagmatic community. The feature of the semio-sphere of a syntagmatic community tells also a lot about the features of the community itself: the more complex the semio-sphere, the more closed the community; on the contrary, the more simple and affordable to newcomers the semio-sphere, the more open the community. This description of the semio-spheres is also close to the distinction between broad and narrow folksonomies (see section 2.1).

1.3 Comparison of different types of knowledge representations used to index resources

Before presenting the different attempts to overcoming the gap between folksonomies and ontologies, let us recall briefly the main types of structured knowledge representations traditionaly used to classify or index resources or documents. These knowledge representations are also called "termino-ontological resources" in the literature and differ mostly from each other in their level of formal structuration, or in their purpose, or in the way in which they are elaborated.

1. **Epistemic classifications** (such as Dewey's classification (Dewey, 1876) used for classifying books in

libraries) consist in defining a vocabularies which can be universally shared. This type of classification (but more flexible than Dewey's classification scheme) is met for instance in the Dmoz[1] initiative to build a directory of Web pages where specialists debate about categories which should be used to classify all the Web pages.

2. The origins of **thesauruses** go back to the 4th century, but the first modern thesaurus is attributed to the British Peter Mark Roget[2]. Modern thesauruses and other types of controlled vocabularies, such as taxonomies, consist in notions or concepts which are defined and hierarchically structured. They provide descriptors used to index documents and are aimed mostly at navigation purposes. The notions composing thesauruses can be contrasted with the concepts of formal ontologies in that they are oriented towards the descriptions of resources, and are not aimed at describing "what something is", but rather "what something is about" (see SKOS specification and the definition of the skos:Concept class[3]). Moreover, the types of semantic relations linking the concepts of thesauruses are usually limited to "broader", "narrower", or "related".

3. Along the expansion of the web, semi-formal and shared knowledge representations have been proposed to organize the information on the Web. Suchapproaches include **Topic maps[4] (Park & Hunting, 2002),** or, with a greater stress on dealing with conflictual views within the communities of users, "semiotic ontologies" Cahier *et al.* (2005). Primarily, semiotic ontologies and Topic Maps can be used by themselves. In some other cases they can also be considered as an intermediary representation to formal ontologies, in that they are not extended by a "referential formalization"[5] but are based on "semiotic expressions" or Topics dealing with another type of semantics which rely mostly on human interpretation. These approaches differ from formal ontologies in their purpose, which is not to obtain a formal and operational scheme, but rather "description networks" used by humans to navigate a corpus of documents and resources.

4. **Formal ontologies** consist in a specification of the conceptualization of a domain of knowledge with the help of formal concepts and properties linking these concepts (Gruber, 1993). They are at the core of the original vision of the Semantic Web proposed by Berners-Lee *et al.* (2001): "The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". Thus, ontologies are at the interface between humans and machines, and can be seen as the formalization of a field of knowledge, given for a specific problem

[1] http://www.dmoz.org/
[2] for an historical review of Roget's thesaurus, see Dolezal (2005)
[3] http://www.w3.org/2004/02/skos/core#Concept
[4] http://topicmaps.org/xtm/
[5] in the sense that their semantics is "referential" (Rastier, 1994), that is, based on objective and measurable features of the objects to which the concepts refer.

or task Bachimont (2000)[6]. This formalization of a domain allows in turn for making inferences and expand greatly the possibility of querying when looking for resources annotated with formal ontologies.

In comparison with all the above mentioned structures of knowledge representation, folksonomies can be seen as semiotic representations of the knowledge of a community, but they do not include any semantic structure. They are not either truly elaborated collaboratively, since they consist merely in a social aggregation of individual knowledge. However, their indisputable advantage over the other types of representations we mentioned above is their simplicity (they require a minimal cognitive cost of elaboration (Sinha, 2005)), that made them adopted by a mass of users.

1.4 Different ways of considering the link between folksonomies and ontologies

The aim of this report is to give the current approaches to reconcile folksonomy-based and ontology-based approaches to support social interactions. The bridging of ontologies and folksonomies can be done in different ways:

Deriving ontologies from folksonomies It is possible to take into account the multiple dimensions of folksonomies as they consist in a triadic structure where tags are associated by people to resources ("who tags what"). This is what Mika (2005), for instance, does in order to extract broader and narrower relationships between tags and to build what he calls "lightweight ontologies", that is, ontologies which consist in an ensemble of terms connected with a limited set of semantic relationships (broader, narrower, related for example).

Synchronizing ontologies life-cycle with folksonomies : Folksonomies, thanks to their versatility and their ability to integrate fresh new vocabularies, are a good opportunity to populate ontologies or to suggest new concepts that could be candidates to be added in an ontology. Passant (2007) exploited this feature of folksonomies to populate a corporate ontology and to support the folksonomy based system he developed to annotate a corporate blog. In his system, ambiguous tags are associated to clearly defined concepts by the users while tagging.

Semantically enriching folksonomies Even if ontologies and folksonomies may remain different entities, several approaches have proposed to semantically enrich folksonomies by adding a semantic layer, or by attempting to semantically structure them with the help of other already available ontologies. For instance, Specia & Motta (2007) have developed a system that apply several semantic treatments to a folksonomy, such as finding equivalent tags or grouping similar tags based on similarity measures computed according to the structure of the folksonomy. Then, they query ontologies on the Semantic Web and try to match the tags

[6] Bachimont (2000) gives the following definition of an ontology, or more precisely, of the "modelling of an ontology": "Defining an ontology for knowledge representation tasks means defining, for a given domain and a given problem, the fonctionnal and relationnal signature of a formal language and its associated semantics".

from these clusters with concepts from ontologies in order to link the tags with semantic relationships. The main limitation of such an approach is the limited coverage of currently available ontologies.

Semantic Web formalism for interoperability Another great benefit of combining ontologies and folksonomies lies in the interoperability brought by the formalism of the Semantic Web. The Linking Open Data project [7] consists in extending the Web with data sources semantically interconnected and which publish varied open data sets in RDF format and following a set of ontologies describing the different types of resources. Ontologies from the Linking Open Data initiative and ontologies include SIOC[8] used to describe online communities exchange, or SKOS [9] used to describe thesauruses.

1.5 Organization of the report

This report is organized as follows. In Section two, we present the different approaches that analyze the nature of folksonomies and tags. Section three is concerned with the analysis of the semantics inherent to the folksonomies and the relationships between the tags which can be extracted in order to build ontologies. Section four will cover methods which semantically enrich folksonomies or which integrate tagging practices in ontology maturing processes. Section five will give an overview of different types of usages of knowledge sharing platforms, and section six will conclude this report with a discussion.

2 Nature and structure of Folksonomies

In this section we focus on research works which analyze the nature and structure of social tagging systems and folksonomies in order to better understand their dynamics and their semantics.

2.1 Folksonomies as collaborative classification means

According to Golder & Huberman (2005), social tagging can be seen as a cognitively lighter alternative system of classification to controlled vocabularies and hierarchical systems, which can be seen in a hierarchy of folders for instance. Social tagging is also about sense making since the goal of a tag for its author is to organize its knowledge sources with labels which are a way of making sense of the resources he tags. Tags are then an important sign of what matters for the users and how he describes it.

But social tagging is also about collaborative sense making, and as such, has the potential of revealing the fuzziness of the manifold individual categories merged under the same tag. In the same trend of ideas, Veres (2006) says that tags are the results of ad hoc categorizations, that is, categories which interface between each

[7] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/
[8] http://sioc-project.org/
[9] http://www.w3.org/2004/02/skos/

user's "world model" in order to achieve a goal. But their linguistic properties reveal that tags can also be similar to standard categories in taxonomies.

Golder & Huberman (2005) detailed seven functions that tags may perform for bookmarks in the context of a typical application of social tagging: (1) "Identifying What (or Who) it is About", that is, the topic of the item tagged; (2) "Identifying What it Is", for example an "article", a "blog" or a "book"; (3) "Identifying Who Owns It", or also to whom this bookmark may be forwarded (see also the "network tags" in delicious.com social bookmarking service); (4) "Refining Categories", that is, tags which refine or qualify existing categories, such as numbers; (5) "Identifying Qualities or Characteristics" such as adjectives characterizing the opinion of the author; (6) "Self Reference", such as tags beginning with "my"; (7) "Task Organizing" which correspond to a particular type of *ad hoc* categories, oriented towards a specific task such as "toread".

Folksonomies have also been characterized by Vanderwal (2004) who distinguish "narrow folksonomies", in which the personal use of tags is predominant, and "broad folksonomies" in which the use of tags is oriented towards more collective and social purposes (which may correspond in some cases to the first three functions given by Golder & Hubermann). Folksonomies are thus a combination of terms which can serve collaborative categorization, and other terms which are only useful for their authors.

2.2 Formal definition

In order to further analyze the structure of folksonomies, we have to model them formally. Hotho *et al.* (2006) thus proposed a formal definition of a folksonomy which they model as a tuple $F := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called users, tags and resources, respectively. Y is a ternary relation between them such that $Y \subseteq U \times T \times R$, and is called tag assignment. As a collection of data provided by a group of individuals, a folksonomy can be seen as the collection of the "personomies" of all the users. Let us call Pu the personomy of a given user $u \in U$, Pu is the restriction of F to u, i. e., $Pu := (Tu, Ru, Yu)$, with $Yu := (t, r) \in T \times R | (u, t, r) \in Y$ that is, the set of all the tag assignments of user u. As Mika (2005) showed it, such a model induces a graph structure of folksonomies. Thus, a folksonomy can also be seen as tripartite hypergraph $H(F) = \langle V, E \rangle$ where the vertices are given by $V = U \cup T \cup R$ and the edges by $E = u, t, r | (u, t, r) \in F$ (see the graphic representation of a folksonomy given by Halpin *et al.* (2007) in figure 1).

2.3 Structure and dynamics of social tagging

Golder & Huberman (2005) proposed one of the earliest quantitative analysis of social tagging in which they discuss its nature as well as the dynamics which can be uncovered with statistical analysis lead on the multi-dimensions structure of folksonomies. Golder & Hubermann give some trends in the use of tags in a social bookmarking system (delicious.com). They remark that users have a tendency to use first more general terms when tagging, the first tag having the greatest frequency of occurrence among all the user's tags, and successive

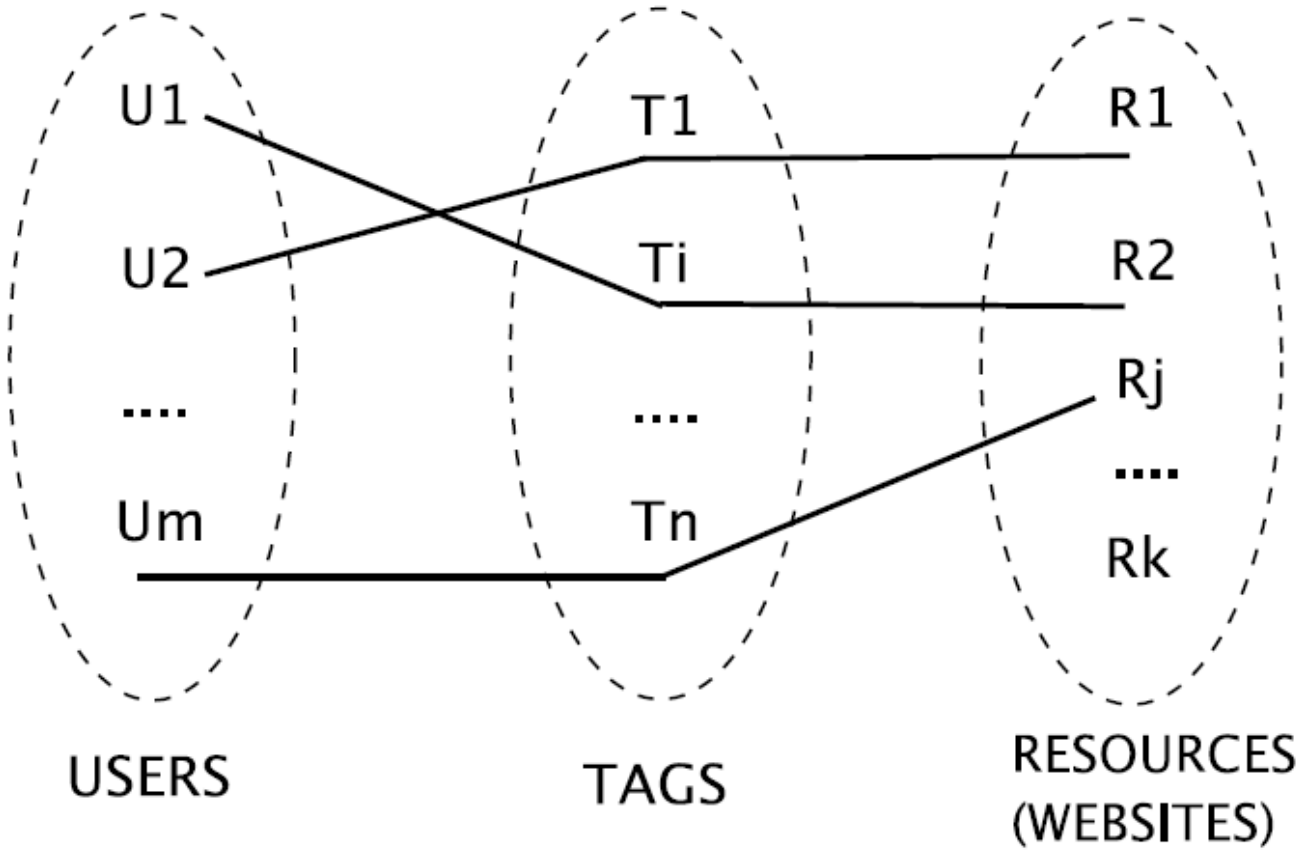


Figure 1 – Tripartite graph structure of a tagging system. An edge linking a user, a tag and a resource (website) represents one tagging instance (Halpin *et al.*, 2007)

tags having generally a smaller frequency. They also observed stable patterns in the distribution of tags for a given resource (URL in delicious.com). Empirically, once a URL has been bookmarked more than 100 times, each tag's frequency remains in a stable ratio of the total frequency of all the other tags used for this URL.

Halpin *et al.* (2007) pursued this analysis of the dynamics of folksonomies and looked for distribution laws in the frequency of use of the tags. They make the same hypothesis that Golder & Hubermann suggest, that the most used tags to annotate a resource remain the same after a certain amount of time, and they show that this distribution follows a power law. They verify that hypothesis for the seven to ten tags most often associated to popular Web resources posted on delicious.com. These observations may be explained by an imitation process, augmented, in the case of delicious.com, by popular tag suggestions while tagging.

But, as Golder & Hubermann suggest, the stability observed in the distribution of the most popular tags persists even for less common tags, which are not shown as suggestions. The choice of the same tags may also be explained by the fact that users share some of the knowledge they express individually when tagging bookmarks. Golder & Hubermann add that this stability in the characterization of some items is linked with the stability of the ideas and characteristics symbolized by the tags; and that, likewise, this stability may no longer persist when a new concept emerges for describing the same items. This was the case, for example, when the concept "ajax" emerged within the realm of Web designers to describe a set of technologies which were all previously known but not named under a single term.

It is also interesting to look at the distribution of tags for smaller folksonomies, as for instance, Passant (2009) in the context of a corporate folksonomy. In this folksonomy, Passant (2009) show that tags follow a distribution in which a lot of tags are used a few times. For example, out of the 12257 tags used to annotate 21614 blog posts, 68% are used at most two times, and only 10% are used more then 10 times. As Hayes *et al.* (2007) showed, it is more difficult to apply classical clustering techniques on this type of distribution in which tags do not neatly partition the annotated data. Indeed, in these cases one should include the content of the annotated data in the analysis of the folksonomy structure.

Concerning the relationships between the tags in a folksonomy, Halpin *et al.* looked for semantic relationships between the most used tags with the help of inter-tag correlation graphs. Each node of these graphs represents a tag and can be seen as a circle whose diameter is weighted by the frequency of occurrence of this tag. The length of the edges of these graphs is weighted by their degree of cooccurrence. The degree of cooccurrence $CoocDegree(T_i, T_j)$ of a pair of tags T_i, T_j is given by :

$$CoocDegree(T_i, T_j) = \frac{N(T_i, T_j)}{\sqrt{N(T_i) * N(T_j)}}$$

Where $N(T_i)$ and $N(T_j)$ denotes the number of times each tag T_i and T_j is used individually to tag all pages, and $N(T_i, T_j)$ denotes the number of times two tags are used to tag the same page, summed over all pages. This visualization (shown in 2) can be seen as a tool for assisting the construction of ontologies out of folksonomies by helping identify visually the most related tags to a given tag.

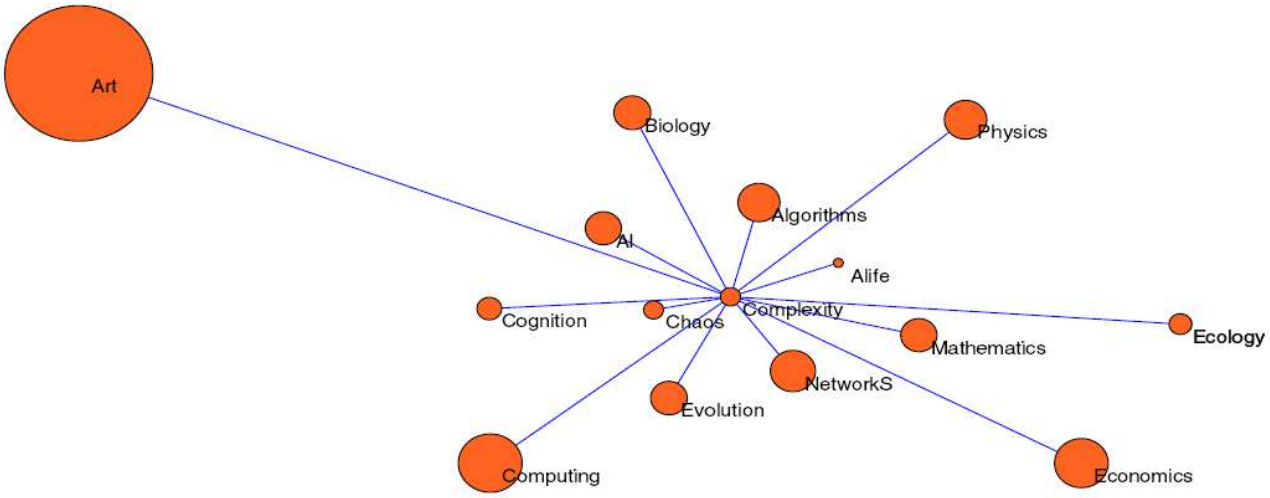


Figure 2 – Visualization of a tag correlation network, considering only the correlations corresponding to one central node "complexity" (data source, delicious.com) (Halpin *et al.*, 2007)

2.4 Looking for common associations in folksonomies

Other works proposed to apply data mining methods to the tripartite model of folksonomies in order to retrieve information in their structure. Jäschke *et al.* (2008) proposed to use formal concept analysis techniques in order to discover the subsets of users sharing the same conceptualizations on the same resources. To do so, they build triples of sets ({R}, {U}, {T}) called tri-concepts where each user of the set {U} has tagged each resource of the set {R} with all the tags of the set {T}. According to the authors, extracting tri-concepts from folksonomies is a first step to build more structured ontologies from folksonomies. Ontologies are thus seen as social constructions where each concept is described by a set of tags which belong to a set of users and are used to characterize a certain kind of resources.

Other data mining techniques have been applied by Schmitz *et al.* (2006) to extract association rules from folksonomies. The first step is to project the tripartite model (Resources, Users, Tags) onto a two-dimension structure called a context in formal concept analysis (Wille, 1982). For instance, one can consider all the tuples (Users, Resources) associated to a set of tags T_x. Then Schmitz *et al.* (2006) apply classical rule mining techniques as proposed by Agrawal & Swami (1993). An example of association rule that may be derived from this projection is: all the users associating tags from the set T_A to a set of resources, often associate the tags from the set T_B to the same set of resources. This kind of association rule may be exploited for example in a recommendation system. Other types of association rules may be a powerful tool to identify sub-groups of users sharing the same tagging practices or interested in the same topics.

	Qualitative study	Quantitative study
Golder & Huberman (2005)	usages of folkso.	
Vanderwal (2004)	broad/narrow folkso.	
Veres (2006)	linguistic nature of tags	
Mika (2005)		graph structure of folkso.
Hotho *et al.* (2006)		formal definition
(Halpin *et al.*, 2007)		power law distribution of tags
Schmitz *et al.* (2006)		association rules mining
Jäschke *et al.* (2008)		formal concept analysis

Table 1 – Comparison table of the approach of section 2 analyzing folksonomies

2.5 Comparison and intermediary conclusions

In table 1, we compare the different approaches presented in this section. We divided these contributions in two categories. First, we can mention the qualitative studies conducted on folksonomies. Golder & Huberman (2005) have analyzed the usages of folksonomies and have proposed seven functions that tags may perform for bookmarks in the context of a typical application of social tagging. Vanderwal (2004) distinguished broad folksonomies (when tags tend to be understandable by numerous users) from narrow folksonomies (when tags are more user-centered). Veres (2006) tried to define the linguistic nature of tags and showed that some tags correspond to taxonomic categories, while other tags correspond to ad hoc categories serving user's purposes.

Second, we distinguished the contributions which focus more on a quantitative analysis of folksonomies. Mika (2005) and Hotho *et al.* (2006) proposed a formal definition of folksonomies, but Mika (2005) pointed out their graph-like properties and defined them as tripartite hypergraphs. Halpin *et al.* (2007) pursued this analysis of the dynamics and usages of folksonomies initiated by Golder & Huberman (2005) and showed that the distribution of most frequent tags of popular web pages on delicious.com follow power laws. Schmitz *et al.* (2006) applied classical rule mining techniques to discover association rules within folksonomies, and Jäschke *et al.* (2008) used formal concept analysis methods to unveil similar conceptualizations in the tagging of resources shared by groups of users of a social bookmarking site.

3 Extracting the semantics of folksonomies

In this section we focus on methodologies and system aimed at deriving ontologies out of folksonomies. The first step in this task is to measure the semantic relatedness between tags. Since usually no explicit semantic relationships are given when users tag, this relatedness have to be first computed by analyzing the tripartite structure of folksonomies (as proposed by Cattuto *et al.* (2008) or Mika (2005)). Then (Cattuto *et al.*, 2008) proposed to semantically ground these measures, while other tried to infer semantic relationships out of this

analysis (see section 3.2). Another type of approach consists in grouping similar tags together, that is tags with close similarity measures, in order to organize tags or to further process these clusters for ontology maturing processes (see section 4.5 for the details of this application of clustering)

3.1 Measuring the relatedness between tags

Cattuto *et al.* (2008), and latter Markines *et al.* (2009), proposed different ways of measuring the similarity between tags and resources in a folksonomy. These approaches can be seen as generalizations of several other approaches (like Mika (2005); Specia & Motta (2007)). The computation of similarity of tags is often the first step to further process the folksonomy data and infer semantic relationships between tags (see 3.2), or to cluster similar tags (see 3.3).

Simple cooccurrence counting

Given a folksonomy $F(U, T, R, Y)$ (see section 2.2) and given a post (u, T_{ur}, r), that is, a subset of the folksonomy corresponding to an annotation of a user u of a ressource r with a set of tags T_{ur}. The similarity measure given by the simple cooccurrence method counts, for a couple of tags t_1 and t_2, belonging to the folksonomy F with $t_1 \neq t_2$, the number of posts which contain both t_1 and t_2

Projection of the tripartite structure of folksonomies

Mika (2005) proposes looking at folksonomies as semantic structures emerging from the usages of the communities. He suggests building out of folksonomies "lightweight ontologies" by unveiling the semantics between tags. To this end, Mika proposed looking at associations via the resources and associations via the users. To achieve this task, Mika projects the tripartite hypergraph of a folksonomy $H(F) = \langle V, E \rangle$ (with $V = U \cup T \cup R$ and $E = u, t, r | (u, t, r) \in F$, and where R is the set of composed of $card(R)$ Resources, U the set composed of $card(U)$ Users, and T the set composed of $card(T)$ of Tags) onto different kinds of two-modes graphs. These two networks correspond to two different ways of projecting the tripartite structure of folksonomies.

The first projection establishes relationships between tags via their pattern of co-occurrence on the resources they are associated with. This projection can be represented by a matrix made of $card(R)$ lines and $card(T)$ columns; when filling up the lines of this matrix, for each resource r_i, we count the number of times each tag t_j has been associated to r_i.

The second projection allows to group similar communities of interest, that is, subsets of users using the same tag. The matrix representation of this projection is made of $card(U)$ lines and $card(T)$ columns; when filling up the lines of this matrix, for each user u_i, we count the number of times each tag t_j has been used by u_i.

Then, Mika extracts from the first projection a weighted one-mode graph connecting tags based on resource associations, and from the second projection a one-mode weighted graph connecting tags based on user

associations. In the case of the user-based association of tags, for a given pair of tags, the weights of the graph are given by the number of users who used both tags at least once. Figure 3 shows an example of such tags graphs build from an excerpt of delicious.com tags and in which a link is drawn between two tags when the weight of the link between these tags is above an arbitrary threshold.

FolkRank based measure of similarity

Hotho *et al.* (2006)developed the FolkRank algorithm which is an adapted version of the PageRank algorithm used for ranking query results and associating a weight to the folksonomy elements (tags, users or resources). Following the main idea of the PageRank algorithm (Brin & Page, 1998), the idea behind the FolkRank algorithm is that a resource tagged by important users with important tags becomes important itself. The same type of relationships being, conversely, true for tags and users, the aim of the FolkRank algorithm in our case is to compute a ranked list of "relevant" tags for a given tag, the most relevant being the most closely related.

The weight spreading computation of the PageRank algorithm cannot be applied directly to the folksonomy since it is a hypergraph (see section 2.2). Thus, the first step is to convert the folksonomy into an undirected graph G_F, where the vertices V consist of the disjoint union of the sets of tags, users and resources so that $V = U \oplus T \oplus R.$, and the edges correspond to all the cooccurrences between the users, tags, or resources (for instance, an edge is drawn between the node corresponding to a user and all the tags he has used at least once). Hotho *et al.* (2006) then apply the weight propagation mechanisms between all the nodes of this undirected graph in order to compute the weight factor $R(v)$ of all the nodes v of the folksonomy graph such that:

$$R \leftarrow c(\alpha R + \beta AR + \gamma P)$$

Where A corresponds to the adjacency matrix of G_F, P is a preference vector where the elements of G_F are given a specific weight, α, β, and γ are constants, and c is a normalization factor such that $\|R\| = 1$. α is a damping factor which is used to avoid oscillation and speed up convergence, while β and γ control the influence of the preference vector P.

In the case of the computation of related tags for a given tag t, belonging to the set of tags T of the folksonomy $F(U, T, R, Y)$, Cattuto *et al.* (2008) apply the above weight propagation with a high weight for t in the preference vector P and compute the vector R_t for all the other tags. Then, the resulting vector is compared to the case where the weight propagation computation is performed without a preference vector P (which corresponds to the case when $\gamma = 0$). Like this, one computes the winners (and losers) that arise when giving preference to a specific tag in the preference vector P. The tags that, for a given tag t, obtain the highest weight are considered to be the most related to t.

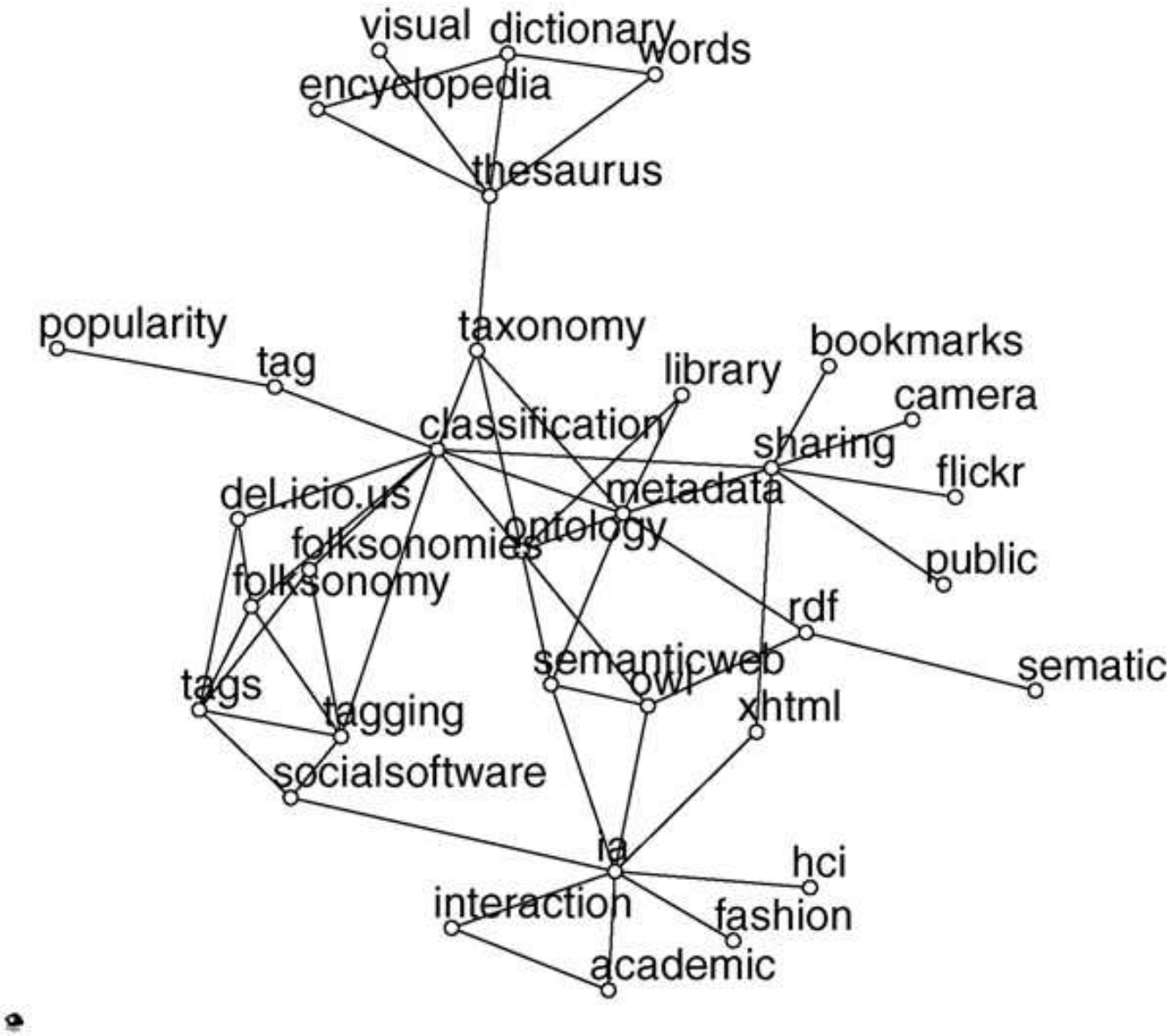


Figure 3 – del.icio.us tags linked thanks to a projection of the folksonomy based on users's association (Mika, 2005)

Distributional aggregation and cosine distance

The following three distributional aggregation of the folksonomy space are based on three different vector space representations of the folksonomy F, named "contexts" by Cattuto et al, and latter generalized as "distributional aggregations" by Markines *et al.* (2009). The idea is to project the tri-partite model of folksonomy into bi-partite representations by aggregating the data according to a given context. For each type of context, we compute the components of the vectors v_t representing each tag in the context:

- **Tag-Tag Context** : the entries of each tag vector v_t corresponds to the cooccurrence with all the other tags as defined above.
- **Tag-Resource Context** . For a tag t, the vector v_t is constructed by counting how often a tag t is used to annotate a certain resource r.
- **Tag-User Context .** For a tag t, the vector v_t is constructed by counting how often a tag t is used by a certain user u.

For instance, in figure 4 we see an example of small folksonomy where two users annotate three resources with three tags. If we pick the Tag-resource context, the matrix representation corresponding to this type of aggregation for the example folksonomy will look like what we give in table 2. For example, the vector of the tag "news" will be $v_{news} = (2, 0, 1)$.

	cnn.com	www2009.org	wired.com
news	2	0	1
web	0	1	1
tech	0	1	1

Table 2 – Example of a distributional aggregation in the tag-resource context of the folksonomy example of Markines *et al.* (2009).

Then the similarity measure between two tags t_1 and t_2 is computed thanks to the cosine distance between the tag vectors v_1 and v_2 representing them:

$$\cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|_2.\|v_2\|_2}$$

Figure 5 provides examples of most related tags using different kind of measures explained above.

Mutual information measure and framework for evaluating similarity measures within folksonomies

Markines *et al.* (2009) proposed a new measure of similarity, the mutual information measure, and a framework to evaluate the different types of similarity measures one can compute within the structure of folksonomies between tags, but also between tagged resources. The first step before measuring similarities is to aggregate

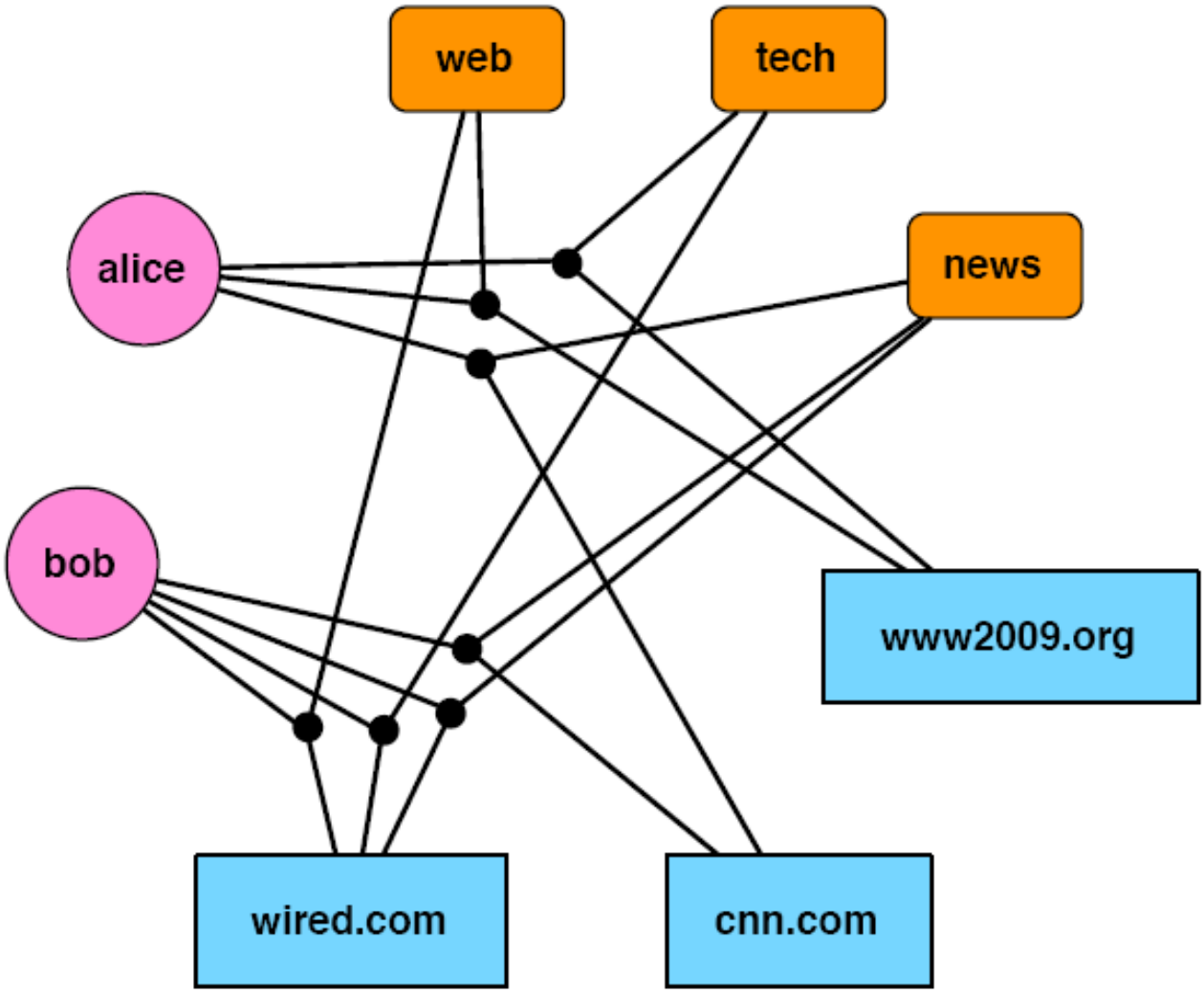


Figure 4 – Example folksonomy proposed by Markines *et al.* (2009). "Two users (alice and bob) annotate three resources (cnn.com, www2009.org, wired.com) using three tags (news, web, tech). The triples (u; r; t) are represented as hyper-edges connecting a user, a resource and a tag. The 7 triples correspond to the following 4 posts: (alice, cnn.com, {news}), (alice, www2009.org, {web, tech}), (bob, cnn.com, {news}), (bob, wired.com, {news, web, tech})."

rank	tag	measure	1	2	3	4	5
13	web2.0	*co-occurrence*	ajax	web	tools	blog	webdesign
		folkrank	web	ajax	tools	design	blog
		tag context	web2	web-2.0	webapp	"web	web_2.0
		resource context	web2	web20	2.0	web_2.0	web-2.0
		user context	ajax	aggregator	rss	google	collaboration
15	howto	*co-occurrence*	tutorial	reference	tips	linux	programming
		folkrank	reference	linux	tutorial	programming	software
		tag context	how-to	guide	tutorials	help	how_to
		resource context	how-to	tutorial	tutorials	tips	diy
		user context	reference	tutorial	tips	hacks	tools
28	games	*co-occurrence*	fun	flash	game	free	software
		folkrank	game	fun	flash	software	programming
		tag context	game	timewaster	spiel	jeu	bored
		resource context	game	gaming	juegos	videogames	fun
		user context	video	reference	fun	books	science
30	java	*co-occurrence*	programming	development	opensource	software	web
		folkrank	programming	development	software	ajax	web
		tag context	python	perl	code	c++	delphi
		resource context	j2ee	j2se	javadoc	development	programming
		user context	eclipse	j2ee	junit	spring	xml
39	opensource	*co-occurrence*	software	linux	programming	tools	free
		folkrank	software	linux	programming	tools	web
		tag context	open_source	open-source	open.source	oss	foss
		resource context	open-source	open	open_source	oss	software
		user context	programming	linux	framework	ajax	windows
1152	tobuy	*co-occurrence*	shopping	books	book	design	toread
		folkrank	toread	shopping	design	books	music
		tag context	wishlist	to_buy	buyme	wish-list	iwant
		resource context	wishlist	shopping	clothing	tshirts	t-shirts
		user context	toread	cdm	todownload	todo	magnet

Figure 5 – Examples of most related tags for different measures (Cattuto *et al.*, 2008)

the tripartite structure of folksonomies onto two-mode views of the data, just as what Mika (2005) and Cattuto *et al.* (2008) did in fact, but Markines *et al.* (2009) propose a generalization of these methods of reduction of the dimensionality of the tagging data.

Non incremental aggregation methods Markines *et al.* (2009) called the aggregation method applied by Mika (2005) (see section 3.1) "projection" aggregation, since the goal is to obtain a one-mode graph view of the three-mode structure of a folksonomy. This method, and the method proposed by Cattuto *et al.* (2008) called "distributional aggregation" and described above in section 3.1, are considered by Markines *et al.* (2009) as non-incremental, since the whole similarity matrix has to be recalculated after each user add a new annotation. Thus, these type of aggregation are not scalable, that is, their computation time does not grow constantly with the growth of the folksonomy.

Incremental aggregation methods To overcome this limitation, Markines *et al.* (2009) propose another type of aggregation, called "macro-aggregation" (in contrast with the distributional measures which can be seen as "micro-aggregations") which consists in (1) considering the tagging of each user separately, and then (2) aggregate across users, that is, to sum the local similarity calculated for each user's data set.

In addition, and in order to take into account the similarity of two resources tagged by the same users but with no tags in common, Markines *et al.* (2009) proposed another way of calculating local similarities, called "collaborative aggregation". The objective of the collaborative aggregation method is achieved by adding a special "user tag" (respectively "user resource") to all resources (respectively tags) of user u. Let us take the example of the tags "news" and "web" for the user "alice" taken from the folksonomy of figure 4. If we add the virtual resource "alice_R" to the binary matrix representing alice's tagging (see table 3) , we will have a non-zero local similarity between the tags "news" and "web" for the user "alice" since these two tags "cooccur" on the virtual resource "alice_R". Then the similarity measure is calculated as in the case of macro-aggregation by summing local similarities across users.

	cnn.com	www2009.org	wired	alice_R
news	1	0	0	1
web	0	1	0	1

Table 3 – Binary matrix representation for the tags "news" and "web" for the user "alice". The last column is the "virtual resource" added to account for the fact that "news" and "web" are used by the same user, but without being cooccurrent. (Markines *et al.*, 2009)

Mutual information similarity measure Markines *et al.* (2009) then give and evaluate several types of similarity measures which can be performed on the four types of aggregation methods mentioned above : matching similarity, overlap similarity, Jacqard similarity, dice coefficient, cosine similarity (as defined above in section 3.1), and mutual information similarity. The detail of the computation of the first four measures

being given in the article, and the cosine similarity being defined similarly to Cattuto *et al.* (2008), we will give here the mutual information similarity measure which outperformed the other in the evaluation conducted by Markines *et al.* (2009) (see below).

Let us take two tags x_1 and x_2, with X_1 and X_2 their vector representation composed of the resource elements w_{xy}. For projection aggregations, the binary vector X can be seen as a set, and $y \in X$ means $w_{xy} = 1$ and $|X| = \sum_y w_{xy}$. Similarly, for a single user u, $y \in X^u$is equivalent to $w_{u,xy} = 1$ and $|X^u| = \sum_y w_{u,xy}$. The mutual information similarity $\sigma(x_1, x_2)$ of two tags x_1 and x_2 is defined for the projection and distributional aggregation as:

$$\sigma(x_1, x_2) = \sum_{y_1 \in X_1} \sum_{y_2 \in X_2} p(y_1, y_2) \log \frac{p(y_1, y_2)}{p(y_1)p(y_2)}$$

where $p(y)$ is the fraction of tags annotating resource y, and the joint probabilities $p(y_1, y_2)$ are given by

$$p(y_1, y_2) = \frac{\sum_x w_{xy_1} w_{xy_2}}{\sum_x 1}$$

With distributional aggregation, one computes fuzzy joint probabilities from the weights composing the vector representation of tags w_{rt} which corresponds to the number of users tagging r with t:

$$p(y) = \frac{\sum_x w_{xy}}{\sum_{r,t} w_{rt}}, \quad p(y_1, y_2) = \frac{\sum_x min(w_{xy_1}, w_{xy_2})}{\sum_{r,t} w_{rt}}$$

where min is a fuzzy equivalent of the intersection operator. In the case of macro and collaborative aggregation we have:

$$\sigma_u(x_1, x_2) = \sum_{y_1 \in X_1^u} \sum_{y_2 \in X_2^u} p(y_1, y_2 \mid u) \log \frac{p(y_1, y_2|u)}{p(y_1|u)p(y_2|u)}$$

where the local simple probabilities $p(y|u)$ are given by

$$p(y|u) = N(u, y)/(N(u) + 1)$$

where $N(u, y)$ is the number of tags used by u to annotate resource y, while $N(u)$ is the total number of tags of u. The joint probabilities are row/column normalized for each user's binary representation.

Evaluation of the similarity measures The evaluation was conducted on the dataset of Bibsonomy.org[10], a social bookmarking service devoted to the annotation of academic works and in which users can define semantic

[10] http://www.bibsonomy.org/faq#faq-dataset-1

relations between tags. The similarity measures have been compared with these user-provided relations by using different threshold values above which a user-provided similarity relation is predicted by the computed similarity. Each similarity measure is thus evaluated by calculating the number of good predictions (true positive) for different values of the threshold. The result of the evaluation showed that mutual information outperform the other types of similarity measures for the case of distributional aggregation, whereas for collaborative aggregation, none of the measures compared gave significantly better results.

Grounding the relatedness of tags using a generic hierarchy of concepts (Wordnet)

Cattuto *et al.* (2008) have proposed a method to semantically ground the relatedness between two tags. To do so, for each tag they (1) use different types of measures, as defined above, to collect similar tags; (2) map these tags into Wordnet (Fellbaum, 1998) synsets; and (3) measure the distance in the Wordnet hierarchy between these terms. In the example depicted in figure 6 (sample data extracted from the 10, 000 most frequent tags of del.icio.us), the original tag is "java". According to the simple cooccurrence measure ("freq" in the figure) and the FolkRank measure, the most related tag to "java" is "programming", and according to the distributional cosine measures, the most related tag is "python". Then, when we look at an excerpt of the Wordnet synset hierarchy containing the original tag and its related tags, we observe (1) that tags given by the cooccurrence and the FolkRank measure correspond to concepts higher in the hierarchy, and (2) that tags given by distributional measures tend to have the same level in the hierarchy. Cattuto *et al.* (2008) repeated this experiment for all of the delicious.com tags which were present in Wordnet, and they draw some qualitative remarks about the semantic relationships each type of measure brings:

- tag context similarity measure tends to give siblings in some suitable concept hierarchy, or to give synonyms.
- context similarities for tags and resources seem to yield equivalent results, especially in terms of synonym identification. The tag context measure, however, seems to be the only one capable of identifying sibling tags.
- user context and tag context measures do not exhibit a strong similarity to any of the other measures.

Giannakidou *et al.* (2008)also proposed to couple statistic-based measures with semantics-based measures of similarity. In their approach they supplement the similarity measure based on cooccurrence (which they call "social" similarity because it reflects the social usage of tags) with a semantic measure based on the distance between the tags in a hierarchy of concepts such as Wordnet. The similarity between tags they compute is thus made of a "social" component and a semantic one, both having a given proportion set as a parameter of the computation of the similarity.

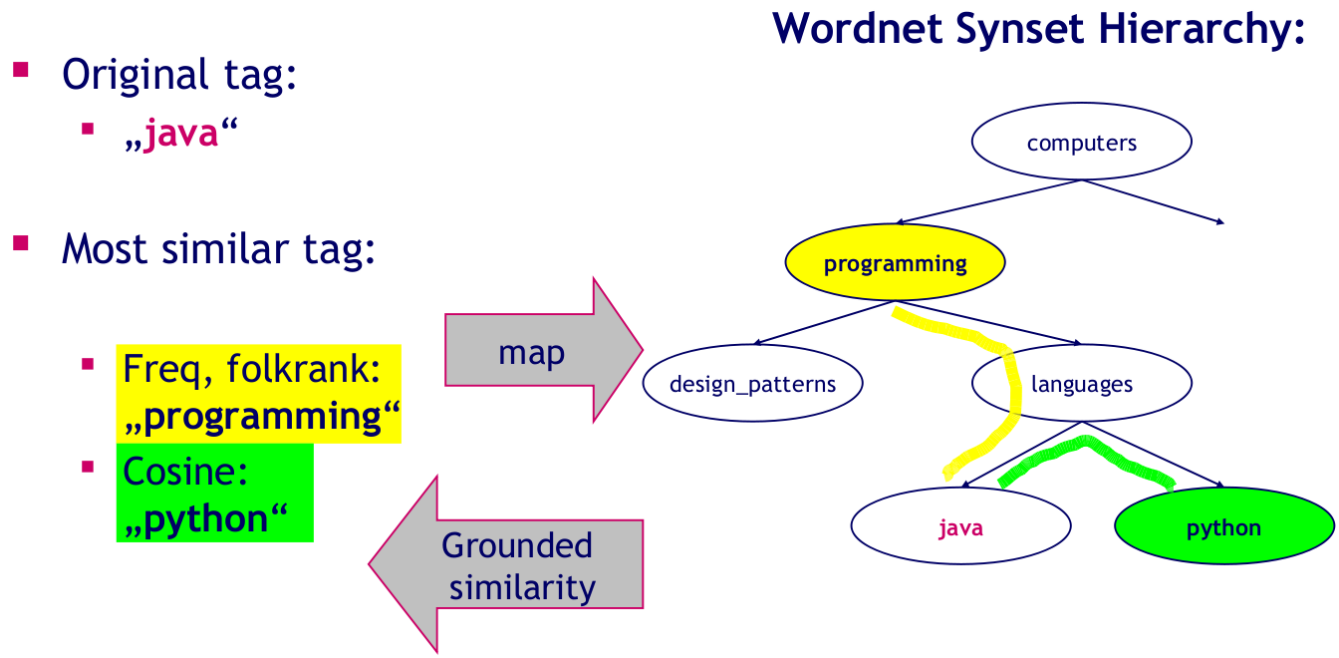


Figure 6 – Semantic grounding of the relatedness of tags using Wordnet (Cattuto *et al.*, 2008)

3.2 Inferring subsumption relations

Several approaches have been proposed to infer subsumption relationships between tags:

Mika (2005) grouped similar communities of interest (as described above in section 3.1) to derive subsumption properties between the tags thanks to the inclusion of communities of interest. In this case, a community of interest may be represented by all the actors who used the tag “fishing”. If the communities of interest “fishing” and “nautic activities” have a number of actors in common, the tags “fishing” and “nautic activities" will be considered as semantically related. Furthermore, if the group of actors using the tag “fishing” is a subset of the group of actors using “nautic activities”, “nautic activities” will be set as a broader term than “fishing”. Mika also shows that subsumption relations are more relevant when derived from inclusions of communities of interest, than when derived from the co-occurrence patterns (given by the Tag-Tag context similarity measure) between tags.

The algorithm proposed by Heymann & Garcia-Molina (2006) takes as input the list of tags in descending order of their centrality in the similarity graph (based here on cosine distance based on the Tag-Resource context). The hierarchy of tags is built starting from the root node, and each tag, taken in order of centrality, is added either as a child of one of the node or the root node (depending on a threshold value of its similarity to these nodes).

Schmitz (2006) used conditional probability to detect subsumption relationships between tags. Given a tag pair (Ti, Tj), let us call the frequency of occurrence of each tag $N(Ti)$ and $N(Tj)$, and the frequency

of cooccurrence of both tags $N(Ti \cap Tj)$. The conditional probability $P(Ti|Tj)$ of having Ti given Tj is calculated as follows:

$$P(Ti|Tj) = \frac{N(Ti \cap Tj)}{N(Tj)}$$

And conversely

$$P(Tj|Ti) = \frac{N(Ti \cap Tj)}{N(Ti)}$$

By comparing both values with each other, we can deduce which of the tags of the pair is more dependent on the other tag. In order to induce a hierarchy from flickr.com tags, Schmitz have adapted the method proposed by Sanderson & Croft (1999), integrating new statistical thresholds to account for the specificity of folksonomies. Thus, tag Ti potentially subsumes tag Tj if :

$$P(Ti|Tj) \geq t \; and \; P(Tj|Ti) < t$$

with

$$N(Ti) \geq Tmin,\; N(Tj) \geq Tmin,\; U(Ti) \geq Umin,\; U(Tj) \geq Umin$$

Where t is a given co-occurrence threshold, $N(Ti)$ and $N(Tj)$ must be greater than a minimum value $Tmin$, and $U(Ti)$ is the number of users that use tag Ti at least once and must be greater than a minimum value $Umin$.

Schwarzkopf *et al.* (2007) also proposed building taxonomies out of folksonomies for user profiling purposes. They pointed out the limitations of the algorithm proposed by Heymann & Garcia-Molina (2006). Indeed, the cosine similarity measure used in this algorithm does not take into account the popularity of tags, while Mika (2005) suggested that relationships between tags established via users (and thus, accounting for the popularity of use, since a tag can subsume another tag only if it is more often used) are more suitable to infer narrower/broader relationships. Schwarzkopf *et al.* (2007) thus proposed exploiting the latter method to infer subsumption relationships between tags, such that: "If resources tagged with t0 are often also tagged with t1 but a large number of resources tagged with t1 are not tagged with t0, t1 can be considered to subsume t0". Schwarzkopf *et al.* (2007) also address the transitivity problem of the inferred subsumption relations and noticed that the "similarity context" of tags is not taken into account when adding a child-tag to a parent-tag, that is, the similarity with all the ancestors of a parent-tag. Thus they combine the users-based associations of tags (Mika, 2005) and the cosine similarity measure to prevent a child-tag from being added to a branch for which it is not related to the origin, such as in `design > web > howto > productivity > business`, where each link makes sense but the whole chain does not.

3.3 Clustering tags

Here we briefly describe different ways of clustering either equivalent tags, which are spelling variants of the same term, or similar tags, which are tags which are considered to be the most closely related with each other.

Finding equivalent tags

The goal here is to detect and group tags that are equivalent in their meanings or in the topic they describe ("new-york" and "newyork", or "folksonomy" and "folksonomies")

- **Stemming algorithms :** They consist in extracting roots from words (e.g. "links" and "linked" become "link") and grouping tags sharing the same roots.
- **String distance metrics:** In this type of method, we measure the difference between the string of characters of the tags. For instance, the Levenshtein algorithm (Levenshtein, 1966) calculates the distance between 2 words by counting the number of letters that have to be replaced, deleted or inserted to turn one word into the other. The threshold value to detect two similar tags may depend on several parameters such as, for instance, the language of the tags. For other algorithm and implementations of this type of methods see the SimMetrics package[11].
- **Exploiting online resources:** (Specia & Motta, 2007; Van Damme *et al.*, 2007) suggest using online resources to check the correct spelling of tags or to find an appropriate representative for a cluster of equivalent tags.

Clustering of similar tags

Here we focus on methods to group similar tags. The similarity can be computed after different measures:

- **Cooccurrence :** between tags that co-occur on the same "resource" (an image, a user's bookmark, an URL, a document, etc.)
- **Specia & Motta (2007)** applied clustering technique to group tags according to the similarity measures within the Tag-Tag context (according to the terminology of Cattuto *et al.* (2008)). During the computation, each cluster starts with a seed tag, and a tag is added only if it has a similarity value above a given threshold with all the other tags of the cluster. Then they apply different heuristic techniques to merge very similar clusters based, for instance, on the percentage of equivalent tags contained in similar clusters.
- **Begelman *et al.* (2006)** first establishes a method to determine strongly related tags. They calculate the cut-off frequency of cooccurrence (as "tags that are used for the same page") between two tags

[11] http://www.dcs.shef.ac.uk/~sam/stringmetrics.html

by looking for a disruption point in the distribution, for each tag, of all the tags co-occurring with it. This method allows to dynamically find, for each tag, the threshold above which its co-occurring tags are strongly related to it, avoiding the use of arbitrary threshold. Then they draw a weighted graph connecting these related tags together. The clustering algorithm takes as input this graph and (1) Uses spectral bisection (Pothen *et al.*, 1990) to split the graph into two clusters, (2) Compares the value of the modularity function[12] Q0 of the original unpartitioned graph to the value of the modularity function Q1 of the partitioned graph. If Q1 $>$ Q0 accept the partitioning, otherwise reject the partitioning, (3) Proceeds recursively on each accepted partition.

3.4 Comparison of the approaches and intermediary conclusions

In this section we have presented several approaches which extract semantic relations between tags by analyzing the structures of folksonomies (in contrast with other types of methods which uses external semantic resources to achieve this task, see section 4.5). All the approaches presented in section 3.2 try to find subsumption relationships between tags. The case of Cattuto *et al.* (2008) is particular in that they characterize different types of similarity measures according to the type of semantic relationships to which they each correspond. Thus, their method can be used to find related tags which share a subsumption relation with a given tag t, however without being sure whether these related tags may subsume or be subsumed by tag t. On the other side, the approaches of Begelman *et al.* (2006) and Specia & Motta (2007) propose a method to cluster similar tags.

The type of similarity measure allows distinguishing all these methods. Mika (2005) applied social network analysis on different projections of the tripartite structure of folksonomies. Hotho *et al.* (2006) adapted the PageRank algorithm to the case of folksonomies in order to find not only relationships between tags, but also between users and resources. Schmitz (2006) used conditional probability methods to induce a hierarchy from Flickr tags. Begelman *et al.* (2006) look closely at the distribution of the cooccurring tags for a given tag, and calculate the threshold above which its cooccurring tags are strongly related to it. Then, several approaches use distributional measures but with different context of aggregation of the folksonomy data, Heymann & Garcia-Molina (2006) using the resource context of association of tags, Specia & Motta (2007) using the tag context of association of tags, Schwarzkopf *et al.* (2007) using a composite measure mixing tag context and user context of association of tags. Finally Cattuto *et al.* (2008) proposed an analysis of the different context of distributional aggregation, while Markines *et al.* (2009) proposed a new type of measure based on mutual information calculus, and a framework to analyze the different types of similarity measures between resources of a folksonomy.

[12] "which measures the quality of a particular clustering of nodes in a graph"(Newman & Girvan, 2003))

	Type of similarity	Subsumption relations	Clustering
Mika (2005)	Network based	yes	no
Hotho *et al.* (2006)	FolkRank	no	no
Schmitz (2006)	conditional probability	yes	no
Begelman *et al.* (2006)	cooccurrence	no	yes
Heymann & Garcia-Molina (2006)	distributionnal (resource context)	yes	no
Specia & Motta (2007)	distributional (tag context)	no	yes
Schwarzkopf *et al.* (2007)	composite	yes	no
Cattuto *et al.* (2008)	distributional (3 contexts)	yes	no
Markines *et al.* (2009)	mutual information	yes	no

Table 4 – Comparison table of the approach extracting semantic relations between tags by analyzing the structure of folksonomies

4 Semantically enriching folksonomies

In this section we present several works which propose to semantically structure folksonomies or to support folksonomy-based social platforms with the formalisms or the tools of the Semantic Web. They either use the tags as attributes of the concepts of an ontology (Passant, 2007), or use ontologies to support the tagging activity (Good *et al.*, 2007; Tesconi *et al.*, 2008) or the semantic structuring of folksonomies (section 4.5), or to represent an extended tagging(Tanasescu & Streibel, 2007).

4.1 Collaborative semantic structuring of folksonomies

Weller & Peters (2008) defines the different aspects of folksonomy improvements taken at a collaborative scale. They define different structural levels on which folksonomies may be improved and edited by the contributors to a folksonomy. (a) Whole document collection vs. single document level. Shall we edit the tags as associated to all the documents, or restrain the editing to tags associated to a single document? (b) Personal vs. collaborative level: should we share the edition of tags or should it be personal? (c) Intra and cross-platform level: depending on the platform we are considering, the treatment applied may differ.

To tackle the problems of ambiguity or misuse of tagging (like spam), Gruber (2005) proposed to "tag the tags". It would then be possible to state that this tag is the synonym of this other tag, or that this tag does not suit this object, integrating mechanisms of regulation like those observed on Wikipedia. Tanasescu & Streibel (2007) applied the idea of Gruber and extended social tagging systems with the possibility to tag the tags themselves and the relationships between them. Indeed, classical tagging systems allow their users to add a "tagging relationship", that is a "is_tagged_by" link between a keyword and a document or a Web resource. But richer information may be obtained from the tagging activity, like the relationships between the tags. These tagging can easily be expressed with triples, such as "car" - "is_a" - "vehicle", all these tags being freely added by the users. This feature allows exploiting the technologies of the Semantic Web to assist navigation

and to suggest to the user other terms semantically related to her query. To prevent irrelevant contributions, the authors proposed solutions based on votes for some tags, in order to appreciate or depreciate them, or solutions based on points that will be granted either to contributors to the tagging task, or to evaluators of the tags of others. Other incentives to contribution could be provided with the "games with a purpose", that is activities presented as games but exploited for a utilitarian purpose, such as categorizing content from the Web Siorpaes & Hepp (2008).

Huynh-Kim Bang *et al.* (2008) also proposed to let the users add semantic information while tagging. The goal is to provide communities of teachers with a tool to organize the educational documents they share, and this tool should merge the flexibility of social tagging and the possibilities of inference brought by semantic formalisms. Thus, they proposed to use structurable tags, that is, tags which can be linked to other tags with a limited set of semantic relationships (in contrast with the openness of the "extreme tagging" of Tanasescu & Streibel (2007)). Two types of semantic relationships are offered to users, each symbolized by a character that users add while tagging : the subsumption of a tag by another tag symbolized by the sign ">" (as in "`plane > airbus`", meaning that tag "`plane`" subsumes tag "`airbus`"), and the synonymy between two tags symbolized by the character "=" (as in "`test = tests`"). Just as all tags are aggregated within a folksonomy, the semantic relationships created by users are also aggregated, meaning that once a user creates a relation between two tags, this relation will be applied to all the users using the same tags.

In the same trend of ideas, Gnizr[13] and Semanlink[14] (Servant, 2006) allow to define semantic relationships between tags. Gnizr describe tags and semantic relationships between them with ontologies presented above, such as SKOS for the subsumption relation, and the TagOntolgy for the tags. Semanlink proposes its own model, but which inherits from SKOS. We should also mention here the "machine tags" in Flickr[15], where users can define enriched tags in the form of predicate:attribute=value, such as `dct:description=New-York` or `geo:lat=42.33`. This type of tags can easily be translated and modeled into RDF triples via the API Flickurl[16].

4.2 Ontologies for modeling folksonomies and online-communities

Gruber (2005) states that there is no opposition between ontologies and folksonomies and proposes constructing an "ontology of folksonomy". The "TagOntology" is a project of an ontology dedicated to formalizing the act of tagging. This model brings in four entities to describe tagging : the tagged object or resource; the term used to tag; the user tagging; and the domain in which the tagging takes place (it can be the service used for instance). Gruber suggests reifying the tagging and to consider each tag as an object as such, and below we will see the different implementation of these ideas.

[13] http://code.google.com/p/gnizr/
[14] http://www.semanlink.net
[15] http://www.flickr.com/groups/mtags/
[16] http://librdf.org/flickcurl/

The Semantically Interlinked On-line Communities (SIOC) project of Breslin *et al.* (2005) provide developers of social Web platforms a formal and technological framework to describe the resources exchanged within and across on-line communities. The formal scheme they propose uses other ontologies like the Simple Knowledge Organization Scheme SKOS[17] which describes systems of organization of knowledge, and Friend Of A Friend FOAF[18] Brickley & Miller (2004) which describes the multiple identities and acquaintances of a user (see figure 7). SIOC describes the most common elements present on Web sites of communities: the concept of "site", the concept of "post" of a Weblog, the concept of "forum", etc. Starting from this vocabulary, the SIOC project proposes tools to automatically annotate the content of some common Web applications (e.g. wordpress.org) according to the SIOC ontology.

The SCOT[19] project proposed by (Kim *et al.*, 2007) aims at representing a folksonomy model with the help of ontologies. This model of tagging is grounded on the Tagging Ontology proposed by Newman *et al.* (2005) and has four main entities: "tagging" itself as an action performed by a user (modeled with `sioc:User` class), "tag" (`scot:Tag`, subclass of the `tags:Tag` class, itself subclass of the skos:Concept class), "cloud of tags" as the containers for the tags of a user, the resource annotated with tags being modeled as `sioc:Item` (see figure 8). SCOT exporter allows mapping content from a given Content Management System (eg. Wordpress) into SCOT ontologies. This offers in turn a better interoperability between different tag spaces and the possibility to form groups of similar or related tag clouds. One of the most direct use case of the SCOT model if the use of meta-search which would allow users to find similar folksonomies, for instance based on the use of tags (number of common tags, that is, the number of tagging using the same `scot:Tag` instance, since all tags spelled the same will be automatically merged).

Other models of tagging have been proposed, such as the one developed by Echarte *et al.* (2007) or TagOnt[20], but none of them seem to have been used contrary to SCOT or SIOC. The Semantic Desktop project NEPOMUK also proposed a class to describe tags through its ontology NEPOMUK Annotation Ontology[21]: the class `nao:Tag` and a property `nao:has_tag`, but without considering the action of tagging as a core element of the model of a folksonomy. Kahan *et al.* (2002)also proposed "Bookmark", a model to describe the infrastructure of the social bookmarking platform Annotea[22]. Even if this model does not include the notion of tags, it allows to link a resource with the terms used to annotate it with the class `bookmark:Topic` and the property `bookmark:Topic`. This model also proposed to organize the topics with the property `bookmark:subTopicOf`, similar to the SKOS property `skos:broader`. We should also mention here the microformat[23] rel:tag. Microformats are the product of a community initiative which defines structured

[17] http://w3.org/2004/02/skos/
[18] http://foaf-project.org/)
[19] http://scot-project.org
[20] http://code.google.com/p/tagont/
[21] http://www.semanticdesktop.org/ontologies/nao/
[22] http://www.w3.org/2001/Annotea
[23] http://microformats.org/

metadata which can be embedded within Web pages via simple html tags attributes [24]. Thanks to GRDDL (Gleaning Resource Descriptions from Dialects of Languages (con, 2007)), which allows to transform XML dialects into plain RDF, we can transform annotations written with the rel:tag microformat into RDF triples based on the `scot:Tag` class for instance.

These ontologies tend to realize the "Web of Linked Data" (now named Linking Open Data[25]) which consists in a vision of the Web where the sources of data and the schema describing them are located with URIs and interconnected in a decentralized way. This project can be realized thanks to ontologies describing the infrastructures where data is stored (such as SIOC, SCOT, and FOAF describe the actors of the social web and the type of data they exchange), or ontologies describing the content or the topics of the data (such as the DBpedia project (Auer *et al.*, 2007)which publishes the Wikipedia content and category structure in a publicly available RDF data store[26]). This project aims at enabling users to access content not only via HTML hyperlinks, but also thanks to the concepts which can be attached to them.

4.3 Infrastructure for linking tags with ontologies

Passant & Laublet (2008) have proposed the MOAT ontology (moat-project.org) which allows users to link the tags they use with a resource (identified with a URI) which represents the meaning of the tag. The MOAT ontology reuses other ontologies such as the FOAF (Brickley & Miller, 2004) ontology to represent the users, or the TagOntology (Newman *et al.*, 2005) to represent the tagging activity, and specifically the "restricted tagging" which corresponds to the link between a tag (defined with MOAT's own class `moat:Tag`), a user, a tagged resource, and a meaning resource (see a graphic representation of MOAT in figure 9). The meaning resources can be any Web pages (such as Wikipedia pages), but also concepts of online semantic resources such as ontologies or thesauruses, the main idea of the MOAT project being to contribute to the elaboration of the Web of Linked Data.

We should also remark that in the approach proposed by Passant & Laublet (2008), the semantic connection is made on the tagging and not merely on the tag itself. The tag is simply taken as a string of characters connecting to an act of tagging and which can be connected to several meanings; tags can be ambiguous, but the act of tagging will be perhaps more imprecise than ambiguous, and thus Passant & Laublet (2008) propose to allow users to precise the intention behind their tagging.

4.4 Linking tags with professional vocabularies

Passant (2007) proposes strengthening the social tagging interface of a corporate Weblog with a centralized ontology. In his approach Passant considers tags as character strings linked with formal concepts with semantic properties. This association of tagging and ontologies is used here to disambiguate the different meanings of

[24] such as `<a href="http://technorati.com/tag/tech" rel="tag">tech</a>`
[25] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/
[26] http://wiki.dbpedia.org/OnlineAccess

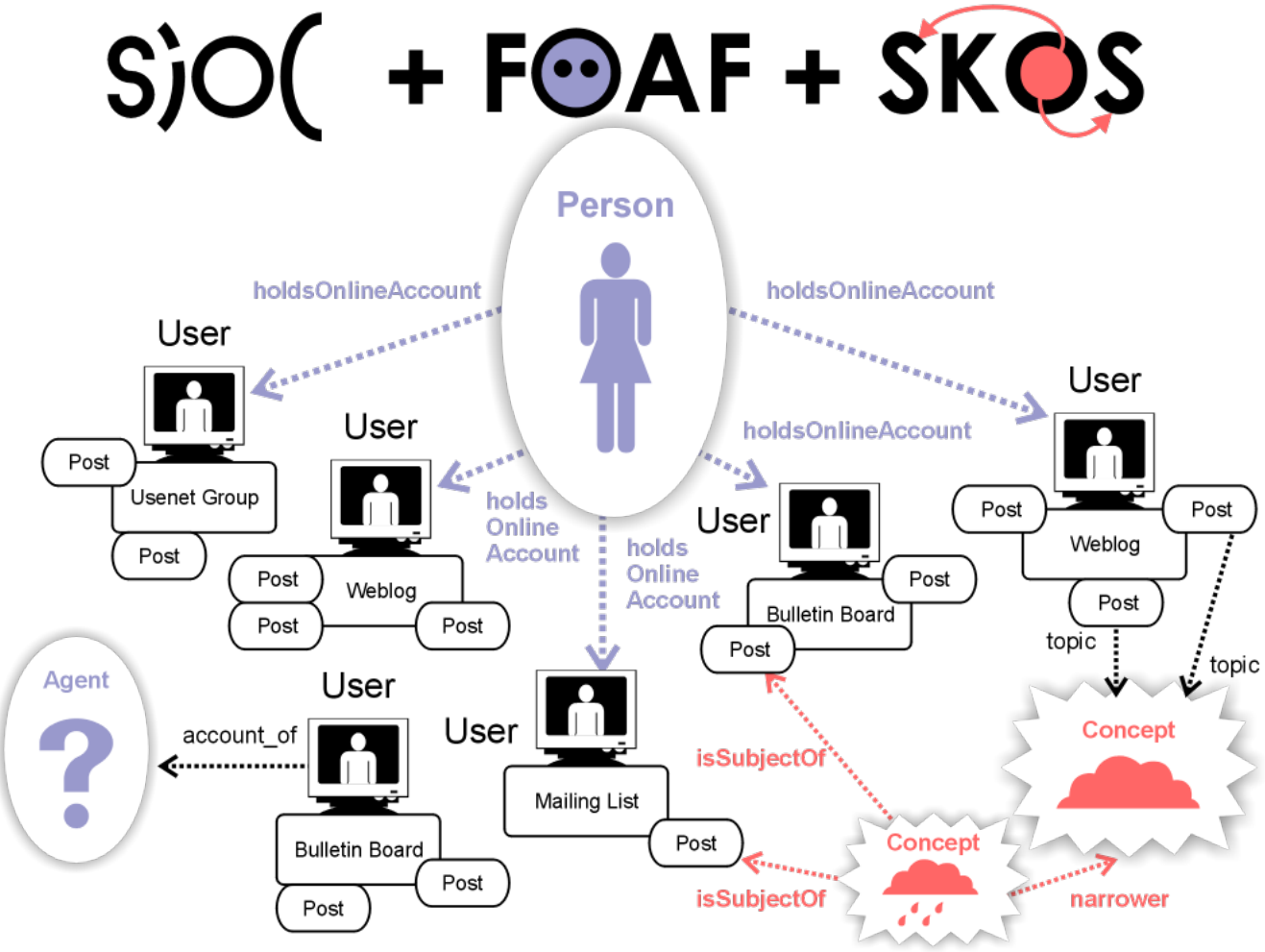


Figure 7 – Modeling online communities: the SIOC model

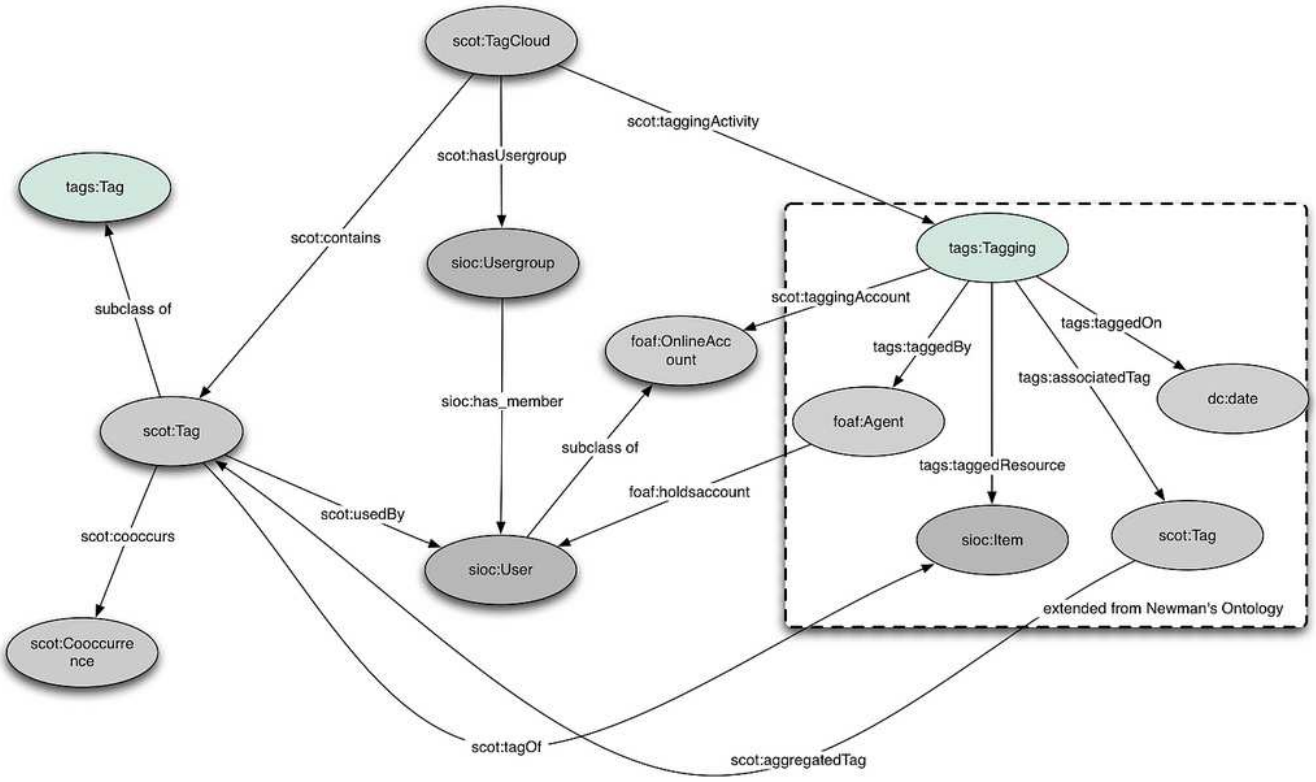


Figure 8 – Modeling tags and folksonomies: the SCOT (scot:) and TagOntology (tags:) models

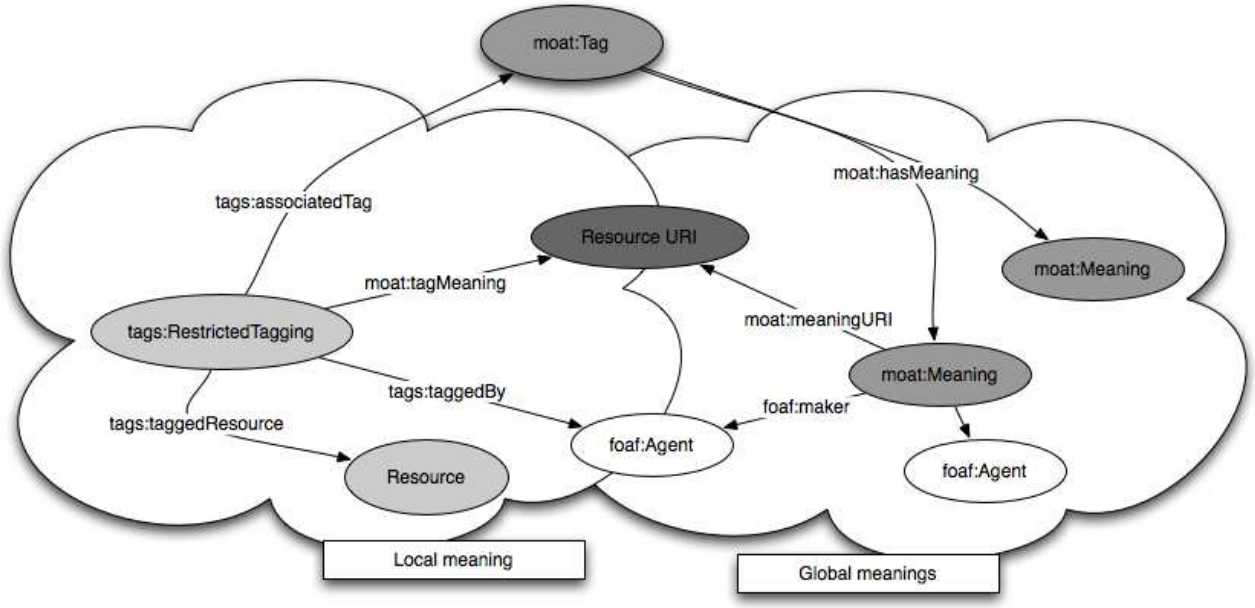


Figure 9 – Description of the MOAT ontology to link tags with unambiguous meanings (Passant & Laublet, 2008)

tags. While tagging, users are suggested to connect the terms with which they are tagging to a controlled vocabulary. Thus, if a tag corresponds to two different concepts (for instance the tag "RDF" may correspond to "Resources Description Framework" or to "Rwanda Defense Forces"), the system asks the user to choose the appropriate concept. When no existing concept matches the user's concept, users are free to propose a new one to the administrators, who in turn will put it in the right place in the ontology. Social tagging is seen here as en empowerment of the construction of an ontology, and ontologies help disambiguating the possible meanings of a tag.

Similar to the approaches described above, the idea of Good *et al.* (2007) is to link tags to professional annotations, by providing an access, while tagging, to pre-defined terminologies organized in formal ontologies. Users can then choose an unambiguous term to use as a tag. Entity Describer uses Connotea, a social bookmarking service specialized in the domain of biology, and enrich the interface (via bookmarklet and a GreaseMonkey Script) by allowing users to access professional terminologies (like MeSH e.g.) in order to use them as a source to select from. When doing so, users can also check the definition of the term they are choosing, which helps them validating or not their choice. The advantages of such an approach are: (1) it does not force users to choose only from a set of controlled terms, but proposes to choose from them. (2) It also allows some extended search capabilities within the collection of bookmarks annotated with controlled terms from ontologies, like finding related terms, narrower or broader terms, etc. (3) It is possible to further exploit the relationships between the terms from the controlled vocabularies and the terms freely chosen by the users: since it is possible to tag an item with both, one may infer semantic relationships between the tags more precisely.

4.5 Assisting semantic enrichment of tagging

The methods we present below differ with the ones above in that they seek to assist the users in the task of linking tags with ontologies. Thus, they do not necessarily make use of the infrastructure described in section 4.3, but focus more on the automatization of the process of semantifying tags.

Clustering and mappings with online semantic resources

The method proposed by Specia & Motta (2007) proposed semantically enriching folksonomies by extending the mere statistical analysis of folksonomies and exploiting online resources such as Wikipedia, Wordnet or ontologies and thesauruses to infer semantic relationships between tags. After solving spelling issues using edit distance measures (Levenshtein) and disambiguating acronyms using Wikipedia, the clustering is done by grouping similar tags using cosine measures computed in the Tag-Tag context (see 3.1). Then, for each cluster, the system looks for elements from ontologies which have the same label as the tags. In case of success, the system is able to map the concepts and their properties to the tags. The result produced by this system is a set of clusters of tags enriched with semantics, but the experimental results show that this type of method

requires that the ontologies used to infer the semantic relations between the tags provide a good coverage of the domain of study.

Currently, several other research works (Angeletou *et al.* (2008) or Van Damme *et al.* (2007)) are trying to pursue the effort of semantically enriching folksonomies. The system developed by Angeletou *et al.* (2008) differs from the approach of Specia and Motta by skipping the phase of clustering similar tags, and by integrating a phase of sense definition and disambiguation of the tags with the help of Wordnet and other terminological resources. Indeed, ontologies available on the Semantic Web are still sparse, and the concepts of these ontologies might not be syntactically equivalent to a given tag of a folksonomy, but rather be labeled with, for instance, a synonym of that tag. Thus, after a first phase of lexical processing of the tags (eliminating isolated tags or user-specific tags which cannot be mapped with already known syntactic categories, such as b&w), each tag is expanded with synonyms or hypernyms found in generic ontologies such as Wordnet, producing a semantically expanded tagset. The next phase, called semantic enrichment, consists in looking within online ontologies for concepts matching one of the terms of each expanded tagset. These matching concepts are called "semantic entities" as they may not belong to the same ontology. The next step in this phase of semantic enrichment consists in discovering relationships between the original tags by exploiting ontology matching techniques to establish semantic relationships between the semantic entities linked with the tags. The result of this approach is a set of semantic entities connected, via the tags, to the annotated resources.

The approach developed by Van Damme *et al.* (2007) aims at building and maintaining ontologies out of folksonomies and their use. One of the differences is that Van Damme *et al.* (2007) are integrating more online resources (such as Wikipedia) and use each resource in several ways. For instance, Wikipedia is used to check spelling or acronyms, but also to map tags with concepts. Furthermore, Van Damme *et al.* (2007) suggest involving the community of users to validate the semantic information previously inferred. Their project can thus be seen as a wish to integrate and extend semantic enrichment of folksonomies, and to involve the users themselves in an ontology engineering process, as proposed by Braun *et al.* (2007) (see section 4.6).

Building a general domain set of semantic tags

Similarly, the TagPedia project proposed by Ronzano *et al.* (2008) and the Tag Disambiguation Algorithm developed by Tesconi *et al.* (2008) aims at achieving the same type of goal, that is, to have tags connected with unambiguous definition of their meaning. Thus, Ronzano *et al.* (2008) propose to assist users in this task by building a "general domain" encyclopedia of terms, TagPedia, which can be then exploited to precise the meanings of tags. By mining the Wikipedia disambiguation pages they connect sets of terms with a unique definition page, representing a concept. The result of their approach is an ensemble of "tag" synsets, that is, sets of synonymous terms linked with a concept defined by a Wikipedia article.

These tag synsets are then utilized by the Tag Disambiguation Algorithm (TDA) developed by Tesconi *et al.* (2008) to connect each tag of a given delicious.com's user to a unique meaning. Indeed, the tag

disambiguation is performed for each user separately, that is, the algorithm is applied to each user's tag set at a time. To achieve this task, the TDA identifies for each tag t a list of candidate meanings for which it computes a sense-rank SR. The higher the rank of a meaning, the better it suits the sense intended by the user for that tag t. In addition, Tesconi *et al.* (2008) assume that the meaning given to a tag does not change across all the taggings of a given user u. To calculate the SR of each possible meaning for a tag t, Tesconi *et al.* (2008) exploits both data from TagPedia, that is sets of different meanings for each tag t and the text of each meaning extracted from the corresponding Wikipedia article, and tagging data given by delicious.com for each bookmark. Thus, given a bookmark where a tag t is associated to a resource r by user u, they count, within the text of each possible meaning m, the number of occurrences of tags related to tag t. Related tags are cooccurring tags on the same bookmark and popular tags also associated by other delicious.com users to resource r (each type of related tags having an arbitrary weight in the computation); Tesconi *et al.* (2008) assume that the higher the number of occurrences of tags related to tag t, the higher the sense-rank of m. The relevance of the results of the DTA has been reviewed by humans, and among 2589 polysemous tags, the DTA has chosen the right meaning of the 89,15% of them.

Once each tag of a user is associated to an unambiguous meaning, it is possible to map these tags to semantically rich structures such as Wordnet hierarchy of synsets, YAGO classes, or Wikipedia categories. YAGO[27] (Yet Another General Ontology) is a generic knowledge representation automatically extracted from Wikipedia which uses Wordnet to organize information. Out of the three semantic resources mentioned above, the Wikipedia categories structure covers the largest part of the disambiguated tags of a sample of 9 delicious.com users. Tesconi *et al.* (2008) also made use of DBpedia[28] (Auer *et al.*, 2007), a publicly available dataset which references each Wikipedia concept with a unique URI and represents the hierarchy of the Wikipedia categories as a thesaurus written in SKOS. Thus, if the disambiguated tags are each connected to a DBpedia URI, the method proposed by Tesconi *et al.* (2008) allows connecting any user's tag with an unambiguous meaning, indentified with a URI and accessible on the Semantic Web, and semantically linked with other concepts from Wikipedia.

Comparison of both methods

The main difference between the methods presented above is that Specia & Motta (2007) apply the mapping of tags with semantic resources on clusters of related tags, whereas Tesconi *et al.* (2008) consider sets of tags belonging to the same user. The semantic enrichment of tags proposed by Specia & Motta (2007) can be used by all the contributors of a folksonomy, and may be useful to a whole community. The tag disambiguation of Tesconi *et al.* (2008) can be applied to different purposes, such as the profiling of the tagging of a user, providing for richer information when consulting the bookmarks database of this user. However, if we apply the algorithm proposed by Tesconi *et al.* (2008) to all the users of a community, we can measure or detect

[27] http://www.mpi-inf.mpg.de/suchanek/downloads/yago/
[28] http://wiki.dbpedia.org

the divergences existing among the users and, for instance, propose them to discuss their points of view in the case of the collaborative construction of an ontology.

4.6 Tagging and collaborative ontology maturing processes

Following the distinctions brought by Weller & Peters (2008) between the individual and the collective level at which folksonomies can be modified, we can distinguish the approaches where the users merely propose new concepts (Passant, 2007), with approaches where users can directly edit the whole shared ontology (Braun *et al.*, 2007), or with approaches where users share their personally maintained ontology (Abbattista *et al.* (2007)). In the latter case, there will be a need to fine-tune sharing strategies or to use ontology mapping techniques (Euzenat & Shvaiko, 2007) in order to efficiently utilize these shared ontologies.

Braun *et al.* (2007) address the problem of collaborative ontology editing and criticize current ontology engineering tools in that they do not integrate the collaborative processes. Individual user-oriented methodologies let each user develop her ontology and then share it with others. Semantic wikis are wikis including semantic functionalities, such as an indexing of pages with formal vocabularies, and can also be useful tools to collaboratively build ontologies. Indeed the ontologies elaborated in such a context can be extracted from the categories used to organize or index the context of the wiki pages, such as what have been developed by Auer *et al.* (2007) with DBpedia.

Braun *et al.* (2007) propose the following description of the ontology maturing process :

1. the first step is the consolidation of the terminology used in the communities (which could be achieved by analyzing the folksonomy for example),
2. the formalization is performed by identifying the concepts and semantic relationships out of the shared terminology,
3. the axiomatizing consists in formalizing more semantic relations between the shared concepts.

This process should also be integrated in current work processes such as information seeking or distribution. The benefit could be a better motivation from the users to participate in ontology-maturing as they wish to retrieve more accurate content in order to be more efficient, or want to make their own publication more visible. Braun *et al.* (2007) implemented a prototype which consists in a bookmarking service with some extra capabilities such as (1) suggestion of tags from the already existing ontology, (2) possibility for all users to add or edit new "semantic" tags , (3) knowledge representation models based on SKOS which includes narrower, broader, and related semantic relationships.

In the corporate blog supported by a centralized ontology proposed by Passant (2007), users who tag their posts do not actually directly participate in the ontology maturing process. There, users propose new instances that should be included in the ontology, the actual ontology design being let to the systems administrators. In the same manner, Torniai *et al.* (2008) propose a method to measure the relatedness between the tags used by

the users of an e-learning platform and the concepts of the ontologies supporting the system and maintained by the teachers. The tool they propose assists the maintainer of the ontologies to integrate new terms and to extend the ontology with new concepts conveyed by the tags of the users.

In a more collaborative approach, Buffa *et al.* (2008a) developed a semantic wiki in which any user can tag the pages and organize the tags of the folksonomy they would for an ontology. The idea is that each action of the user benefits to all the other users. To this respect, Braun *et al.* (2007) remark that current collaborative tagging systems offer few functionalities to structure the vocabularies, and when they do, the structuring is not shared among users (for instance in delicious.com, the "super tags", which are used to subsume a bundle of tags, are not shared).

In the same trend of sharing the semantic individual actions, Abbattista *et al.* (2007)proposed an approach to assist the construction and the evolution of ontologies using collaborative tagging principles. Each user is thus seen as a "knowledge organizer" which contributes to the construction of a collective knowledge base by sharing her structured data. The tool they developed seeks to assist the users in this organization process by (1) providing, for a selected resource, relevant metadata from several repositories, (2) assisting the user in disambiguating the chosen terms using lexical resources (Wordnet (Miller *et al.*, 1990)), (3) suggesting the user to place the terms in relevant location within a personal taxonomy. The user then choose to share parts of her knowledge base, called "binders", that is, groups of annotated resources and the corresponding portion of her personal taxonomy, the result being a shared information space.

4.7 Comparison and intermediary conclusions

In table 5 we compare the approach presented above. Gruber (2005) suggested constructing collaboratively an ontology of folksonomy to support more advanced use of tagging. This idea has been implemented by Newman *et al.* (2005), and further improved by Kim *et al.* (2007) which integrated their SCOT ontology with SIOC Breslin *et al.* (2005), another ontology modeling users' interaction on social Web platforms. Later, Passant & Laublet (2008) have extended these interconnected schemas with MOAT, an ontology allowing to link tags with online resources to define precisely the meaning of tags and to tie them with the "Web of Linked Data"[29], a vision of the Web where resources are linked with each other thanks to the concepts which can be attached to them.

Other approaches focus on user intervention in the process of semantically enriching folksonomies. Huynh-Kim Bang *et al.* (2008) proposes the concept of structurable tags where users can define semantic relations between tags, and Tanasescu & Streibel suggest letting the users to tag the links existing between tags. The two latter approaches do not make direct use of semantic Web formalisms, as they focus more on the flexibility of the system than on logical consistency of the knowledge structure obtained. Passant (2007) developed a semantically augmented corporate blog where users can attach their tags to the concepts of centrally maintained

[29] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/

	User intervention	Ext. resources	Automatic	Sem. Web
Gruber (2005)	-	no	no	yes
Newman *et al.* (2005)	-	no	no	yes
(Tanasescu & Streibel, 2007)	yes	no	no	no
Huynh-Kim Bang *et al.* (2008)	yes	no	no	no
Breslin *et al.* (2005)(Kim *et al.*, 2007)	-	no	no	yes
Passant (2007)	yes	yes	no	yes
Good *et al.* (2007)	yes	yes	no	yes
Specia & Motta (2007) Angeletou *et al.* (2008)	no	yes	yes	yes
Tesconi *et al.* (2008)Ronzano *et al.* (2008)	no	yes	yes	yes
Van Damme *et al.* (2007)	yes	yes	yes	yes
section 4.6	yes	no	no	yes

Table 5 – Comparison table of the approach enriching folksonomies which (1) exploit users intervention, and/or (2) make use of external semantic resources, and/or (3) seek the automatization of the process, and/or (4) are based on Semantic Web formalisms

ontology, while Good *et al.* (2007) suggest terms from professional vocabularies fetched online at tagging time. Specia & Motta (2007) and Angeletou *et al.* (2008) proposed automatic methods to link tags to online ontologies, similarly to Tesconi *et al.* (2008) and Ronzano *et al.* (2008) who, first, build sets of terms-meaning by mining Wikipedia, and then link each tag of delicious.com users to a unique meaning. Van Damme *et al.* (2007), in the same trend, suggest integrating as many semantic online resources as possible, and, at the same time, integrating also users intervention to build, at a reasonable cost, genuine "folks-ontologies". Finally, the approaches presented in section 4.6 focus on the ontology maturing processes and exploit Web 2.0 tools to achieve this task like wikis (Buffa *et al.* (2008a)), blogs (Passant (2007)), e-learning platforms (Torniai *et al.* (2008)), personal knowledge organizers (Abbattista *et al.* (2007)), or social bookmarking sites (Braun *et al.*)

5 Knowledge sharing in the social and semantic Web

In this section we give an brief overview of different cases where online interactions and shared knowledge representation play a central role for the exchange of knowledge on the social and semantic Web. Delalonde & Soulier (2007) seek to assist the task of experts finding and show that structured vocabularies may help in this task. Then we focus on knowledge sharing platforms (section5.2) and semantic wikis (section 5.3) which take the benefit of semantic formalisms.

5.1 Collaborative information and experts seeking

A problem often posed by collaborative work is expert seeking: how to know "who does what"? The study and the system proposed by Delalonde & Soulier (2007) addresses this problem in the context of a big organization. Delalonde & Soulier (2007) developed "DemonD", a system which aims at creating the conditions of social

interactions which yields to capitalized knowledge. DemonD is grounded on personal profiles filled in by the users who state their field of expertise and interests with keywords (which can be seen as tags) and by attaching relevant documents. Then the process starts when one of the user asks a question to the system which selects a list of persons and documents relevant to this question. The selection depends on four main criteria (1) matching keywords in the resource, (2) connectivity with other resources, (3) participation of the person in past interactions, and (4) the reputation evaluated by other peers. Then the system automatically creates a forum of discussion to which the selected persons are invited to participate. The system also includes a step of knowledge capitalization as soon as the original question is answered and that this answer is validated. Thus, this approach includes a collaborative elaboration of knowledge which is partly based on folksonomy-like annotation of the resources. To this respect, Delalonde & Soulier (2007) suggest that the system could be enhanced by suggesting tags when the users build their profiles, and that semi-structured vocabularies could also support the annotation process and help more accurate and more relevant selection of resources.

5.2 Sharing on the semantic Web

Other works propose integrating several ontologies to assist the sharing of data. Hausenblas & Rehatschek (2007) designed "mle", a system which automatically treats mailing lists in order to map the structure of email to appropriate concepts of an ontology (SIOC). These annotations, generated in RDF, allow this database to be queried with the language of the Semantic Web SPARQL (www.w3.org/TR/rdf-sparql-query/).

Revyu.com Heath & Motta (2007) proposes applying the principles of the "Web of Linked Data" (see section 4.2) to organize the sharing of reviews of cultural items (books, movies, etc.). Revyu.com includes these principles by (1) allowing anyone to access data stored on other databases in order to prevent redundancies; (2) utilizing RDF to annotate the resources; and (3) keeping open the field of knowledge which can be covered since Revyu.com uses multiple ontologies and other types of knowledge bases to categorize items.

Other approaches allow to semantically structuring the tags in order to enrich social bookmarking services, like GroupeMe![30] Abel *et al.* (2007) or inter.est[31] Kim *et al.* (2007). "inter.est" uses the SCOT ontology which describes the structure of the tag clouds of the users. The goal of inter.est is to help users find groups sharing the same interests by allowing users to aggregate tag clouds, to form groups of exchange and to facilitate the search of similar tag clouds.

5.3 Semantic Wikis

Semantic wikis were among the first applications to exploit the potential of ontologies to support collaborative practices. Gaved (2006) thus proposed to develop wikis supporting physical rather than virtual communities, and aimed at providing local information guides which could serve as a community memory for a geographical

[30] http://groupme.org/
[31] http://int.ere.st/

area. The Open Guides project aims at highlighting the different types of usages and future uses, and to provide a theoretical framework about wikis of locality. The Open Guides were developed after an adaptation of generic wiki principles in order to describe items with locative elements : latitude and longitude, address, opening time, name of the area. These wikis can be considered as semantic wikis since each entry can be exported in RDF/XML, and all the info of each entry is structured following concepts from several vocabularies devoted to the sharing of online resources (FOAF, DublinCore, ChefMoz). Gaved also identified common tasks performed by users of wikis, such as locating, exploring, grazing, monitoring, sharing, and asserting about the information described in each entry of the wiki, leading to truly collaborative semantic processes. The analysis of the usages lead to make some other observations concerning the interface which should empower non-technical experts to contribute, the sustainability of the system which can be enhanced by providing more machine-readable metadata, and the spam of diverse kind which tended to pollute the content of the wikis. This return on experiment is of great usefulness for a designer of collaborative tools and addresses the main problems arising from the use of collaborative semantic tools.

SweetWiki (Buffa *et al.*, 2008b) is another example of semantic wikis: users can edit and modify pages, and also tag any document published on the wiki. The tags are tied together in a folksonomy which is expressed with the languages of the Semantic Web. All the new tags are collected as the labels of new classes which are, by default, subsumed by the class "new concept". All the users are then able to organize the tags of the folksonomy, and to edit them, to add new labels in other languages, to create relations of synonyms, to merge classes, etc. The author of pages can also use tags to keep an eye on the activity of other contributors in a targeted manner: each user can specify in her homepage her topic of interest in the form of tags. For instance, a user interested in wikis will put a tag "wiki" in the field "interested by". Then, whenever a page is tagged with "wiki" or a subclass of "wiki", the user will be notified. This function allows watching content that does not yet exist. By keeping track of created or modified pages, and by analyzing over time the behavior of users, it is possible to detect acquaintance networks or communities of interest. This reveals several possibilities: finding the most active person on a given topic, finding the users using similar tags as others, inferring relationships between tags when they are used by the same users, etc.

5.4 Comparison and intermediary conclusions

To conclude this brief overview we can see that, except from Delalonde & Soulier (2007) which propose to assist users in finding experts in the social context of corporate organizations, all the other approaches integrate Semantic Web formalisms to describe their data model. In table 6, we can distinguish these approaches with the type of content they organize or with the type of services they offer. While some applications target no specific social context (Hausenblas & Rehatschek, 2007), some others are set in the Web 2.0 by dealing with the sharing of cultural items (Heath & Motta, 2007) or simply by providing semantically enriched social bookmarking services (Kim *et al.* (2007) and Abel *et al.* (2007)). Finally, semantic wikis have been developed

	Type of platform	social context
Delalonde & Soulier (2007)	Expert finding	organization
Hausenblas & Rehatschek (2007)	mailing list	generic
Heath & Motta (2007)	reviews sharing	web 2.0
Kim *et al.* (2007).Abel *et al.* (2007)	social bookmark	web 2.0
Gaved (2006)	wiki	city
Buffa *et al.* (2008b)	wiki	organization

Table 6 – Comparison table of the approach of section 5.

to assist the communities of the inhabitants of cities (Gaved, 2006), or to assist the activity of organizations in a broad sense.

6 Discussion

6.1 The best of both worlds

We have seen that it is possible to describe a folksonomy and all the activities occurring on social Web sites with an ontology. In this report we have compared different approaches which aim at bridging ontologies and folksonomies to support and leverage the exchange of knowledge over the social Web. In this regard, these research works are relevant to the design of social Web platforms in that their methods or algorithms can greatly benefit to the final user's experience, by proposing more precise tools to navigate within and across the different platforms. Interoperability is a critical factor for the future of on-line social software, and once adapted to fit the usages, technologies and standards of the Semantic Web can significantly improve the current situation.

The approaches we presented above often complement each other and they can be distinguished against different criteria:

Analysis versus formalization: First, we can extract out of the folksonomies a "lightweight ontology" thanks to statistical analysis (Specia & Motta, 2007; Mika, 2005; Halpin *et al.*, 2007), or we can directly formalize the tags and their usage among communities of users as with the SIOC and the SCOT projects, or as Gruber (2005) suggested it. Both types of approaches aim at improving information retrieval in folksonomy-based systems.

Type of resources annotated: Second, we can distinguish the different types of resources annotated. Breslin *et al.* (2005) seek to assist the exchange of resources on weblogs and forums, while Heath & Motta

(2007) treats the case of reviews. In the same trend, Buffa *et al.* (2008b) enhanced the collaborative edition of wiki pages with social tagging functionalities and formalisms of the Semantic Web.

Social context: Third, we can distinguish different kinds of social contexts. A centralized system works well with a clearly defined field of knowledge (Passant, 2007), while, for instance, the collection of reviews of cultural items or bookmarks will require an open field of knowledge Heath & Motta (2007), and in some cases, the degree of formalism in the knowledge structure would have to be adapted to the level of sharing of the knowledge among the members of the community (Zacklad, 2007).

Integration and design: Fourth, we can distinguish the systems with respect to the design aspects. Some approaches can seamlessly integrate current social platforms such as the SIOC plug-ins, which generate meta-data about the content organized by some popular Content Management Systems (Wordpress, Drupal). Other works can also simply exploit the data already available Mika (2005); Halpin *et al.* (2007) and infer extra semantic information which can in turn be used to describe more precisely the users data. Finally, other works propose reconsidering the design of social platforms by embedding in them technologies or formalisms of the Semantic Web Abel *et al.* (2007); Heath & Motta (2007); Buffa *et al.* (2008b).

6.2 Adapting the models and tools to the usages

It is also necessary to keep in mind the social aspects of knowledge sharing, and to strive to design models fitting actual usages. For example, Sinha (2006) proposed a social and cognitive analysis of tagging where she shows that annotating a resource with several keywords requires less cognitive effort than choosing a unique category. Tagging is thus simpler since it allows picking up all the concepts first activated in the mind.

Cahier *et al.* (2007) distinguish between "information seeking", in which the user does not already and exactly know what she is looking for, and "information retrieval", in which the user knows exactly what she wants to find. In the information retrieval case, a high precision is expected in the results of a query. In the "information seeking" case, the user does not expect a great precision in the result, and she is refining her request along the navigation within each consequent results. Here, a high recall may be favored, and serendipity may also be appreciated, or at least not fought against. Cahier shows how ontologies, in the sense of structured vocabularies, allow for better recall by connecting terms semantically.

Zacklad (2007) proposed a comparison of the different types of classification with respect to their adequacy to the needs of the communities. To this regard, ontologies allow for a richer representation of the knowledge of a community, and the comparison of the different levels of formality of the ontologies (see section 1.3) is highly relevant to the information seeking processes. When the queries deal with information related to places or to resources having clearly and neatly defined properties (such as in the field of natural or physical sciences) formal ontologies provide for the most powerful assistance with inference mechanisms. However, in some other fields where more subjective and disputable criteria are at stake, semi-formal knowledge representations such

as semiotic ontologies, Topic-Maps, or thesauruses can be interesting alternatives to help navigate within a corpus of knowledge and to represent the different point of views of the members of the community.

In a corporate context, Van Damme *et al.* (2008) applied coocurrence-based similarity measures and conditional probability to retrieve a hierarchy between tags used to annotate several types of contents produced within a company (messages, internal notes, etc.). They show that companies may be interested in the outputs of such methods as a management tool, which would help them in decision making (by looking at recurrent patterns within their terminology), in revealing them the emergence of new terms (technological watch and monitoring purposes), and in helping them creating new teams (by detecting proficiencies and communities of interest).

6.3 Perspectives

Gruber (2008) differentiates collective intelligence from collected intelligence. He gives three characteristics of the current systems which collect knowledge: (1) the production of content performed by the users, (2) a synergy between users and the system, (3) increasing benefit with the size of the domain covered. In order to upgrade this type of system towards a collective intelligence, Gruber proposes adding another feature: the emergence of knowledge beyond the mere collection of each contributor's knowledge. He suggests that this fourth feature directly benefits from the integration of the technologies of the Semantic Web.

Thus, the potential of hybrid systems which exploit the benefit of both the ease of use of folksonomies and the support of the formalisms and the methods of the Semantic Web opens new perspectives for assisting knowledge exchange on the social Web. But several challenges remain. Specia & Motta (2007) showed the efficiency of combining statistical techniques with extra knowledge from the ontologies on the Semantic Web, but since the fields of knowledge that could be appropriate is potentially infinite, we need methods to efficiently select the source of information to help structuring the folksonomies. For instance, Review.com (Heath & Motta, 2007) uses that kind of technique to clearly identify whether the provided Web link is about a movie by querying the IMDB.com database, but identifying concepts dealing with the content of the reviews may be more complex and poses the problem of the selection of the sources of additional information. These issues, plus the need to find similarities between groups of tags or to match tags with elements from other ontologies could also benefit from exploiting some of the "ontology matching" field's methods (Euzenat & Shvaiko, 2007), such as the identification of similar tags thanks to more or less strict string matching, or the evaluation of the similarity between two tags according to their relative positions within a graph.

The other challenges that social on-line platforms may be faced with, is the workload needed to administrate or contribute to the system. The current approaches to add semantic information to the resources exchanged in the social Web are: (1) organizing tag data *a posteriori*, that is, analyzing the tags and their usage Mika (2005); Specia & Motta (2007), or proposing the users organizing the tags Buffa *et al.* (2008b) or tagging the tags themselves Tanasescu & Streibel (2007); (2) asking users to raise the ambiguity at tagging time

Passant (2007), or to provide more detailed information when submitting content Heath & Motta (2007). The question social Web platforms designers may ask at this moment is how much effort they can expect from their users. And this question is not simple since the social context plays a role: incentives to contribute to an enterprise weblog or to a platform of shared reviews may largely differs in the amount of effort users may put in providing precise information (workmates may be rewarded for good quality contributions), and (even more complex) in the agreement they may find when dealing with non-consensual knowledge (when commenting on a movie, different and contradictory views may emerge). One of the key to these questions may rely on a balance between top-down-style administration of the knowledge base and bottom-up-style auto-regulation. But both of these components of social Web platforms will need appropriate tools and methods (1) that seamlessly integrate within the everyday tasks of the users, without any extra burden and produce useful extra information as a side effect of the natural use of the systems; and (2) achieve an appropriate combination of precision, diversity, and representativeness of the knowledge representations supporting the activities of the "interconnected on-line communities" of the social Web.

References

(2007). Gleaning resource descriptions from dialects of languages (GRDDL), W3C recommendation.

Abbattista F., Gendarmi D. & Lanubile F. (2007). Fostering knowledge evolution through community-based participation. In *Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007) at WWW 2007*, Banff, Canada.

Abel F., Frank M., Henze N., Krause D., Plappert D. & Siehndel P. (2007). Groupme! - Where Semantic Web meets Web 2.0. In *ISWC/ASWC*, volume 4825 of *LNCS*, p. 871–878: Springer.

Agrawal, R. I. T. & Swami A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD1993, ACM Press*.

Angeletou S., Sabou M. & Motta E. (2008). Semantically enriching folksonomies with flor. In *CISWeb Workshop at Europ. Semantic Web Conf.*

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R. & Ives Z. G. (2007). Dbpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber & P. Cudré-Mauroux, Eds., *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, p. 722–735: Springer.

Bachimont B. (2000). *Ingénierie des connaissances: Evolutions récentes et nouveaux défis*, chapter Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances. Eyrolles.

Begelman G., Keller P. & Smadja F. (2006). Automated tag clustering: Improving search and exploration in the tag space.

Berners-Lee T., Hendler J. & Lassila O. (2001). The Semantic Web. *Scientific American*, **284**(5), 34–44.

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

Breslin J., Harth A., Bojars U. & Decker S. (2005). Towards Semantically-Interlinked Online Communities. In *ESWC 2005*.

Brickley D. & Miller L. (2004). *FOAF Vocabulary Specification*. Namespace Document 2 Sept 2004, FOAF Project. http://xmlns.com/foaf/0.1/.

Brin S. & Page L. (1998). The anatomy of a large–scale hypertextual web search engine. *Computer Networks and ISDN Systems*, **30**(1–7), 107–117.

Buffa M., ErÃ©tÃ©o G., Faron-Zucker C., Gandon F. & Sander P. (2008a). SweetWiki: A Semantic Wiki. *Journal of Web Semantics, special issue on Web 2.0 and the Semantic Web*, **6**(1).

Buffa M., Gandon F., Ereteo G., Sander P. & Faron C. (2008b). SweetWiki: A semantic Wiki. *J. Web Sem.*, **6**(1), 84–97.

Cahier J.-P., Zaher L., Pétard X., Leboeuf J.-P. & Guittard C. (2005). Experimentation of a Socially Constructed "Topic Map" by the OSS Community. *workshop on Knowledge Management and Organizational Memories, IJCAI-05*.

Cahier J.-P., Zaher L. & Zacklad M. (2007). Information seeking in a "socio-semantic web" application. In *ICPW07: Proceedings of the 2nd international conference on Pragmatic web*, p. 91–95, New York, NY, USA: ACM.

Cattuto C., Benz D., Hotho A. & Stumme G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. *7th International Semantic Web Conference*.

Delalonde C. & Soulier E. (2007). DemonD: Leveraging social participation for Collaborative Information Retrieval. In *1st Workshop on Adaptation and Personalisation in Social Systems: Groups, Teams, Communities, Corfu, Greece*.

Dewey M. (1876). *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*.

Dolezal F. (2005). *A History of Roget's Thesaurus by Werner Hullen*, volume 32. John Benjamins Publishing Company.

Echarte F., Astrain J. J., Córdoba A. & Villadangos J. E. (2007). Ontology of folksonomy: A new modelling method. In S. Handschuh, N. Collier, T. Groza, R. Dieng, M. Sintek & A. de Waard, Eds., *SAAKM*, volume 289 of *CEUR Workshop Proceedings*: CEUR-WS.org.

Euzenat J. & Shvaiko P. (2007). *Ontology Matching*. Berlin, Heidelberg: Springer.

C. Fellbaum, Ed. (1998). *WordNet An Electronic Lexical Database*. Cambridge, MA ; London: The MIT Press.

Gaved M. (2006). Wikis of locality: Insights from the open guides. In *WikiSym06, Odense, Denmark*.

Giannakidou E., Koutsonikola V., Vakali A. & Kompatsiaris Y. (2008). Co-clustering tags and social data sources. *Web-Age Information Management, 2008. WAIM '08. The Ninth International Conference on*, p. 317–324.

Golder S. & Huberman B. A. (2005). The structure of collaborative tagging systems.

Good B., Kawas E. & Wilkinson M. (2007). Bridging the gap between Social Tagging and Semantic Annotation: E.D. the Entity Describer. Available from Nature Precedings.

Gruber T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**(2), 199–220.

Gruber T. (2005). Ontology of Folksonomy: A Mash-up of Apples and Oranges. In *Conference on Metadata and Semantics Research (MTSR)*.

Gruber T. (2008). Collective knowledge systems: Where the Social Web meets the Semantic Web. *J. Web Sem.*, **6**(1), 4–13.

Halpin H., Robu V. & Shepherd H. (2007). The Complex Dynamics of Collaborative Tagging. In *WWW*: ACM Press.

Hausenblas M. & Rehatschek H. (2007). mle: Enhancing the Exploration of Mailing List Archives Through Making Semantics Explicit. In *Semantic Web Challenge, ISWC*.

Hayes C., Avesani P. & Veeramachaneni S. (2007). An analysis of the use of tags in a blog recommender system. In *Twentieth International Joint Conferences on Artificial Intelligence*.

Heath T. & Motta E. (2007). Revyu.com: a Reviewing and Rating Site for the Web of Data. In *ISWC/ASWC*, volume 4825 of *LNCS*, p. 895–902: Springer.

Heymann P. & Garcia-Molina H. (2006). *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Rapport interne, Stanford InfoLab.

Hjelmslev L. (1963.). *Prolegomena to a theory of language*. University of Wisconsin Press, Madis.

Hotho A., Jäschke R., Schmitz C. & Stumme G. (2006). Information Retrieval in Folksonomies: Search and Ranking.

Huynh-Kim Bang B., Dané E. & Grandbastien M. (2008). Merging semantic and participative approaches for organising teachers' documents. In *Proceedings of ED-Media 08 ED-MEDIA 08 - World Conference on Educational Multimedia, Hypermedia & Telecommunications*, p. p. 4959–4966, Vienna France.

Jäschke R., Hotho A., Schmitz C., Ganter B. & Stumme G. (2008). Discovering Shared Conceptualizations in Folksonomies. *J. Web Sem.*, **6**(1), 38–53.

Kahan J., Koivunen, Prud'hommeaux E. & Swick R. R. (2002). Annotea: an open rdf infrastructure for shared web annotations. *Computer Networks*, **39**(5), 589–608.

Kim H.-L., Yang S.-K., Song S.-J., Breslin J. G. & Kim H.-G. (2007). Tag Mediated Society with SCOT Ontology. In *Semantic Web Challenge, ISWC*.

Lave J. & Wenger E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.

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

Markines B., Cattuto C., Menczer F., Benz D., Hotho A. & Stumme G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *18th International World Wide Web Conference*, p. 641–641.

Mathes A. (2004). *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. Rapport interne, GSLIS, Univ. Illinois Urbana-Champaign.

Mika P. (2005). Ontologies are Us: a Unified Model of Social Networks and Semantics. In *ISWC*, volume 3729 of *LNCS*, p. 522–536: Springer.

Miller G. A., Beckwith R., Fellbaum C., Gross D. & Miller K. J. (1990). Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, **3**(4), 235–244.

Newman M. E. J. & Girvan M. (2003). Finding and evaluating community structure in networks.

Newman R., Ayers D. & Russell S. (2005). Tag Ontology Design. http://www.holygoat.co.uk/owl/redwood/0.1/tags/.

Park J. & Hunting S. (2002). *XML Topic Maps: Creating and Using Topic Maps for the Web*. Addison-Wesley Professional.

Passant A. (2007). Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. In *International Conference on Weblogs and Social Media*.

Passant A. (2009). *Technologies du Web Sémantique pour l'Entreprise 2.0*. PhD thesis, Université Paris IV - Sorbonne.

Passant A. & Laublet P. (2008). Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*.

Pothen A., Simon H. D. & Liou K.-P. (1990). Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, **11**(3), 430–452.

Rastier F. (1994). *Sémantique pour l'analyse*, chapter Interprétation et compréhension, p. 1–22. Masson, Paris.

Ronzano F., Marchetti A. & Tesconi M. (2008). Tagpedia: a semantic reference to describe and search for web resources. In *WWW 2008 Workshop on Social Web and Knowledge Management, Beijing, China*.

Sanderson M. & Croft B. (1999). Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, p. 206–213, New York, NY, USA: ACM.

Saussure F. D. (1916). *Cours de linguistique générale*. Paris: Bayot.

Schmitz C., Hotho A., Jäschke R. & Stumme G. (2006). Mining Association Rules in Folksonomies. *Data Science and Classification*, p. 261–270.

Schmitz P. (2006). Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW06)*.

Schwarzkopf E., Heckmann D., Dengler D. & Kroner A. (2007). Mining the structure of tag spaces for user modeling. *Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling*, p. 63–75.

Servant F.-P. (2006). Semanlink. In *Jena User Conference (JUC)*.

Sinha R. (2005). A cognitive analysis of tagging. http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html.

Sinha R. (2006). Tagging from Personnal to Social : Observations and Design Principles. In *Tagging Workshop, WWW*.

Siorpaes K. & Hepp M. (2008). Ontogame: Weaving the semantic web by online gaming. In M. Hauswirth, M. Koubarakis & S. Bechhofer, Eds., *Proceedings of the 5th European Semantic Web Conference*, LNCS, Berlin, Heidelberg: Springer Verlag.

Specia L. & Motta E. (2007). Integrating folksonomies with the semantic web. *4th European Semantic Web Conference*.

Tanasescu V. & Streibel O. (2007). ExtremeTagging: Emergent Semantics through the Tagging of Tags. In *ESOE at ISWC*.

Tardini S. & Cantoni L. (2005). A semiotic approach to online communities: Belonging, interest and identity in websites' and videogames' communities. In *IADIS Intl. Conf. e-Society*, p. 371–378: IADIS.

Tesconi M., Ronzano F., Marchetti A. & Minutoli S. (2008). Semantify del.icio.us: Automatically turn your tags into senses. In *Proceedings of the First Social Data on the Web Workshop (SDoW2008)*.

Torniai C., Jovanovic J., Bateman S., Gaševic D. & Hatala M. (2008). Leveraging folksonomies for ontology evolution in e-learning environments. In *ICSC '08: Proceedings of the 2008 IEEE International Conference on Semantic Computing*, p. 206–213, Washington, DC, USA: IEEE Computer Society.

Van Damme C., Coenen T. & Vandijck E. (2008). Turning a corporate folksonomy into a lightweight corporate ontology. In W. Abramowicz & D. Fensel, Eds., *BIS*, volume 7 of *Lecture Notes in Business Information Processing*, p. 36–47: Springer.

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

Vanderwal T. (2004). Folksonomy Coinage and Definition. http://www.vanderwal.net/folksonomy.html.

Veres C. (2006). The language of folksonomies: What tags reveal about user classification. In *Natural Language Processing and Information Systems*, volume 3999/2006 of *Lecture Notes in Computer Science*, p. 58–69, Berlin / Heidelberg: Springer.

Weller K. & Peters I. (2008). Seeding, weeding, fertilizing. different tag gardening activities for folksonomy maintenance and enrichment. In S. Auer, S. Schaffert & T. Pellegrini, Eds., *Proceedings of I-Semantics08, International Conference on Semantic Systems. Graz, Austria, September 3-5*, p. 10–117.

Wille R. (1982). *Ordered Sets*, chapter Restructuring lattices theory : An approach based on hierarchies of concepts, p. 445–470. Reidel, Dordrecht-Boston.

Zacklad M. (2007). Classification, Thésaurus, Ontologies, Folksonomies : Comparaisons du Point de vue de la Recherche Ouverte d'Information (ROI). In *CAIS/ACSI*.