



A peer-reviewed student publication  
of the University at Buffalo  
Department of Library and Information Studies

## **Collaborative tagging, folksonomies, distributed classification or ethnoclassification: a literature review**

Edith Speller  
Assistant Librarian  
Royal College of Music  
London, United Kingdom

Library Student Journal,  
February 2007

### **Abstract**

Tagging, folksonomy, distributed classification, ethnoclassification—however it is labelled, the concept of users creating and aggregating their own metadata is gaining ground on the internet. This literature review briefly defines the topic at hand, looking at current implementations and summarizing key advantages and disadvantages of distributed classification systems with reference to prominent folksonomy commentators.

After considering whether distributed classification can replace expert catalogers entirely, it concludes that distributed classification can make an important contribution to digital information organisation, but that it may need to be integrated with more traditional organisation tools to overcome its current weaknesses.

### **Introduction**

The literature of tagging is largely opinion-based, almost entirely online, and the topic is largely absent from academic literature as it has only emerged as a system in the last two years. This review aims to make some sense of the subject by drawing out some dominant themes from the literature, with reference to a range of interesting and/or significant papers.

### **Definitions**

Metadata creation, and who does it, is an important issue with the current explosion of textual and non-textual information available on the internet. In his seminal introduction to tagging and folksonomies, Mathes (2004) summarises three different agents who may create metadata: professionals (e.g. cataloguers), authors, and users. Tagging is an approach to user metadata creation where users describe information objects with a free-form list of keywords ('tags'), often to allow them to organise and retrieve the objects at a later date. It is important to remember that objects can be tagged with as many or as few words as desired; there is no restriction to placing objects in one category. The tags produced by many users can be aggregated to form a non-hierarchical group of terms christened by Vander Wal as a 'folksonomy' (Porter, 2005; Smith, 2004). This can be searched or browsed by users, either to retrieve items

they previously tagged themselves, or to discover items tagged by other people.

There is some argument over the most appropriate term to use for this type of system—as well as ‘folksonomy’, suggestions have included ‘ethnolocation’ (Merholz, 2004b), and ‘distributed classification’ (Mejias, 2004). ‘Folksonomy’ is emerging as the dominant term, with Google results for the word in the tens of millions rather than the tens or hundreds of thousands found for the other terms. However, it is argued that folksonomy is a misleading term as the systems in question bear very little relation to taxonomies or ontologies (as clarified by Vander Wal, 2005), and that distributed classification is more descriptive of the tagging process (Hammond, Hannay, Lund, & Scott, 2005; Merholz, 2004a). Therefore, this article will refer to this type of system as distributed classification.

## **Precursors**

For some time, there has been awareness that existing models of information organisation needed revision and development to cope with both digital and multimedia information. Bates (1998) states that users performing unknown-item searches have a different level of knowledge from the indexers describing the objects for retrieval. She also contends that indexing and term selection is a deeply subjective process, affected by morphological, syntactic and semantic term variation, which means that imposing one person's view of an item in a retrieval system may hinder retrieval by others. The idea of asking users to index items was already being developed by Hilderley and Rafferty (1997), who saw the process as a possible way of indexing particularly subjective forms of information where full-text searching is either not possible or not useful, such as multimedia objects and fiction. They developed the idea of aggregating users' indexing terms to create

a generalised overall view of the resources, which is being adapted by some working systems: for example, Del.icio.us (<http://del.icio.us>) search results display the most common tags used for each result.

## **Current implementations**

The two best-known implementations of distributed classification are Del.icio.us and Flickr (<http://www.flickr.com>). The former is a social bookmarking system where users tag and store web links which they want to find again in the future, while the latter is a photo-sharing website where users upload and tag their own photos to aid retrieval by themselves and others. Multiple users tagging the same resource is rare in Flickr, although image owners can permit others to add new tags to their photos if they wish.

There are numerous other interesting implementations of the idea. For example, CiteULike (<http://www.citeulike.org>) has a similar approach to Del.icio.us but is more focused on scholarly writing and journal articles in particular, allowing academics to keep a record of reading, and share useful references with others. Some sites are developing distributed classification of multimedia resources, such as Last.fm (<http://www.last.fm>) for music and YouTube (<http://www.youtube.com>) for video (often user-authored, like the photography in Flickr).

Interestingly, there are also emerging attempts to organise offline resources using distributed classification. One example is LibraryThing (<http://www.librarything.com>) which enables users to create metadata records for their book collections based on MARC library catalogue records (or Amazon records) as well as tagging them. It currently boasts nine million (non-unique) catalogued items, and is becoming an impressive source of book information.

A prototype library catalogue incorporating tagging came about after Levine (2005)

mentioned the idea on her weblog. Pattern (2005a, 2005b) developed the idea and created both a 'tag cloud' (a visual representation of how popular different tags or subjects are) from his catalogue's professionally-added subject headings, and an experimental interface allowing users to tag books in the library catalogue. Casey Bisson has also been experimenting with alternative ways of displaying OPAC information, and has been awarded for his WPopac project (Bisson, 2006; Guinee, 2006). The WPopac involves catalogue records being posted as blog entries, which allows records to be tagged and enables a variety of other enhancements to the OPAC, such as increasing the possibilities of finding items through alternative search systems, including Google.

## **Advantages of distributed classification**

### **The consensus viewpoint**

One of the most evangelical supporters of distributed classification, Clay Shirky (2005b), gave a talk and wrote an essay entitled 'Ontology is Overrated' in early 2005. One of his key claims about distributed classification is that it can be harnessed to create a bottom-up consensus view of the world, which is more valid than any one view imposed from the top down. This ties in with the idea of 'desire lines' described by Merholz (2004b) and Kroski (2005). They argued that if index terms are created by users, other users are more likely to find what they need. This in turn corresponds with the ideas articulated by Bates mentioned above. A related idea is the theory of 'wisdom of crowds' by James Surowiecki (explained by Sinha, 2006). This suggests that there are four steps needed to harness the wisdom of crowds: diverse opinions, independent decision-making, decentralisation of power, and a way of aggregating opinions. All of these are usually possessed in distributed classification systems to differing degrees.

The ease of participation described by Mathes (2004) and Kroski (2005) should help to enable anyone who wants to contribute to do so. One argument against these ideas, particularly desire lines, is given by Powers (2005), who extends the metaphor to contend that this system gives disproportionate power to the few, saying paths are worn by young people who don't care about the common good. It is debatable whether giving most of the power to some opinionated people is better than giving all the power to an educated few, but at least the former allows input from a broad spectrum of information users.

### **Pseudo-faceted classification and analogue classification: 'this is probably about x, but might be about y too'**

Pseudo-faceted classification, or the classification of objects using several aspects of their 'nature' simultaneously, is a feature shared by many thesauri, such as those used for indexing journal articles. However, at present the indexing terms generated by distributed classification are not grouped together in any way like a true faceted scheme, which usually consists of multiple navigable hierarchies (Weinberger, 2006). A few attempts to aggregate tags into facets are underway at present, such as Fac.etio.us (<http://demo.siderean.com/facetious/facetio.us.jsp>; using data from Del.icio.us), and the tag section of Mefeedia (<http://mefeedia.com/tags>). Golder and Huberman (2006) point out that tags can be seen as filters which can be used to create combinable sets of search results.

Analogue classification is tangentially related to faceting. When many users have classified the same object, the degree of term overlap can be used to tell users that the object is thought to be about subject x by (say) 75% of users, while 30% of users see it as being about subject y. Hammond et al. (2005) thinks this will be useful when search results are ranked by relevance,

while Shirky (2005b) explains that this can also be used to automatically determine which terms are perceived as synonymic.

### **Rapid adaptability to changing vocabularies**

A commonly cited advantage of distributed classification is its flexibility and ability to reflect changing terminologies, without the substantial effort involved in choosing and adding terms to a controlled vocabulary (Hammond, et al., 2005; Kroski, 2005; Mathes, 2004; Merholz, 2004b; Shirky, 2005b). This is helpful in the rapidly evolving technology field but may also be relevant more generally, especially with more subjective information: Veltman (2004) contends that cultural information is very context-dependent and its meaning may well change over time. On the other hand, Smith (2005) points out in his critique of Shirky's essay that unless people reclassify older objects after shifts in vocabulary, distributed classification systems will face the same problems as traditional classification schemes. This means many items will become difficult to retrieve or may be indexed confusingly. Perhaps there is a need for some sort of time-weighting or an expiry date for tags, although this may cause its own problems: for example, older tags with current relevance could receive a lower weighting than newer and less useful tags, simply because of their age.

### **Serendipitous browsing**

Some feel that distributed classification is an excellent tool for serendipitous browsing: the flat structure of the metadata can be beneficial as it allows for more chance discovery through exploring related tags, which may have been widely separated in a thesaurus (Kroski, 2005; Mathes, 2004). This aspect of the concept helps to support this method of information-seeking online and make up for the reduction of users browsing a physical library collection, especially in the context of OPAC

searching. Tagging can be used to make keyword searches more effective and more closely linked to users' own views of the material in question. This may enable more fruitful searching than keyword searching of more traditional metadata such as title and imposed subject headings.

## **Disadvantages of distributed classification**

### **Synonyms and homonyms**

Synonyms and homonyms present different but related problems to distributed classification systems, as pointed out by many writers (Golder & Huberman, 2006; Guy & Tonkin, 2006; Kroski, 2005; Mathes, 2004; Merholz, 2004b; Powers, 2005). Distributed classification usually makes no allowance for formal identification of synonyms, which leads to the fracturing of collections and reduced recall when searching the system (Kroski, 2005). Del.icio.us aims to aid retrieval by automatically finding 'common tags', i.e. those often used in combination with the search term, and suggesting the user searches for these tags too. LibraryThing (2007) takes this a step further and allows users to select synonymous tags from the list of related terms, which the system then groups together: for an example see the 'library science' tag.

The presence of homonyms in the system means that search precision is reduced: one example is that a search for 'apple' may give results about both fruit and the technology company, one of which is bound to be irrelevant to the searcher's needs (Weinberger, 2006). Flickr (2007) tries to improve precision by automatically generating 'clusters' of terms to disambiguate different meanings of the search term, which can work well: for example, look at the clusters formed around the 'apple' tag. Of course, this increase in precision results in a loss of recall as some photos may be tagged 'apple' only and so

will not be found but, as Kroski (2005) asserts, this is a necessary trade-off. Golder and Huberman (2006) suggest the same solution but performed manually—searching for more than one tag simultaneously will reduce the homonymy problem.

### **‘Sloppy’ tagging: morphological, syntactic and semantic term variation**

Indexing terms selected without guidance by a large group of untrained users are almost inevitably going to be inconsistent and even inaccurate. Two morphological problems are the use of singular versus plural nouns, and the varying capitalisation of terms, which can be partly solved using stemming and treating tags in different cases as synonymous. However, Shirky (2005b) feels that any form of aggregating terms dilutes the strength of distributed classification, and Meijas (2005) points out that users may use ‘apples’ instead of ‘Apple’ to distinguish the fruit from the technology company. Guy and Tonkin (2006) found that a substantial proportion of tags were misspelled, although this included many compound-word tags, which are not yet handled consistently; users often insert punctuation to separate words. This may not matter if items are tagged many times by different users and reach a ‘critical mass,’ allowing cross-referencing between different spellings, but in a system like Flickr where photos tend only to be indexed once, a misspelling can lead to the item being very difficult to find.

Multilingual user groups also cause problems for distributed classification, as the lack of structure makes it difficult for tags in one language to be translated into another, and some systems cannot even handle non-Latin alphabets adequately (Guy & Tonkin, 2006). Powers (2005) takes this point further and argues that, as some languages are less suitable for keyword indexing of any sort, distributed classification as a system is "based on bias,

formed from a specific culture, which tends to be male, western, with English as a native language" (¶ 45) The implication is that it is little better than the inherent structural biases of a system like the Dewey Decimal Classification. However, Shirky (2005a) argues that at least with a distributed classification system the people creating metadata tend to be visible members of the system which allows their possible bias to be understood and taken into account. Guy and Tonkin (2006) hope that user profiles will become more extensive and that discussion facilities can be integrated into the systems to allow people to explain why they have tagged objects in a particular way. Several writers also hope that systems can be developed further to encourage better tagging through suggesting popular tags (which Del.icio.us already does when users tag links), suggesting different facets of the object to describe, or giving feedback (Guy & Tonkin, 2006; Hammond, et al., 2005; Mathes, 2004; Pind, 2005). An alternative is the user community deciding on voluntary tagging guidelines (perhaps a new netiquette), for example whether to use singular or plural nouns and how to handle compound terms. But creating these guidelines may in itself hinder the system's ease of use which is one of its great benefits (Guy & Tonkin, 2006; Meijas, 2005).

A thorny semantic problem is the issue of ‘basic level variation’ described by Kroski (2005), and Golder and Huberman (2006). This relates to how much detail an individual will go into when tagging a resource: Golder and Huberman (2006) give the example of an average person naturally tagging a photo as ‘bird’ when a birdwatcher would have naturally tagged it ‘robin’. This difference means that the two users may find each other's tags next-to-useless because they are at the wrong level of specificity for their needs. Hierarchy and structured vocabularies come into their own when facing this problem: this is one of the strongest arguments in favour of finding a way to synthesise hierarchical structures

into distributed classification without destroying its essence.

## Can distributed classification replace expert cataloguers?

Some commentators feel that distributed classification is inevitable and essential if we are to successfully organise online information, due to the money and labour required to catalogue objects using other methods (Kroski, 2005; Shirky, 2005a). Most agree that distributed classification is useful and that it is unlikely to go away, but they still see important functions that other organisation systems should fulfil. Lawley (2005) says folksonomies are extremely useful, but she does not think "all expertise can be replicated through repeated and amplified non-expert input". Smith (2005) concurs, saying "One can be enthusiastic about tags and folksonomies (I am) and still confront the serious problems that face them as a stand-alone tool for organizing information." Hammond et al. (2005) and Guy and Tonkin (2006) agree that tagging can play a complimentary role alongside more formal types of organisation.

There are many other areas relating to distributed classification which there is not space to discuss here: for example, the social issues surrounding tagging including spamming, the issue of privacy and the growth of social networks, as well as the topic of trust and authority (Blood, 2005; Hammond, Hannay, Lund, & Scott, 2005; Kroski, 2005; Lawley, 2005; Sinha, 2006). The cognitive process behind tagging is another interesting area, which Sinha (2005) is exploring.

## Conclusion

Distributed classification is in its infancy both as a theory and as a system in practice. The emergent viewpoint is that the system may work best when organising non-physical information objects in tandem

with 'traditional' information organisation tools such as controlled thesauri, but how this could happen in practice has not yet been worked out. Research and experimentation in this area is necessary; one promising area appears to be the integration of tagging into OPACs, working in conjunction with traditional cataloguing; another may be the use of user-generated indexing to devise a thesaurus more relevant to users' needs and use of language. Raising the quality of tagging is an important challenge unless aggregation of tags can be used to get around the problems of 'sloppy' tagging. The level of specificity used when indexing may well become a greater issue as collections grow: users are already discovering that the level of specificity they chose to use when first indexing items may be too general once their collection grows. The folksonomy will not eliminate traditional forms of classification, as its greatest cheerleaders have claimed, but instead will perform a complimentary role, and will become a useful tool for information professionals and others aiming to organise and enable access to information objects.

## References

- Bates, M. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13), 1185–1205.
- Bisson, C. (2006). *WPopac: An OPAC 2.0 Testbed*. Retrieved 17 January, 2007, from <http://maisonbisson.com/blog/post/11133/>
- Blood, R. (2005). *I've noticed a slight problem with the Technorati tagging system*. Retrieved 17 January, 2007, from <http://www.rebeccablood.net/archive/2005/01.html#11technorati>
- Flickr. (2007b). Clusters for the 'apple' tag. Retrieved 17 January, 2007, from

<http://www.flickr.com/photos/tags/apple/clusters/>

Golder, S., & Huberman, B. (2006). Usage patterns of collaborative tagging Systems. *Journal of Information Science*, 32(2), 198–208.

Guinee, E. (2006). *Bisson's WPopac wins Mellon Award*. Retrieved 17 January, 2007, from <http://librarystudentjournal.blogspot.com/2006/12/bissons-wpopac-wins-mellon-award.html>

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up tags? [Electronic Version]. *D-Lib Magazine*, 12. Retrieved 17 January, 2007, from <http://www.dlib.org/dlib/january06/guy/01guy.html>

Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005). Social bookmarking tools (I): A general review [Electronic Version]. *D-Lib Magazine*, 11. Retrieved 17 January, 2007, from <http://www.dlib.org/dlib/april05/hammond/04hammond.html>

Hidderley, R., & Rafferty, P. (1997). Democratic indexing: An approach to the retrieval of fiction. *Information Services & Use*, 17(2/3), 101–109.

Kroski, E. (2005). *The Hive Mind: Folksonomies and user-based tagging*. Retrieved 17 January, 2007, from <http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/>

Lawley, L. (2005). *Social consequences of social tagging*. Retrieved 17 January, 2007, from [http://many.corante.com/archives/2005/01/20/social\\_consequences\\_of\\_social\\_tagging.php](http://many.corante.com/archives/2005/01/20/social_consequences_of_social_tagging.php)

Levine, J. (2005). *MLS February Tech Summit on Social Bookmark Services*. Retrieved 17 January, 2007, from [http://www.theshiftedlibrarian.com/archives/2005/01/18/mls\\_february\\_tech\\_summit\\_on\\_social\\_bookmark\\_services.html](http://www.theshiftedlibrarian.com/archives/2005/01/18/mls_february_tech_summit_on_social_bookmark_services.html)

LibraryThing. (2007). *Library Science tag*. Retrieved 17 January, 2007, from <http://www.librarything.com/tag/library+science>

Mathes, A. (2004). *Folksonomies - cooperative classification and communication through shared metadata*. Retrieved 17 January, 2007, from <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>

Mejias, U. A. (2004). *Bookmark, classify and share: A mini-ethnography of social practices in a distributed classification community*. Retrieved 17 January, 2007, from [http://ideant.typepad.com/ideant/2004/12/a\\_delicious\\_stu.html](http://ideant.typepad.com/ideant/2004/12/a_delicious_stu.html)

Mejias, U. A. (2005). *Tag literacy*. Retrieved 17 January, 2007, from [http://ideant.typepad.com/ideant/2005/04/tag\\_literacy.html](http://ideant.typepad.com/ideant/2005/04/tag_literacy.html)

Merholz, P. (2004a). *Ethnoclassification and vernacular vocabularies*. Retrieved 17 January, 2007, from <http://www.peterme.com/archives/000387.html>

Merholz, P. (2004b). *Metadata for the masses*. Retrieved 17 January, 2007, from <http://www.adaptivepath.com/publications/essays/archives/000361.php>

Pattern, D. (2005a). *Taggytastic!* Retrieved 17 January, 2007, from <http://www.daveyp.com/blog/index.php/archives/46/>

Pattern, D. (2005b). *Taggytastic - part 3*. Retrieved 17 January, 2007, from <http://www.daveyp.com/blog/index.php/archives/57/>

Pind, L. (2005). *Folksonomies: How we can improve the tags*. Retrieved 17 January, 2007, from <http://pinds.com/articles/2005/01/23/folksonomies-how-we-can-improve-the-tags>

Porter, J. (2005). *I've heard of folksonomies, Now how do I apply them to my site?* Retrieved 17 January, 2007, from [http://www.bokardo.com/archives/applying\\_folksonomies/](http://www.bokardo.com/archives/applying_folksonomies/)

Powers, S. (2005). *Accidental smarts á la mode (a response to just about about any body who is interested)*. Retrieved 8 January, 2007, from <http://weblog.burningbird.net/2005/02/10/accidentalsmarts/>

Shirky, C. (2005a). *Folksonomies are a forced move: A response to Liz*. Retrieved 17 January, 2007, from [http://many.corante.com/archives/2005/01/22/folksonomies\\_are\\_a\\_forced\\_move\\_a\\_response\\_to\\_liz.php](http://many.corante.com/archives/2005/01/22/folksonomies_are_a_forced_move_a_response_to_liz.php)

Shirky, C. (2005b). *Ontology is Overrated: Categories, Links, and Tags*. Retrieved 17 January, 2007, from [http://www.shirky.com/writings/ontology\\_overrated.html](http://www.shirky.com/writings/ontology_overrated.html)

Sinha, R. (2005). *A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular)*. Retrieved 17 January, 2007, from [http://www.rashmisinha.com/archives/05\\_09/tagging-cognitive.html](http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html)

Sinha, R. (2006). *A social analysis of tagging (or how tagging transforms the solitary browsing experience into a social one)*. Retrieved 17 January, 2007, from [http://www.rashmisinha.com/archives/06\\_01/social-tagging.html](http://www.rashmisinha.com/archives/06_01/social-tagging.html)

Smith, G. (2004). *Folksonomy: social classification*. Retrieved 17 January, 2007, from [http://atomiq.org/archives/2004/08/folksonomy\\_social\\_classification.html](http://atomiq.org/archives/2004/08/folksonomy_social_classification.html)

Smith, G. (2005). *Market populism in the folksonomies debate*. Retrieved 17 January, 2007, from [http://atomiq.org/archives/2005/04/market\\_populism\\_in\\_the\\_folksonomies\\_debate.html](http://atomiq.org/archives/2005/04/market_populism_in_the_folksonomies_debate.html)

Vander Wal, T. (2005). *Folksonomy definition and Wikipedia*. Retrieved 17 January, 2007, from <http://www.vanderwal.net/random/entrysel.php?blog=1750>

Veltman, K. H. (2004). Towards a semantic web for culture [Electronic Version]. *Journal of Digital Information*, 4. Retrieved 17 January, 2007, from <http://jodi.tamu.edu/Articles/v04/i04/Veltman/>

Weinberger, D. (2006). *Taxonomies and Tags: From Trees to Piles of Leaves*. Retrieved 17 January, 2007, from [http://www.hyperorg.com/blogger/misc/taxonomies\\_and\\_tags.html](http://www.hyperorg.com/blogger/misc/taxonomies_and_tags.html)

## Author's Bio

Edith Speller works as an Assistant Librarian at London's Royal College of Music, and recently completed an MSc in Library and Information Studies at City University. Her dissertation researched affective (emotional) dimension user indexing of pop music. Edith's pre-qualification work experience includes a graduate traineeship at the London Library. Edith maintains a personal weblog (<http://elgg.net/edith/weblog/>) in which she writes about her studies as well as more general issues relating to librarianship.

