Social Tagging/indexing

Jeroen Hoppenbrouwers, Vrije Universiteit Brussel
<jhoppenb@vub.ac.be>

Workshop at the 31th ELAG conference, Barcelona, May 9-11, 2007

This document is meant to be a starting point for discussion during the workshop. Actual workshop targets will be decided on during the first part of the session. It is intended to be a true workshop, so participants will be asked to actively approach the subject and work together on the final results and presentation. However, nobody will be assumed to be an a priori specialist on the topic, and this includes the author of this introduction.

Asking end users to classify content and generate metadata within online knowledge sharing systems by using a predefined, fixed taxonomy can improve the findability of content, but it has two main problem areas:

1. The taxonomy or metadata structure (thesaurus) may be too rigid to support user needs or too complex to be easily learned; this leads to frustration and abandon.
2. The overheads of classification are borne by the user, but the group reaps the benefits; this leads to demotivation "to just work for others and getting nothing back in return."

A possible solution to these problems might be the introduction of *social tagging*, *folksonomies* or other forms of far less restricted, simple, and uncontrolled classification mechanisms which are all based on the principle of allowing each end user to add random words or very small phrases to any content in the system. Before the large-scale introduction of computer and network technology, this was practically impossible. But today's (inter)networked world offers plenty of opportunities for social tagging, which together are often dubbed *Web 2.0*.

The obvious disadvantage of this approach is its lack of precision (synonym/antonym control, related terms, context, etc.), but in many practical usage scenarios the trade-off between simplicity and precision can make sense.

The following is an adaptation from Wikipedia

> A folksonomy is a user-generated taxonomy used to categorize and retrieve content using open-ended labels called *tags*. The process of folksonomic tagging is intended to make a body of information increasingly easy to search, discover, and navigate over time. A well-developed folksonomy is ideally accessible as a shared vocabulary that is both originated by, familiar, and useful to its primary user community.

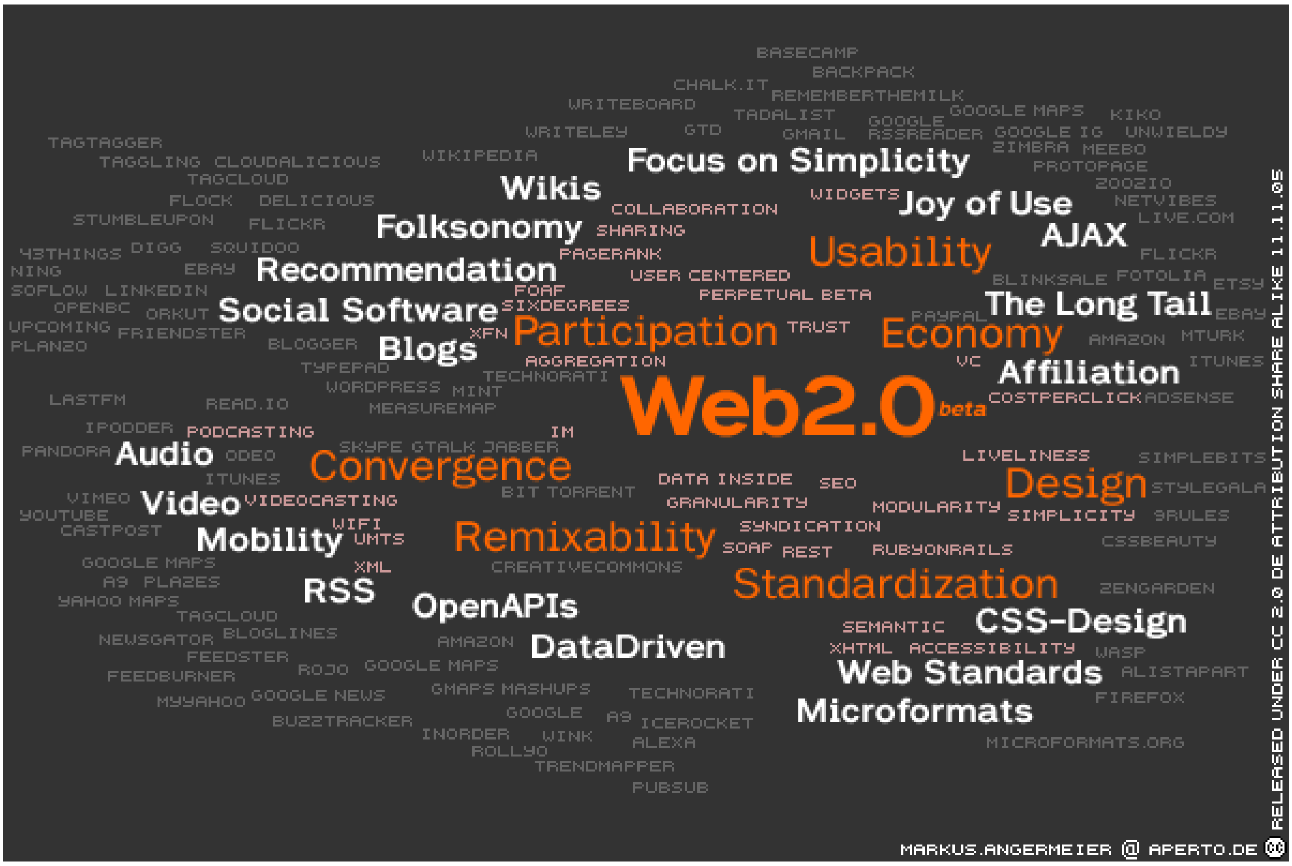
The following is an adaptation from Del.icio.us

> Tags are one-word descriptors that you can assign to content to help you organize, remember, and retrieve it. Tags are a little bit like keywords, but they're chosen by you, and they do not form a hierarchy. You can assign as many tags as you like and rename or delete the tags later. So, tagging can be a lot easier and more flexible than fitting your information into preconceived categories or folders.

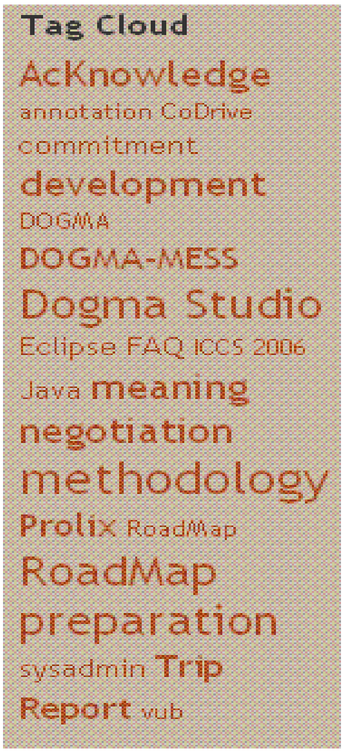
> For example, if you have an article about how to make a certain kind of cake, you can tag it with *recipes sweets yogurt* or whatever other tags you might use to find it again. You don't have to rely on the designer of a system to provide you with a category for French cake recipes. You make up tags as you need them, and use the tags that make the most sense to you.
>
> This is great for organizing and finding personal data, but it goes even further when someone else tags related content using the same tags. You begin building a collaborative repository of related information, driven by personal interests and creative organization.
>
> There are no wrong tags.

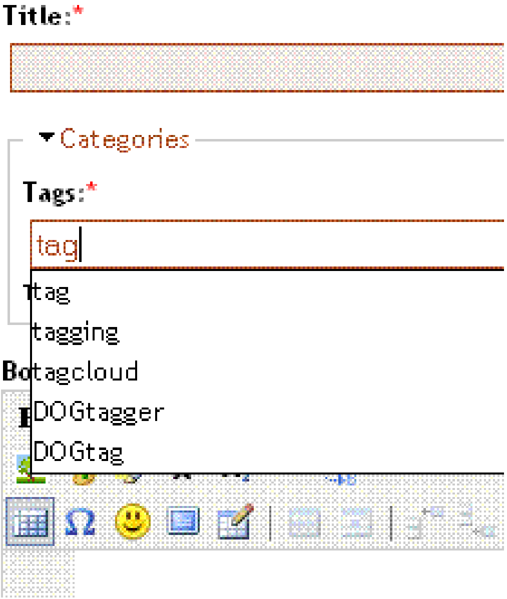
A clear difference between the two descriptions of 'tagging' is that folksonomies use the notion of a taxonomy, whereas pure tagging systems do not have any explicit relationship between tags at all. Nonetheless, even pure tagging systems can supply users with relationships between tags, by using the tagged content as implicit relationship: if many content objects get tagged with the same pair (group) of tags, there should be some relationship between these tags, though it often cannot be easily verbalised.

Other common ways of using the tags to display some global structure in the content organisation are the *tag cloud* and the immediate display of orthographically related tags (e.g., tags starting with the same characters) during the typing of new tags.



Tag clouds come in many forms, from mainly graphical to purely textual, and can be organised in many different ways (semantically, popularity, alphabetically...). Especially the many possible ways to visualise tag clouds, some interactive and 'living,' have received a lot of attention. Sometimes however, the graphical presentation appeared to be considered more important than the added value or the actual tag content.




Title:*

Tags:* tag

Immediately showing related existing tags while a user enters a new one can contribute to a more coherent way of adding tags to content, as typical spelling variations including morphological variants (tag, tags) can be caught early and users might be tempted to reuse more specific existing tags instead of adding the one they had in mind. Changing tags later on usually is possible, which also opens the way to tag merging and consolidation.

Another effect of folksonomies or social tagging, which comes directly from the fact that there is no central vocabulary that, by necessity, is limited to 'most popular items,' is that many niche areas can get proper coverage as well. As the costs to create a new tag are near-zero, it becomes feasible to freely add many tags to many content objects. This leads to an effect dubbed *The Long Tail* in recent literature: a few items may be highly polular, but the vast majority of the collection is not popular -- yet the total access of these items exceeds the total access of the popular items.

The combination of social tagging and formal, centralised taxonomy building is also feasible. The less popular items come by with end-user tagging, while the more popular tags slowly get moulded into a nice framework of relationships. This needs serious work, so resource prioritisation is mandatory.

Efforts to get more semantics out of people

Many of the existing semantic annotation systems that have been in use reveal that people seem to have a preference to organise tags (or anything) into taxonomies, or type hierarchies. There is nothing against this, but often the type hierarchy becomes the target and end point of the effort. Relationships between elements of the semantic collection, or *ontology*, do not always get the attention they deserve. It compares a bit to the classical thesaurus relationships: 'broader' and 'narrower' are well-used, and nearly all the rest is lumped into 'related'.

In semantics research (which encompasses a large set of disciplines, from philosophy via logic and linguistics to computer science and artificial intelligence), relationships beyond 'is a' have had a prominent place for a long time. Although some kind of common set of basic relationships has been emerging, including 'part of' and other well-known but sometimes cryptic or very theoretical relationships, the majority of useful relationships between concepts needs to be built and named by the ontology constuctors, just as the concepts themselves. In this view, concept and relationship go hand in hand.

The down side of this is that relationships are very rarely associated with content, unlike concepts. It is unlikely that end users who focus on content and content tagging will ever get seriously involved with ontology editing at the relationship-between-tags level. This does not mean that these efforts should be abandoned. Relationships between tags (beyond common usage by content objects) are essential to produce meaningful tag clouds and associative hints when searching for content. This could point to a natural distinction between the end user involvement (tags) and the professional/community involvement (relationships between tags).

Community-driven semantics

At STARLab, Vrije Universiteit Brussel, research on community-driven semantics has produced an approach which leverages the shared interest of communities in establishing shared meaning about shared concepts. Although this methodology originally focused on typical ICT-centred domain description models, it has been expanding into more general applications. One of the possible situations to which this methodology, called DOGMA-MESS,[1] can be applied is the descriptive ontology construction and maintenance required to successfully enable semantic searching and indexing on collections of content objects.

The following is an extract out of the original description of the DOGMA-MESS methodology

> In DOGMA-MESS, there are three user roles; (1) Knowledge Engineer, (2) Core Domain Expert and (3) Domain Expert. The task of the Knowledge Engineer is to assist the (Core) Domain Experts in their tasks. The major chunk of knowledge is captured by the Domain Experts themselves. The Core Domain Expert builds high-level templates in the so-called Upper Common Ontology. The Domain Experts specialize these templates to reflect the perspective of their organization in their Organizational Ontologies. The Domain Experts are shielded from complexity issues by assigning specific tasks in the elicitation process. In every iteration of the process, common semantics are captured in the Lower Common Ontology whilst organizational differences are kept in the Organizational Ontologies. Information in the Lower Common Ontology is distilled from both the Upper Common Ontology and the Organizational Ontologies using meaning negotiation between (Core) Domain Experts. The Lower Common Ontology is then used as input for future versions in the process.

It is a challenge to extrapolate the DOGMA-MESS approach to library situations. Not all recognised roles immediately map to standard library roles, such as a normal library user. It is not common for library users to actively leave tags or other annotations in the system while browsing for literature. However, with increasingly online library access and most systems moving to a web-based technology, such interactivity may become fully accepted, especially with the growing amounts of users familiar with 'Web 2.0' sites where social tagging is the norm. Add a community of people which has an established interest in a given domain, and the core domain experts might be available. The ontology engineers then can be the current library staff doing the indexing work and maintaining the subject heading languages and other vocabularies used for indexing.

[1] DOGMA: Developing Ontology-Grounded Methods for Applications
MESS: Meaning Evolution Support System

Towards accepted ontologies

DOGMA-MESS's Upper Common Ontology usually aims to span as large a group of organisations or end users as possible within a given domain. As such, it may become a prime candidate for formal standardisation. This does not mean that it will replace the Lower Common Ontologies or even the Organizational Ontologies -- far from this! But it may serve as the standard reference for other ontologies, enabling them to be semantically linked up and becoming interoperable.

It is a known fact that high-level, abstract ontologies span the widest range while specialised, low-level ontologies are the most useful. Driving towards one Master Ontology therefore seems a bad idea. DOGMA-MESS might open the door to a mesh of ontologies that leverages both the high-level standardised ontology and the low-level specialised ontologies, without restricting people in going ahead and tagging along.

Existing systems which link up subject heading languages, such as MACS,[2] may form the infrastructure under such a community-approach to establish shared semantics. Currently, MACS does not provide for explicit meaning negotiation support, but the framework is there, and the national libraries of Germany, France, the UK, and Switzerland have been actively cross-linking their subject heading languages for a few years. With a more explicit meaning negotiation layer on top of a system such as MACS, and a broader community to feed the ontology, we might be able to leverage all existing systems based on the subject heading languages.

It would be a great day for the library world if they can use their century-long experience in semantic annotation to do what the scientific world has until now not yet been able to do: to actually create the Semantic Web, instead of only building tools to create and use it if it were there.

References and Suggested Further Reading

http://en.wikipedia.org/wiki/Folksonomy
http://dublincore.org/groups/social-tagging/
http://del.icio.us/help/tags
http://www.personalinfocloud.com/2005/02/explaining_and_.html

DOGMA-MESS background: http://starlab.vub.ac.be/website/biblio

MACS project: http://macs.cenl.org/

Scott Golder, Bernardo A. Huberman (2005): “The Structure of Collaborative Tagging Systems”.
PDF available from http://arxiv.org/abs/cs.DL/0508082
Thomas van der Wal (2007): “Tagging That Works”.
PDF available from http://www.vanderwal.net/random/category.php?cat=153

[2] MACS: Multilingual ACcess to Subjects