KNOWLEDGE MEDIA
KMi
INSTITUTE

Semantic Enrichment of Folksonomies

Technical Report kmi-08-06
October 2008

Sofia Angeletou

Angeletou, S., Sabou, M., and Motta, E. (2008) Semantically Enriching Folksonomies with FLOR, Workshop: 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) at European Semantic Web Conference

Angeletou, S., Sabou, M., Specia, L., and Motta, E. (2007) Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report, Workshop: Bridging the Gap between Semantic Web and Web 2.0 at European Semantic Web Conference

The Open University

Semantic Enrichment of Folksonomies

Sofia Angeletou

September 16, 2008

Contents

Chapter 1

Introduction

The goal of this research is to explore the potential of combining the heterogeneous technologies of Semantic Web and Web2.0 in order to contribute to an open and intelligent World Wide Web. This document is the **second year research report of this PhD study** on integrating Web2.0 and the Semantic Web.

World Wide Web has significantly contributed to the dissemination of information to the world and has changed the way people work, communicate and spend their free time. Since it's first rise in 1990, the WWW has critically evolved. The early practices were based on the publication of static web pages composed in plain HTML, followed by the generation of dynamic web pages populated at runtime and on request by databases according to the user queries, specifications and needs. The content discovery has been significantly based on powerful web search engines such as GOOGLE, YAHOO and more.

As circumstances and requirements change, the need for intelligent applications that perform knowledge, rather than document, retrieval becomes an important challenge. Until today the web content has not been exploited into intelligent applications, partly because 1. **the intelligent applications were not mature enough** and partly because 2. **the content was not described in a machine processable form**. The evolution for both the above issues has already begun. In particular, the large scale content annotation and metadata generation has became reality as the applications of Web2.0 have gained momentum allowing users to annotate their content in a personal choice. On the other hand the advances in the Semantic Web technologies are promising there is potential for intelligent applications capable to integrate distributed content and knowledge from various heterogeneous resources.

Although there is significant discussion on the combination of the two trends([11, 26, 37, 28, 54, 46]), and there are real world applications combining Semantic Web and Web2.0 ([29]), the two “forms of the Web” have not managed to converge, hampering their evolution towards an interpretable, intelligent Web. This work explores the potential of combining Semantic Web and Web2.0 practices in order to achieve better **information description and facilitate intelligent applications**.

Chapter 1 explains the background and the motivation for this work as well as define the research problem and the partial research questions. Chapter 2 presents the relevant work and the open issues in the area. In Chapter 3 we detail our approach and methods as a result of our work during the first 24 months.

In the following section we describe our research problem. As previously mentioned there is a high interest over the combination of the Semantic Web and Web2.0 aiming at the full potential of the web. In the following sections we describe what is Web2.0 and Semantic Web and how our work on the combination of the two is motivated.

1.1 Web2.0

Web2.0 is a term introduced by Tim O’Reilly [49] as an attempt to describe new practices in comparison to the conventional web practices. Web2.0, which is also mentioned as **Social Web** due to its plethora of social dimensions, has become an issue of high interest among web experts and web users and is considered to be the next generation of World Wide Web. The basic asset of Web2.0 is its strong focus on the user satisfaction. Web2.0 has an intense social dimension which allows users to create networks of trusted users within the same areas of interest. In addition to that, the Web2.0 applications are very usable and don’t require specific skills, motivating in that way the users to upload their content and annotate it with absolutely freely chosen labels (no controlled vocabulary, no guidelines, no restrictions).

The outcomes of this open and user-focused guideline are very useful. Primarily, users frequently contribute with more content, enriching in this way the already data-intensive web. The second, most important outcome is the effortless user annotation of the content. The result, thus, is a **data**

intensive and annotated web. This, despite its preliminary and unstructured state, is already an advance compared to the conventional web, as the current Web2.0 sites are annotated with a primitive ”semantic layer” even if this is weak and uncontrollable.

1.1.1 Folksonomies

A typical and very popular example of Web2.0 systems is a folksonomy [59], a system that allows users to upload content, annotate it with labels they have selected themselves and share it with other folksonomy users and, in some cases, the whole web users. Following we describe the basic characteristics of a folksonomy:

- The **content** or **resource** depending on the folksonomy can be images in Flickr [3] audio in Last.fm [4], video clips in YouTube [6], bookmarks in del.icio.us [2] and more.
- The annotation is consisted of one or more **tags**, which are words freely chosen by users to describe a resource.
- The **user** of a folksonomy can create resources and annotate them with tags he thinks are appropriate, but depending on the folksonomy the user may also tag content he did not upload or create.
- The basic activities carried out in a folksonomy is the **tagging** of content and the **search** for content. The users can do both activities, so depending on the role we call them either **taggers** or **searchers**.

Folksonomies have been rapidly adopted because they don’t expect users neither to have prior knowledge or specific skills to use the system nor need to rely on a priori agreed structure or shared vocabulary, thus **not imposing any constraint to the users regarding the tagging process**. While this is the main advantage of folksonomies it also introduces some issues that limit the full potential of folksonomies.

For example, a zoologist can tag a photograph of a lion with {`felidae, pantherinae, mammal`}, while a non-expert in zoology can use {`lion, king, animal, jungle`} for the same purpose. Unfortunately, the simplistic tag-based search used by folksonomies is agnostic to the way tags relate to each other although they annotate the same or similar resources. For example, a search for {`mammal`} ignores all resources that have not been tagged with this specific word, even if they are tagged with related concepts such as {lion, cow, cat}. As a result, content retrieval activities such as searching, subscription and exploration are limited , they provide low-recall and hardly lend themselves to query-refinement. Therefore, to obtain satisfactory results, a searcher needs to build multiple complex queries to cover all the possible tags that could have been used by taggers. As searchers rely on their own view about what inter-related tags best describe the resource they are looking for, it follows that content retrieval could be enhanced if folksonomies were aware of the relations between their tags.

As the user tagging reflect their different perceptions of the world, this leads to phenomena such as annotating the same resource with completely different annotations. While this contributes to the richness, globality and completeness of views about the world in the system, it hampers the content retrieval mechanisms. For example consider a user who annotates the web page of a budget airline with the tags budget airlines. Another user who looks for cheap flights will not encounter this web page although it is exactly what he is looking for; this is caused because there is no semantic association between the tags of these resources.

But the problems of content retrieval in folksonomies is not limited to the above example. Consider a Flickr user who knows what he is looking for but doesn't know how to express it. For example, wishes to search for pictures of animals that live in the water. The majority of results from search of water animals in Flickr are images of animals (including dogs, cows and tigers) close to water (**irrelevant results**), rather than aquatic animals such as dolphins, seals and so on (**missed results**). The system is not able to provide the user with some kind of recommendation such as ”marine animals” or ”aquatic animals” because the notions of “lives in”, “water” and “animal” are not associated in any manner. Finally in the case where a user wants to browse through the content of folksonomies because he doesn't know what he is looking for, the topic based browsing is not available in folksonomies.

Our belief is that restricting the user and posing any kind of tagging rules would dramatically decrease the popularity of folksonomies. Our intuition is that the effort needs to be focused on studying and exploiting the implicit, emergent semantics generated by the user contribution and add a semantic layer on top of the tags rather than controlling their generation process.

1.2 Semantic Web

The vision of the Semantic web was introduced almost seven years ago by Berners Lee, Lassila and Hendler in [12] as a web of data. The vision for the Semantic Web is a universally understood and processable description for the web resources allowing in that way their exploitation in intelligent applications able to do reasoning, its usability across domains, and so on.

The Semantic Web realisation, in contrast to Web2.0, strongly depends on knowledge representation experts, domain experts and the collaboration among them. This is because in order to have the web content formally annotated there is a need for formal conceptual models. These conceptual models that describe the objects of a domain and their relations, are called ontologies and due to their importance in the overall application of the models they need to be created through the collaboration of a group of domain experts and knowledge representation experts.

Ontologiesare the backbone of the Semantic Web. They are conceptual structures formally specifying objects, their "behaviours" and their relations. The vision of Semantic Web is to describe the web resources with formally derived metadata from these ontologies making the resource metadata machine processable, widely available and allowing the resources to participate in intelligent cross domain applications.

1.3 Motivation and Research Problem

Despite the progress noted in the area of the Semantic Web and the Web2.0 some of the problems of the web still remain. Brachman in [13] notes that one of the main problems of the web is the poor retrieval of search engines and the lack of semantics behind the queries. According to Brachman search engines do search but the users are not always interested in getting results of documents, they are interested in finding information and knowledge. He claims that this can be achieved by applying semantics in the content retrieval process. By doing so, the discovery of the “appropriate” content from various, heterogeneous and distributed resources, the context definition, the reasoning over it, the integration of the content to an understandable processable way, the recommendation proposal to the users and many more would be feasible.

[47] claim that the new semantic web applications are geared to exploit the vast amount of data available on the web without creating their own and, in

addition, they have to cope with the heterogeneity of web resources and adopt Web2.0 practices. [26] and [37] suggest that the Semantic Web technical capabilities can fulfill the Semantic requirements of Web 2.0 applications, compliment the Web2.0 business models and enhance the Web2.0 ecosystems.

[26] also claims that "*the basic subsumption reasoning of the Semantic Web is able to extend, enrich and structure the flat tag systems, allow suggestions and increase the precision and recall of the current tag systems in a way that non-experts can create and maintain adequate mapping functions between large amounts of constantly changing ontology and instance information or to manage their periodic versioning and maintenance ... The ability to support sophisticated long-tail queries over the dynamic, user-contributed content of Web 2.0 applications is one such capability* (of the Semantic Web)".

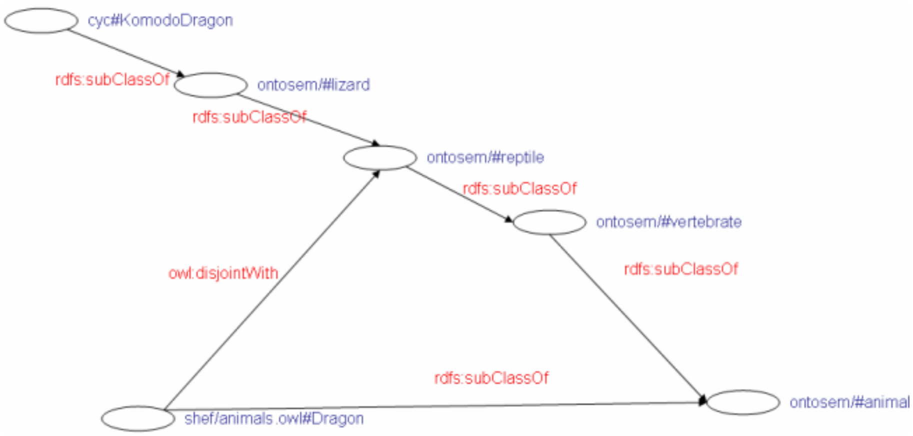


Figure 1.1: Komodo Dragon in the Semantic Web

The focal point of this work is to investigate how **can the Semantic Web contribute towards the semantic enrichment of folksonomy content** and **how can it support sophisticated queries over the enriched folksonomies**. On the other hand to explore **how this can support the Semantic Web evolution** according to the actual needs of the web users. The short experiment described next implied that Semantic Web can actually be a very strong basis for explicitly stating the implicit semantics that already lie within folksonomies. We searched for the tags from the tagset {`lizard, reptile, monitor, komodo, dragon, dangerous, carnivore`}, describing a Flickr photo of a komodo dragon, within ontologies and the results are presented in Fig. 1.1. From the statements found in ontologies, **freely available on the web**, we can explicitly deduce that lizard is an animal and the photograph depicts an animal, despite the fact that this word is not contained in its describing tagset.

Hypothesis 1: The problem of irrelevant results and exclusion of results while searching in folksonomies is caused because the tags are treated as raw text, ignoring completely their meaning and their context. **Our hypothesis is that enriching folksonomies, with explicit semantics will enhance their content retrieval mechanisms**. With the help of this we can formulate the following specific research questions:

- **RQ1: How can folksonomies' tagspaces be semantically enriched automatically?** This research question can be further analysed into the following questions. How to discover automatically the meaning of tags based on their context? How can the Semantic Web be exploited for the semantic enrichment of the tags and what other resources are required in case the Semantic Web falls short of that task?

- **RQ2: How can the enriched tagspaces be evaluated in terms of content retrieval** against the non enriched tagspaces? What performance measures should be established to measure content retrieval in folksonomies before and after the semantic enrichment?

Chapter 2

State of the Art

In this section we present the related work conducted in the areas we base our research. In the literature there are various approaches aiming to tackle the problems of folksonomies and to combine the Semantic Web and Web2.0 technologies. There is a variety of works spanning from generic, motivational works to more specific task based research on **Combining Semantic and Web2.0**. These are presented in Section 2.1.

Next, in Section 2.2 we present more specific **works focused on folksonomy understanding and enhancement**. There are works studying folksonomies in a macroscopic but also microscopic level, aiming to understand the user behaviour and their usage patterns. Further there is research done aiming to enhance the functionalities of folksonomies with regards to content retrieval, annotation and organisation of resources. In the same line we present the works that try to leverage folksonomies to ontologies.

2.1 Semantic Web 2.0

The majority of works referenced in Section 1.3 as well as the works described here are the evangelists of **Semantic Web 2.0** or **Semantic Social Web** or **Web3.0**. They support the argument, which also formulates the hypothesis of this work, that **the convergence of Semantic technologies and Web2.0 technologies and content are going to lead the Web to its full potential**.

With the exception of Clay Shirky ([55]), who supports social software and is opposed to the use of traditional organisation schemes for content organisation, the following works support the above argument. The combination of the Semantic Web and the Web2.0 technologies to reach the web's full potential is vision of many web experts ([11, 26, 37, 28, 54, 46]). They claim that the Semantic Web standards have matured to support open data and a new view for information processing that emphasises to information rather than processing, creating a new way to content interoperability. The same time the Web2.0 provide ways to more democratic, personalised and complete web. However, the intelligent applications "mashing-up" socially generated Web2.0 content require a shared meaning in order to connect, integrate and organise the various components and data sources. This is where the Semantic Web technologies can contribute to increase precision and recall in social network search especially by providing powerful query and reasoning capabilities.

Further to the above, there are examples of works utilising the Semantic and Social aspects of the web in order to benefit either. An early approach on the exploitation of social network generated semantics towards the creation of ontologies and the further utilisation of these ontologies to enhance the functionalities of the social networks is given by [45], where the author claims that the social web and the Semantic Web do co-exist already and can benefit from each-other. He introduces an interesting approach of extending the bipartite ontology model of concepts and instances to a tripartite with three different classes of nodes actors, concepts, instances. He induces ontologies applying techniques on network analysis based on the co-occurrence of tags and actors and the co-occurrence of tags and resources.

Later works such as, [15, 38] demonstrate how how the Semantic Web can represent the community networks, and facilitate the data and knowledge sharing. Also [24, 56, 16] present the idea of using the online communities and social tagging concept to advance the Semantic Web and describe a framework for collaborative ontology evolution in order to tackle the problems of ontology development and ontology population.

2.2 Folksonomy Studies

Since the term *folksonomy* was coined, research has focused on comprehending the inherent characteristics of folksonomies and investigating their emergent semantics. The primer works explore and analyse the structure, types of tags and user incentives in tagging using various methods.

[25] analyze the structure of collaborative tagging systems and their dynamics and prove that tagging activity follows a stable pattern after a specific amount of time. They also identify the semantic and cognitive aspects of classification which are polysemy, synonymy and basic level variation problem, issues which appear also in folksonomies. Another important contribution of the study is the identification of distinct types of tags according to the users intention on future reuse. They prove their hypothesis by conducting a case study on Del.icio.us and show how tags and resources tend to stabilize their relations after some time utilising also use Cloudalicio.us ([51]) to prove their argument.

Along the same line with [25], [43] present an analytical model of the collaborative tagging systems, as well as a comprehensive analysis of the of the tagging systems to date, based on system design and user incentives. Their aim is to provide the framework for further research on folksonomies by defining a conceptual model for social tagging systems comprised by users, resources and tags. They also identify two main groups of parameters that play an important role to the tagging process, the system design and user incentives which affect the overall evolution and dynamics of folksonomies.

Another work studying the emergent semantics from folksonomies and more specifically the association and differences of peer to peer systems with collaborative tagging systems is presented in [44]. The author identifies the importance of interest based locality in folksonomies and performs a case study on Del.icio.us in order to compare the metadata provided by Del.icio.us and the metadata provided by the Open Directory Project.

In [19] the authors study the problem of visualisation of the tag evolution which also implies the evolution of the community focus. They present a system which is applicable to a wide type of timescales (daily, weekly, and so on). They present their infrastructure and five algorithms to deal with the partial tasks of their system. Applying their methods on a dataset from Flickr they provide their results including the identification of different categories of interesting tags.

Cloudalicio.us, mentioned previously, providing a visualisation of how folksonomy tag clouds evolve over time is presented in [51]. Their system generates a graph depicting the tag usage for a specific URL (this work is focused on the Del.icio.us folksonomy) identifying patterns of stabilisation over time.

The authors of [8] attempt to respond to further questions such as: "Where do folksonomies fit in the spectrum from professionally manually assigned keywords to machine context based keyword extraction?". The experiment with a set of bookmarks extracted from delicious and they perform term extraction on their textual content. This set of terms constitutes the first set (A) of terms for the respective resource. The second set (B) is the respective tags of each resource and the third (C) is composed by a human evaluator. When the latter is then asked to compare the sets of keywords A and B as to which is more descriptive of the resource, the evaluator is closer to set B rather than A. This translates to folksonomy tags carry more semantics (or have higher semantic value) than the automatically extracted keywords from the text of the resource. This is justified by the fact that keyword sets B and C are closer than C and A. Finally the authors claim that folksonomies have added contextual dimensions as the tags come from different agents, thus there is more variety. Finally they give some insights on the expertise of delicious users and the variable background of them.

[41] analyse statistical properties of broad folksonomies aiming to identify laws and characteristics underlying properties for folksonomy based retrieval. This work deduces that the tag based search on folksonomies as opposed to full text search, including other lexical values such as title, description, notes and so on, performs better on recall and precision.

In the following we describe and categorise the works according to their primary goal, i.e. the problem they aim to solve. The three main problems we identified, the literature is trying to solve are folksonomy organisation, folksonomy search and navigation and folksonomy ontological modeling. However, there is a fair amount of overlap within these categories as many works attempt to perform first organisation or ontological modeling of folksonomies and then provide search and browsing mechanisms.

2.2.1 Folksonomy Organisation

The works that present research on folksonomy organisation include organisation of resources into a structure (a posteriori, utilising the existing status of folksonomies), proposition of schemas for enriched annotation (a priori, proposing a novel "semantified" way to tag and annotate), and ways to "publish" folksonomy data in a semantic way according to the Linked Data principles [5].

One of the works that aims at providing a semantic infrastructure for folksonomy organisation and description according to the Linked Data principles is described in [50]. The authors propose a modular architecture system that allows the users to assign explicit meaning for their tags during the tagging. They present their architecture and propose an ontology modeling the tagging process as well as the linking between the tags and their meanings. Finally they analytically explain all tasks and functions involved in the meaning assignment process during the tagging and the publication of Semantic Data on the web.

Using a different method [42] focuses on the semantic definition of tags, primarily by using WordNet. For example they try to identify the meaning of tags in order to enrich the relevant resources with RDF descriptions. The authors distinguish six conceptual categories of tags in Flickr. Using WordNet and other knowledge resources for these conceptual categories they organise the tags accordingly. Then they enrich the Flickr photos with RDF triples created for each of the tag categories. These triples are generated either by predefined predicates or from WordNet signatures depending on the categories they belong to.

T-ORG ([7]) performs ontology based organisation of Flickr photos into a set of predefined categories according to the tags describing them. At first the user selects an ontology of interest. Then, the system extracts the concepts and tries to identify semantic relatedness between these concepts and the tags by querying the web with various linguistic patterns between them. Then each tag is categorised under a superclass of the concept to which was more related by the web search.

[30] aim at organising the resources of folksonomies in a hierarchical taxonomy by automatically building a hierarchy of tags from the data in a tagging system. Their algorithm leverages notions of similarity and generality that are implicitly present in the data generated by users as they annotate objects.

2.2.2 Folksonomy Search and Navigation

In this section we describe the works that aim to enhance the folksonomy search and navigation capabilities. [62] aim to statistically infer a global semantic model to enhance the annotation and search in folksonomies. They discover semantically connected resources within the scope of a folksonomy using graph theory. They represent the users, resources and tags as multidimensional vectors and place them in a multidimensional space of domains. They create links between domains and items according to the relations of tags, users and resources to these domains. The domains are identified through tag clustering which, along with the positioning of items within the domain space, is carried out dynamically.

The authors of [36] describe a method that expands the related tags clusters of Del.icio.us with more related tags based on co-occurrence. The expanded clusters are presented as navigable hierarchical structures or semantic trees. These semantic trees are derived from WordNet. Using a combination of WordNet based metrics they identify the possible WordNet sense for each tag. Then they extract the path of this tag from the WordNet hierarchy and they integrate it into the semantic tree of the tag's cluster.

The TagPlus system described in [39] uses WordNet to disambiguate the senses of Flickr tags by performing a two step query. First a user queries for a tag, then the system returns all the possible WordNet senses that define the tag and the user selects (disambiguates) which sense he initially meant. Finally their system looks for all the Flickr photos tagged with this tag and its synonyms. The authors also describe a similar system in [40] where they present the SynTag "sense interface" (generated by WordNet) where the user selects the intended sense for his tag while tagging and while querying. They store this link in their local server and perform the mapping of tags-senses and photos through this intermediate resource.

2.2.3 Folksonomy Ontological Modeling

As the value of user power became apparent through folksonomies, many researchers try to exploit it for the benefit of the Semantic Web, whose main problem is the high cost and effort for ontology creation and metadata generation. The majority of studies try to export ontologies from folksonomies but recently there has been quite a few works following the path of Tom Gruber [27].

He proposes the TagOntology, a common ontology formalizing the tagging activity in folksonomies. He proposes the expression "tagging (object, tag, tagger, source, +/-)" as a formal representation of the tagging activity. The meaning of the above is that, the tagger tags the object with the tag in the context of the source system, further empowering (+) or weakening (-) the latter association with his tagging activity. With this model Gruber is trying to provide an formalization of the tagging activity in order to exploit the data created in the scope of folksonomies.

Apart from the tagging ontologies there has been some early works on bottom-up ontology creation. One representative example of this effort is given by[53]who provides a subsumption based model on the Flickr tag set for ontology generation. In this work the generated model is more a taxonomy of related tags rather than an ontology from the formal Semantic Web perspective. [61] identify some key parameters in folksonomies exploration which are the identification of field leaders and local communities of interest. They also address the issue of ontology generation by approaching it as a hierarchical clustering problem. [60] tries to export categories from folksonomic data in order to create formal ontologies. He is paying special attention to the linguistic attributes of the tags as a means to identify the users perception.

In [32] and [31] the authors propose a folksonomy contextualization method based on Formal Concept Analysis aiming to provide shared meaning for the tags and create concept hierarchies from tags of blogosphere. They are based on the assumption that if a blog has relationships with others, they would use the similar set of tags. They deduce that contextualised folksonomies are able to provide context-centric and shared collections of tags to semantically-interlinked online communities.

Finally in [34] and [33] the authors present a review and alignment respectively of the most popular tag ontologies two of which a have already mentioned previously [50, 27]. Additionally they discuss the tag ontologies presented in [35, 48, 23] and conclude that tags, resources and taggers are the elementary entities described in all the ontologies. Another conclusion were that despite the similar goals of all the tagging ontologies the results were very variable. The final conclusion of the comparison and alignment was that the combination of MOAT [50] and SCOT [35] is capable to support folksonomy modeling, as well as data reuse across different domains and applications.

2.3 Defining the Gap

Despite the significant amount of research carried out in different aspects of this area there are some still open issues. Referring to Section 1.3, the goal of this work is to enrich folksonomies in order to enable the efficient information retrieval. Further more a side effect of the semantic enrichment of folksonomies would be the publication and the availability of this semantically content to the web according to the Linked Data principles.

One of the most interesting and partialy overlapping works to ours is described in [33] as the alignment of the two tagging ontologies MOAT and SCOT. Although they define and demonstrate a case of online presence and interaction their main drawback is that they expect the users to explicitly define the meaning of their tag based on a selection of senses from DBpedia [1]. As they also claim, this project is not about automatically defining the meaning of tags as they are based on the collaborative meaning definition, but can support a method that does.

Furthermore, other works that explicitly define the meaning of tags using other resources such as WordNet still require human intervention. Other works require some initialising from the user's side (e.g., a priori selecting ontology or knowledge resources for the relevant categories of tags) and utilise a single knowledge resource to define the meaning of their tags. Finally many of the works propose frameworks and solutions that would require folksonomies to be rebuilt on their principles and thus don't experiment and evaluate their solutions in the real world scenario.

Our work aims at providing a solution that A) automatically defines the tags that already exist in folksonomies, thus B) be independently plugged and perform on existing systems. Also we aim to C) utilise more than one knowledge source (i.e., all the online available ontologies, WordNet, DBpedia, e.t.c) in order to achieve higher coverage of enrichment from and finally our goal is to D) evaluate this enrichment in terms of real tagging data extracted from folksonomies.

Chapter 3

Folksonomy Enrichment

The main objective and the expected contribution of our work is to **automatically enrich folksonomies tagspaces with a semantic layer** and then **utilise this semantic layer to query the tagspaces**. Our secondary goal is to **identify measures, for the evaluation of content retrieval in folksonomies before and after the semantic enrichment** in a way that we can decide the contribution of the enrichment to the folksonomy itself on a content retrieval basis. Additionally we aim to generate semantic data in a manner and form that is in line with the Linked Data principles [5].

This chapter presents our work in a chronological manner as well as a short description of our future work. Our preliminary experiments with Folksonomies and our first results, that served as a feasibility study of our research and contributed to the formulation of our methodology, are presented in Section 3.1 and were published in [10]. In Section 3.2 we discuss our generic approach and overall methodology on folksonomy enrichment and evaluation of the enriched folksonomy content and, finally, in Section 3.3 we present our work on folksonomy enrichment, published in [9].

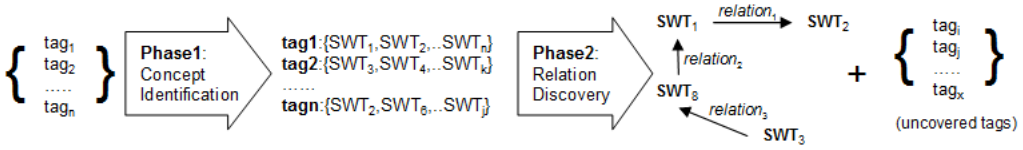
3.1 Preliminary Experiments

We focus our experiments on the tags of the resources rather than other lexical attributes such as title, description and notes. This is because tags offer higher precision and recall to the query in comparison to the all the lexical information of the resource [41]. Our method is based on [57], which describes a hybrid approach that combines harvesting the Semantic Web with using other Web resources such as Wikipedia and Google. As the goal of our work is to understand the potential and limitations of the Semantic Web when used to semantically enrich folksonomies, we have modified their algorithm so that it only relies on online ontologies. Our algorithm, presented next, takes as input a cluster of implicitly related tags and returns 1) a knowledge structure obtained by making explicit the semantic relations among them and 2) a set of tags which could not be semantically related to any other tag in their cluster or were not covered by the Semantic Web.

The goal of our experiments is twofold. On the one hand, we wish to reveal how much of the semantic enrichment of folksonomy tags can already be automated by using the software developed in [52] which partially implements the current version of our envisioned algorithm (the part described in Section 3.1.2). On the other hand, we wish to understand any problematic issues so that they can be addressed in the design of the final, complete algorithm. At a higher level, these issues give an insight in how folksonomies and the Semantic Web relate. In a first experiment (Section 3.1.3) we applied the software developed in [52] to Flickr and Del.icio.us clusters generated by [57]. This experiment lead to valuable insights into issues that hamper the enrichment and prompted us to repeat the experiments with another set of clusters selected directly from Flickr. We discuss the second set of experiments in Section 3.1.4 and our preliminary conclusions in Section 3.1.5.

3.1.1 Semantic Enrichment Method

The semantic enrichment of each cluster is depicted in Fig. 3.1 and consists of two phases: Phase 1, concept definition for each tag (i.e., linking tags to ontology concepts) and Phase 2, relation discovery between all the possible pairs of tags.



$\{tag_1, tag_2, ..., tag_n\}$
Phase1: Concept Identification
tag1:{SWT$_1$,SWT$_2$,..SWT$_n$}
tag2:{SWT$_3$,SWT$_4$,..SWT$_k$}
......
tagn:{SWT$_2$,SWT$_6$,..SWT$_j$}
Phase2: Relation Discovery
SWT$_1$
relation$_1$
SWT$_2$
relation$_2$
SWT$_8$
relation$_3$
SWT$_3$
+
$\{tag_i, tag_j, ..., tag_x\}$
(uncovered tags)

Figure 3.1: Semantic Enrichment Method

Phase 1. Concept Identification: The first step *explicitly defines the meaning* of each tag by extracting all Semantic Web Terms (SWT) whose label or localname match with the tag. The matching between the tag and the SWT can be achieved using anchoring techniques ranging from strict to flexible string matching as described in [52].

Using the Semantic Web for extracting concepts is proposed in the work of [20] as a first step to query disambiguation. The authors search for candidate senses in online ontologies and then perform disambiguation based on the semantic similarity of the retrieved senses (e.g., `bass` can either refer to a fish or to musical notes depending on the context in which it is used). While we use the same technique for SWT identification we do not explicitly disambiguate between them. In our case, disambiguation is a side effect of relation discovery (Phase 2).

The disambiguation of the tag sense (i.e., finding the right concept for a tag given its context) is approached differently in [57]. The authors rely on the heuristic that if pairs of tags from a cluster appear in the same ontology then this leads to an implicit disambiguation (i.e., searching for `apple` and `fruit` leads to ontologies about fruits, while when searching for `apple` and `computer` they identify ontologies about computers). While this intuition holds in the case of domain-specific ontologies, it is problematic when the tags appear in broad, cross-domain ontologies such as WordNet [22] or TAP[1]. Also, by considering only ontologies that contain both tags, this approach potentially misses important information that might be declared in ontologies defining only one of the tags. This information can prove to be useful when combined with information from other ontologies. For example, an ontology containing *Apple* and *Mac*, can be combined with information from another ontology containing information about *Mac* and *Computer*. For these reasons, we retrieve all the potential SWTs for each tag and discover relations between them in Phase 2.

Phase 2. Relation Discovery: This step identifies *explicit semantic relations* among all the pairs of SWTs (`T1` and `T2`) discovered in the previous phase:

- **Subsumption Relations:** when one of the two SWTs is a subclass of the other, `T1 subClassOf T2`. This relation can be either declared in an ontology or derived by different levels of inference (no inference, basic transitivity, Description Logics reasoning). An example of inferred relation is: if `T1 subClassOf T2` and `T2 subClassOf T3` then `T1 subClassOf T3`.

[1] `http://tap.stanford.edu/data/`

- **Disjointness Relations:** when `T1` and `T2` are disjoint, `T1 disjointWith T2`. Again this relation can be declared or inferred. We use the algorithm described in Section 3.1.2 to discover disjointness and subsumption relations.

- **Generic Relations:** when a generic relation holds between the two SWTs, e.g., `Property1 hasDomain T1` and `Property1 hasRange T2` or inversely.

- **Sibling Relations:** when the two SWTs share a common ancestor, which can be either a direct or an indirect parent. Note that our definition covers the three sibling definitions described in [57].

- **Instance Of Relations:** such as `T1 instanceOf T2` or inversely. Unlike the previous relations, this relation is not considered by [57].

The identification of these relations can be made in two ways. First, a relation between SWT's might be declared **within a single ontology**. Second, if no single ontology mentions both SWT's, then a **cross-ontology relation discovery** can be performed by combining knowledge from several ontologies.

Cross-ontology relation discovery has been successfully implemented in the case of ontology matching [52]. An important issue to be considered is how to deal with potential contradictory relations, e.g., `T1 subClassOf T2` and `T1 disjointWith T2`. This remains a future work topic.

The semantically connected tags form the knowledge structures mentioned in the beginning of Phase1 and the tags not linked to SWTs or not related to other tags compose the set of uncovered tags. The study of the latter is expected to provide hints about how to evolve the Semantic Web, as described in Sections 3.1.3 and 3.1.4. Next we describe the current implementation of our approach which identifies only subsumption and disjointness relations found in single ontologies.

3.1.2 Subsumption/Disjointness Discovery Based on One Ontology

The discovery of subsumption and disjointness relations between two terms within one ontology has been described and implemented on Swoogle'05 ([18])

in [52]. Given two candidate concept names (`A` and `B`) as an input, corresponding concepts are selected in online ontologies (`A'` and `B'`) by using strict string based anchoring. The possible semantic relations occurring between concepts in an ontology are shown using description logic syntax, e.g., `A'` $\sqsubseteq$ `B'` means that `A'` is a sub-concept of `B'`. The returned relations are expressed with arrows such as, e.g., $\mathtt{A} \xrightarrow{\sqsubseteq} \mathtt{B}$. The steps of this strategy in detail are:

1. Select ontologies containing concepts `A'` and `B'` corresponding to `A` and `B`;

2. If no such ontology is found, then `A` and `B` do not relate;

3. If there are returned ontologies, for each:

   - if `A'` $\equiv$ `B'` then derive $\mathtt{A} \xrightarrow{\equiv} \mathtt{B}$;
   - if `A'` $\sqsubseteq$ `B'` then derive $\mathtt{A} \xrightarrow{\sqsubseteq} \mathtt{B}$;
   - if `A'` $\sqsupseteq$ `B'` then derive $\mathtt{A} \xrightarrow{\sqsupseteq} \mathtt{B}$;
   - if `A'` $\perp$ `B'` then derive $\mathtt{A} \xrightarrow{\perp} \mathtt{B}$;

In the simplest implementation, we can rely on *direct* and *declared* relations between `A'` and `B'` in the selected ontology. But, for better results, *indirect* and *inferred* relations should also be exploited. For our experiments, we used an implementation relying on basic transitivity reasoning (i.e., taking into account all parents of `A'` and `B'`) and stopping as soon as a relation is found.

3.1.3 Experiment 1

The number of results obtained by running our algorithm with the clusters generated in [57] were surprisingly low. Two major reasons explain this. First, our implementation only searches for `subClassOf` and `disjointWith` relations. Unfortunately, the majority of tags in the clusters we work with are not related by these relations but by generic relations. The second major reason is that few of the tags in the analysed clusters could be identified in ontologies in the Semantic Web. Taking a closer look to the tags that were not found we individuated the following cases:

Novel terminology. Folksonomies are social artifacts, built by large masses of people and dynamically change to reflect the latest terminology in several domains. As such, they greatly differ from ontologies which are generaly developed by small groups of people and evolve much slower. Therefore, it is not surprising that many of the tags used in folksonomies, e.g., {`ajax, css`}, have not yet been integrated into ontologies. Identifying frequent folksonomy tags that are missing from ontologies has a great potential for the Semantic Web as it can provide the first step towards enriching existing ontologies with these novel terms.

Instances. When people tag resources, especially pictures, they more often tend to tag them with specific names rather than more abstract concepts. In particular, we frequently found names of people {`monica, luke, stephanie`}, names of places {`japan, california, italy`} and particular dates {`august2005, aug292005`}. Unfortunately, the current version of our system only works at terminological level (it deals only with concepts and not with ontology instances), so we did not identify any of these instances in the experiments. Apart from that limitation it is unlikely that instances related to people and specific dates can be reliably identified in ontologies anyway.

Photographic jargon. Given the scope of Flickr as a photo annotation and sharing site, many of the tags that are used reflect terms used in photography, such as {`nikon, canon, d50, cameraphone, closeup, macro`}. Unfortunately, this domain is weakly covered in the Semantic Web.

Multilingual tags. Both Flickr and Del.icio.us (but especially Flickr) contain tags from a variety of languages and not only English. These tags are usually hard to find on the Semantic Web because the language coverage of the existing ontologies is rather low. Indeed, statistics[2] performed on a large collection of online ontologies (1177) in the context of the OntoSelect library indicate that 63% of these ontolgies contain English labels, while a much smaller percentage contains labels in other languages (German 13.25%, French 6.02%, Portuguese 3.61%, Spanish 3.01%).

[2] `http://olp.dfki.de/OntoSelect/w/index.php?mode=stats`

Concatenated tags such as {`christmasornament, xmlhttprequest, librariesandlibrarians`} appear frequently but obviously it is hard to identify concepts with the same spelling.

Given the very low coverage of the Semantic Web for the above mentioned categories of tags, we decided to repeat the experiments for clusters of tags that are well-covered in the Semantic Web. Also, since at this stage our system only discovers subsumption and disjoint relations, we decided that the experiments should consider significantly larger clusters than those provided by [57].

3.1.4 Experiment 2

In the second set of experiments we relied on the lessons learnt from the first experiment to identify clusters of tags that would be more appropriate for our goal. To address the first conclusion (i.e., that clusters should be potentially well covered in the Semantic Web), we relied on the results of previous work in the context of ontology matching [52]. Follow up experiments covered in the Semantic Web. Therefore, we selected a couple of tags from these domains, based on the concepts for which the most mappings were found during the matching experiments. We selected the tags: `mushroom`, `fruit`, `beverage` and `mammal`.

The next step was to identify clusters of tags related to each of these tags. As we said, we were looking for large clusters that would be more likely to accommodate subsumption relations and not just generic relations between tags. We chose the cluster generator provided by Flickr, since it returns much larger clusters of related tags than Del.icio.us and Technorati (moreover, since Del.icio.us and Technorati are mostly oriented towards news, business and web technologies, the clusters they provide for our tags in the food and animal domains are quite small).

The same algorithm as in Experiment 1 was then applied to these clusters. As expected, we found several relations among tags as depicted in the figures below (directed arrows represent `subClassOf` relations, dotted lines depict `disjointWith` relations). 23% of the investigated tags was discovered in ontologies. Besides the tags between which we found relations, there were also sets of tags that could not be linked with any other tag in their cluster. We analyze these tag sets and describe possible causes that led to this failure.

Type	Tags
Not covered by the SW	{amanitamuscaria, toadstool, flyagaric}
Generic relation (location)	{nature, forest, garden, grass, moss}
Generic relation (seasons)	{autumn, fall, herfst}
Generic relation (usage)	{cooking, dinner, pasta, lunch}
Colors	{green, white, yellow}
Photo jargon	{macro, nikon, closeup}

Table 3.1: mushroom related tags that could not be connected semantically

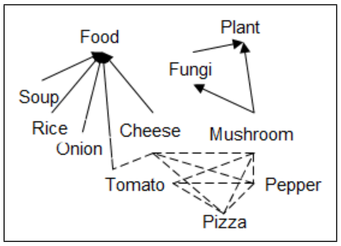


Figure 3.2: Mushroom in the Semantic Web.

The case of Mushroom. The semantic relations identified among the 21% of the tags related to mushroom by using online ontologies are depicted in Fig. 3.2. *Mushroom* was identified as a kind of *Fungi* and a kind of *Plant*. Also, we have learnt that it is disjunct with *Pizza*, *Pepper*, *Cheese* and *Tomato* and so are these with each other. *Mushroom* also co-occurs with *Soup*, *Rice* and *Onion*. As expected, there is no subsumption relation between these concepts and *Mushroom*. However, they are all subclasses of *Food*, as are *Tomato* and *Cheese* as well.

Table 3.1 shows some of the tags in the cluster of mushroom that could not be related semantically to any other tag, grouped according to the reason why they could not be linked. These are:

Tags that are not covered by the Semantic Web. These tags refer to kinds of mushrooms or scientific names that are not described in the Semantic Web. Generally, our experience is that currently very few online ontologies cover scientific labels.

Tags generically related to mushroom. The next three sets of tags are related to mushroom through other generic relations than subsumption or disjunction and describe locations, time and potential ways to use mushrooms.

Tags about colors. This set of tags is not surprising reflecting the fact that we retrieved the tag clusters from a photo-sharing system where users add color names to describe the image content of their photos. Note, however, that these colors might be meant to describe the rest of the tags associated to a resource, e.g., {`green pepper, white mushroom, yellow cheese`}. Unfortunately, because the creation of compound tags such as these is not well handled by folksonomies, users have to add each tag separately, thus loosing the relationship between them.

Photo jargon. The remaining group of tags are Flickr related tags, as we discussed in Experiment 1, and are not covered in the Semantic Web. Also, given the fact that they describe the photographs rather than their content, even if they were covered it is quite unlikely that they could be related to mushrooms or any other tag describing image content.

The case of Fruit. We obtained interesting results for the cluster of `fruit` (Fig. 3.3) and the highest percentage of related tags, 29%. As fruits are well-covered by the Semantic Web, the generated semantic structure contains much more information than a single relation between the tags of the cluster. For example the multiple relations that exist between *Fruit* and *Vegetable*, and how this affects their common subclass, *Tomato*. In a biological context, a tomato is indeed the fruit of a tomato plant, however, normally one would classify tomatoes as types of vegetables. While such different views can co-exist, the fact that *Fruit* and *Vegetable* are disjoint makes this bit of knowledge inconsistent. Therefore, once such structures are derived from multiple ontologies, their consistency should be verified. Also, according to online ontologies, *Fruit* is disjoint with *Dessert*. The validity of this statement depends on the point of view we adopt: some would argue that fruits are desserts, while others might consider desserts generally inappropriate catogorisation for fruits. Finally *Strawberry* and *Watermelon* were also found as subclasses of *Fruit*, but declaring them as subclasses of *Berry* and *Melon*, respectively, automatically infers they are also subclasses of*Fruit*.

The tags that could not be connected to *Fruit* fall into five categories (see Table 3.2), two of which are related to colors and photo jargons, as discussed before. A new set of interesting tags describes attributes generally related to fruits: {`juicy, yummy, delicious, fresh, sweet`}. Unfortunately, most concepts in ontologies model nouns. Attributes are often modeled as properties (geneneric relations). Finally, the other two sets of interesting tags refer to fruit cultivation methods and possibly best seasons for consumption of specific fruits, which again share generic relations with fruits, currently not in the scope of our software.

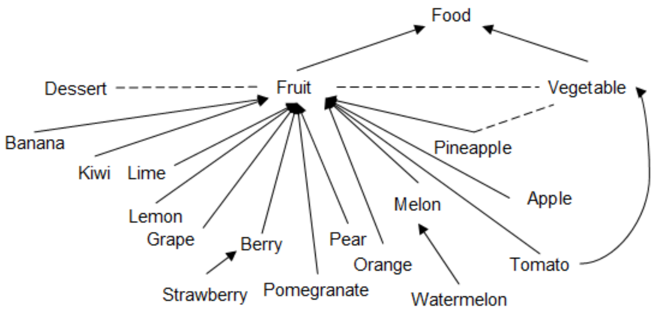


Figure 3.3: *Fruit* in the Semantic Web

Type	**Tags**
Attributes	{`juicy, yummy, delicious, fresh, sweet`}
Generic relation (cultivation)	{`tree, nature, plant, seeds, leaves`}
Generic relation (seasons)	{`summer, autumn, fall, red, pink`}
Colors	{`brown, green, white, red, pink`}
Photo jargon	{`closeup, macro, canon`}

Table 3.2: `fruit` related tags that could not be connected semantically

The case of Beverage. Beverage is the least covered tag with 18% of its related tags found to be connected in the Semantic Web. The knowledge structure that emerged from the semantic enrichment of the cluster related to `beverage` is shown in Fig. 3.4. As in the case of `fruit`, the cluster for `beverage` contains many concepts that were more specific than *Beverage*. Accordingly, these were identified to be in a subsumption relation with *Beverage* by our system. The two most interesting cases are of *White* being a subclass of *Beer* (white beer as a type of beer) and *Water* not being connected to *Liquid*. *Water*, though, was found to be related with *Fluid* which doesn't belong to the related tags of `beverage`. The tags that could not be related fall under the types of categories that we have already discussed in the previous cases and are presented in Table 3.3.

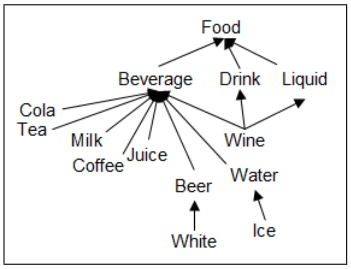


Figure 3.4: Beverage in the Semantic Web.

Type	**Tags**
Not covered by the SW	`{energy_drink, soda, martini, latte}`
Generic relation (container)	`{straw, mug, can, bottle, glass, cup}`
Generic relation (event/place)	`{breakfast, restaurant, party, starbucks}`
Generic relation(ingredient)	`{lemon, fruit, cream, orange}`
Attributes	`{hot, delicious, refreshing}`
Colors	`{brown, black, orange, green, red, pink}`
Photo jargon	`{closeup, macro, canon}`

Table 3.3: `beverage` related tags that could not be connected semantically

Some types of beverages are not covered by the Semantic Web. It is interesting to note here that `latte` is not just an English word for a type of coffee, but also Italian for milk. The fact that it is not covered can be a side-effect of the low level of multilinguality in online ontologies, as we discussed in Experiment 1. Additionally, certain tags could be related to *Beverage* by generic relations, but these are not discovered by the current version of our system. These tags express types of containers, events and locations where beverages are served, as well as the ingredients of drinks. It is worth noticing that `orange` could belong both to the categories representing colors and ingredients. The final set of tags that could not be related refer to attributes which, as discussed before, have generally a weak coverage on the Semantic Web.

The case of Mammal The last tag that was investigated is `mammal`. Relations for the 25% of its tags were found in the Semantic Web. Fig. 3.5 shows the structure derived from its cluster. It is interesting to observe that the subclasses of *Mammal* do not represent the same level of abstraction. We note many common names of animals like *Horse*, *Monkey*, *Rabbit*, but also two subclasses of higher abstraction, *Rodent* and *Feline*. This is another evidence that users annotate their content with a variable level of generality: although *Squirrel* and *Rabbit* appear in the graph as subclasses of *Mammal*, their superclass, *Rodent*, appears as well. This confirms the hypothesis put forward by [25] according to which different users will settle at different "basic levels" depending on their level of expertise.

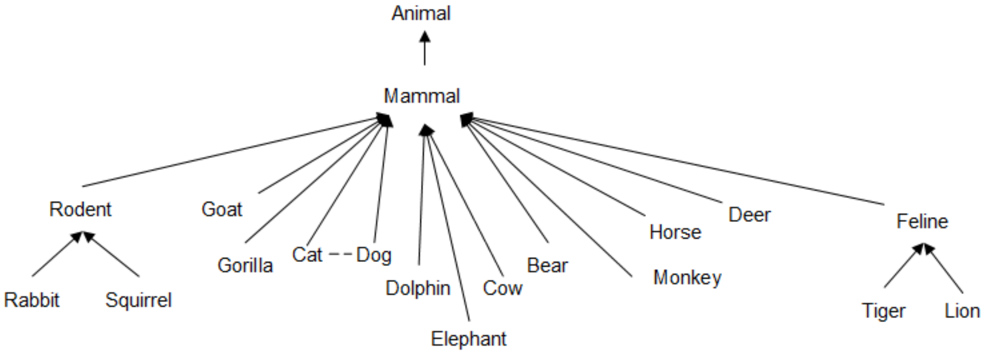


Figure 3.5: *Mammal* in the Semantic Web

The tags that could not be related are displayed in Table 3.4. Most of these categories have been discussed previously, along with a set of tags that could have been related by generic relations indicating the location or habitat of mammals. Two tags were found to describe the state of the mammal when it was shot {`eating, sleeping`}. Finally, an interesting set of tags depicts body parts which should be related to mammals through a part-of relation.

Finally, it is worth pointing out that in all of the above cases we identified certain tags, which were also found in Experiment 1, describing the places shown in the images, such as `barcelona`, `japan`, or the interests of the users, such as `ilovenature`, `stilllife` (we found 84.077 pictures annotated with `ilovenature` and 39.320 with `stilllife`).

Type	**Tags**
Not covered by the SW	{giraffe, seal, zebra}
Generic relation (location)	{zoo, nature, water, ocean, wild, farm, outdoors}
Generic relation (action)	{eating, sleeping}
Part-of	{fur, whiskers, eyes, face, nose}
Attributes	{cute, pet, funny, bunny}
Photo jargon	{portrait, closeup, macro, canon}

Table 3.4: mammal related tags that could not be connected semantically

3.1.5 Conclusions

The preliminary experiments provided results that could serve as an answer to our main research question, which is to explore whether folksonomies can be automatically enriched by harvesting the Semantic Web. Based on the results of the preliminary experiments presented above, we can already conclude that it is indeed possible to automate the semantic enrichment of folksonomy tag spaces by harvesting online ontologies. By using these ontologies, we were able to automatically obtain semantic relations between the tags of several clusters of related tags. An immediate goal of our future work is to apply our approach on folksonomies and evaluate it in terms of Information Retrieval performance values (recall and precision). As an answer to our second research question, which is to identify the inherent characteristics of folksonomies and the Semantic Web and how they should be approached, the experiments also yielded relevant observations about these characteristics which have an impact on folksonomy enrichment process:

1. Folksonomy Characteristics. Our experiments show that many folksonomy tags fall in specific categories that require special attention. First, by being dynamically updated by large masses of people, folksonomies reflect the newest terminology within several domains (**novel terminology**). Second, many folksonomy tags refer to specific **instances** (names of people, places, dates). Third, folksonomies contain tags representing words in a variety of languages (**multilinguality**). Fourth, some of the tags that are frequently used depend on the purpose of the folksonomy and usually describe the resource itself rather than its content (**folksonomy jargon**). Fifth, folksonomy tags often describe **attributes** of the content, for example, colors (especially in Flickr). Sixth, there are many **concatenated tags**

which describe a large number of photographs and need to be exploited. Finally, a **broad range of semantic relations** can exist between tags, including subsumption, disjointness, meronymy and many generic relations (e.g., location).

2. Semantic Web Characteristics. The most important observation regarding the Semantic Web is that even if it is growing fast it still suffers from *knowledge sparseness* (i.e., it presents good coverage for certain topics, but very low coverage for others). Due to this limitation, we needed to restrict our experiments to domains that are well-covered (related to animals and food). Also, some of the categories of tags that appear frequently in folksonomies are difficult to find in online ontologies. First, **novel terminology** that emerges from folksonomies is often missing from ontologies. Second, the majority of **specific instances** that appear in folksonomies cannot be found (e.g., `aug2004`) or are difficult to reliably map to ontology instances (e.g., `monica`). Place names are an exception to this. Third, few of the online ontologies contain **multilingual labels**, therefore tags in languages other than English are unlikely to be found in ontologies. Fourth, **specific jargons**, such as those related to photography are weakly covered as well. Fifth, online ontologies are rather **poor in describing generic attributes** such as color. One of the reason for this is that attributes are most often modeled as part of properties rather than concepts.

We are confident, however, that surpassing some of the current limitations is a matter of time as many of them will be solved as more ontologies will appear online. For example, the AGROVOC [21] ontology contains roughly 16000 concepts and their labels in 12 different languages. Making this single ontology available online will positively impact on the issue of anchoring multilingual tags. Nevertheless the appearance of more online ontologies can also be seen as a potential risk for this work as different ontologies reflect different views which often lead to contradictory bits of knowledge. Combining these bits may result in inconsistencies in the derived semantic structures. However, existing reasoning techniques can be used to filter out and eliminate possible inconsistencies.

Being aware of these characteristics help us to identify the **current limitations of our software**. Our software only implements a subset of the functionality envisioned for the enrichment algorithm. First, it is currently implemented on Swoogle'05 which lags behind in ontological content. Our final algorithm will be built on top of up-to-date semantic search engines

[17]. Second, the anchoring mechanism is based on strict string matching and therefore needs to be extended to more flexible anchoring. Third, from the broad range of semantic relations that can exist between tags, our software only identifies subsumption and disjointness. Obviously, extensions are needed that can discover the other types of relations as well. Finally, note that we have only experimented with finding relations within a single ontology and excluded cases when knowledge can be derived by combining facts from multiple ontologies. Another important future work will be to implement this cross-ontology relation derivation.

The experimental work reported in this section indicates that the proposed enrichment process has the potential to benefit both folksonomies and the Semantic Web, thus answering our third research question. On the one hand, even using a software with limited functionality we were able to derive *explicit* semantic relations between tags, thus going beyond existing methods that identify *implicitly* inter-related tags. We believe this could considerably enhance content retrieval in folksonomies. On the other hand, the differences between folksonomies and ontologies (such as novel terminologies emerging in several languages) can be used to evolve the Semantic Web. This valuable knowledge available in folksonomies could allow keeping online ontologies up to date, extending them with multi-lingual information and evolving them towards being truly *shared* conceptualisations of a much broader range of domains.

3.2 Methodology

Our approach is to create a semantic layer on top of folksonomies as an intermediate interface between the user queries and the tagspaces. We aim for this layer to provide the additional semantics that folksonomies lack and to enhance the content retrieval. As demonstrated in Fig. 3.6, our work is divided in two phases, each addressing one of two research questions defined in Section 1.3.

Currently, the content stored in folksonomies is in XML format similar to the one described in Fig. 3.7. There is, apart from the title, description and other attributes of the resource, the user(s) that have tagged or uploaded this resource, depending on the folksonomy, the URL where this can be accessible and finally a set of tags given to this resource by the user(s).

The Semantic Enrichment (see Fig. 3.6) generates the semantic layer on

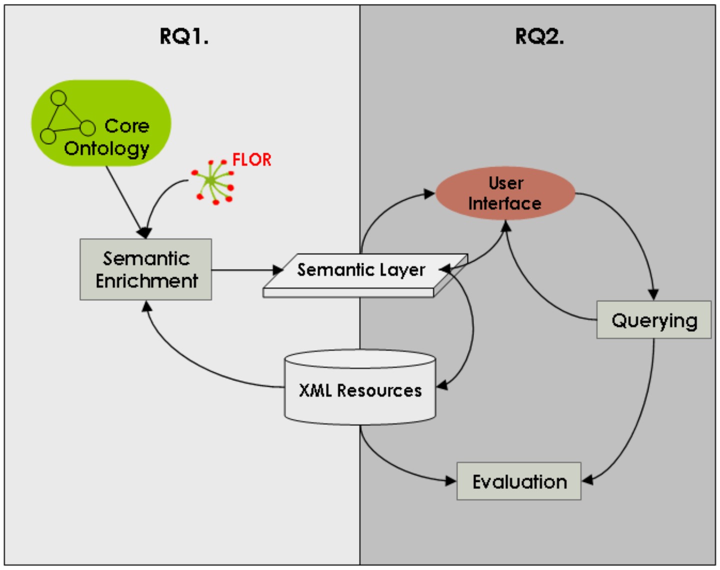
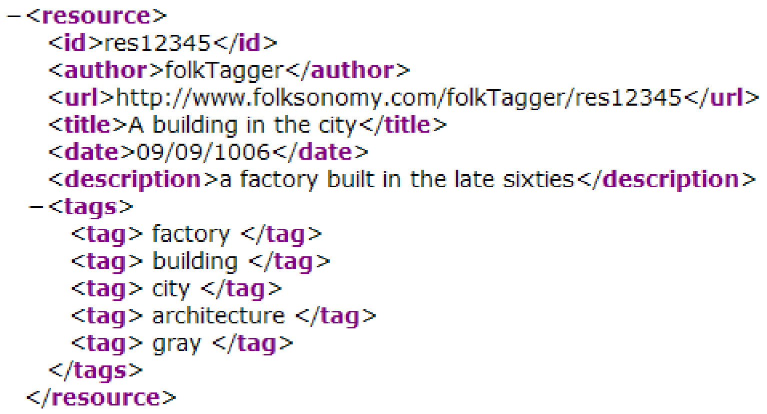


Figure 3.6: Semantic layer on a folksonomy tagspace

top of folksonomy resources and requires apart from the folksonomy data, a Core Ontology and a Semantic Enrichment mechanism/tool, FLOR. We describe our work on FLOR in detail in Section 3.3.

The Core Ontology is necessary to describe relations between the folksonomy entities (users, resources and tags) provided by the XML schema of the input data. Also, the Core Ontology is the schema of the semantic layer, defining how the semantics extracted from online ontologies with FLOR will be integrated in the semantic layer. The Core Ontology schema is presented in Fig. 3.8.

The relations *User* `Tags` *Resource* and *Resource* `isTaggedWith` *Specific_Tag* are already provided as in Fig. 3.7. (The rest of the information regarding the resource, such as title, date, description, e.t.c. can also be included in the Core Ontology, if necessary, in the same way). We define as a *Specific_Tag* the class to represent the existence of a tag in the context of a resource and a user and *Global_Tag* the class to represent the tag as a unique entity in the system. This is done because we aim to further assign the *Specific_Tag* and the *Global_Tag* a specific definition, represented by the class *Semantic_Definition*. Furthermore the *Semantic_Definition* `isDefinedBy` *Semantic_Web_Entity*.



```
-<resource>
   <id>res12345</id>
   <author>folkTagger</author>
   <url>http://www.folksonomy.com/folkTagger/res12345</url>
   <title>A building in the city</title>
   <date>09/09/1006</date>
   <description>a factory built in the late sixties</description>
 -<tags>
     <tag> factory </tag>
     <tag> building </tag>
     <tag> city </tag>
     <tag> architecture </tag>
     <tag> gray </tag>
   </tags>
 </resource>
```

Figure 3.7: Folksonomy Data

For example if we want to enrich two resources tagged with the tag {`jaguar`}, one meaning the animal and the other the car then we would have two instances of *Specific_Tag* for each resource and one instance of the *Global_Tag*, which incidentally is going to be unique and the same for all the resources tagged with this tag. Furthermore, each *Specific_Tag* would be defined by a respective *Semantic_Definition*, one defined by a Semantic Web Entity declaring an animal and the other defined by a SWE defining the car. Finally the *Global_Tag* would relate to both Semantic Definitions with the relation `hasPotentialDefinition`.

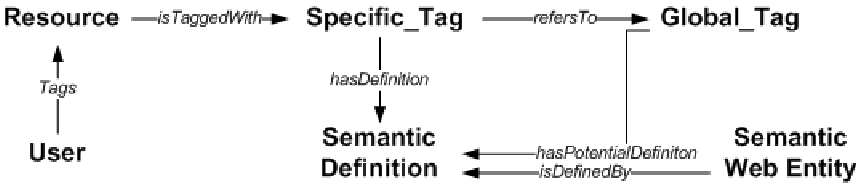


Figure 3.8: Core Ontology

Referring to Fig. 3.6, with regards to the first phase that addresses the Research Question 1, we shortly described the Core Ontology and in Section 3.3 we describe the Semantic Enrichment tool, FLOR, which essentially creates the Semantic Definitions/Tag Meanings-Senses and links them to the

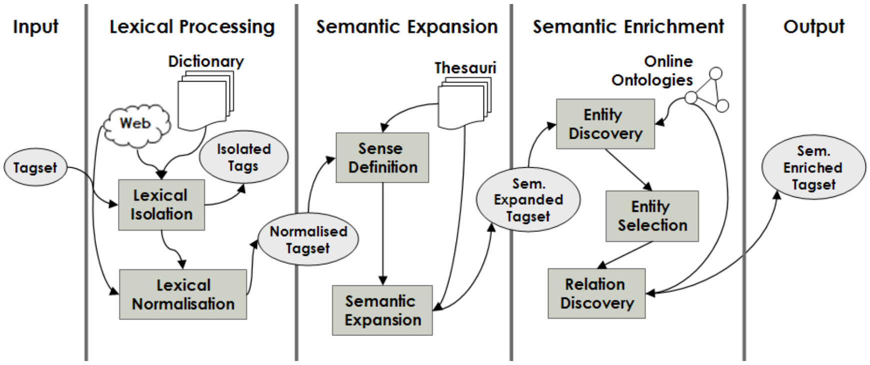


Figure 3.9: FLOR Phases

appropriate Semantic Web Entity, automatically.

3.3 FLOR: a tool for folksonomy enrichment

Our preliminary experiments on folksonomy enrichment were presented in [9] and described the FoLksonomy Ontology enRichment tool, FLOR. The goal of FLOR is to transform a flat folksonomy tagspace into a rich semantic representation by assigning relevant Semantic Web Entities (SWEs) to each tag. A SWE is an ontological entity (class, relation, instance) defined in an online available ontology. While in this section we describe the process of enriching a set of tags with SWEs, the ultimate goal of our system is not just to connect to SWE's but also to bring in other knowledge related to these SWE's. An example of the inputs and expected outcomes to FLOR is demonstrated in Fig. 3.10. The input consists a set of tags and the output is a set of semantically enriched FlorTags. Note that FLOR is agnostic to the way in which this tagset was obtained. It can either be the set of all tags associated to a resource, or a cluster of related tags obtained through co-occurrence based clustering. The experiments reported in this section used sets of tags associated with a given resource.

Intuitively, FLOR performs three basic steps (see Fig. 3.10). First, during the **Lexical Processing** the input tagset is cleaned and all ptentially meaningless tags are excluded. We rely on a set of heuristics to decide which tags are likely to be meaningless. Second, during the **Sense Definition and Semantic Expansion** we attempt to assign a WordNet sense to each tag

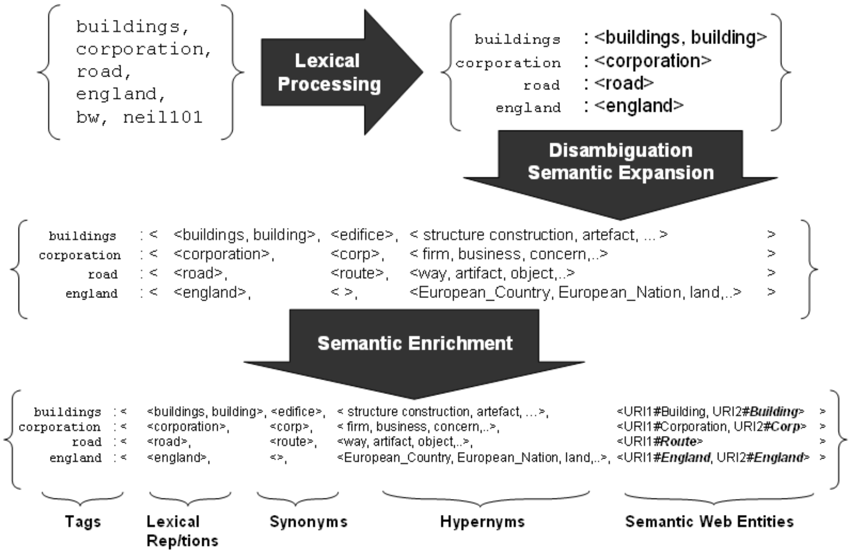

buildings,
corporation,
road,
england,
bw, neil101
Lexical Processing
buildings : <buildings, building>
corporation : <corporation>
road : <road>
england : <england>
Disambiguation Semantic Expansion
buildings : < <buildings, building>, <edifice>, < structure construction, artefact, ... > >
corporation : < <corporation>, <corp>, < firm, business, concern,..> >
road : < <road>, <route>, <way, artifact, object,..> >
england : < <england>, < >, <European_Country, European_Nation, land,..> >
Semantic Enrichment
buildings : < <buildings, building>, <edifice>, < structure construction, artefact, ...>, <URI1#Building, URI2#Building> >
corporation : < <corporation>, <corp>, < firm, business, concern,..>, <URI1#Corporation, URI2#Corp> >
road : < <road>, <route>, <way, artifact, object,..>, <URI1#Route> >
england : < <england>, <>, <European_Country, European_Nation, land,..>, <URI1#England, URI2#England> >
Tags
Lexical Rep/tions
Synonyms
Hypernyms
Semantic Web Entities

Figure 3.10: FLOR Methodology

based on its context (i.e., the other tags in its cluster) and to extract all relevant synonyms and hypernyms so that we migrate to a richer representation of the tag. Finally, during the **Semantic Enrichment** step each tag is associated to the appropriate SWE. Note that there is a strong correlation between the steps of FLOR and the components of the final FlorTag structure. The first step results in the **Lexical Representations** which is a list of lexical forms for the tag, such as plural and singular forms for nouns, or various delimited types of compound tags (sanFrancisco, san.Francisco, e.t.c). The second step identifies **Synonyms** and **Hypernyms** for each tag.

The last step generates the list of **Entities** containing the associated SWE's. Note that a tag can be associated to several relevant SWE's.

3.3.1 PHASE1: Lexical Processing

Due to the freedom of tagging as a basic rule of folksonomies, a wide variety of different tag types are in use. Understanding the types of tags used is the first step in deciding which of them are meaningful and should be taken into account as a basis of a semantic enrichment process. Previous work ([10, 25, 42]) has identified different conceptual categories of tags (event, location, person), as well as tag categories that can be described by syntactic characteristics.

For example, there are many tags containing special characters (e.g., `:P`), numbers (e.g., `aug07`), plurals as well as singular forms of the same word (e.g., `building`, `buildings`), concatenated tags (e.g., `littlegirl`) or tags with spaces (e.g., `little girl`) and a big number of non-English tags (e.g., `sillon`). The role of the lexical processing step is to identify these different categories of tags and exclude those that are meaningless and should not be further included in the semantic enrichment process. This is done in two steps.

The Lexical Isolation

phase idenfies sets of tags that should be excluded as well as those that can be further processed. Currently we isolate and exclude all tags with numbers, special characters and non English tags. The reason for excluding non-English tags is that our method explores various external knowledge sources (WordNet, Semantic Web ontologies) that are primarily in English. As future work, we will extend FLOR to isolate additional types of tags as well and deal with non-English tags.

The Lexical Normalisation

phase aims to solve the incompatibility between different naming conventions used in folksonomies, ontologies and thesauri such as WordNet. This phase produces a list of possible **Lexical Representations** for each tag aiming to maximise the coverage of this tag by different resources. For example, the compound tag `santabarbara` in folksonomies appears as *Santa-Barbara* or *Santa+Barbara* in various ontologies and as ***Santa Barbara*** in WordNet. However, as the lexical anchoring to these resources is a quite complex problem, we try to address it by producing all the possible lexical representations for each tag such as: {santaBarbara, santa.barbara, santa_barbara, santa barbara, santa-barbara, santa+barbara, ...}.

3.3.2 PHASE2: Sense Definition and Semantic Expansion

Due to polysemy, the same tag can have different meanings in different contexts. For example, the tag `jaguar` can describe either a car or an animal depending on the context in which it appears. Before connecting a tag with a relevant SWE, it is important to determine its intended sense in the given context. This task is performed in the first step of this phase.

Another issue to take into account is that, despite its significant growth, the Semantic Web is still sparse. A direct implication is that while online ontologies might not contain concepts that are syntactically equivalent to a given tag, they might contain concepts that are labeled with one of its synonyms. To overcome this limitation, we perform a semantic expansion for each tag, based on its previously identified sense, in the final step of this phase.

The Sense Definition and Disambiguation

phase deals with discovering the intended sense of a tag in the context it appears. As context we consider the set of tags with which the given tag co-occurs when describing a resource. For example, in the tagset: {`panther`, `jaguar`, `jungle`, `wild`} the context of `jaguar` is {`panther`, `jungle`, `wild`}. We use WordNet as a sense repository and rely on its hierarchy of senses to compute the similarities between the senses of all tags in the tagset and thus achieve their disambiguation. WordNet also provides rich sense definitions which facilitate the semantic expansion in the next step.

To define the senses of the tags in a tagset, we identify all the lexical representations for each tag in WordNet. In the cases that a tag has more than one senses in WordNet (synsets) we exploit the contextual information of the tagset to identify the most relevant sense. For this, we calculate the similarity between all the combinations of tags in the tagset using the Wu and Palmer similarity formula ([63]) on the WordNet graph. The similarity degree between two senses is calculated based on the number of common ancestors between them in the WordNet hierarchy and the length of their connecting path. The result for each calculation is a couple of senses and a similarity degree for these senses. We select the two senses of the tags that return the highest similarity degree provided that this is higher than a specified threshold. If a tag has low similarities when compared to all the other tags in its cluster, then it is assigned to the most popular WordNet sense.

We currently use a threshold value of 0.8 which we observed to correctly indicate relatedness in most of the cases. Indeed, as high values as 0.7 are often assigned to unrelated tags. For example, in the tagset: {`girl`, `eating`, `red`, `apple`} the similarity between `red` and `girl` is 0.7 for the senses:

Bolshevik, Marxist, Pinko, Red, Bolshie (emotionally charged terms used to refer to extreme radicals or revolutionaries)

Girlfriend, Girl, Lady_friend (a girl or young woman with whom a man is romantically involved)

These two senses are connected through the concept ***Person*** in the WordNet hierarchy, however the two tags are unrelated in the context of this tag cluster. While this empirically established 0.8 value lead to reasonable results and was sufficient for this proof of concept prototype, we plan to establish an optimal value through systematic experiments.

Thanks to the modular architecture of FLOR, the disambiguation and sense selection method can be replaced by other methods (e.g., such as those used in [58] and [64]). Or our current method could be modified to exploit a different similarity measure between two concepts such as the Google Similarity Distance [14].

Another possible improvement could be achieved by further expanding the resource tagset with more related tags. These can be discovered with statistical measures based on tag co-occurrence as described in [57]. For example, the expanded tagset of {`apple`, `mac`} could be {`apple`, `mac`, `computer`, `macOs`}. So instead of trying to disambiguate with two tags we increase the possibilities of finding the correct sense by disambiguating with a more specific context.

The Semantic Expansion

includes the synonyms and hypernyms of a tag in the FlorTag (see Fig. 3.10). For the purpose of this work we used WordNet to extract the synonyms of the correct sense and the synonyms of this sense's hypernym in WordNet. For example, if in the specific context the tag `jaguar` refers to an animal then the semantic expansion would include a list of synonyms: {***Panther***, ***Panthera onca***, ***Felis onca***} and a list of hypernyms: {***Big cat***, ***Feline***, ***Carnivore***}.

3.3.3 PHASE3: Semantic Enrichment

This phase of FLOR identifies the SWEs that are relevant for each tag by leveraging the results of lexical cleaning and semantic expansion performed in the previous two phases. The final output of FLOR is produced by this phase (see Fig. 3.10) and it is a set of FlorTags enriched with relevant SWEs and their semantic neighbourhood (e.g., parents, children, relations).

The relevant SWEs are selected by querying the WATSON Semantic Web gateway[17], which gives access to all online ontologies. We search for all ontological entities (Classes, Properties, Individuals) that contain in their local name or in their label(s) one of the lexical representations or the synonyms of a tag.

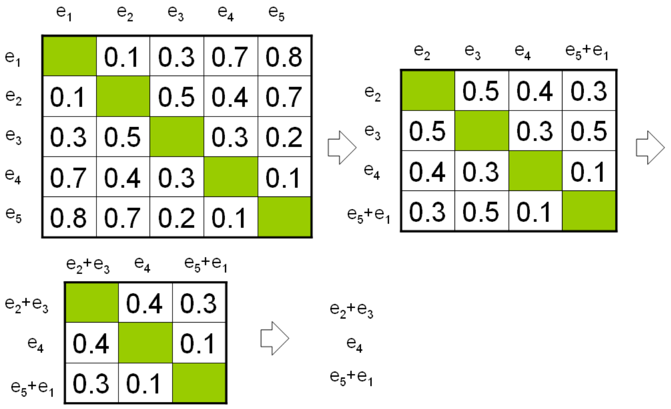
Such queries often result in several SWEs some of which are very similar (or the same when they appear in ontologies that are versions of each other). To reduce the number of SWEs, we perform an entity integration process similar to the one described in [58]. The goal of this process is to "collapse" entities that have a high similarity into a single semantic object, thus reducing redundancy. To compute similarity between two entities we compare their semantic neighbourhoods (superclasses, subclasses, disjoint classes for classes; domain, range, superproperties, subproperties for properties) and their localnames and labels. The similarity $simDgr$ for two SWEs e_1 and e_2 is calculated as:

$$simDgr = W_l * simLexical(e_1, e_2) + W_g * simGraph(e_1, e_2)$$

$simLexical(e_1, e_2)$ is the similarity between the lexical information of two entities, i.e., their labels and localnames, computed with the Levenshtein distance metric. $simGraph(e_1, e_2)$ is the similarity of the entities' neighbourhoods, where the similarity of each neighbourhood element is computed based on string similarity. Because we consider the similarity of the semantic neighbourhoods more important than the similarity of the labels, we set the weights as $W_l = 0.3$ and $W_g = 0.7$. Note that these weights will be fine-tuned through systematic experiments.

If the similarity between two entities is higher than a threshold we merge them in one entity by integrating their neighbourhoods into one. Then we repeat the process until all entities are sufficiently different from each other, i.e., their similarity falls under a chosen threshold.

Consider for example Fig. 3.11 where five SWEs $e_{1,5}$ are compared against a threshold value of 0.5. We start by performing their pair-wise comparison and observe that the pairs (e_1, e_4), (e_1, e_5), (e_2, e_3) and (e_2, e_5) have a similarity equal or above the set threshold. We proceed by merging the first two entities with the highest similarity, e_1 and e_5, to one entity e_1+e_5 and compute the similarities between the new entity and the remaining ones. This



	e_1	e_2	e_3	e_4	e_5
e_1	🟩	0.1	0.3	0.7	0.8
e_2	0.1	🟩	0.5	0.4	0.7
e_3	0.3	0.5	🟩	0.3	0.2
e_4	0.7	0.4	0.3	🟩	0.1
e_5	0.8	0.7	0.2	0.1	🟩

	e_2	e_3	e_4	e_5+e_1
e_2	🟩	0.5	0.4	0.3
e_3	0.5	🟩	0.3	0.5
e_4	0.4	0.3	🟩	0.1
e_5+e_1	0.3	0.5	0.1	🟩

	e_2+e_3	e_4	e_5+e_1
e_2+e_3	🟩	0.4	0.3
e_4	0.4	🟩	0.1
e_5+e_1	0.3	0.1	🟩

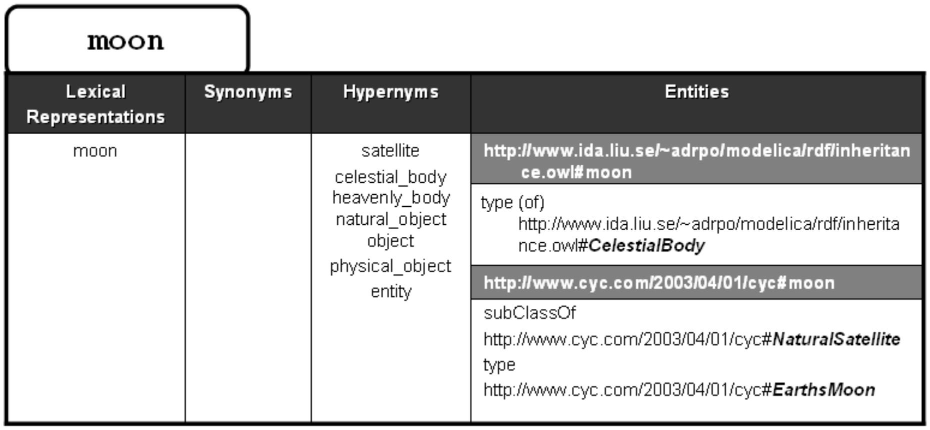
Figure 3.11: Merging Strategy with threshold 0.5

process continues until all similarities are lower than the set threshold, which implies that the obtained entities are sufficiently different.

Once the merged entities are created we enrich the tag with the relevant entities. This is done by comparing the ontological parents of the merged entity with the hypernyms retrieved from WordNet. The ontological parents are the superclasses of classes, the superproperties of properties and the classes of individuals. For example, as shown in Fig. 3.12, the tag `moon` is enriched with two entities. The superclasses of both the entities have as localname one of the hypernyms extracted from the WordNet sense of `moon`. Also, apart from the semantic definition of the tag with the respective entity, we further enrich the tag with the information carried by the entity, *EarthsMoon* `TypeOf` *Moon*.

3.3.4 An Enrichment Example

In this section we present a full cycle of the FLOR semantic enrichment method for the tag `lake`, which was found in the following five tagsets: {`rush`, `lake`, `pakistan`, `rakaposhi`, `mountain`, `asia`, `kashmir`, `snow`, `glacier`, `green`, `white`, `sky`, `blue`, `clouds`, `water`}, {`moraine`, `alberta`, `banff`, `canada`, `lake`, `lac`, `rockies`, `scan`}, {`rising`, `sunlight`, `lake`, `quality`, `bravo`}, {`lake`, `nature`, `landscape`, `sunset`, `water`, `organisms`} and {`lake`, `finland`, `suomi`, `beach`, `bubbles`, `blue`, `sunlight`, `kids`, `natural`}. Note that these tagsets contain the tags that remain after the lexical processing performed in the first phase. Fig. 3.13 shows the information contained in the automat-



moon

Lexical Representations	Synonyms	Hypernyms	Entities
moon		satellite celestial_body heavenly_body natural_object object physical_object entity	**http://www.ida.liu.se/~adrpo/modelica/rdf/inheritance.owl#moon** type (of) http://www.ida.liu.se/~adrpo/modelica/rdf/inheritance.owl#***CelestialBody*** **http://www.cyc.com/2003/04/01/cyc#moon** subClassOf http://www.cyc.com/2003/04/01/cyc#***NaturalSatellite*** type http://www.cyc.com/2003/04/01/cyc#***EarthsMoon***

Figure 3.12: Enriched FlorTag `moon`

ically obtained FlorTag.

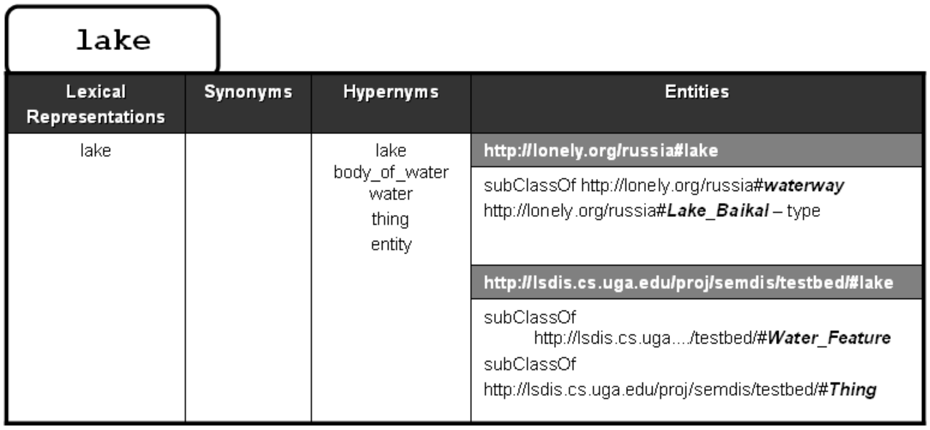
For the second phase of FLOR, Sense Definition and Semantic Expansion using WordNet, the available WordNet senses for ***Lake*** are considered. These are the following:

WordNet 1: ***Lake→Body of water***, ***Water→Thing→Entity***
(a body of (usually fresh) water surrounded by land)

WordNet 2: ***Lake→Pigment→Coloring material→Material → Substance→Entity***
(a purplish red pigment prepared from lac or cochineal)

WordNet 3: ***Lake→Pigment→Coloring material→Material →Substance→Entity***
(any of numerous bright translucent organic pigments)

Applying the Wu and Palmer formula for the senses of `lake` and the senses of the rest of the tags in these tagsets we obtained variable similarities from 0 to 0.86. The zero similarities were obtained for location names such as `banf`, `pakistan`, `suomi` and for generally unrelated tags such as `quality`, `scan`, `sunlight`, `sunset`. Interestingly, `lake` returned zero similarity for the tags `glacier` and `mountain` while they should be related. This is due to the fact that, in WordNet, ***Glacier*** and ***Mountain*** are hyponyms of ***Geological formation*** which is a hyponym of ***Natural object*** while ***Lake***



lake

Lexical Representations	Synonyms	Hypernyms	Entities
lake		lake body_of_water water thing entity	**http://lonely.org/russia#lake** subClassOf http://lonely.org/russia#***waterway*** http://lonely.org/russia#***Lake_Baikal*** – type **http://lsdis.cs.uga.edu/proj/semdis/testbed/#lake** subClassOf http://lsdis.cs.uga..../testbed/#***Water_Feature*** subClassOf http://lsdis.cs.uga.edu/proj/semdis/testbed/#***Thing***

Figure 3.13: Enriched FlorTag `lake`

is a hyponym of ***Body of water*** which is a direct hyponym of ***Thing***. Furthermore ***Glacier*** is a hyponym of ***Ice mass*** but there is no subsumption relation between ***Ice mass*** and ***Ice*** or ***Water*** that would allow for a connecting path between ***Lake*** and ***Glacier***. This fact motivates further research on how to identify similarities between tags of a tagset beyond the subsumption relations provided by WordNet.

The highest similarity, 0.86, for `lake` was obtained with the tag `water`, because Sense 1 of ***Lake*** is related to ***Body of water*** (Sense 2 of ***Water***) with a direct hyponymy relation. Note that, in most of tagsets the first sense of ***Water***, ***Liquid***, is selected as this is the most common sense in which the tag is used. Therefore, this is a nice example of phase 2 identifying a non-trivial correlation.

Sense 1. ***Water***, ***H2O***: (binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid) → ***Binary Compound*** AND → ***Liquid***

Sense 2. ***Body of water***, ***Water***: (the part of the earth's surface covered with water) → ***Thing***

Once the correct sense is selected and the tag is semantically expanded with hypernyms (there are no synonyms for this sense of ***Lake***) then the third phase of FLOR queries the online ontologies through WATSON and selects the SWEs that correspond to this sense. As shown in Fig. 3.13 both selected entities have the term *Lake* in their localname and their superclass in the ontology contains one or more of the hypernyms returned by WordNet, *Water* and *Thing*, as a whole or as a compound. This example shows that our anchoring to ontologies is strict for the tags to be defined (their lexical representations and synonyms) and the localnames and labels of the entities and flexible for the ontological parents and hypernyms. Note also that the selected SWEs carry additional information about two superclasses of *Lake* (*Waterway*, *Waterfeature*) and an instance of *Lake* (*Lake Baikal*) thus further enriching the tag.

3.3.5 Experiments and Results

To assess the correctness of FLOR enrichment (i.e., whether tags were linked to relevant SWEs) we applied FLOR on a Flickr data set comprised of 250 randomly selected photos with a total of 2819 individual tags. During the Lexical Isolation we removed 59% of the initial tags resulting to 1146 tags in total. We isolated 45 tags with two characters (e.g., `pb`, `ak`), 333 tags with numbers (e.g., `356days`, `tag1`), 86 tags with special characters (e.g., `:P`, `(raw → jpg)`), and 818 non English tags (e.g., `turdus`,`arbol`). Then we filtered out the photos that exclusively contained the isolated tags (24 photos) and obtained a dataset of 226 photos with a total of 1146 tags. After running the FLOR enrichment algorithm for these 226 photos, one of the authors manually checked all the assignments between tags and SWE's.

The assignment of a SWE to a tag is considered correct if the concept described by the SWE is the same as the concept of the tag in the context of its tagset. To decide that the evaluator was given a tagset and the SWEs linked to its tags. She evaluated each tag enrichment as CORRECT if the tag was linked to the appropriate SWE and INCORRECT otherwise. In cases when she was not sure about the intended meaning of the tag, she rated the enrichment as UNDETERMINED. Finally, a NON ENRICHED value was assigned to tags that were not associated to any SWE.

The results are displayed in in Table 3.5.

Out of the individual 1146 lexically processed tags, FLOR correctly enriched 281 tags and incorrectly enriched 20 tags thus leading to precision results of 93%. An example of incorrect enrichment is that of `square` in the context {`street`, `square`, `film`, `color`, `documentary`}. While its intended meaning is ***Geographical area***, because during the disambiguation phase `square` did not return high similarity with any of the rest of the tags,

Enrichment Result	# of Tags	Percentage
CORRECT	281	24.5%
INCORRECT	20	1.7%
UNDETERMINED	4	0.3%
NON ENRICHED	841	73.4%
Total	1146	100%

Table 3.5: Evaluation of semantic enrichment for individual tags.

the WordNet sense assigned to it was the most popular one, ***Geometrical shape***. This lead to the assignment of non-relevant SWE's namely, *Square* `SubClassOf` *Rectangle* and *Square* `SubClassOf` *RegularPolygonShaped*. Despite this error, the rest of the tags in this tagset were correctly enriched.

FLOR failed to enrich 841 tags, i.e., 73.4% of the tags (see Table 3.5). Because this is a significant amount of tags, we wished to understand whether the enrichment failed because of FLOR's recall or because most of the tags have no equivalent coverage in online ontologies. To that end we selected a random 10% of the 841 tags (85 tags) and manually identified appropriate SWE(s) using WATSON and taking into account the context(s) of the tags in the tagset(s) they appear. Out of the 85 tags we manually enriched 29. We therefore estimate that the number of tags that could have been enriched by FLOR (i.e., those for which an appropriate SWE exists) is approximately 287. Thus, taking into account that the overall number of tags that should be correctly enriched was 568 (281+287) but only 281 were enriched by FLOR this leads to an approximate recall rate of 49%. While this is quite a low recall, these results are highly superior to the ones we have obtained in previous experiments where phase 2 was not part of FLOR, i.e., we directly searched for SWEs for the tags without relying on WordNet as an intermediary step. Indeed, the WordNet sense definition and expansion of the tags with synonyms and hypernyms (FLOR phase 2) increased the tag discovery in the Semantic Web thus having a positive effect on recall.

FLOR failed to enrich the above 29 tags due to the following reasons. The majority of the failures (55%) was due to **different definition** in terms of superclasses in WordNet and in online ontologies For example, the definition of `love` in WordNet and the relevant entity found in the Semantic Web are:

WordNet: *Love→Emotion→Feeling→Psychological feature*
(a strong positive emotion of regard and affection)

Semantic Web: *Love* `SubClassOf` *Affection*

Although both these definitions refer to the same sense, and additionally the superclass *Affection* belongs to the gloss of ***Love*** in WordNet, they were not matched because *Affection* does not appear as a hypernym of *Love.* Current work investigates alternative ways of Semantic Expansion.

A further 24% of the tags not connected to any SWE were assigned to the **wrong sense** during phase 2. For example, `bulb` referring to `light bulb` in its tagset is assigned the incorrect sense ***Bulb*** → ***Stalk*** → ***Stem*** → ***Plant organ***. The rest of the unenriched tags are due to failures in anchoring them into appropriate SWE's. For example, the sense of `butterfly` was correctly identified, but non of its lexical forms matched the label of the appropriate SWE (*Butterfly_Insect*):

WordNet: ***Butterfly→Lepidopterous insect*** → ***Lepidopteron*** → ***Lepidopteran*** → ***Insect***

Semantic Web: Identified entity with localname *Butterfly_Insect*

In the case of 4 tags the evaluator could not determine whether the enrichment was correct or incorrect (Table 3.5). This is because the meaning of the tag was unclear even when considering its context and the actual photo. For example, in the photo of Fig. 3.14 the meaning of the tag `volume` is unclear. In the second phase of FLOR the tag was expanded with the hypernyms ***Measure*** and ***Abstraction***. Then, it was related to the SWE *Volume* `SubClassOf` *Measure*. As the meaning of the tag was not clear for the evaluator, she evaluated it as {UNDETERMINED}. More generally, there are several cases when tags only make sense to their author (and maybe to his social group) and thus will be difficult to enrich by FLOR.

After evaluating the individual tag enrichments the evaluator was able to draw conclusions on the overall enrichment of the tagset i.e., by photo. The evaluation output is displayed in Table 3.6. This would result to {CORRECT, INCORRECT, MIXED, UNDETERMINED, NON ENRICHED}. According to this table, 179 enrichments (about 80%) were {CORRECT}, i.e., all the enriched tags of the photo are enriched correctly. Note that the {CORRECT} enrichment results are much higher from a photo-centric perspective as many tags may appear in many photos. For the total of 20 {INCORRECT} and {MIXED} enrichments, 3 of the photos had all enriched tags incorrect and 17 had at least one tag incorrectly enriched. Finally the

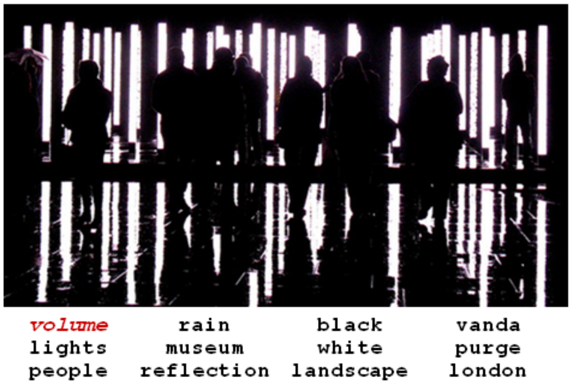


Figure 3.14: UNDETERMINED Enrichment

Enrichment Result	**# of Photos**	**Percentage**
CORRECT	179	79.2%
INCORRECT	3	1.3%
MIXED	17	7.5%
UNDETERMINED	4	1.8%
NON ENRICHED	23	10.2%
Total	226	100%

Table 3.6: Evaluation of SWE assignment to photos.

above 4 {UNDETERMINED} tags resulted to 4 {UNDETERMINED} enrichments one of which is displayed in Fig. 3.14. Finally if no enriched tag appears in the photo then the result for the photo is {NON ENRICHED}.

3.3.6 Conclusions and Future Work on Folksonomy Enrichment

We presented the methodology and the experiments we performed to test the hypothesis that **enrichment of folksonomy tagsets with ontological entities can be performed automatically**. We selected a subset of Flickr photos and after performing lexical processing and semantic expansion we correctly enriched the 72% (179 of 250) of them with at least one Semantic Web Entity. We enriched approximately the 49% of the tags with a precision of 93%. Compared to our previous efforts to define the tags with Semantic Web Entities without previously expanding them with synonyms and hypernyms, this is a significant improvement. Analysing the results we identified a number of issues to be resolved to enhance the performance of FLOR.

The **Lexical Processing** phase requires supplementary methods to identify and isolate additional special cases of tags (e.g., photography jargon, dates). Furthermore, the understanding of the impact of excluding these tags from the overall process, the implementation of strategies to deal with them and their integration in FLOR will be addressed by our future work.

As indicated by the results in Section 3.3.5, the cases of incorrect enrichment and lack of enrichment were mainly caused due to the failure of the **Sense Definition and Semantic Expansion** phase. The following issues are currently investigated in order to correct the errors and enhance the performance of this phase. First, it is essential to extend the tag similarity measure to also identify generic relations rather than only subsumption relations. This flaw was exemplified in the case of `lake` and `glacier` which were considered unrelated based the hierarchical structure of WordNet (Section 3.3.4). Also, in the example of `square` co-occurring with `street`, the incorrect sense definition for `square` caused further incorrect enrichment (Section 3.3.5) . One of the possible solutions to this is the context expansion based on tag co-occurrence. For example, expanding the {`square`, `street`} tagset with their frequently co-occurring tags e.g., {`building`, `park`} can increase the semantic relatedness between the tags and potentially lead to mapping the tags to the correct sense. Finally, to solve cases where the WordNet sense and the SWE are the same but with different hypernyms (see the example of `love`) the goal is to identify more relevant words as hypernyms or synonyms in order to achieve higher coverage in the Semantic Web.

The quality of the results returned from the **Semantic Enrichment** phase depends on (1) the input provided to this phase by the Semantic Expansion step and (2) on the anchoring of the tags' lexical representations and synonyms into online ontologies (see the case of `butterfly`). Alternative strategies for flexible anchoring to increase the number of successful enrichments and the same time keep the number of irrelevant matches low, are investigated by our current work. Also, we aim to experimentally identify optimal values for the thresholds and weight used in the second and third phases.

Finally, we aim to evaluate FLOR in large scale experiments and to assess the usefulness of the semantic enrichment in a real content retrieval application. This is to identify the possible implications of the overall process that are not apparent in a small scale study like the current one.

To conclude, we demonstrated that the **automatic enrichment of folksonomy tagsets using a combination of WordNet and online ontologies is possible** without user intervention in any step of the methodology and by using straightforward methods for lexical isolation, disambiguation, semantic expansion and semantic enrichment. The goal is to create a semantic layer on top of the flat folksonomy tagspaces, that allows intelligent annotation, search and navigation as well as the integration of resources from distinct, heterogeneous systems.

Bibliography

[1] http://.dbpedia.org.

[2] Del.icio.us: Social bookmarking. http://delicious.com.

[3] Flickr: Photo sharing. http://Flickr.com.

[4] Last.fm: The social music revolution.

[5] Linking open data. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/Linking

[6] Youtube: Video broadcasting. http://youtube.com.

[7] R. Abbasi, S. Staab, and P. Cimiano. Organizing resources on tagging systems using T-ORG. In *Proc. of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, pages 97–110, Innsbruck, Austria, 2007.

[8] H. S Al-Khalifa and H. C Davis. Measuring the semantic value of folksonomies. In *Innovations in Information Technology, 2006*, pages 1–5, 2006.

[9] S. Angeletou, M. Sabou, and E. Motta. Semantically enriching folksonomies with FLOR. In *Proc of the 5th ESWC. workshop: Collective Intelligence & the Semantic Web*, Tenerife, Spain, 2008.

[10] S. Angeletou, M. Sabou, L. Specia, and E. Motta. Bridging the gap between folksonomies and the semantic web: An experience report. In *Proc. of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, pages 30–43, Innsbruck, Austria, 2007.

[11] R. Benjamins, J. Davies, R. Baeza-Yates, P. Mika, H. Zaragoza, M. Greaves, J. Gomez-Perez, J. Contreras, J. Domingue, and D. Fensel. Near-term prospects for semantic technologies. *IEEE Intelligent Systems*, 23:76–88, 2008.

[12] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web, 2001.

[13] Ron Brachman. Emerging sciences of the internet: Some new opportunities. In *4th Eur. Semantic Web Conf.*, pages 1–3, Innsbruck, Austria, June 2007.

[14] R. Cilibrasi and P. Vitanyi. The google similarity distance. *Transactions on Knowledge and Data Engineering, IEEE*, 19(3):370–383, 2007.

[15] Tanguy Coenen, Dirk Kenis, Céline Van Damme, and Eiblin Matthys. Knowledge sharing over social networking systems: Architecture, usage patterns and their application. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 189–198, 2006.

[16] Céline Van Damme, Martin Hepp, and Katharina Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Proc. of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, 2007.

[17] M. dAquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. Watson: A gateway for the semantic web. In *4th Eur. Semantic Web Conf.*, Innsbruck, Austria, 2007.

[18] Li Ding, T. Finin, A. Joshi, Yun Peng, Rong Pan, and P. Reddivari. Search on the semantic web. *Computer*, 38:62–69, 2005.

[19] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *15th Int. WWW Conf.*, pages 193–202, Edinburgh, Scotland, 2006. ACM Press. TY - CONF.

[20] Mauricio Espinoza, Jorge Gracia, Raquel Trillo, and Eduardo Mena. Discovering the semantics of keywords: An ontology-based approach. In *In Proc. of 2006 Int. Conf. on Semantic Web and Web Services (SWWS06)*, pages 193–, Las Vegas, Nevada, USA,, 2006.

[21] FAO. Agrovoc, a multilingual agricultural thesaurus. http://www.fao.org/agrovoc.

[22] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[23] Alberto Crdoba Jess Villadangos Francisco Echarte, Jos Javier Astrain. ontology of folksonomy a new modeling method. In *Proc. of Semantic Authoring, Annotation and Knowledge Markup*, 2007.

[24] Domenico Gendarmi and Filippo Lanubile. Community-driven ontology evolution based on folksonomies. In *Community Informatics 2006*, France, 2006.

[25] Scott Golder and Bernardo Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32:198–208, 2006.

[26] M. Greaves. Semantic Web 2.0. *IEEE Intelligent Systems*, 22(2):94–96, 2007.

[27] Tom Gruber. Ontology of folksonomy: A mash-up of apples and oranges, 2005. TY - ELEC.

[28] J. Hendler. The dark side of the semantic web. *IEEE Intelligent Systems*, 22(1):2–4, 2007.

[29] J. Hendler. Web 3.0: Chicken farms on the semantic web. *Computer*, 41:106–108, 2008.

[30] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, 2006.

[31] H. Kim, H. Hwang, and H.G. Kim. Fca-based approach for mining contextualized folksonomy. In *Proc. of the 2007 ACM symposium on Applied computing*, pages 1340–1345, Seoul, Korea, 2007. ACM.

[32] H. Kim, S. Hwang, Y. Kang, and H. Yang. An agent environment for contextualizing folksonomies in a triadic context. In *1st KES Int. Symposium*, pages 728–737, Wroclaw, Poland, 2007.

[33] H. Kim, A. Passant, J.G. Breslin, S. Scerri, and S. Decker. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In *Semantic Computing, 2008 IEEE Int. Conf. on*, pages 315–322, 2008.

[34] H. Kim, S. Scerri, J. Breslin, S. Decker, and H.G. Kim. The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In *Proc. of the 2007 ACM symposium on Applied computing*, 2008.

[35] H. Kim, S. Yang, J. Breslin, and H.G. Kim. Simple algorithms for representing tag frequencies in the SCOT exporter. In *IAT '07: Proc. of the 2007 IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology*, pages 536–539, Washington, DC, USA, 2007. IEEE Computer Society.

[36] D. Laniado, D. Eynard, and M. Colombetti. Using WordNet to turn a folksonomy into a hierarchy of concepts. In *Proc.of Semantic Web Application and Perspectives - 4th Italian Semantic Web Workshop*, pages 192–201, Bari, Italy, 2007.

[37] O. Lassila and J. Hendler. Embracing "Web 3.0". *Internet Computing, IEEE*, 11(3):90–93, 2007.

[38] Faith Lawrence and Schraefel. Bringing communities to the semantic web and the semantic web to communities. In *www06*, pages 153–162. ACM, 2006.

[39] S. Lee and H. Yong. Tagplus: A retrieval system using synonym tag in folksonomy. In *Int. Conf. on Multimedia and Ubiquitous Engineering*, pages 294–298, Seoul, Korea, 2007.

[40] Sun-Sook Lee and Hwan-Seung Yong. Component based approach to handle synonym and polysemy in folksonomy. In *Computer and Information Technology, 2007. CIT 2007. 7th IEEE Int. Conf. on*, pages 200–205, 2007.

[41] M. Lux, M. Granitzer, and R. Kern. Aspects of broad folksonomies. In *Database and Expert Systems Applications, 2007. DEXA '07. 18th Int. Conf. on*, pages 283–287, 2007.

[42] M. Zied Maala, A. Delteil, and A. Azough. A conversion process from flickr tags to rdf descriptions. In *10th Int. Conf. on Business Information Systems*, Poznan, Poland, 2007.

[43] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *Collaborative Web Tagging Workshop (WWW '06)*, 2006.

[44] Elke Michlmayr. A case study on emergent semantics in communities. In *Workshop on Semantic Network Analysis, Int. Semantic Web Conf. (ISWC2005)*, November 2005.

[45] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *4th Int. Semantic Web Conf.*, pages 522–536, 2005.

[46] A. Mikroyannidis. Toward a social semantic web. *Computer*, 40:113–115, 2007.

[47] Enrico Motta and Marta Sabou. Next generation semantic web applications. In *1st Asian Semantic Web Conf.*, Beijing, China, 2006.

[48] R. Newman, D. Ayers, and S. Russell. Tag ontology, 2005.

[49] Tim O'Reilly. What is web2.0?, September 2005.

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

[51] T. Russell. cloudalicious: folksonomy over time. In *Proc. of the 6th ACM/IEEE-CS Joint Conf. on Digital Libraries*, page 364, 2006.

[52] Marta Sabou, Mathieu d'Aquin, and Enrico Motta. Using the semantic web as background knowledge for ontology mapping. In *Int. Workshop on Ontology Matching (OM-2006), collocated with ISWC'06*, Athens, Georgia, USA, 2006.

[53] Patrick Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop (WWW '06)*, 2006. TY - CONF.

[54] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21:96–101, 2006.

[55] Clay Shirky. Ontology is overrated: Categories, links, and tags, 2005.

[56] Katharina Siorpaes. myontology: The marriage of ontology engineering and collective intelligence. In *Proc. of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, pages 127–138, 2007.

[57] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *Proc. of the 4th Eur. Semantic Web Conf.*, pages 624–639, 2007.

[58] R. Trillo, J.Gracia, M. Espinoza, and E. Mena. Discovering the semantics of user keywords. *Journal of Universal Computer Science*, 13(12):1908–1935, 2007.

[59] Thomas Vanderwal. Folksonomy coinage and definition, February 2007.

[60] Csaba Veres. The semantics of folksonomies: The meaning in social tagging. *Proc of the 12th Americas Conf. on Information Systems*, other:paper 478, 2006.

[61] Harris Wu, Mohammad Zubair, and Kurt Maly. Harvesting social knowledge from folksonomies. In *Proc. of the 17th Conf. on Hypertext and hypermedia*, pages 111–114. ACM, 2006.

[62] Xian Wu, Lei Zhang, and Yong Yu. Exploring social annotations for the semantic web. In *Proc. of the 15th Int. Conf. on WWW*, pages 417–426. ACM, 2006.

[63] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico, USA, 1994.

[64] C. Yeung, N. Gibbins, and N. Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In *Int. Semantic Web Conf.*, Busan, South Korea, 2007.