Folksonomy: the New Way to Serendipity

Nicolas AURAY
Ecole Nationale Supérieure des Télécommunications, Paris

Abstract: Folksonomy expands the collaborative process by allowing contributors to index content. It rests on three powerful properties: the absence of a prior taxonomy, multi-indexation and the absence of thesaurus. It concerns a more exploratory search than an entry in a search engine. Its original relationship-based structure (the three-way relationship between users, content and tags) means that folksonomy allows various modalities of curious explorations: a cultural exploration and a social exploration. The paper has two goals. Firstly, it tries to draw a general picture of the various folksonomy websites. Secundly, since labelling lacks any standardisation, folksonomies are often under threat of invasion by noise. This paper consequently tries to explore the different possible ways of regulating the self-generated indexation process.
Key words: taxonomy, indexation, innovation and user-created content.

On December 17th 2006 Time chose *you* as personality of the year. *You*, that is, you the internet user in the sense that you contributed to the history of community and collaboration on a greater scale than ever before. This is the symbol of a turning point in the information society, marked by an ongoing convergence between the professional and amateur world.

From one point of view the domains where the amateur productions begin to rival the level of the content of professional and institutional productions are on the increase: public corporations recognise that the blog and its anonymous aficionados contain more information than the 'official' medium produced by their distributors or agents, (BEAUDOIN & LICOPPE, 2002), or created by their political party (BEAUVALLET, 2007); while new artists are appearing who first earned reputations in the blogosphere (CARDON & DELAUNAY, 2006). This wealth of amateur content represents a generalisation throughout all cultural life of the phenomenon known as the 'democratisation of innovation' (von HIPPEL, 2006); with amateurs injecting radical innovations into that realm undreamt of by engineers and whose existence specialists had previously primarily observed in cutting edge areas confined to IT, biotech or chemistry.

There are structural causes behind this massive and remarkable appearance of amateurs on the cultural stage by means of the internet. They include the significant increase in cultural capital on the part of internet uses, supported by upgraded education, the generalisation of intellectual work and the spread of the use of the internet among white-collar workers. Thus LEADBEATER & MILLER (2004) explain the development of the 'professionalized amateurs' (labelled as "pro-ams") by the generalised articulation we are now seeing of the vocational model of voluntary leisure and creative investment in paid work. This movement, backed by sustainable causes, seems to be further supported by the growing interest shown by the economic world in stimulating and organising this enormous business of the decentralised production of content. Apart from the sporadic attempts made by some companies to use amateur content (for the Netscape example, see AURAY, 2000; on crowdsourcing in general, see HOWE, 2006), what we observe is a continuous reprocessing of amateur content by the cultural industries (as JENKINS, 2006 has shown in the sector on video games), or the development of a musical remix culture that endows objects, which have been paired with others or removed from their original environment, with greater fame than the original products.

This huge inflow of amateur content has overturned the organisation of data on the internet. It multiplies referencing problems. The volatility and novelty of content is giving rise to congestion in search engines that are poorly suited to seeking non-textual data, blog commentaries or forum posts. But at a broader level, in line with a principle often observed by science sociologists, and which Merton has dubbed the "Saint Matthew effect", search engines depend on algorithms that favour content, which has already been massively quoted; they therefore focus all their attention on a limited number of sites that match the tastes of the majority. It turns out, however, that a major feature of the growing power of amateurs is the emergence of a new distribution model characterised by a 'long train' (ANDERSON, 2004). Demand is spread out over a large number of products. Products previously reserved for niche markets may see-saw into mass popularity, driven by word-of-mouth and web-user recommendations. Distribution circuits are pluralistic and horizontal, with growth in the number of cultural middle-men, where the judgment of the public and the commentary of web-user communities are added to the verdict of critical opinion and promotion appearing on the mass media. This gives rise to a requirement expressed by various web-users, ill-served by search engines, for a finer referencing of the complex space of cultural tastes based on the multiplication of subcultures and niches.

Definition of 'folksonomy'

To provide a remedy for this problem, users have come up with the idea of importing the spirit of collaboration by allowing the users themselves to create and share their key words. The term 'folksonomy' refers to these forms of self-generated references [1]. It is usually attributed to Thomas van der Wal, an information architect and senior consultant at Infocloud. The basis of the word is 'folk', meaning people in general and taxonomy, from the Greek taxis (arrangement, name, law, used here in the sense of systematic classification). Leaving their differences aside, folksonomies display three marked characteristics:

They depend on 'popular' self-indexation, in the sense that it is web users themselves, the contributors or readers of contents, who label them by allocating key words to them. Unlike classical classification systems, such as the Dewey universal classification, contributors to a folksonomy are not bound by a pre-defined terminology, but can adopt whatever terms they like to classify their resources. These terms are often known as key words or tags.

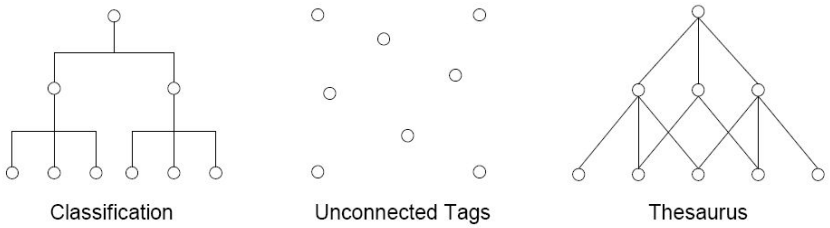
While in classical classification theory a taxonomy designates systems whereby categories are connected to each other by the inclusion of classes, the more inclusive a category is in a taxonomy, the higher its level of abstraction, folksonomies depend on multi-indexation – the same item may appear under a number of key words. This leads to a flexibility of the hierarchical relationship between the taxons: the same item may be placed in a subordinate category or terminal (which has no sub-category) and throughout the length of the category tree. The picture of a Siamese cat, for example, may be labelled under the key words Siamese and cat, but also under feline or animal, which is the generic category. But folksonomy also allows an object to be labelled by its characteristic attributes or features, and not by the categories to which it belongs. A photograph of a swallow is as likely to be labelled wings or feathers, as bird. Unlike, say, the Linnaean system for classifying animals, which is an exemplary system, folksonomies are more horizontal, labelling systems organised on a grid and open to characteristics that are known in linguistics as sememes.

Folksonomy is based on a relaxation of the relationships between the term and the index. VOSS (2005) has also thrown light on the difference

[1] Some writers also use the terms "potonomy" and popular "taxonomy".

between this and the thesaurus categorisation system. The thesaurus, in the sense in which it was developed by Peter Luhn in 1957, is a tagging system 'controlled' by the fact that the a priori relationships between terms in the index are rendered explicit by the fact that they are connected by three types of relationships – equivalence , hierarchy and association. Folksonomy releases this type of restriction between the indices of the items, by rendering them uncorrelated or non-connected.

Fig. 1 – "Structure of indexing systems" (from VOSS, 2006)



The first shared sites which used the principle of 'popular taxonomies' were del.icio.us , a shared site of web page bookmarks, set up in 2001, and FlickR, a shared photo site, created in 2000. Reaching a critical mass seems to have been the essential parameter for the success of these sites. They attracted their first users by offering them tools to help them store individual content. In other words, del.icio.us and FlickR were successful (GUY & TONKIN, 2006) because it was easy to enter tags on them (APIs on the browsers allowed exportation onto the site), and particularly because they offered an easy way for users to check their photographs or bookmarks from anywhere. What attracted the first users was more like a raft of 'selfish' motivations (GUY & TONKIN, 2006), such as simplified access to their own files [2].

These taxonomy systems (in the broadest sense) then became applied, via self-generated indexation, to other types of content:

[2] As evidence of their success, clones of each of these sites developed, shown by the numerous offspring spawned by *delicious*, of which the most well-known, del.irio.us, an open source version of del.icio.us, has not survived. But others have, such as BlogMarks, a similar service but restricted to French language content, or Connotea, dedicated to a precise subject area, in this case, that of science.

Blogs

Technorati, one the of the largest blog indexation engines has offered the provision in its pages of tags connected to content since January 2005. By August 2005, there were no less than 25 million labelled tags, and this figure has now risen to 34 million, with 12,000 new tags appearing every day.

Current events

Slashdot, a technical/scientific current events site, was the first to initiate the trend in this type of content. It produced a number of clones, of which the largest were Newsvine, offering current events in the form of a blend of professional sources (Associated Press) and web-user publications, and Ohmynews!, a current events site with 41,000 'citizen' contributors over six years, and where 30% of the site is edited by an editorial team.

Human skills

The Tagalag site allows for the tagging of individuals referenced by their names and geographical location or the terms that best characterise them. The service, which also exists in a beta version, would make it possible to find, for example, animated film fans in Manhattan. With the same type of content, the http://www.43things.com site, inspired by a local exchange system, makes it possible to establish relationships between people with an offer of 43 things to do, and people who either want to discover what the others want to do, or if they are in the same area, want some help on a reciprocal basis thanks to self-produced tags.

Image and sound

Music has been added to the sharing of videos (dailymotion and YouTube) and photographic images. In 2006, Yahoo launched a podcast directory, which also included tags, which tended to demonstrate that all new services had to include this browser principle from then on.

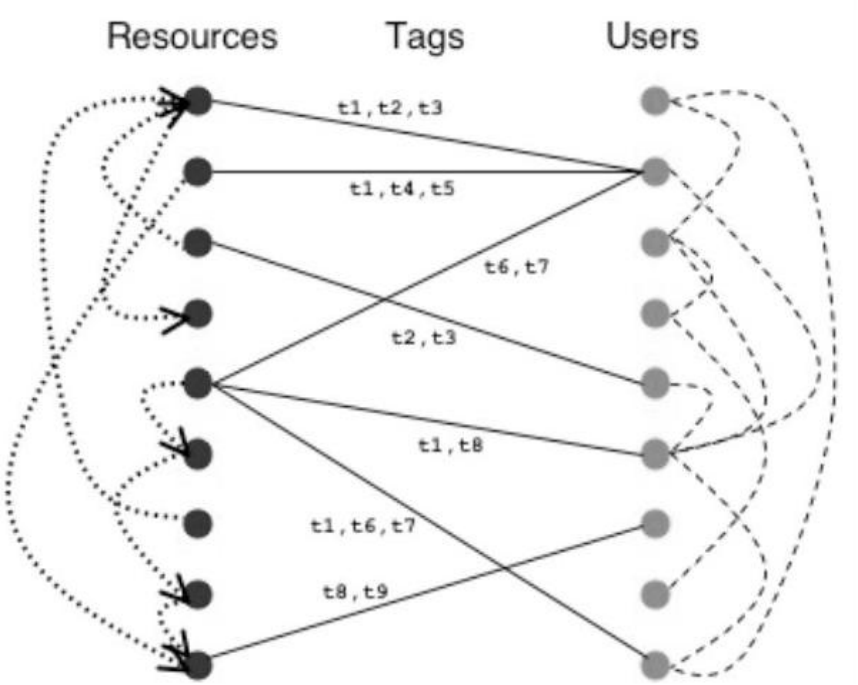
Folksonomy can therefore be seen as a fundamental movement that expands the collaborative process, by granting users the power to index content, hence returning a certain degree of control to users. It also represents a somewhat paradoxical phenomenon: on the one hand users, customers or end users are increasingly concerned with being located at the heart of the system as individuals, while striving to register themselves within a communal or network initiative on the other. Even so, as a means whereby co-operative tagging can be set up, folksonomy rests on very precise and

specific methods. Three powerful properties can be discerned: the absence of a prior taxonomy, multi-indexation and the absence of thesaurus. These features are not, however, unique to the system. BEAUVISAGE (2005) showed that some internet directories had already opted for the structural choice of multi-indexation prior to the emergence of folksonomies.

Folksonomies are polarized and focus on two social objectives

An original aspect of the relationship structure of folksonomy site is the fact that it is three-way: there are the resources (the content), tags and users.

Fig.2 - A model of tagging system



The specific nature of this structure as resting on a number of bases has given rise to the drawing of two-way graphics to code the data. Kleinberg, for example, has grasped the way in which the social graphic of writers and the social graphic of content are connected (*via* links). CHAKRABARTI *et al.* (1998) have also tacked this process, by making a distinction between writers, content and anchor text.

The opposition between broad and narrow folksonomies

Thomas van der WAL (2005) distinguishes two types of folksonomies, the broad and the narrow. This distinction is important. The first category, of which del.icio.us is an example, is characterised by the breadth of the number of tags authorised to reference an item, and the fact that the user may re-allocate an already issued tag to existing content. The second category, of which FlickR is an example, is characterised by a limitation in the number of tags used to reference an item. This limitation starts from a restriction in the number of users who have the right to apply a label, and moves to a quota system for the number of tags that can be entered. Each system has its advantages and disadvantages. 'Broad' folksonomy, for example, has the plus that it is possible to view the various ways in which people describe a collection of shared content, and thus of identifying emerging vocabularies and trends. An indirect consequence of this advantage is that 'broad' folksonomy allows a population to be broken down into affinity groups according to the descriptive vocabulary they use; this is a convenient tool for matching similar individuals, that is, people who have similar systems for perceiving and categorising the world. But the downside of 'broad' folksonomy is the dispersal of the description of the object into a large number of different key words; it becomes harder to find a specific piece of content. 'Narrow' folksonomy, however, because it is less casual as regards the attachment of a key word to a piece of content, reveals its strength in finding precise content from a key word search. It is particularly useful when it comes to building databases on content, which cannot be easily found by text-based searches using the standard tools. An indirect advantage is that it allows for the grouping of content on a basis of the co-occurrence of key words within the groups by ascending classification methods [3]. A high level of importance is attributed to grouping by FlickR, for example, allowing photographs with a similar content to be tracked down by ascending classification [4].

This central opposition suggests that there is a distinction between social usages. Vander Wal emphasises the fact that broad folksonomy systems are based, above all, on the placing of importance on social grouping properties, where connections between individuals are achieved on a basis of breaking

[3] It is, of course, possible to proceed in the same way in 'broad' folksonomy systems, but the results do not have the same level of relevance.

[4] The grouping would in particular appear to be the attraction of the killer interactive environment which Ludicorp has created for FlickR.

down the population into groups with similar perception concepts. These types of folksonomy, resting as they do on social networks, can only classify information and share it; they put together users who share the same centres of interest. The indexer-user becomes in turn himself indexed to a certain degree and placed in relationship with other key words. Narrow folksonomy systems, on the other hand, are above all focussed on the development of the properties of the exploration of the corpus, little by little, thanks to the relevance of the descriptors used [5].

More generally, the opposition between the two systems of folksonomy emphasises, in an underlying way, a feature common to both – the re-evaluation of exploratory usages. This opposition therefore reveals two different modalities of these curious usages: the first corresponding to cultural curiosity, the second to social curiosity.

The assumption of cultural curiosity

The folksonomy would be a solution for a more exploratory web search than that using a search engine: the little by little quest. It is a more random way of exploring the blackboard. This gives rise to the theory behind Adam MATHERS' work (2004): "Browsing versus Finding," namely that there is a fundamental difference between direct searching with a query and browsing to find interesting content. The primary virtue of folksonomy is '**serendipity**' (in the sense of lucky chance). This is a solution that encourages browsing, and, via a collection of interlinked tags, constitutes a fantastic source for identifying unexpected finds, which would never have been revealed without it. It is the same difference between exploring a problem space to formulate questions and seeking effective answers to precisely formulated questions. On this point it is not obvious whether the use of folksonomy sites is a response to an explicit prior preference for exploration, or if, on the contrary, folksonomy has provided the occasion and influenced the development of exploratory uses of the web without it having been planned in advance (lucky chance). What is certain, however, is that folksonomy sites allow a more open, more random, exploration of content than the use of a search

[5] By way of an extension of Vander Wal's argument, certain interpreters have forced the opposition: thus, LE DEUFF (2006) infers from Vander Wal's dichotomy a different distinction, namely that narrow folksonomies, focused on content, would be primarily used in an individual objective (to describe them he uses the term personomy), while the broad, focused on other users, would favour the collective and collaborative aspect. This extremist dichotomy is open to doubt.

engine. In this way, folksonomy music sites will show you the playlists of those people who like the music you listen to, which leads us to constantly vary our music. The Pandora site boasts that it allows web users to "create a random radio", by discovering artists by making use of the users' playlists.

Figure 3 - the Pandora Music website

Similarly, del.icio.us mentions the importance of assistance in exploration, in curious sifting. Visual aids make it possible to browse from tag to tag by navigating around the graphic of co-occurring tags such as "Revealicious", a tool developed by Sébastien Pierre de Ivy (software architecture and design); hublog, a tool developed by Alf Eaton (see EATON, 2005). More sophisticated plug-ins make it possible to browse from a tag to all the people who tagged the tag [6]. In order for this to be possible, all folksonomy sites depend on mechanisms, which reinforce the relevance of this cultural browsing: the uniqueness of the tagging process and collaborative filtering (MALTZ, 2005).

On the one hand, in order not to crush the diversity of content under a small number of dominant categories used to index them, folksonomy sites strive to maintain diversity in labelling as an objective. All categorisation rests on deciding between two limitations: (see ROSS, 1978 for a more detailed presentation): providing the maximum of information with the

6 http://www.hyperorg.com/blogger/mtarchive/003702.html#comments

smallest cognitive effort; and staying as close as possible to the perceived world so that the categories represent (map) the structure of the actual world as accurately as possible. It would appear that in navigating between these two limitations, folksonomy sites place the cursor closer to the second limitation, by increasing the number of categories. This is made clear on del.icio.us.

What makes del.icio.us powerful is the **uniqueness of the tagging process**: to retain the contextual specificity of the tagging, the tagger is not offered all the tags available for tagging the same context, but only the intersection between the tags already issued and his own collection of tags (the term used to designate this intersection is "Recommended Tags"). This encourages the tagger to reuse his own tags; this favours a large number of categories. It has also been deemed fundamental by del.icio.us that the possibility of personalised and potentially idiosyncratic tagging behaviour be allowed for and encouraged. The Recommended Tags in del.icio.us thus replace the traditional compromise in classification theory between the need for simplifying harmonisations and truth to reality. By retaining a contextual and idiosyncratic anchorage for the tagging activity, the somewhat mysterious process of "emergence" has been relegated: as the site's founder noted, "classification will emerge from the totality of the tags, which have been imposed by the users."

The second characteristic is **collaborative filtering** (MALTZ, 2005). Instead of receiving a random guide to pages by simple co-occurrences calculated on the whole of the panel, the user receives first the content contributed by other users with a similar cultural profile. Thus, on the bookmark sharing site del.icio.us, the problem of tag distribution has been taken into account in accordance with a law of power. A user, Kiddphunk, has therefore set up a plug-in called "delicious discover" [7], which is designed to facilitate the connecting of the user with users who favour the same content, i.e., who share with him not popular links (links which many users have), but rare links. At the outer limit, his tool allows connection to be made with elective affinities, that is, the totality of the users who are the only ones who share precisely this link with him. In order to achieve this, a weighting system is used, according to rarity, for the content which each user has in common with him. His method aims to highlight, for each user, the type of content, known as the 'sweet spot', which is both widely unpopular yet really matter to him.

7 http://www.mandalabrot.net/delicious/

Social exploration: the discovery of the other and the meeting

Collaborative sharing does not have a lucid acceptation of cultural curiosity as its only interest. It sometimes spills over into a number of functions leading to connection with other individuals. This is sometimes explicit on certain folksonomy sites dedicated to meeting up. Hence, **http://www.43things.com** offers, on the basis of keywords entered in turn by skills seekers and suppliers (see above), to create a list of 43 things to do in your life, to discover what others want to do, to be in the same geographical area and to help each other. From this point of view taxonomy is an icebreaker (Brown), a support in social exploration, in the sense that it provides a filter, so that users can be sure of finding themselves in a relationship pre-defined on the basis of a reliable cultural profile. However, this possibility of meeting exists in more or less extensive ways on all the sites: on FlickR, users can associate themselves with interest groups; they can thus invite friends to see their private photographs. The link is not reciprocal: inviting a friend to view your photos that does not mean you can, in turn, access their private photographs.

The possibility of mutual contact is thus favoured by the alert device. We have just seen how, with FlickR, it is possible to subscribe to the updated collections of other users (only by invitation): this means that a user can keep up to speed with someone else's photographic activity. This is also the model used by Pandora, the music listening site:

> "As you listen (with iTunes, Winamp, Windows Media Player, or others), your tracks automatically appear in your online musical profile – we call this 'scrobbling'. Explore custom recommendations and personalised radio, find your musical soulmates, discuss your favourite bands, and share your musical insights with friends, family, and the world!"

These devices for making social exploration easier sometimes take on a comical appearance. For example, the creation of "Chinese portraits" of users has been implemented on del.icio.us, where they are identified by means of a mosaic of images. Extisp.icio.us [8] (from the Latin word for divination by the inspection of entrails) allows you to view a mosaic of images drawn from Yahoo images on the basis of tag words used several times by a del.icio.us user.

8 http://kevan.org/extispicious

As an extension, to facilitate meetings between users, certain folksonomy sites have set up a fun modalisation of their functions. Certain folksonomy sites are explicitly based on "modalisation" (in the sense used by GOFFMAN, 1995) in the form of riddles of the labelling experience: the idea is not to apply tags to a resource, but to guess the tags which have been applied by someone else. For instance, Luis von Ahn developed a game [9] in which two people are simultaneously given an image of the same picture, with no way to communicate. The game has been licensed by Google in the form of the Google Image Labeler. When the user's label matches the partner's label, both will earn points and move on to the next image until the 90-second period runs out. The game keeps the high scores of registered users and these are displayed both for the day and for "all time"; Google is betting on users' competitiveness to rack up high scores to swell the number of images ranked. It is a clever way for Google to build an accurate database used when using the image search. Sometimes more sophisticated forms of fun framing of the relationship exist: on a site like Odeo, a podcast site, users face each other in the framework of a competition in which they try to build sentences in a cloud of tags, by shifting the tags and adding others and by changing the general order.

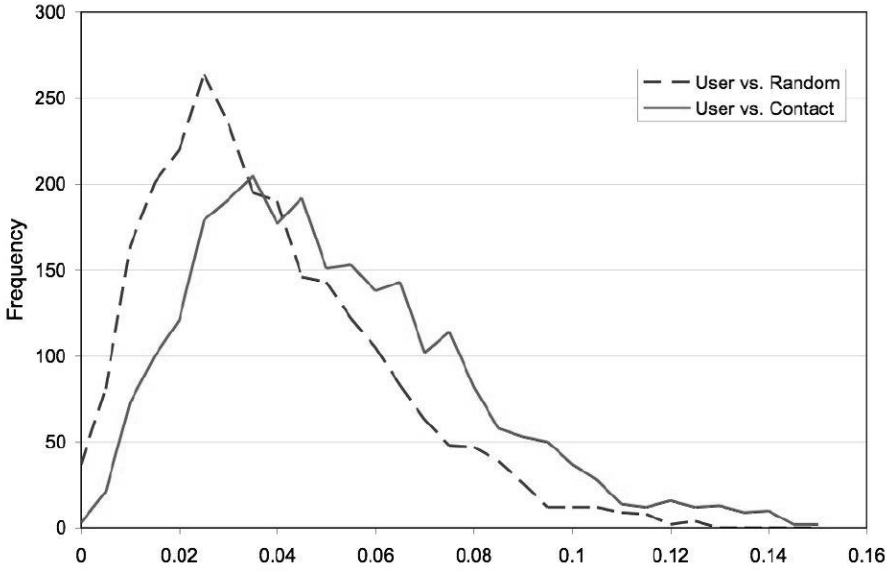
The articulation between social connections and semantic proximity

A number of studies have striven towards a better understanding of the articulation between social affiliations and the creation of a new vocabulary. For example, Marlow and Naarman chose at random 2,500 of the most intensive users on FlickR (those who had sent a number of over 100 tags). Using this sample of users, they tried to test out whether the fact of having contacts (they worked on contacts and not on membership of a group) influences the process of the formation of a vocabulary. In order to do this, they attempted to calculate the rate of shared tags (the number of common tags/the sum of their two groups of tags) between two users, a rate which could be called "overlap". They compared the overlap between two users chosen at random from this list, and the overlap between two users in contact. The overlap between two users in contact is higher than between two users chosen at random. Other observations can be made: the frequency of considerable overlap is significantly higher for users in contact

9 http://www.espgame.org/

than for two users chosen at random. There is therefore a relationship between social affiliation and the formation of the vocabulary of the tags.

Figure 4 – Vocabulary overlap distribution for random users and contacts (n=2500)



This type of study leads to the question of whether membership of a group of contacts is correlated to the existence of a common "sociolect". It again throws doubt on drawing too clear-cut a distinction between cultural exploration and social exploration. It would appear that cultural proximity is an important factor in social affiliation. Some additional studies would be appreciated, particularly to test whether it is pure resemblance which is a socially connective factor, or if, on the contrary, a small degree of dissimilarity helps guarantee a relationship. What would be a satisfactory level of dissimilarity supporting the best possibility of a sustainable relationship? From a broader viewpoint, diachronic studies would be needed to provide a better understanding of the direction of the influence between social relationship-establishing phenomena and cultural exploration phenomena: is it merely the case that an overlap between semantic proximity and social affiliation simply bears out the fact that it is easier to connect up with people who share your tastes? Would it be impossible to reveal a connection between a sustainable relationship and the diversification of the partners' semantic fields?

Even if they remain focussed on the provision of a variety of cultural and social exploration functions, folksonomies have more recently witnessed the development of unexpected functions. Aside from the diversity of their exploratory uses, a third type of unforeseen use appears to have established itself, deriving from users: the opinion poll. Thus Technorati displays as a matter of course the 250 most popular tags of the moment. Tag aggregators like Guten Tag [10] reveal the popularity of the words used as descriptors on other services. Using them, it is possible to calculate statistics on the popularity of tags deriving from a range of different sites. Del.icio.us displays the most popular sites of the day, or, *via* connected services (LiveMarks), the list of novelties published on del.icio.us. From a broader viewpoint, with viewing tools such as cloud of tags, the majority of these sites make it possible to show in a flash what a given grouping is thinking about at a given moment. The larger the tag, the more it is likely to mean that a large number of users have entered the tag. The folksonomy sites are therefore also indicators of opinion. However, there is a lack of reflexion on the amount of past time to consider so as to include an item in "dominant cultural climate." The difference between a transitory mode and a value is unclear in this counting of tags, so that we disagree with the aim of some folksonomies websites to qualify their results as a genius seculi, or to label their "cloud of most popular tags in the last week" with the hegelian word "Zeitgeist".

The governance of self-generated taxonomies

The folksonomy is often described in the literature as 'feral', that is, uncontrolled: WALKER, 2004; MARLOW *et al*., 2005. The main characteristic of these works is to highlight the emergency of taxonomies on the basis of the unique and unchecked entry of each user. A number of studies, which have often given rise to implementation in websites, have striven towards remedies with a view to the governance of self-generated taxonomies. They begin by identifying the main problems, then move on to the solutions intended to solve them.

Folksonomy suffers first and foremost from basic problems of polysemy and homonym management [11]: does a page indexed by 'java' refer to a style

---

[10] http://gutentag.viabloga.com/

[11] It is not possible to talk about spelling problems or a lack of coherence in the choice made by users in the use of concatenations (some use a dash, others a slash, others nothing at all) as it

of music, a programming language, or a town in Wyoming? Is a page bearing the tag 'glass' talking about the material, an object, or the expression 'to have a glass'? These homonymy problems are aggravated by the fact that the key words are often mixed up, as on Technorati. Folksonomy also suffers from the problem of synonymy. Multilingualism makes both problems worse: if a folksonomy site sees itself as international, not only does the number of tags explode because of the different translations, but at the same time very puzzling homonyms appear – the word pain means 'bread' in French! Does the key word 'cap', a cape in French, refer to the most beautiful geographical capes in the world, or to the English headgear? Language mixing, practised by Technorati, FlickR or del.icio.us, gives rise to a major problem, which can be regulated by adaptive behaviour on the part of the users. Moreover, will French web users, faced with the predominance of English terms in their indexes, be tempted to use English-language markers to avoid the problems of multilingualism, thus reinforcing the predominance of descriptors in English?

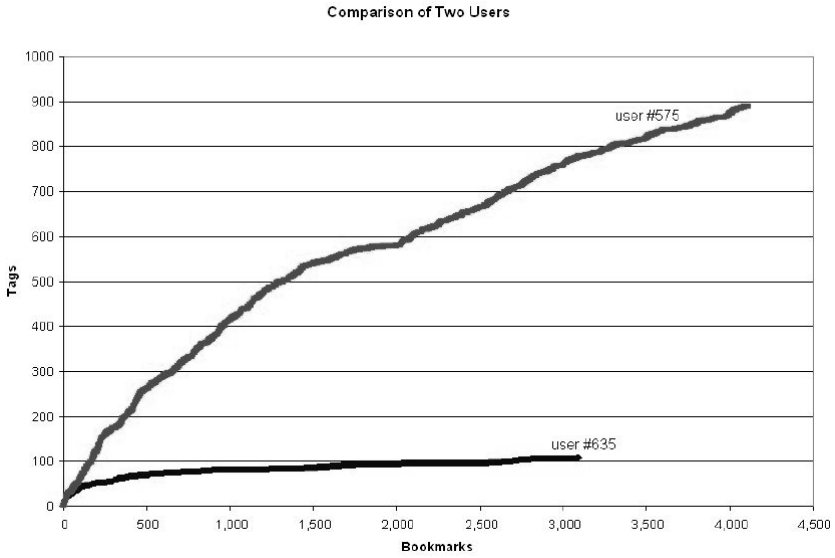
But leaving aside these basic problems, folksonomy also suffers from a crucial problem: the lack of standardisation in the tagging process. What should one think of the tag 'music'? Does this refer to content of a professional nature (a studio, a label, a music vendor, etc), the lyrics of a song, a historic site, or just a music fan's site? And more generally, folksonomy says as much about people's ways of seeing and classification as it does about the content thus classified. There is no universality in the definition of 'basic terms', which would allow content to be labelled: two different people will have different opinions about the 'basic terms' of a piece of information [12]. A history of cats, for example, could be tagged differently by the participants according to their expertise: as cat, or as Persian or as Felis silvestris catus longhair Persian. Sub-populations of experts can operate at a more specific level than the basic level in their area of expertise. Another problem is that, depending on the organising principles for categories that people have in their heads, some will choose different markers to designate an object: thus, a photograph of a piano could have a tag 'grand' for an individual classifying the piano as a musical instrument, while it will be different for a person classifying it as furniture. The position of the user in the social space of taste leads to perception and categorisation

would appear that standardisation faults can easily be rectified by a careful reading of the instructions.

12 For further light on the term universalist used to distinguish Rosch and Lloyd's theory of categorisation (1978), which postulates the existence of a 'basic level', see RASTIER (1991).

systems (BOURDIEU, 1979), which therefore transforms labelling. A final problem is that some users will want to apply a large number of tags, others few. GOLDER & HUBERMAN (2006) revealed a sharp contrast between two sub-populations, one of which labels with on average very few bookmarks, while the other is more lax. It revealed a very high level of dissimilarity of practice.

Fig. 5 - Two extreme users (#565, #635) tag growth (*)

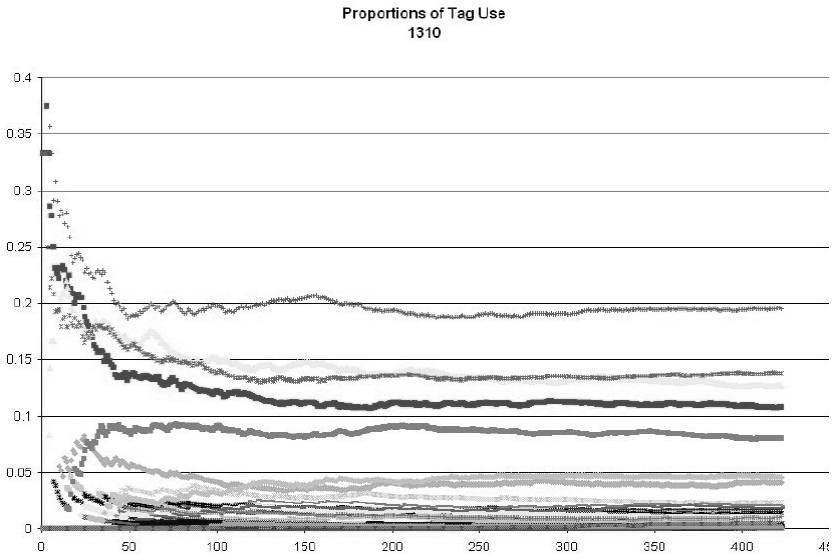


(*) As they add more bookmarks, the number of tags they use increases, but at very different rates

However, these governance problems should not be overestimated. The folksonomy sites are betting on emergence. They are counting on the fact that by self-organised aggregation, taxonomies will naturally emerge on the basis of these decentralised contributions. Some studies have revealed empirical evidence in this sense. GOLDER & HUBERMAN (2006), for example, have shown how swiftly a tag system can stabilise itself, with the frequency of appearance of a tag increasing when a piece of content is frequently referred to a fixed value. This is a very important phenomenon: consensus is tacitly formed without the direct wish of a higher authority. Tag convergence takes place when a reference becomes increasingly popular. GOLDER & HUBERMAN (2006) observe that, on the basis of a relatively low number of citations for the same URL (100 bookmarks), the relative proportion of the different tags used to describe it stabilises, and becomes

very stable. This is a very interesting result, which means that the reliability of labelling obtained via folksonomy is rather more optimistic. In actual fact, in the case of URLs cited more than 100 times, which is a low number, relative objectivity and labelling soundness already exists.

Figure 6 - The stabilization of tag's relative proportions (from GOLDER & HUBERMAN, 2006)



Attempts at governance have been driven by the need to domesticate these feral taxonomies without throwing more doubt on what makes these folksonomies interesting: the principle of faithfulness to the real world and to systems of perception. How are the significance of a bottom-up (self-organising) model and the need for homogeneity to be reconciled? As long as folksonomy sites still retained community homogeneity they could do without such rules since the existence of the community was posited on the existence of similar thought and systems and hence the use of identical terms to designate identical things: the problem of the explosion in the number of tags, for a programmer community, means that the different meanings of the term 'java' have little importance when practically all requests for this tag relate to IT. And the negative 'meta noise' elements for the conoscenti are less to be found within communities, and where they are negative are easily identified and excluded.

A governance requirement

As these sites have become more popular, however, attempts at governance have been developed. MARLOWE *et al*. have identified a classification in the types of governance existing. Three can be distinguished:

- Opening up labelling rights: 'self tagging' and 'free-for-all tagging' are opposed. In the first case, the tags are made by the posters of content themselves (readers have no right to write: this is the case with FlickR). In the second case, the tags are open to the totality of the readers (readers have the right to write: this is the case with del.icio.us). Between these two extremes a variety of levels of compromise exist. Some systems, for example, may select the resources which the users are permitted to tag (the images in ESP Game, for example), while other systems grant permission according to category (friends, family or contacts). There are also systems that authorise certain sub-populations to remove a tag. The management of tagging rights may explain the large differences between these systems.

- Support for tag writing: 'Blind tagging' is opposed (users are unable to see the tags assigned to the same resource by other users, as in del.icio.us) and 'viewable tagging' (users are able to see the tags already associated with a resource). Between these extreme positions lies 'suggestive tagging' (possible tags are suggested to a user). The creation of suggestions is a powerful convergence factor (GOLDER & HUBERMAN, 2006). FlickR assists the user when he enters his second tag on the same photograph, by informing him of the tags occurring along with the first tag he entered Del.icio.us offers an intersection between others' tags and the tags already placed by the tagger.

- A third important characteristic is the aggregation of tags for a given resource. Naturally, the two models can be placed in opposition, the bag model in which users are authorised to place a large number of tags for one resource; and the set (type FlickR) model, in which syntactic constraints and number limitations exist.

These types of practice make for a clear distinction from the self-managed taxonomy sites. Thus it can be seen that del.icio.us is a free-for-all system, suggestive, and resting on the model of the bag. Conversely, FlickR is a self-tagging system, blind to the first tags and viewable for the second tags, and above all based on the overall restriction model. What heuristics can be suggested to introduce greater objectivity into tagging practices? Three classes of heuristic can be distinguished.

An initial category of studies seeks to allow the users to operate at their liberty and to organise the viewing of their results by clustering by displaying the hierarchies between the tags. For this reason FlickR has recently introduced the idea of 'tag clusters', a first step towards a way of hierarchically arranging key words. For example, the term 'jaguar' brings up a page offering several groups, one a collection of cats, another of British cars and a third of French fighter aircraft. In the same vein is the site fac.etio.us, a search tool which offers an alternative presentation of the content of del.icio.us, organised in line with a more structured classification system. Fac.etio.us is a plug-in whereby it is possible to add some degree of organisation into the 'flat' system of collaborative tags; it introduces facets according to the genre of the tag: fac.etio.us is a reworking of the del.icio.us database, which makes use of faceted classification, grouping tags under headings such as 'by place' (Iraq, USA, Australia), 'by technology' (blog, wiki, website) and 'by attribute' (red, cool, retro). HEARST (2006) is working on faceted taxonomy systems (which blend hierarchical categorisation with clustering).

Other studies emphasize the need to train users. This work concentrates on the need to orient them in the direction of a syntax. One result from Ulises Ali MEJIAS (a student who undertook an interview-based study, although based on too small a panel) is that the mixing of ultra-specific tags and general tags is an effective means of retaining the uniqueness, which lies at the heart of the richness of del.icio.us and ensures that the taggings remain comprehensible. To improve the pertinence of the emission of tags, most websites let their users to pass through funny helps, which let them integrate the minimal rules of tagging.

Algorithms for quality recommendation and convergence

A third modality for regulating the self-generated taxonomies exists. This is by the design of the recommendations (viewed by the writer of a tag before he enters its own) that the system should be led towards the production of more relevant tags. SHEN *et al.* (2006) have undertaken a study on a folksonomy site dedicated to cinema, which indexes movies [13]. In order to fully appreciate the results of their work, we must understand that they were provided in advance with a quality criterion of a folksonomy

[13] The website is stored on the server of their original institution, the University of Michigan: http://movielens.umn.edu

website. This criterion has the following definition. Following the classification made by MARLOW *et al.* (2006), they arrange the tags in three classes: factual tags (item topics, kinds of item, category refinements), subjective tags (express user opinions, evaluations), and personal tags (which have an intended audience of the tag applier himself: task organisations, self-reference, item ownership). On the basis of this categorisation system, they construct quality criterions for a folksonomy system. A first quality criterion – which we could name a "substantial" criterion- relates to the proportion of factual tags on the whole: the factual tags are more easily understood and indicate a respect of objectivity criterions; the higher the proportion of them, the better the system is able to generate common knowledge. A second quality criterion, which could be called procedural, concerns the stability over time of the distribution between the three classes of tags: when the proportion stabilizes over time, the authors consider that this indicates a system reaching quality.

The interest of the study of SHEN *et al* (2006) is related to the fact that the authors are also the conceptors of the movie indexation website, which led them to program different recommandations vizualisation systems. They presented four different experimental panels: the unshared group (no tags are shared between members), the shared group (saw tags applied by other members of their group to a given movie), the shared-pop group (saw only most popular tags, i.e. those applied to the item by the greatest number of persons) and the shared-rec group (saw only tags applied to both the target item and to the most similar items to the target item; similarity between a pair of items was defined by the cosine similarity of the ratings provided by the users). The final tag class distributions between those four experimental groups were very different. The shared-pop and the shared-rec groups were dominated by factual tags, the shared group by subjective tags while the unshared group was divided more evenly. The most interesting result is that the tag selection algorithms have a major impact on tag class distribution.

Moreover, the authors looked at whether tag class distributions converged quickly or slowly: again, the shared-pop and the shared-rec converged, when the shared and unshared had less visual evidence of convergence: they were drifting. As a matter of fact, the shared-rec and shared-pop displayed algorithms that favoured tags applied by many different people, and those tend to be factual in nature (80% of tags applied by five or more people are factual). The third result was that the shared-pop and shared-rec groups had the greater number of tags. It can thus be seen that the choice of a precise selection algorithm for proximity criteria and popularity criteria is better if the group is oriented towards quality.

Conclusion

Folksonomy is a fundamental movement that expands the collaborative process, by allowing contributors to index content. It rests on three powerful properties: the absence of a prior taxonomy, multi-indexation and the absence of thesaurus. It concerns a more exploratory search than an entry in a search engine. Its original relationship-based structure (the three-way relationship between users, content and tags) means that two forms of results can be used that relate back to different social usages: cultural exploration and social exploration. Since labelling lacks any standardisation, folksonomies are often under threat of invasion by noise. Systems of governance have been set up to combat this threat based on three different formats. Firstly, systems of visualisation have attempted to reinstate a form of hierarchy-driven labelling by ascending breakdown. Secondly, user training, sometimes by way of a game, has aimed to lead users towards greater objectivity. However, it would appear that the most advanced forms of governance are by way of controlling recommendation algorithms. A judicious choice of these algorithms allows for great objectivity, or at least greater stability, in the self-organised aggregation of the individual labels.

Bibliography

ANDERSON Chris (2004): "The Long Trail", *Wired*, October.

AURAY N. (2000): "Le savoir en réseaux et l'empreinte inventive", *Alice*, no. 3, pp. 78-97.

BEAUDOUIN V. & LICOPPE C. (2002): "La construction électronique du lien social : les sites personnels. L'exemple de la musique", *Réseaux*, vol. 20, no. 116, pp. 53-96.

BEAUVALLET G. (2007): "Parties de champagne. Militer en ligne au sein de Désirs d'avenir", Hermès.

BEAUVISAGE T. (2004): *Sémantique des parcours des utilisateurs sur le Web*, thèse de sciences du langage, sous la direction de F.Rastier (Université Paris X Nanterre).

BOURDIEU P. (1979): *La Distinction. Critique sociale du jugement*, Editions de Minuit, Paris.

CARDON D. (2006): "La production de soi comme technique relationnelle. Un essai de typologie des blogs par leurs publics", *Réseaux*.

EATON A.: "Graph del.icio.us related tags": http://hublog.hubmed.org/archives/001049.html

GOFFMAN, E. (1995) : *Les Cadres de l'expérience*, éds. de Minuit, Paris.

GOLDER S. & HUBERMAN B. (2006): "Usage Patterns of Collaborative Tagging Systems", *Journal of Information Science*, 32(2), pp. 198-208.

GUY M. & TONKIN E. (2006): "Folksonomies – Tidying up Tags?", *D-Lib Magazine* 12, 1.

GYONGI Z., GARCIA MOLINA H. & PEDERSON J. (2004): "Combating spam with trustrank", Proceedings of the 30th International Conference on Very Large Databases (VLDB).

HEARST M. (2006): "Clustering versus Faceted Categories for Information Exploration", in *Communications of the ACM* 49(4), April.

(von) HIPPEL (2005): *Democratizing innovation*, MIT Press (Ca., USA), 208 p.

HOWE J. (2006): "The rise of crowdsourcing", *Wired*, no. 14.06

JENKINS H. (2006): *Convergence Culture: Where Old and New Media Collide*, MIT Press, Ca, USA.

Kipp Campbell Pattens and Inconsistencies in Collaborative Tagging Systems : an Examination of Tagging Practices, American Society for Information Science and Technology, 2006

LEADBEATER C. & MILLER P. (2004): *The Pro-Am Revolution. How enthusiasts are changing our economy and society,* Demos. See: [http://www.demos.co.uk/publications/proameconomy].

LE DEUFF O. (2006): "Folksonomies: Les usagers indexent le web", *BBF*, no. 4, pp. 66-70. See: <http://bbf.enssib.fr> (consulted February 16th 2007).

MAHLZ C. & EHRLICH K. (1995): "Pointing the way: Active collaborative filtering", Proceedings of CHI.

MARLOW C., NAAMAN M., BOYD D. & DAVIS M.,"Position paper, Tagging, Taxonomy, Flickr, Article, ToRead".

MATHES A. (2004): "Folksonomy – Cooperative Classification and Communication througs shared Metadata", working paper, Computer Mediated Communication, Graduate School of Library and Information Science, University of Illinois Urbana – Champaig

MEJIAS U. A. (2004): "Bookmark, Classify and Share: A mini-ethnography of social practices in a distributed classification community"

RASTIER F. (1991): *Sémantique et recheches cognitives*, PUF, Paris.

ROSCH E. & LLOYD B. (1978): *Cognition and Categorisation*, Hillsdale, Erlbaum.

SHEN S., LAM S., RASHID A., COSLEY D. & FRANKOWSKI D. (2006): "Tagging, communities, vocabulary, evolution", *Proceedings of Computer Supported Cooperative Work*, Banff.

SHEN K. & WU L. (2004): "Folksonomy as a Complex Network", Fudan University, Shangaï.

VOSS J. (2006): "Collaborative thesaurus tagging the Wikipedia way", Wikipetrics research papers, Wikimedia Deutschland e.V.

(van der) WAHL (2005): "Explaining and Showing Broad and Narrow Folksonomies". See: http://www.personalinfocloud.com/2005/02/explaining_and_.html