

When Taxonomy Meets Folksonomy: Towards Hybrid Classification of Knowledge?

Dr. Olivier Glassey
olivier.glassey@unil.ch
Observatoire Science, Politique et Société (OSPS)
Université de Lausanne

Introduction

The last five years have witnessed a sustained evolution of information technologies, coupled with a massive democratization of the production, the publication and the diffusion of electronic contents. The wide distribution of simplified editing tools and user-friendly electronic exchange platforms (blogs, wikis, etc) has led to a steep rise of online available information¹. The exponential growth of the number of information sources, as well as their strong heterogeneity in terms of quality and relevance raises the question of the readability and the accessibility of the online contents. In this context, we could make the hypothesis that the retrievability and the usability of information will rank among the critical issues related to knowledge diffusion.

This article is focused on the socio-technical strategies, which are designed to cope with the problem of informational entropy. To address this issue the mainstream approach has usually been to define and circulate new layers of information that encapsulate and frame the existing data in order to be able to classify them. Many efforts have been deployed to produce data about data (metadata) which could provide a generic common ground in order to facilitate an overall classification and use of content. The Dublin Core Metadata Initiative² is an example of such a cooperative effort towards the production of a stable set of attributes used to describe any kind of data. Such initiatives belong to a wider trend encompassing developments of a variety of data interchange languages supporting extended data models (ontologies) in computer science³. They are part of the semantic web approach which aim is to provide both interoperability among different technological platforms and easier way to find, share and integrate data.

Large scale private and public organisations have invested sizeable resources in order to cope with the heterogeneity of the content they produce⁴. The generalized use of such strategies as

¹ The evolution of the Blogosphere size (i.e. the number of active blogs, or personal web journals) could be used as an illustration of such evolutions. There were over 70 million blogs in early 2007 a number which tends to doubles every 320 days.

² <http://dublincore.org/>

³ As the exemple of such language, one can cite XML (Extensible Markup Language) which is derived like HTML from SGML (Standard Generalized Markup Language).

⁴ For an exemple in the public sphere of this kind of effort eGOV An Intergrated Platform for Realising Online One-Stop Government (Glassey, 2001).

a means to significantly reduce the informational entropy is nonetheless confronted with some strong limitations:

- The production of additional levels of classification is an expensive solution in terms of resources. It postulates an intensive collaborative process the role of which is to define and legitimate the classification categories, protocol and tools to be used.
- Even once negotiated and validated, the use of a new taxonomy requires the existence of some kind of control committee and training processes for the future users in order to ensure the consistency of this classification.
- The implementation of these classifications requires much time and is subject, consequently, to some incompressible inertia. This inertia could be a problem if we consider the fuzzy and fast evolving Internet content.

To confront these difficulties, a set of “new” classification strategies were worked out and implemented at large scale through many distinct initiatives. These strategies, known as “folksonomies”, connect the supposed philosophy of ancestral classification techniques with new technologies supporting collective referencing. Basically, they replace the top-down method proposed by taxonomic classification by bottom-up approach in which the classification categories are built directly by and for the users.

In this paper, our aim is to document some of the major stakes which underline the current cohabitation situation between taxonomy and folksonomy. In the first part we will briefly present the strengths and the weaknesses of folksonomy in order to analyze how they are constructed. We will study the role an intermediary category of actors whose main activity is to build bridges between taxonomy and folksonomy. Our main hypothesis is that these actors, whose tools and strategies have yet to be sociologically technologically and economically investigated, are central to an understanding of the future shape of the emerging knowledge society.

What are folksonomies?

Folksonomy belongs to the family of what could be described as “distributed classifications”.

Categories in a folksonomic process are not produced by a group of experts which control the whole classifying structure and who validate a common thesaurus to use. Folk classification emerge directly from users contributions. This means that in folksonomies users are free to label according to their own preferences the contents they produce and/or access. The million of photographs or videos which are available online on sites like Flickr, Youtube or Dailymotion have a series of labels which characterize them in a non-systematic way. These labels, known under the name of tags, are in these cases proposed by the people who provide the content⁵. To establish these tags, users can tap into a natural thesaurus: their own imagination (or other users’ imagination). The core principle of this classification process consists in offering to users pragmatic means of contents navigation that are close to their concerns and their own spontaneous classifying practices. As stated by one of the site promoting tags use, the advantage to rely on tags lays in the fact that they 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 to a bookmark as you like and rename or delete the tags later. So, tagging can be a

⁵ A distinction could be made when only content producers are allowed to put tags on their contribution (narrow folksonomy) or when all visitors could flag content with tags (broad folksonomy) (Vander Wal, 2005).

lot easier and more flexible than fitting your information into preconceived categories or folders”⁶

The analyses led on the nature of this user constructed thesaurus underline the heterogeneity of the utilized notions. They cover a broad array constituted by subjective concepts (“super”, “cool”, “interesting”), phonetic diminutives, misspelled words and even by multilingual linguistic innovations of Creole type (Golder and Hubermann, 2006). These elements form a kind of “semantic soup” that has no hierarchy. This lack of structure is obviously one of the main weaknesses of the folksonomic systems, which fail to generate any kind of guiding principles for user orientation within the classification items. This limit concerning a structuring framework is particularly sensitive at the level of synonyms management. Indeed, in open folksonomies there are no tools to manage semantic ambiguities. Consequently, the queries results based on this type of categorizations are not very precise. Folksonomies offer, at best, an approximate result which is not too far away and have “something to do” with the expected results. Another issue is the lack of lasting consistency due to the unending evolution of tags which are in state of constant (re)edition and thus make it impossible to guarantee the meaning of used tags. The absence of lexical control has also for consequence that tags could be subject to intentional manipulation or inappropriate ambivalences due to conflicting interests among users groups.

Considering these limits it’s legitimate to wonder whether, folksonomies are not, indeed, more a factor of entropy increase than a means to reduce it. The examination of the perceived benefits of these classification modes, however, could bring some nuances to this early assessment.

In fact, some of folksonomies identified weaknesses are also among the key reasons of their popularity. For example, the absence of hierarchy is undoubtedly problematic but it also lowers the barriers of access to the creation and use of classification. To make a query within a folksonomy, the user doesn’t have to go through the process of acquiring a mental map of the hierarchal structure of some classification tree. He/She doesn’t need to think about the semantic accuracy and deal with possible synonyms to have access to results that could turn out as meaningful. The simplicity of free typing and loose associations is just a lot easier than making a decision about the degree of match (Mathes, 2004). In practice classical taxonomies are very limited when it comes to deal with a very large and fast evolving content pool. In these cases, which covers many basic every day life Internet uses, the obligation to learn a complex, hierarchical controlled vocabulary would just be too costly for many users in term of time and cognitive effort (Mathes, 2004). In stark contrast, within large tags pools, basically any word (even misspelled or contracted like in SMS language) will bring some results more or less loosely related to the intended query.

To narrow down uncertainties linked to folksonomy’s messy classification people often use a combination of several tags which could led to more precise results. Tags combinations are powerful and flexible tools which allow building original hybrid categories.

In many cases, however, the lack of precision is not perceived as a defect. First of all, as very few people do go beyond the firsts pages of Google results; users usually do not have the time nor the interest to do exhaustive data mining. In most cases, what most users seek is not an exhaustive survey of what is online but a few good results to choose from. The added value of folksonomy comes from the fact it allow to produce, as a query results, different contents which links are made by other users subjectivity. By doing so, folksonomies produces a type of results especially favourable to exploration and unexpected discoveries (Merholz, 2004;

⁶ <http://del.icio.us/help/tags>

Kroski, 2005). The online radio *Last fm* illustrates this kind of exploration. On the website, the user is simply asked to introduce the name of an artist he/she likes into the system. Using this base and compiling thousands of other users tagged tastes the musical program is automatically built. The radio proposes to each listener a personalized musical selection of a dozen of artists (among several thousand) whose artistic works are somehow related, according to tags, to the selected artist. The evaluation of the obtained result remains, obviously, very subjective but the success met by this web radio leaves to think that numerous auditors find with it an enjoyable way to explore music.

Another example of the use of distributed tagging system in the field of culture could be found in the experiment led by eight art museums in the USA in order to allow their visitors to build their own tags⁷. The goal of the project is to propose alternative ways to describe and thus to navigate through the art collections. Such alternatives opens work of art classification to visitors own experiences. The project's aim is to explore original ways to visit and also to share the visits experience among people (Trant and Wyman, 2006). As stated by one of the involved actors " There is no way that you would or you could have museum professionals-academics art historians – sit down and look at every work of art and come up with every way you could possibly describe them" (Fitzgerald, 2006).

The folksonomy weaknesses, as we have mentioned them (lack of precision, lack of synonym control and lack of hierarchy of contents), led many of the promoters of such classification to explore new interfaces to help the user among the mass of tags. Some interfaces are basic and just offer the possibility to type tags or search for them but many are more original in the way they propose to interact with information related to content. Hierarchy info clouds (Fig.1) are one of the best known example of this type of effort. It provides a representation in clusters of tags which font size is proportional to tags popularity. The idea is to help the user to understand the main shared interest within a specific folksonomy.



Fig.1 The Tag cloud of the 150 most use tags on Metafilter⁸

Other interfaces propose the construction of dynamic semantic clusters (Fig. 2). From these clusters the system draws a network where the selected tag is the core and in which nodes made by tags quantitatively the most related to this core. The topography of the network is created accordingly to the numbers of tags co-occurrences. The user could navigate in this virtual landscape by selecting a new node from which the system will automatically build a new network. This kind of interface is aimed to offer a better understanding of the lexical neighbourhood of any notion within a folksonomy and could to help the users to cope with the lack of synonym control.

⁷ The Art Museum Community Cataloging Project : URL : www.steve.museum

⁸ Source Metafilter community weblog : <http://www.metafilter.com/tags/>

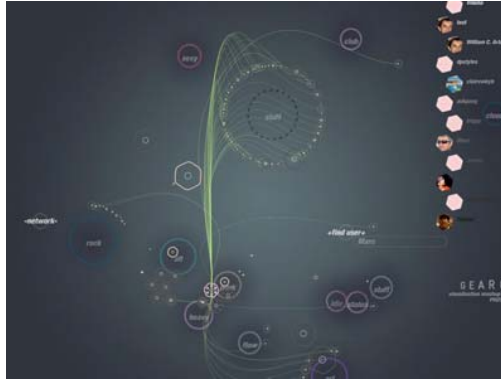


Fig.2 Fidg't' a prototype of a dynamic map blending social networks and tags⁹

In conclusion of this very brief overview of distributed classification main weaknesses and strengths we would argue that folksonomies, due to their fuzzy nature and proliferating multiplicity, take objectively part in the increasing complexity of online contents management but, at the same time, they do also offer means to reduce the perception of this complexity at individuals subjective level. Before further exploring folksonomies, it would be necessary to explore what kind of theoretical resources are available to tackle their analysis.

Old practice and new tools

Folksonomy has strong connection with other terms such as ethno-classification, social classification, and folk classification¹⁰. It belongs to the long tradition of researches investigating formal and informal classification modes that finds its modern roots in the works of the first ethnologists and sociologists. Folksonomy could be related with most ethnographical studies and theories which were interested to find out how any human culture organizes the description and the understanding of the world which surrounds it.

It is far beyond the range of this contribution to be able to provide a satisfying account of the richness and the diversity of the theories built to explain these classification processes. More modestly, we would like to stress here that, since the early age of these disciplines the question arose whether if there is an ontological difference between, on the one hand, the spontaneous classification modes existing in any human traditional society, and, on the other hand, scientific classification. Authors like Mauss and Durkheim (1903) defended the plurality of the classification modes and make them specific to each human group. In contrast, later works like those of Berlin (1992) attempted to attenuate this cultural relativism by showing that a small number of dimensions could be enough to describe the large majority of ethnobiological traditional knowledge. The hypothesis of an ontological differences between taxonomy and folksonomy which leads to an opposition between these two classification modes is largely dominant when one considers the literature related to the application of these processes within the field of information technologies. However, many authors refuse to oppose in a strict and dichotomist way these modes of organizing the world. While recognizing differences, they consider that these two classification processes are indeed overlapping and they prefer to evoke a continuous transformation which leads from one to the other (Descola, 2005; Ghiselin and Landa, 2006).

⁹ Source : <http://www.fidgt.com/visualize>

¹⁰ In the field of information technologies terms like *free tagging* or *open tagging* are also used (Lund et ali., 2005).

These theoretical approaches, which are not present in the current analyses of the IT- driven folksonomy, suggest, in our opinion, the way for further investigations on the articulations between these two classification modes.

Recent researches linked to the sociology of sciences and technology offers an interesting contribution to carry out such type of research. Among others, Bowker and Star (2000) provided a theoretical frame that dynamically binds sociological, economical and technical dimensions of what could be called “classification engineering”. They pay a special attention to the negotiation processes among the actors who design any classification. These approaches take into account the trajectory of the produced classification and seek to better understand how classifications re-enact themselves once delegated to technical devices that convey them. Along with other studies, they also explore how the identity of an “object” is constructed through the various levels of categorization which are applied to it and how this identity carry the representations and the cognitive models brought in by all the implied social actors and their institutions (Star and Griesemer, 1989; Boltanski, 2004).

In our view such an approach allows to bypass the binary opposition of classification schemes and to focus more precisely on the production mode of these categories. Consequently, it authorizes us to raise the question about the types of actors who produce, maintain and develop the infrastructure required by those systems.

The wealth of bridge builders

From the point of view of the complementary characteristic offered by taxonomies and folksonomies, we have suggested the existence of a new category of actors who are able to establish a link between those classification modes. The question to address is how to distinguish those actors who are able to both use the systematic power of taxonomies and to harness the spontaneous creativity, reactivity and ease of use of folksonomies? The first obvious answer would consist in pointing out that each individual, in his/her every day practice operates and combines simultaneously these two levels. This obviousness, if undeniable, accounts only for a part of the observable reality. Indeed, the scope of large scale societal knowledge diffusion requires to pay a specific attention to groups of individuals and private or public organisations which have made the navigation between these two levels an essential component of their development strategy.

Besides the expertise of professionals who create, organise and sometimes control classification categories and, on the other hand, the user who produce his/her own set of classification templates one could find the emergence of a new category of intermediary actors. These actors have in common the fact that they do not produce, directly, classification but cultivate other people skills to do so. Many among the most successful (in term of popularity) websites, rely heavily on some kind of folksonomy. As mentioned above, commentators often present a binary distinction between the top-down classification and bottom-up initiatives. Real cases and in depth analysis often offer a somewhat less clear dichotomy. Taxonomies are rarely pure in the sense that even the more comprehensive classification systems are subject to local interpretation and are not immune to vernacular developments. In the other end, folksonomies are not as free from taxonomy as suggested by most commentators. The infrastructures of the tagging systems are often controlled on technological and even at semantic levels by small groups of technicians/supervisors who act as moderators.

As one of the most obvious examples the now well publicized Google success could partially be attributed to the setup of folksonomic type strategy. Along with the algorithms used to

perform information filtering and massive automated indexation processes, one of the key components of their strategy is to collect how people use hyperlinks to point to other online contents. These links could be considered as a basic tag functioning like a token of appreciation linked to a specific content. The automated aggregation of these tags is used to construct the final ranking of the results. A “popular tags vote” structures the query results. It is therefore not a surprise that one systematically finds Wikipedia’s articles (which belong to some of the most referred to websites) among the first results related to any topic. Such a system raised many criticisms and it indeed authorizes to manipulate the results by artificially creating a large number of links in order to promote to the top of Google’s results a specific content (Google bombing). The originality of the tactic adopted by in Google is due to the fact that this company did not try to build taxonomies to predict users’ expectations. On the contrary, they adopted architecture for results restitution which was directly based on the monitoring of other users behaviours (Shirky, 2005). The successes met by this kind of hybrid methods in various contexts caused many followers. It also gives a hint of the scope and the stakes involved in the use of folksonomy.

Folksonomies aren’t stand alone products. The building of a folksonomic “space” engenders the creation of a community of users which, in turns, offers many advantages not only to the users but also to the project promoters. Popular websites like *del.icio.us* or *Digg* are basically built as a huge folksonomic thesaurus for what is considered by each user as “interesting” on the web in relation with a specific field. *de.li.cious* functions at a first level as personal repository for Internet bookmarks which could be organised by tags. On a second level it allows tags sharing with other members as resources to find more easily “interesting” content. Whereas we know few studies conducted on this topic, it appears nevertheless rather obvious that the consistency of the distributed thesaurus rely, at least partly, on the homogeneity of the users community involved in the process (Guy and Tokin, 2006). Thus, while quantitative evidence is still lacking it’s possible to observe an evolution of the nature of what is regarded as worth tagging in *de.li.cious* in parallel with the evolution the composition of users group who were, initially, mostly Internet specialist. The service offered became so popular that countless Internet sites propose to their visitors a dedicated button on which they could click in order to promote this particular site within *de.li.cious*. The advantages of such tagging system are multiple. Obviously, an active user community generates a sustained traffic which is still one the predominant stick yards for assessing the economic value of websites. Moreover, an active folksonomic community also provides something which could be even more valuable: a real-time update of what is considered as innovative and valuable by a large panel of internet users.

In the same orientation but with the specific public of scientists as a targeted audience, the revue *Nature* launched in 2006 its own service, “Connotea”¹¹, which offers to store articles references and to let the users arrange them with tags. In this case also, provided the fact that it will be a success (i.e. adopted by a significant number of scientists), such services could give an extraordinary insight into how scientists map their knowledge and build/update their references list and even, ultimately, how they work. This service could also provide the base for an early detection tool which could identify fast evolving fields within any scientific domains.

Many commentators present folksonomies as being more “democratic”, more respectful to individual freedom and thus less oriented than taxonomies (Kroski, 2005). Contrary to these positions, we would defend the idea that, harbouring a folksonomy is not more neutral than

¹¹ <http://www.connotea.org/>

building a taxonomy. While direct control is often impossible and some freedoms are granted to users, all the maintenance and the technical support of a folksonomy often lays in the hand of a small group of people. The study of folksonomies should not leave in the shade the actors who, on the operational level, implement these open classifications and are able to exploit them. Such study should not undermine either the fact that folksonomies produce a lot of information about data and users behaviours. These informations are an asset of strategic importance in a knowledge society as they could be declined in many different services. The interests of folksonomy providers could be, in many cases altruistic. Nonetheless, they occupy a privileged place to hold and defend an epistemological advantage which researchers need to analyze and understand.

Conclusion

We consider the field of folksonomy study as promising at different levels. At the first level there is a long history of research and theories related to ethnoclassification, which provide an interesting background to the ongoing socio-technical evolutions. As many authors pointed out, the relationships between social and technical dimensions are deeply embedded when we consider modern classification. Their mutual shaping processes are subtle and tightly knitted. Information technologies enhanced folksonomies offer a challenging field to investigate how such combinations could play a role in the diffusion and the access to information and, ultimately, to knowledge. As we have presented it, folksonomies are indeed powerful tools to lower cognitive barriers required to access online information navigation. As such, and even considering their limits, they will play an original role in users learning strategies. Moreover, the practice of the distributed classification draws the outline of the possible emergence of a new type of sporadic epistemic communities. Folksonomies are produced by a continual iterative process of validation and invention of the categories considered as relevant by groups of users. It is possible to look at them as a form of semantic virtual community, whose coherence; promptness and lifespan are strongly variable accordingly to the group of users and intermediary users they depend upon.

It is undoubtedly too early to have a precise idea of the future of these communities of classification. Among the preliminary interrogations about their future, we can, for example, wonder to who belongs the knowledge resulting from folksonomic processes? Whereas the intermediate users we have previously identified are strongly interested in capitalizing on their folksonomic inheritance, controversies are brewing on the ethical and legal grounds for the legitimate use of these classifications. Another legitimate question could also be to assess if the folksonomies multiple instances will find ways to overcome their insulation and profit from one to each other. In other words, will folksonomies be able to implement, stabilize and generalize some forms of horizontal collaboration?

In all ways, the practices related to folksonomy are now tightly incorporated in online navigation interfaces and consequently in the Internet's every day use. We would argue that this type of approach to data classification will have a lasting impact in the understanding of both users' representations about knowledge and the future shapes of knowledge's access and diffusion in our societies.

References

- Berlin B., (2002), *Ethnobiological classification: principles of categorization of plants and animals in traditional societies*, Princeton University Press, Princeton.
- Boltanski L., (2004), *La Condition fœtale : une sociologie de l'engendrement et de l'avortement*, Paris, Gallimard, 2004.

- Bowker G.C., Star S.L., (2000), *Sorting Things Out, Classification and Its Consequences*, The MIT Press, Cambridge Ma.
- Descola P., (2005) « Anthropologie de la nature », cours au collège de France. http://www.college-de-france.fr/media/anthrop/UPL19764_descolares0405.pdf2005
- Fitzgerald M., (2006), «The Name game», in *CIO Magazine*, April 2006.
- Ghiselin M.T., Landa J., (2005), «The Economics and Bioeconomic of Folk and Scientific Classification», in *Journal of Bioeconomics*, n°7.
- Glasse O. (2001), «Une plate-forme intégrée pour réaliser un guichet virtuel unique en ligne : le projet e-GOV», in *Isp.ch newsletter (Lettre d'information du groupe de coordination société de l'information*, n°11, 2001.
- Golder S. and Huberman B. A., (2006), «Usage Patterns of Collaborative Tagging Systems», in *Journal of Information Science*, 32(2).
- Guy M., Tonkin E., (2006), «Folksonomies, Tidying up Tags?», in *D-Lib Magazine*, vol. 12, n°1, January 2005.
- Kroski E., (2005), « The Hive Mind: Folksonomies and User-Based Tagging », blog InfoTangle <http://infotangle.blogspot.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/>
- Lund B., Hammond T., Flack M. and Hannay T., (2005), «Social Bookmarking Tools (II), A Case Study - Connotea» in *D-Lib Magazine*, vol. 11, N° 4, April 2005.
- Mathes A., (2004), «Folksonomies – Cooperative Classification and Communication through Shared Metadata», <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>
- Mauss, M., & Durkheim, E., (1903), « De quelques formes primitives de classification. Contribution à l'étude des représentations collectives », in *Année sociologique*, VI, pp. 1-72.
- Merholz P., (2004), «Metadata for the Masses». <http://www.adaptivepath.com/publications/essays/archives/000361.php>
- Star, S., L. & Griesemer, J. R. (1989). «Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology 1907-39», in *Social Studies of Science*, n°19, 387-420.
- Shirky's C., (2005), «Folksonomies & Tags: The rise of user-developed classification», IMCExpo conference, April 2005.
- Trant J., Wyman B. (2006), «Investigating social tagging and folksonomy in art museums with steve.museum» <http://www.archimuse.com/research/www2006-tagging-steve.pdf>
- Vander Wal T., (2005) « Folksonomy », presented at the Online Information Conference, London, December 2005.