FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies

Céline Van Damme[1], Martin Hepp[2], and Katharina Siorpaes[2]

[1]Vakgroep MOSI, Vrije Universiteit Brussel, Brussels, Belgium
[2]Digital Enterprise Research Institute (DERI), University of Innsbruck, Innsbruck, Austria
celine.van.damme@vub.ac.be, mhepp@computer.org,
katharina.siorpaes@deri.org

Abstract. We can observe that the amount of non-toy domain ontologies is still very limited for many areas of interest. In contrast, folksonomies are widely in use for (1) tagging Web pages (e.g. del.icio.us), (2) annotating pictures (e.g. flickr), or (3) classifying scholarly publications (e.g. bibsonomy). However, such folksonomies cannot offer the expressivity of ontologies, and the respective tags often lack a context-independent and intersubjective definition of meaning. Also, folksonomies and other unsupervised vocabularies frequently suffer from inconsistencies and redundancies. In this paper, we argue that the social interaction manifested in folksonomies and in their usage should be exploited for building and maintaining ontologies. Then, we sketch a comprehensive approach for deriving ontologies from folksonomies by integrating multiple resources and techniques. In detail, we suggest combining (1) the statistical analysis of folksonomies, associated usage data, and their implicit social networks, (2) online lexical resources like dictionaries, Wordnet, Google and Wikipedia, (3) ontologies and Semantic Web resources, (4) ontology mapping and matching approaches, and (5) functionality that helps human actors in achieving and maintaining consensus over ontology element suggestions resulting from the preceding steps.

1. Introduction

It has been argued e.g. in [1] that the insufficient involvement of users in the construction of ontologies is a significant cause for the current shortage of and the unsatisfying coverage found in domain ontologies. One of the reasons for this deficiency is that there are high barriers for laymen users for suggesting new conceptual elements. For example, a new concept, instance or property is added to the ontology only by a privileged group. This requires that ontology users with domain expertise take the burden and have the skills to make respective suggestions, which is different from the evolution of a natural language, where a new word can be invented on the spot when needed and immediately added to the vocabulary [1, 2].

Also, since ontology specifications are expressed in a formal language, potential users face difficulties in understanding the formal specifications of the ontology [1, 2]. This is important, since the inferences authorized by using a given ontology are represented only in its formal semantics, i.e. to what one commits to when adopting a particular ontology is not obvious from the human-readable labels of ontology elements but only from the associated axioms. In addition to that, we can observe that the detachment of ontology *usage* (e.g. creating annotations) from ontology

construction and maintenance in current practice cuts off valuable feedback and actually makes the social agreement over ontology elements brittle and vague.

Tagging, i.e., users describing objects with freely chosen keywords (tags) in order to retrieve content more easily, avoids these limitations, since new tags can be introduced on the spot when needed and the construction and maintenance of the tags is closely linked to their actual usage.

While the resulting tag sets and their assignment to objects are at first only reflecting subjective conceptualizations, many of those *subjective* representations can be used to derive *intersubjective* representations. Such aggregation of raw tag data leads to a flat bottom-up categorization or folksonomy [3]. Popular examples of the tagging/folksonomy mechanism are found in the social bookmark manager deli.cio.us (http://del.icio.us), the image sharing system Flickr (http://www.flickr.com), and the blog search engine Technorati (http://technorati.com).

Tagging features create a wealth of data that reflects (1) subjective assignments between words and categories of objects, (2) intersubjective patterns in these associations, and (3) implicit information on social networks.

However, tags are flat and no relationships or conceptual meanings are formally attached to them. This causes problems such as (1) lexical ambiguity; for instance, the tag “bank” can mean a financial institution or it can be used in the context of a river edge; (2) different tags (e.g. “NY” and “big_apple”) may refer to the same concept (e.g. the city New York), and (3) specialized (e.g. “seagull”) and more general tags (e.g. “bird”) may be attributed to the same object (e.g. a picture of a seagull on Flickr) [4].

Also, the same tag may be used for very different objects in clearly distinct contexts. For example, the tag “Italy” can be used to categorize *pictures taken in Italy* (in a picture database) or *customers living in Italy* (in a tagged address data base). Ontologies, on the contrary, require a clear and context-independent notion of what it means to be an instance of a respective class.

In this paper, we suggest taking an integrated approach of combining five types of resources and techniques for improving the construction of domain ontologies. We propose to exploit (1) the statistical analysis of folksonomies and the wealth of data resulting from their construction, usage, and the underlying social relationships between actors by providing a set of tools and techniques that identify structural patterns in folksonomies, (2) on-line lexical resources like dictionaries, Wordnet, Google, and Wikipedia; (3) ontologies and Semantic Web resources, (4) ontology mapping and matching approaches, and (5) functionality that helps the community in achieving and maintaining consensus.

The structure of the paper is as follows. In section 2, we give an overview of potential resources and techniques that are available for lifting folksonomies to the level of ontologies. In section 3, we explain the FolksOntology approach that is based on the integration of these elements and the involvement of the community. In section 4, we give a preliminary assessment of the possible contribution of each resource and technique. In section 5, we discuss our proposal in the light of related work, identify future research challenges, and summarize the main findings.

2. Resources for Lifting Folksonomies to the Level of Ontologies

In this section, we give an overview of promising resources that can be exploited for deriving ontologies from folksonomies. There exist at least three groups of such resources: First, folksonomies and their associated data (subsection 2.1); second, online lexical resources (subsection 2.2); and third, ontologies and other Semantic Web resources (subsection 2.3). In subsection 2.4, we discuss how mapping and matching techniques can support the process.

2.1. Folksonomies and Associated Data

Quite clearly, tagging generates more data than merely tags. When we look at Web sites that have an inherent tagging feature, we can see that there are four groups of entities involved in the tagging process: (1) tags, (2) objects, like images or bibliographic references, (3) actors, and (4) the folksonomy-driven Web sites or systems[1] themselves [5]. There is interaction between those entities, which generates a large amount of potentially valuable data, as described in the subsections below.

2.1.1. Folksonomies and Social Networks in One System

During the tagging process, actors are assigning tags to objects (figure 1). The actors describe an object using their own, freely chosen keywords, usually in order to facilitate a later retrieval process. As a consequence, the tags are expressing and reflecting the actors' subjective level of knowledge on and their interest in the respective object.

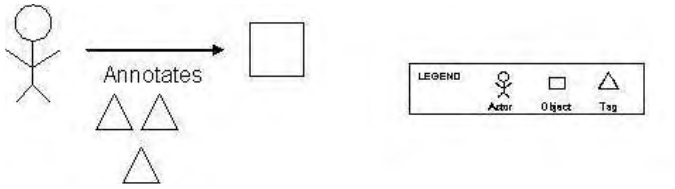

Fig. 1. The Tagging Process

In the past few years, there have been successful attempts of enriching tags with hierarchical relations [6] and the creation of faceted ontologies [7] through studying the use of objects and tags in a system. However, more information is available than merely tags, as explained e.g. in [8], in which the social dimension of actors was introduced. Out of a tripartite model of tags, objects, and actors, three bipartite graphs were generated based on the co-occurrence of its elements: the AC (actor-tag) graph, AI (actor-object) graph, and the CI (tag-object) graph. The folding of these graphs into one-mode networks generates implicit social networks, a network of instances and lightweight ontologies. [8] examines these two lightweight ontologies (one based on sub-communities of interest and another on object overlaps) on a data set of the deli.cio.us system and reveals broader/narrower relations. The authors concluded that analyzing a lightweight ontology of a sub-community is a good mean for discovering

[1] In the rest of the paper, we will use the term systems.

the emergent semantics of a community. Therefore, consolidating and analyzing the user-created data of sub-communities seems a valuable start data set for the creation of ontologies of this sub-group.

We argue that the implicit social networks in a system, which are not studied in [8], may return additional significant information. In particular, one can safely assume that actors are indirectly linked with others by sharing the same tags and/or objects. For example, as shown in figure 2, actors A and B are linked by tag3 and actors B and C are related because they both have tagged object5. In the first case, the social binding is the common language, in the second case, it is the interest in the same objects.

Analyzing such data might reveal relevant relations that can help us in reconstructing an ontology for the respective domain of interest. For instance, there might be a significant relation between object1 (annotated by actor A) and object5 (annotated by actor B): and maybe the tags should be consolidated. Furthermore, a relation might exist between tag3 and the tag set (tag4, tag5, tag6) since they are all used to annotate object5.

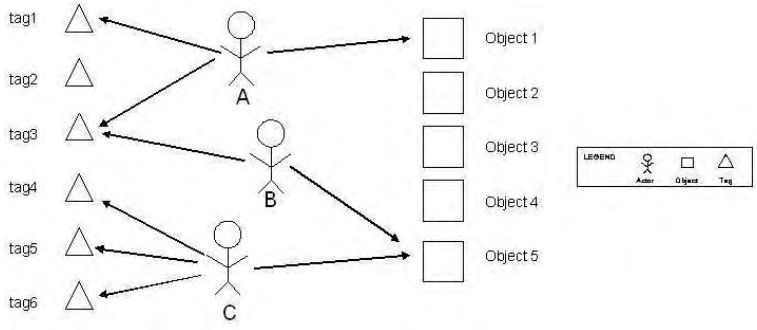

Fig. 2. The Collective Tagging Process

Sometimes, actors have already made explicit their area of interest or expertise, e.g. by joining one or more user groups on the system, which is a feature in some systems (e.g. Bibsonomy, Flickr, YouTube). By that, actors with similar interests can share their objects and tags. However, since everyone may create a new group, reduntant groups and a topic overlap between groups is likely. On Flickr, many groups are discussing and generating tags on similar kind of subjects - there exist, e.g., more than 1290 public groups on wine[2]. Therefore, aggregating the data from those groups may reveal valuable data for the creation of wine ontologies.

Actors can also make their relations and interests public by inviting other actors to their network, as is supported e.g. by deli.cio.us. Adding an actor to your network implies you are having the same interests as this actor, or that there exist some other social bonds. When all the actors are making their interests public, more information can be extracted.

[2] http://www.flickr.com/search/groups/?q=wines retrieved on April 1, 2007.

2.1.2. Folksonomies and Social Networks in Several Systems

As already mentioned, there is a fourth type of entities involved in the tagging process, i.e., *systems*. Since more and more systems are emerging, we believe that tagging data on similar topics and objects is created in parallel on different systems.

Systems are *implicitly* connected through shared sub-communities of interest or common objects. Sub-communities are not exclusively related to just one system. For instance, a sub-community on wines may exist on Flickr as on deli.cio.us. However, we have to be careful when comparing data from different kinds of systems, since a folksonomy can be broad or narrow [3]. In case the actor and creator are both the same, as is the case on Flickr, the consolidated tags constitute a narrow folksonomy. On deli.cio.us every object is tagged by, depending on the popularity of the object, several actors and the aggregation of the tags lead to a broad folksonomy. On the other hand, there may exist implicit links between systems because the actors are annotating the same sets (or kinds) of objects. For instance, the same scholarly publications are tagged on different systems (e.g. Bibsonomy and CiteULike). Consolidating the entire user-created data of similar kinds of objects, which is dispersed on several systems, may generate a more complete overview on the meta data of overlapping objects.

On the other hand, some systems are also *explicitly* connected through explicit social networks of their actors. Information on a person can be given e.g. using FOAF. FOAF allows everyone to describe him/herself (e.g. name, family name, friends), online accounts, groups and documents in a lightweight formal way[3]. Extracting the information that is stored in FOAF profiles can unveil the explicit social networks. The explicit social networks can be used for determining people with shared objects and tags. In [18] a system is proposed where actors can next to tagging their bookmarks, explicitly describe their relations with other people by FOAF. Then, they can import the tags of their friends and establish mappings between their tags and those of their peers. Doing this implies a certain level of trust and can enhance the feedback functionality in the bookmark system. In that way, [18] are trying to create a community-based ontology that is based on explicitly described relations and trust.

We can conclude that this tagging process produces several kinds of data sets that can be analyzed to exploit the information hidden in these systems. It is obvious that the design of proper tools for exploiting structural patterns in folksonomies is a core challenge for tapping this potential.

2.2. Online Lexical Resources

The data sets obtained from the previous resource can be complemented with information from lexical or terminological resources such as Leo Dictionary, Wordnet, Google, and Wikipedia.

Dictionaries are generally considered as a valuable and reliable resource containing definitions of several common words. Nowadays, several dictionaries are online accessible such as Leo Dictionary and the lexical database Wordnet. However, it is not sufficient to rely solely on these resources. For example, rather new or very specific words such as *folksonomy* can not be retrieved although the latter is an established term on the Web. Thus, we should exploit other lexical resources the Web

[3] http://xmlns.com/foaf/0.1/#sec-foafvocab retrieved on April 1, 2007.

is offering, e.g. *Google* and *Wikipedia.* Google is providing some kind of dictionary functions. Each time the user is entering a search key word, Google tries to find similar key words [15]. The search results for both queries are compared (the original one entered by the user and the similar ones). In case the alternative spelling has more hits, a suggestion is made to the user. For instance when typing in the query *occurence*, Google will make the suggestion *occurrence* since the number of results for the user key word *occurence* are significant lower. This suggestion feature is based on the principle of collective wisdom: if the majority of the Web community is using this key word, it is accepted as an existing and well-spelled word. The principle of collected wisdom can also be used for checking the proper usage of language, e.g. for finding proper prepositions. It can be futher improved by considering the region of origin and the authority of the returned Web pages (the page http://www.bbc.co.uk will have a higher credibility than on http://yahoo.com/users/pmiller.htm). The Google dictionary function can be complemented with Wikipedia, the online collaborative encyclopedia, for the identification of words. Everyone can edit and make a new Web page in this user-created encyclopedia. For instance, for "folksonomy", a Wikipedia article was already created in November 2004, whereas the respective word does still not exist in regular dictionaries[4]. With more than 5,300,000 articles [9] in various languages, Wikipedia constitutes a huge corpus of knowledge. In the English language, 1,710,088[5] articles can be identified by a URI; plus it has been shown in [2] that the conceptual meaning of the articles does not change in most cases and thus Wikipedia URIs can be regarded as authoritative identifiers for many concepts.

2.3. Ontologies and Semantic Web Resources

After consulting all the lexical resources, ontologies and Semantic Web resources can be employed as the second level of resources. Freely available ontologies can be retrieved e.g. through the Semantic Web search engine Swoogle. This search engine is searching and indexing Semantic Web documents written in RDF and OWL. It indexes the metadata of the documents and computes relationships between them [10].

Wordnet, which we mentioned in the previous section, can also be exploited as a freely available thesaurus, for which an OWL transcript is available[6]. Wordnet provides an overview of terms and their relationships (e.g. synonyms, meronyms and homonyms). It is often suggested and applied in research papers for extracting semantic information (e.g. in [11], Wordnet is employed for finding synonyms and related terms in order to reduce the communication obstruction between intelligent agents with different ontologies, and [12] use Wordnet to add a conceptual meaning to the tags when annotating a bookmark) .

2.4. Ontology Mapping and Matching Approaches

Next to resources, we can build on established techniques for ontology matching and mapping. In principle, matching of conceptual elements in two ontologies can be

[4] Merriam Webster Online, Leo Dictionaries
[5] http://en.wikipedia.org, retrieved on March 27, 2007
[6] http://www.w3.org/TR/wordnet-rdf/, retrieved May 9, 2007

based either on the labels or on the ontology structure, or both. For deriving ontologies from folksonomies, those techniques may be used in particular for identifying relationships between tags, between tags and lexical resources, and between tags and elements in existing ontologies. [13] describe the theory of formal classification, where labels are translated to a propositional concept language. Each node is associated to a normal form formula that describes the content of the node. This approach is able to capture knowledge that exists implicitly within simple classification hierarchies. [14] describe semantic matching, an approach to matching classification hierarchies. This approach is focused to the graph representation of ontologies, which means it cannot be directly applied to tag data. [15] present the FCA-Merge method, where the input to the method is a set of documents from which concepts and the ontologies to be merged are extracted using natural language techniques. These documents should be representative of the domain at question and should be related to the ontologies. They also have to cover all concepts from both ontologies as well as separating them well enough.

3. The FolksOntology Approach

In this section, we describe (1) *how* the resources from the previous section can be fully exploited for making ontologies out of folksonomies and (2) *how* the community can be involved as a mechanism to validate all the information extracted from the resources.

3.1. Fully Exploiting the Resources

A first principle of our approach is that we try to integrate every reasonable data resource and invokable functionality from the Web that can help us construct ontologies from the social interaction taking place on the Web. In other words, we want to take the vast amount of evidence created by users contributing to the Web and extract consensual conceptualizations from that.

3.1.1. Cleansing and Preparation of Tags

Before analyzing all the data sets of folksonomies, we must clean tag sets. Since actors can choose any keyword for categorizing their content, they are applying their own spelling and tagging rules (e.g. singular or plural nouns, conjugated verbs). As a consequence, tags are polluted and need to be cleansed. This can be performed through stemming algorithms. These algorithms are reducing tags to their stem or root. It is important not to loose the context of the tags, therefore the stemming process of tags should be limited to plural nouns and conjugated verbs. After this stemming algorithm, it has to be checked whether all the tags are spelled correctly. We can use the four lexical resources Leo Dictionary, Wordnet, Google, and Wikipedia to check whether or not the tags are misspelled. In case a tag is not retrieved in any of these resources, the frequency of this tag should be counted. A low frequency may indicate that the tag is misspelled and a high frequency can be an indication of the offset of a new word created in the tagging community. This word should be added to the list of new words that has to be examined by the community (subsection 3.2).

3.1.2. Statistical Analysis of Folksonomies, Usage Data, and Social Networks

In this paragraph we give an overview of data sets described in section 2.1 and explain the objective, input, output, and techniques that can be employed.

Table 1. Statistical analysis of tagging data on a single system

Step	Objective	Input	Output	Techniques
1	Determining pairs of tags	Tags, tag/object data	Pairs of tags	Co-occurrence technique: each time two tags are used to tag the same object, the tie strength between two tags is increased [19].
2	Enriching tags	Objects and Tags	a) Hierarchical relations between tags b) faceted ontology	a) [7] presents an algorithm based on the cosine similarities between tags. Tags are aggregated in tag vectors and the cosine similarity calculates the angle between two tag vectors. The smaller the angle, the more similar the tags are. The tags are consequently placed as a node in a similarity graph. If the similarity of two tags exceeds a threshold value, the two nodes are connected with an edge. A hierarchical taxonomy can be deducted from the similarity graph. b) A combination of co-occurrence between tags and a subsumption-based model is presented in [6].
3	Analyzing and creating sub-communities	Actors and tags	Lightweight ontologies based on community overlap	1) [8] folds the AC Graph (actor tags Graph) into a network based on tags. The weights of tags are calculated by the number of times the actors have used the tags in combination. [8] uses social network analysis measures (such as degree, closeness and betweenness centrality) to determine the general and specialized tags. General tags are used to bridge two clusters and specialized tags are parts of a specific cluster. Clustering techniques are used to determine the synonyms of the specialized tags. [8] uses set theory to determine the broader/narrow relations in the subcommunity
4	Analyzing social networks based on shared objects	Actors and objects	Clusters of actors with shared objects	1) Analyzing a social network. The tie strength between actors is measured by the number of times the actors have tagged the same object. Social network measures and/or clustering techniques can be used for determining the clusters of actors with similar tagged objects. 2) Analyzing the objects of the actors in each cluster: text mining techniques, digital photo similarity analysis
5	Analyzing social networks based on shared tags	Actors, tags, and objects	Clusters of actors with shared tags	1) Analyzing a social network. The tie strength between actors is measured on the number of times the actors have used the same tag. Social network measures and/or clustering techniques can be used for determining the clusters of actors using the same tags. 2) All the tags used by the actors of a cluster can be further analyzed by using the technique described in step 1
6	Merging similar	Groups (+tags,	Clusters of similar groups	1) The groups can be clustered by setting up a network analysis with groups instead of actors.

	groups	objects, actors)		However, the analysis has to be performed on data sets of equal size. This means if the size of the different groups (=number of tags) are differring, the frequency of tags has to be adjusted in proportion. The tie strength between two groups is calculated on the basis of shared tags. Social network measures and/or clustering techniques can be used for determining the clusters. 2) These clusters can be further analyzed by using the technique described in data set 1
7	Analyzing explicit social network	Actors and their relations	Clusters of actors	1) Analyzing the social network. The tie strength between actors can be 0, 1 or 2 depending on the fact of two persons have linked to each other. 2) These clusters can be further analyzed by using the technique described in step 1

Table 2. Statistical analysis of tagging data across multiple systems

Step	Objective	Input	Output	Method
1	Analyzing and creating sub-communities	Actors and tags of different systems	Clusters of communities with similar interests	1) The same techniques as described above can be employed. However, the analysis has to be performed on data sets of equal size. This means if the tags "size" of the different systems are differing, the frequency of tags has to be adjusted in proportion. 2) These clusters can be further analyzed by using one of the techniques described in step 1 in Table 1.
2	Analyzing communities of shared objects	Actors and objects of systems with the same annotated objects	Clusters of communities on overlapping objects	1) The same techniques as described above can be employed, except that the weights of the objects are calculated by the number of times the actors have used the objects in combination. However, the analysis has to be performed on data sets of equal size. This means that if the size of the different systems is differring, the proportions have to be adjusted. 2) These clusters can be further analyzed by using the technique described in step 1 in Table 1.
3	Analyzing the explicit social network	Actors (FOAF)	Clusters of actors	We can take the direct RDF data for determining social proximity.

3.1.3. Exploiting Online Lexical Resources

The tag data set obtained from the previous steps can be enriched by using the online lexical resources as described in section 2.2. However, these lexical resources can also be used for other purposes than merely spelling checks (except for Google). Tags can be replaced by concepts and homonyms, or translated from a foreign language into English as is elaborated in the following paragraphs.

Wikipedia: Wikipedia articles are identified by URIs which can be regarded as reliable identifiers for conceptual entities [2]. The meaning of those entities is

described in natural language and augmented by multimedia elements and agreed upon by a large community. Hence, Wikipedia is the biggest available collection of conceptual entities that are described with natural language and identified by URIs. Already having unique identifiers (e.g. URIs) assigned to concepts defined only in natural language is very beneficial, for it helps improve recall and precision in information retrieval by avoiding synonyms and homonyms. Additionally, Wikipedia contains disambiguation pages in order to deal with homonyms. When one word has several meanings, the meanings are collected on a disambiguation page in order to lists articles associated with the same title. This feature can be used to identify and deal with homonyms. Wikipedia also contains an implicit and evolving multilingual dictionary, since a Wikipedia page can have links that refer to the same topic in another language. These links can be retrieved in an XML format easily with the Wikipedia export function[7].

Leo dictionaries: Leo (Link everything online) provides a translation service for German, English, French, and Spanish. This functionality can be used for dealing with different languages. Additionally, Leo contains a definition of terms in German.

Wordnet can be used to deal with synonyms and homonyms: words with similar or identical meaning must be mapped to each other (e.g. baby and infant). Furthermore, words that have different conceptual meanings (e.g. Jaguar as the car and the animal) can be identified with Wordnet as well.

3.1.4. Ontologies and Semantic Web Resources

The tag sets obtained in subsection 3.1.2 can also be enriched by trying to establish mappings to elements in existing ontologies. Also, the explicit relationships in existing ontologies may be reused, e.g. for determining whether a hierarchical relation holds between two terms. In particular, the Swoogle engine can be used to query for ontologies and ontology usage data.

3.1.5. Mapping and Matching approaches

The formal classification theory of [13] can be employed for mapping the labels of existing classifications with the tags obtained from the folksonomies. Consequently, we can also use the lexical resource Wordnet to create a mapping with an existing ontology.

3.2. Mechanisms for Involving the Community

Instead of aiming at the fully automated creation of ontologies from folksonomies, we suggest a semi-automated approach, in which the aforementioned techniques are combined with collective human intelligence. In other words, we propose that (1) the results from the previous stages have to be confirmed by the community and (2) information that could not be retrieved from the resources (e.g. relations between tags) may be contributed by the community on demand. For this, we can combine visualization techniques and implicit and explicit voting mechanisms on conceptual choices. For example, a concept hierarchy reconstructed from data could be presented

[7] http://de.wikipedia.org/wiki/Spezial:Exportieren retrieved on April, 1 2007.

to the users on a separate Web page, but respective subClassOf relations would only be created if the community approves this.

4. Overview of the Contribution of Each Resource and Technique

In this section, we give a preliminary evaluation of the potential contribution of the various resources and techniques. In Table 3, we summarize the type of contribution that available techniques can provide. In Table 4, we assess the size of lexical and structural data sources that we propose to exploit. While the mere size of a resource is not always and advantage, we assume that in here, a large size makes a resource more attractive for our approach.

Table 3. Type of contribution of each technique

Technique	**Type of Contribution**
Ontology matching algorithms	Finding equivalences between labels or between conceptual elements in graphs
Co-occurrence technique	Finding tag pairs
Co-occurrence technique + Subsumption model	Creating a faceted ontology of tags
Social Network Analysis techniques + set theory	Lightweight ontologies based on community overlap
Social network techniques	Creating a) Clusters of actors with shared objects b) Clusters of actors with shared tags c) Clusters of similar groups d) Clusters of actors that have explicitly indicated their relationship
Visualizations	Visualization of ontologies helps user to grasp the intention of concepts.
Discussion and voting	Like on Wikipedia, users can remove disputes by performing discussions and then vote on the result.

Table 4. Type of contribution and size of available resources

Resource	Type of Contribution	Size
Wikipedia entries in multiple languages	Since Wikipedia contains a wealth of mutual links between pages in multiple languages that cover the same topic, we can exploit this for the unique identification of conceptual entities and for spelling checks.	5,300,000 [9] entries in total 1,710,088[8] English articles
Wikipedia disambiguation pages	Indicators for homonyms	6,67% of English articles [2]
Overlap of multiple folksonomy-driven websites targeting at the same type of objects	Finding similar tagged objects, e.g. tags referring to the same scholarly publication.	No information available
Tags	Raw set of candidate concepts	We were unable to get information on the total amount of deli.cio.us and Flickr tags – for Technorati, we at least know that there exist more than 81 Million posts[9].
Annotations	Finding tag-object patterns	Delicious: 53.000.000[10] Flickr: No data available Technorati: : 27 million weblogs[11]
Actors	Finding users with similar interests and vocabulary	Delicious: 90.000[12] Flickr: No data available Technorati: also about 27 million (if we assume that every actor has, on average, only one weblog)
Google suggestions	Spell checks	No information available
WordNet	Mapping synonyms, retrieving descriptions of terms, ancestors	27 semantic properties[13]
Swoogle	Finding related ontologies and annotations	More than 10,000 ontologies[14], though many of questionable maturity
Leo Dictionaries	Translation of terms	453,994 entries[15]

[8] http://en.wikipedia.org, retrieved on March 27, 2007
[9] http://technorati.com/weblog/2006/02/81.html , retrieved on April 3, 2007.
[10] http://www.techcrunch.com/2006/08/04/more-stats-on-delicious-this-time-positive/#comments, retrieved on April 2, 2007.
[11] http://technorati.com/weblog/2006/02/81.html , retrieved on April 3, 2007.
[12] http://www.pui.ch/phred/archives/2005/05/delicious-statistics-that-is-extrapolation.html, retrieved April 3, 2007.
[13] http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/#details, retrieved on April 2, 2007.
[14] http://swoogle.umbc.edu/, retrieved on March 29, 2007.
[15] http://dict.leo.org, retrieved on March 29, 2007

Of these resources, Google, Wikipedia, and Wordnet can be accessed either by APIs or straightforward screen-scraping techniques. For Leo, there is currently no API access supported.

5. Discussion and Conclusion

In this section, we compare our proposal to previous works, evaluate the added value, and identify future research steps.

Our work is closely related to [15]. In [15] the authors are presenting an approach to enrich tags with semantics in order to integrate folksonomies and the Semantic Web. Similarly to our approach, they are also using online lexical resources, ontologies and Semantic Web resources for amending the tags. Our approach extends this direction, since, first, we suggest deriving actual ontologies out of folksonomies, while [15] focuses on using existing resources and ontologies to map tags into concepts, properties or instances and determine the relations between these mapped tags. Second, we suggest to consider the varying resources not only as an isolated source helping in a single step of the tag processing but to channel all the social interaction manifested on the Web in such resources as the main input for automatically creating and maintaining domain ontologies. Third, we suggest to continuously involve human intelligence in the form of community approval of the resulting conceptualization, in order to confirm the semantics obtained from existing ontologies and resources.

Putting the community in the center of the ontology engineering process has already been proposed in [16] and [17], and other work on collaborative ontology engineering. In [16] and [17], the authors are generating a community-driven ontology based an ontology maturing process. This process contains the following steps: 1) community members are generating new ideas and related terminology through the tagging process, 2) the new tags and their concept definitions are presented and discussed in the whole community: everyone can change a definition, add synonyms etc., 3) the textual concept definitions created in the second phase, are formalized and hierarchical relations are added. In [17], axiomatization as a fourth phase is added to the process. During this step, additional semantics are added. In [17] two tools are presented that are based on this ontology maturing process. One of the discussed tools is using visualizations and another is using wiki technology to support the formation of consensus in the community. This approach differs from ours since they are not relying on existing resources for reuse. They are generating ontologies from scratch. In a nutshell, our approach aims at combining the strengths of [16, 17] and [15] and fully using a “mash-up” of available lexical, semantic, and social data sources for producing and maintaining domain ontologies

Acknowledgements: The work presented in this paper has been supported by the European Commission under the projects SUPER (FP6-026850) and MUSING (FP6-027097), and by the Austrian BMVIT/FFG under the FIT-IT project myOntology (Grant no. 812515/9284). Martin Hepp is also supported by a Young Researcher’s Grant (Nachwuchsförderung 2005-2006) from the Leopold-Franzens-Universität Innsbruck.

References

[1] Hepp, M.: *Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies*. IEEE Internet Computing, 2007. **11**(7): pp. 96-102.

[2] Hepp, M., D. Bachlechner, and K. Siorpaes: *Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements*, in: *Proceedings of the Workshop on Semantic Wikis at the ESWC2006 (ESWC2006)*, Budva, Montenegro, 2006.

[3] Vander Wal, T.: *Folksonomy*. Available from http://vanderwal.net/folksonomy.html, retrieved March 30, 2007.

[4] Golder, S. and B.A.Huberman: *Usage Patterns of Collaborative Tagging Systems*. Journal of Information Science, 2006. **32**(2): pp. 198–208.

[5] Gruber, T.: *Folksonomy of Ontology: A Mash-up of Apples and Oranges*. *First On-Line conference on Metadata and Semantics Research (MTSR2005).* Available from: http://tomgruber.org/writing/mtsr05-ontology-of-folksonomy.htm, retrieved March 30, 2007.

[6] Schmitz, P.: *Inducing Ontology from Flickr Tags*, in: Proceedings of the *Collaborative Web Tagging Workshop at the 15th WWW Conference (WWW2006)*, Edinburgh, Scotland, 2006.

[7] Heymann, P. and Hector Garcia-Molina: *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Stanford InfoLab Technical Report 2006-10.

[8] Mika, P.: *Ontologies Are Us: A Unified Model of Social Networks and Semantics*, in: *Proceedings of the 4th International Semantic Web Conference (ISWC2005).* LNCS 3729, Springer-Verlag, 2005.

[9] Wikipedia Foundation: *About Wikipedia*. Available from: http://en.wikipedia.org/wiki/Wikipedia:About#Contributing_to_Wikipedia, retrieved March 30, 2007.

[10] Ding, L., et al.: *Swoogle: A Search and Meta Data Engine for the Semantic Web*, in: *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM'04)*, Washington D.C., USA, 2004.

[11] Bailin S. and W. Truszkowski: *Ontology Negotiation between Agents Supporting Intelligent Information Management*, in: *Proceedings of the Workshop on Ontologies in Agent-based Systems* at the Fifth International Conference on Autonomous Agents (Agents 2001). Montreal, Canada,2001.

[12] Spyns, S., et al.: *From Folksologies to Ontologies: How the Twain Meet*, in: *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA and ODBASE (OTM2006)*, Montpellier, France: Springer, 2006.

[13] Giunchiglia, F., M. Marchese, and I. Zaihrayeu: *Towards a Theory of Formal Classification*, in: *Proceedings of the AAAI-05 Workshop on Contexts and Ontologies: Theory, Practice and Applications (C&O-2005)*, Pittsburgh, Pennsylvania, USA, 2005.

[14] Giunchiglia, F. and P. Shvaiko: *Semantic Matching.* The Knowledge Engineering Review, 2004. **18**(3): pp. 265-280.

[15] Stumme, G. and A. Maedche: *Ontology Merging for Federated Ontologies on the Semantic Web*, in: *Proceedings of the International Workshop for Foundations of Models for Information Integration (FMII2001)*, Viterbo, Italy, 2001.

[15] Specia, L. and E. Motta: *Integrating Folksonomies with the Semantic Web*, in: *Proceedings of the European Semantic Web Conference (ESWC 2007)*, Innsbruck, Austria: Springer, 2007.

[16] Maier, R and Schmidt, A.: *Characterizing Knowledge Maturing: A Conceptual Process Model for Integrating E-Learning and Knowledge Management***,** in: *Proceedings of the 4th Conference Professional Knowledge Management - Experiences and Visions (WM '07)*, Potsdam, Germany, 2007.

[17] Braun, S., et al.: *Ontology Maturing: A Collaborative Web 2.0 Approach to Ontology Engineering*, in: *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge (CKC) at the 16th International World Wide Web Conference (WWW 2007)*, Banff, Alberta, Canada, 2007.

[18] I. Ohmukai, Hamasaki, M., and Takeda, H. *A Proposal of Community-based Folksonomy with RDF Metadata*, in: *Proceedings of the Workshop on End User Semantic Web Interaction*, co-located with the Fourth International Semantic Web Conference (ISWC2005), Galway, Ireland, 2005.

[19] T. Coenen, et al.: *Knowledge Sharing over Social Networking Systems: Architecture, Usage Patterns and their Application*, in: *Proceedings of the OTM Workshops 2006 (OTM2006)*, LNCS 4277, pp. 189–198, Berlin, Heidelberg, Springer-Verlag, 2006.