# Folksonomies and science communication

*A mash-up of professional science databases and Web 2.0 services*

Wolfgang G. Stock
*Department of Information Science, Heinrich-Heine-University Düsseldorf, Universitätsstraße 1, D-40225 Düsseldorf, Germany*
*E-mail: stock@phil-fak.uni-duesseldorf.de*

**Abstract.** Folksonomies complete the methods of indexing scientific documents. Now scientists in their function as readers may play an active role in science communication as well, since they can tag documents with terms taken from their professional or personal environment. Folksonomies allow the indexing of documents by everyone without following any rules. Besides the benefits of folksonomies there are severe problems, e.g. the tags' lack of precision. In order to overcome the shortcomings of this collaborative indexing method we introduce natural language processing of tags and a relevance ranking algorithm which is based on specific tag distributions, on aspects of collaboration and on the actions of the "prosumers". This article is a plea for the combination of the "old" science databases and the benefits of the folksonomies.

Keywords: Folksonomy, tags, online information suppliers, science communication, Web 2.0, science databases, indexing, relevance ranking, tag distribution

## 1. Introduction

In popular Web 2.0 services [22] like *Del.icio.us*, *Flickr* or *YouTube* people collaborate while creating content and indexing the documents. Now the users can play an active role in Web communication. In the sense of Toffler [35] the user turns into a "prosumer", who combines the producer's and consumer's properties. In order to index documents the prosumers apply the method of folksonomy [20], which is a kind of collaborative free keyword indexing. There are no indexing rules, everyone can tag a document with his or her favourite words. In science communication we also see some Web 2.0 services, e.g. *CiteULike* or *Connotea* [13,18]. Currently professional science databases of the "old" information industry [32] (e.g., *CAS* for chemistry, *INSPEC* for physics, *MEDLINE* or *EMBASE* for medicine, *BIOSIS* for the life sciences and *Web of Science* or *Scopus* for academic sciences in general) do not work with folksonomies. To my knowledge, only Elsevier's service *Engineering Village* supports the tagging of records. Today the user – in this case the scientist as a reader – is mostly passive. He can participate actively in his or her social science communication system by following two options: (a) using informal channels (personal communication) or (b) preparing a critical article and publishing it via a formal channel, i.e. a scientific journal. (a) is effective in deed, but limited, and (b) is a time-consuming process. Is it possible to formalize the informal channels? Can we use folksonomies to turn the scientist as reader into a scientific prosumer? Science databases are part of the formal system of scientific communication. Is it useful to improve the current databases with user-created tags? If the answer is "yes", what are the benefits and what are the shortcomings of folksonomies in science communication, what is the nature of the distribution of tags assigned to a document, and – finally – what specific information retrieval methods are necessary to process tagged documents?

## 2. Indexing scientific documents

Today's science databases make use of two indexing methods. Discipline-specific databases apply documentation languages [17] with controlled terms (classification systems, thesauri, ontologies), e.g. the Medical Subject Headings (MeSH) in *MEDLINE* or the International Patent Classification (IPC) in most of the world's patent databases. The actor whose task is to represent the document's content by using documentation languages is an interpreter, both a domain and an information specialist. The result is *one single* interpretation (surrogate) of the article in the database. Multidisciplinary science databases make use of the method of citation indexing [6]. Here the user sees the references of an article and can track both the references (backwards) and the citations (forwards). Additionally, some databases offer its users the full texts (with linked references and citations) of the documents. Here the "indexing" actors are the authors themselves, more precisely their texts containing the references. Till today, the user of scientific information is only a passive consumer; he or she does not play an active role regarding indexing scientific documents. With the aid of folksonomies the end user in formal science communication, the reader, is able to contribute to the indexing of scientific documents (Fig. 1). The user-created tags are searchable for everyone beside the interpreter-created controlled terms and the author-created text words and references. Following the basic idea of collaboration in Web 2.0 services, readers should be given the opportunity to write comments and give recommendations to single articles.

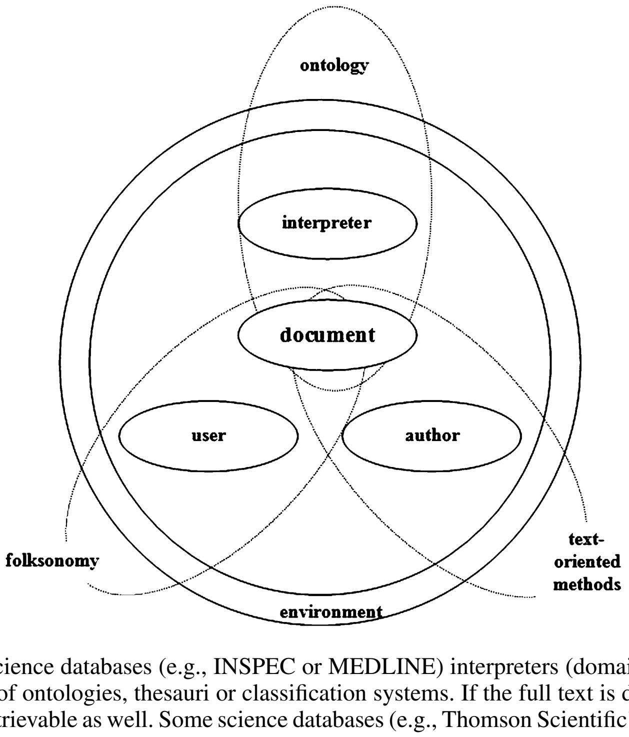


Fig. 1. In "old" professional science databases (e.g., INSPEC or MEDLINE) interpreters (domain and information specialists) index documents with the aid of ontologies, thesauri or classification systems. If the full text is digitally available, the author's text and his/her citations are retrievable as well. Some science databases (e.g., Thomson Scientific's Web of Science or Elsevier's Scopus) allow tracking the documents' references (backward) and citations (forward). Folksonomies complete the indexing environment of the scientific document. Here the users interpret the document's content and attach tags taken from his or her occupational area. Discipline- and user-specific interpretations become possible.

## 3. Folksonomy

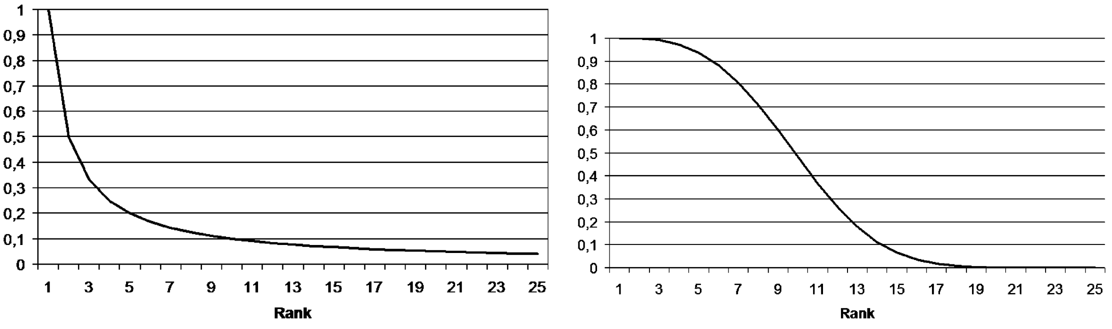
According to Smith [30] the term “folksonomy” was coined by Vander Wal in 2004 as a combination of “folk” and “taxonomy”. The reference to “taxonomy” is misleading, as folksonomies do not work with notations and relations as taxonomies, but with a flat list of uncontrolled terms. Folksonomy shows a lot of benefits: Tagging represents authentically the use of language inside scientific communities [28], it allows for multiple interpretations [27] from different disciplines or different schools, it can help to recognize neologisms and new scientific results fast. Folksonomy is by no means opposing controlled vocabularies [11]; it should be clear that the development and the updating of ontologies will profit from tagging, because tagging provides a rich source of authentic term material. If we consider documents, tags and users as nodes in a network, we can identify scientific communities which share the same topic [38]. Concerning to Shirky [29] there is an inherent kind of quality control: The more scientists tag a document, the more relevance does this article seem to have for this people. This would lead to a new scientific “currency” besides citations: The number of tags an article receives. Not unimportant is that folksonomies are cheap methods of indexing.

Folksonomies have some shortcomings. One main problem is the tags’ lack of precision [12]. In contemporary popular databases which allow for tagging we find different word forms, nouns in singular, nouns in plural, abbreviations, and misspelled words. Because there are people in different countries tagging documents, we see terms in English and in several other languages as well. There is no control of synonymy and homonymy. The users act in different contexts, have different tasks and different motivations, so there is no common basic level of indexing [9]. Most of the tags identify what the document is about, but we also find tags that describe formal aspects (“book”), give a judgment (“stupid”), are syncategorematic (“me” on the photo service *Flickr*) or plan an action (“to read”). According to Peters [23–25] it is not advisable to work with folksonomies in professional environments exclusively, but in combination with other indexing methods (which are given in science databases).

## 4. The long tail and the long trunk

Referring to Vander Wal [36] there are two kinds of folksonomies. In a broad folksonomy (applied for example to the social bookmarking service *Del.icio.us*) different readers may tag the same piece of information, so there is a document-specific distribution of tags. In narrow folksonomies (applied for example to *Technorati*, *Flickr* and *YouTube*) attached tags are recorded just once. Here we are able to count frequencies how often a document has been found by using a certain tag as a query term, so there is a document-specific distribution of (query) tags as well. This option was given in “old” databases as well (replace “tag” by “controlled term”, “class notation”, etc.), but to my knowledge it has never been used.

In the literature we find a large amount of examples for a power-law distribution of tags (and other information units) with its typical “long tail” [5, p. 293]. But that is not the only “typical” distribution. Another kind, called “inverse logistic” [33], consists of a long tail as well, but there are some tags on the left-hand side of the distribution (the “long trunk”) with (more or less) the same high values. Both distributions (Fig. 2) offer the possibility to create a new retrieval tool: in an inverse logistic distribution, all tags of the long trunk, and in a power-law distribution, the first $n$ (e.g., 3) tags are considered as “power tags”. This could be a feasible way to overcome the information overload in formal science communication.



(a) (b)

Fig. 2. Distributions of tags to a given document. The power law distribution (a) is a well-known distribution of information objects: There are only few very relevant tags and a "long tail" of more or less irrelevant or unspecific tags. Distribution (a) follows the formula $f(x) = C/x^a$, where $C$ is a constant (in the example: 1), $x$ is the rank of the tag sorted by frequency and a is a value ranging normally between about 1 and about 2 (here it is 1). The inverse logistic distribution (b) consists of a lot of relevant tags (the "long trunk") and the known "long tail". Distribution (b) follows the formula $f(x) = \mathrm{e}^{C'(x-1)^b}$ (e is the Euler number, $x$ is the rank of the tag, $C'$ is a constant and the exponent $b$ is approximately 3; in our example $b$ is 3 and $C'$ is 0.1). In most cases the "long trunk" is shorter than the "long tail". (a) Power law distribution ("long tail"). (b) Inverse logistic distribution ("long trunk – long tail").

## 5. Processing of tags for retrieval purposes

Folksonomies have shortcomings. In order to solve (some of) the problems, we should consider tags as elements of natural language and treat them by means of automatic methods of natural language processing (NLP) [34]. The algorithm follows typical NLP tasks: language identification, word identification, error detection and correction, identification of named entities, word form conflation, merging synonyms and separating homonyms (by using a thesaurus like *WordNet* [21]), and application of hierarchical relations merging the folksonomy with a classification system, a discipline-specific thesaurus or an ontology [10]. It is possible to suggest tags to users in order to avoid misspellings [19]. Not all scientific articles may be tagged already. In this case we have to create a substitution by calculating important terms with TF*IDF (term frequency, inverse document frequency) [2,3].

In current science databases the user can sort hit sets by the number of citations (e.g. in *Web of Science* and in *Scopus*) and occasionally by the frequency of the search terms in the document surrogates (again in *Web of Science*), but there is no elaborated relevance ranking [31]. Some databases make use of NLP techniques. E.g., *Web of Science* consists of KeyWords Plus [7], using "derivative indexing" [8] by processing title information in cited references.

How can we sort tagged scientific articles by relevance [26] or – concerning to Yahoo!/Flickr – by "interestingness" [4]? There are three aspects for the creation of ranking criteria (Fig. 3): (1) the tags and their distribution, (2) indicators of collaboration and (3) recommendations of the readers. The processing of tags will work adequately in a vector space model, in which the dimensions are the different tags in the database, the value of the dimension is determined by TF*IDF (1a), the documents are represented by vectors and finally the similarity of a query to a document is calculated by the cosine (1b). The term frequency TF within broad folksonomies depends on the number of readers tagging the document with the specific term, TF within narrow folksonomies is calculated by the number of queries which found the article with the assigned tag. Perhaps some of the users will be "super posters" [14,15], who index many articles. May be it could be a good idea to weight documents indexed by super posters higher than

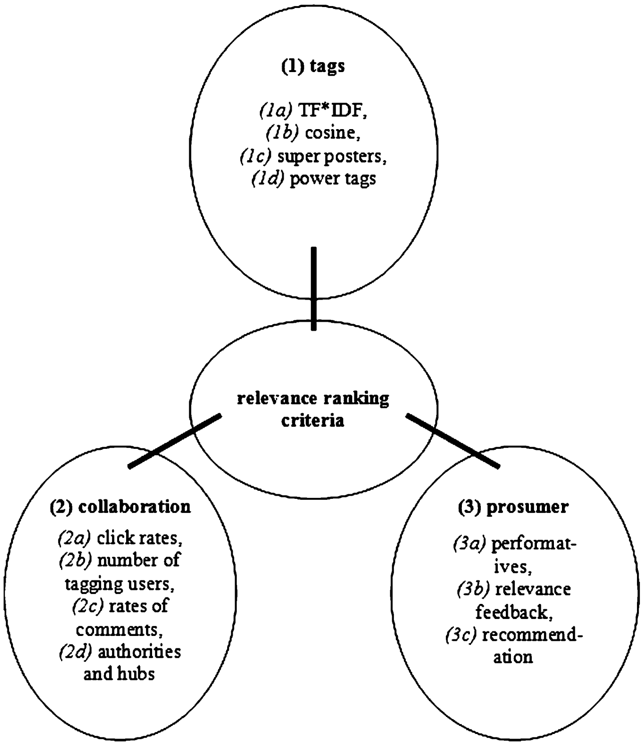


Fig. 3. The criteria for relevance ranking of folksonomy-like tagged documents consist of aspects of the assigned tags, of the Web-2.0-specific collaboration and of actions of the prosumers.

documents indexed by others (1c). Depending on the nature of the distribution, the amount of power tags is different. Such power tags should be weighted higher than all other terms (1d).

In Web 2.0 services collaboration is essential (2). So click rates (and additionally download rates) (2a), the number of tagging users (2b) and the number of comments linked to articles (2c) become criteria for relevance ranking. As scientific articles are linked by citations, it is possible to calculate hub- and authority-weights [16] or PageRanks [1] for such documents (2d).

With regard to the prosumer (3) the ranking algorithm has to weight such articles higher, to which users assign positive performative tags (e.g. “to read”) (3a). The remaining ranking criteria are relevance feedback (document marked positive or negative by readers) (3b) and explicit recommendations (3c).

All aspects try to formalize the former informal science communication. Of course, the system designers must consider that the user can choose to sort by relevance according to the “collective intelligence” [37], to pick out one specific relevance aspect or to switch off relevance ranking and sort by citation rates, by date, by author, etc.

This short articles ends with a proposal to mash-up the benefits of the “old” science databases (professional indexing, citation indexing, full-text processing) and the benefits of folksonomies (authentic language use of the readers, multiple interpretations, and new ranking options).

## Acknowledgement

The author likes to thank Isabella Peters and Jasmin Schmitz (Heinrich-Heine University Düsseldorf) for assistance.

## References

[1] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems* **30** (1998), 107–117.

[2] C.H. Brooks and N. Montanez, An analysis of the effectiveness of tagging in blogs, in: *Computation Approaches to Analyzing Weblogs. Papers from the 2006 AAAI Spring Symposium*, N. Nicolov, F. Salvetti, M. Liberman and J.H. Martin, eds, AAAI Press, Menlo Park, CA, 2006, pp. 9–15. (Technical Report SS-06-03. American Association for Artificial Intelligence.)

[3] C.H. Brooks and N. Montanez, Improved annotation of the blogosphere via autotagging and hierarchical clustering, in: *Proceedings of the 15th International World Wide Web Conference*, ACM, New York, 2006, pp. 625–632.

[4] D.S. Butterfield et al., *Interestingness Ranking of Media Objects.* Patent Application No. US 2006/0242139 A1, assignee: Yahoo! (filed: Feb. 8, 2006).

[5] L. Egghe and R. Rousseau, *Introduction to Informetrics*, Elsevier, Amsterdam, 1990.

[6] E. Garfield, *Citation Indexing*, Wiley, New York, 1979.

[7] E. Garfield, Keywords Plus™, *Current Contents* **32** (1990), 5–9; **33** (1990), 5–9.

[8] E. Garfield and I.H. Sher, Keywords Plus™ algorithmic derivative indexing, *Journal of the American Society for Information Science* **44**(5) (1993), 298–299.

[9] S.A. Golder and B.A. Huberman, Usage patterns of collaborative tagging systems, *Journal of Information Science* **32**(2) (2006), 198–208.

[10] T. Gruber, Ontology of folksonomy: A mash-up of apples and oranges, in: *1st On-Line Conference on Metadata and Semantics Research (MTSR '05)*, 2005. Online: http://tomgruber.org/writing/ontology-of-folksonomy.htm.

[11] T. Gruber, Where the social Web meets the semantic Web, *Lecture Notes in Computer Science* **4273** (2006), 994.

[12] M. Guy and E. Tonkin, Folksonomies: Tidying up tags? *D-Lib Magazine* **12**(1) (2006). Online: http://www.dlib.org/dlib/january06/guy/01guy.html.

[13] T. Hammond, T. Hannay, B. Lund and J. Scott, Social bookmarking tools. A general review. Part 1, *D-Lib Magazine* **12**(1) (2005). Online: http://www.dlib.org/dlib/april05/hammond/04hammond.html.

[14] A. Hotho, R. Jäschke, C. Schmitz and G. Stumme, Information retrieval in folksonomies: Search and ranking, *Lecture Notes in Computer Science* **4011** (2006), 411–426.

[15] A. Hotho, R. Jäschke, C. Schmitz and G. Stumme, Trend detection in folksonomies, *Lecture Notes in Computer Science* **4306** (2006), 56–70.

[16] J. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM* **46**(5) (1999), 604–632.

[17] F.W. Lancaster, *Indexing and Abstracting in Theory and Practice*, 3rd edn, University of Illinois, Champaigne, 2003.

[18] B. Lund, T. Hammond, M. Flack and T. Hannay, Social bookmarking tools (II). A case study – *Connotea*, *D-Lib Magazine* **11**(4) (2005). Online: http://www.dlib.org//dlib/april05/lund/04lund.html.

[19] M.B. MacLaurin, *Selection-Based Item Tagging.* Patent Application No. US 2007/002871 A1, assignee: Microsoft (filed: Jul. 29, 2005).

[20] A. Mathes, *Folksonomies – Cooperative Classification and Communication Through Shared Metadata*, University of Illinois Urbana-Campaign/Graduate School of Library and Information Science, Urbana, IL, 2004. Online: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[21] G.A. Miller, Nouns in WordNet, in: *WordNet. An Electronic Lexical Database*, C. Fellbaum, ed., MIT Press, Cambridge, MA, London, 1998, pp. 23–46.

[22] T. O'Reilly, *What is Web 2.0. Design patterns and business models for the next generation of software*, 2005. Online: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[23] I. Peters, Inhaltserschließung von Blogs und Podcasts im betrieblichen Wissensmanagement (Indexing of blogs and podcasts for the corporate knowledge management), in: *Content. 28. Online-Tagung der DGI, 58. Jahrestagung der DGI. Proceedings*, M. Ockenfeld, ed., DGI, Frankfurt, 2006, pp. 143–151.

[24] I. Peters, Against folksonomies: Indexing blogs and podcasts for corporate knowledge management, in: *Preparing for Information 2.0. Online Information 2006. Proceedings*, H. Jezzard, ed., Learned Information Europe, London, 2006, pp. 93–97.

[25] I. Peters and W.G. Stock, Corporate Blogs im Wissensmanagement (Corporate blogs used in knowledge management), *Wissensmanagement* **6** (2006), 40–41.

[26] I. Peters and W.G. Stock, Folksonomy and information retrieval, in: *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology (Vol. 45)*, 2007, CD-ROM.

[27] E. Peterson, Beneath the metadata. Some philosophical problems with folksonomies, *D-Lib Magazine* **12**(11) (2006). Online: http://www.dlib.org/dlib/november06/peterson/11peterson.html.

[28] E. Quintarelli, Folksonomies: Power to the people. Paper presented at the *ISKO Italy UniMIB meeting*, Milan, June 24, 2005. Online: http://www.iskoi.org/doc/folksonomies.htm.

[29] C. Shirky, *Ontology is Overrated: Categories, Links, and Tags*, 2005. Online: http://www.shirky.com/writings/ontology_overrated.html.

[30] G. Smith, Folksonomy: Social classification. [Blog post; 2004-08-03.] Online: http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

[31] M. Stock and W.G. Stock, Online-Hosts für Wissenschaft, Technik und Medizin auf dem deutschen Informationsmarkt [Online-suppliers for science, technology and medicine on the German information market], *Password* **2** (2005), 18–23. Online: http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/1/11109682372_2005_wtm.pdf.

[32] W.G. Stock, *Informationswirtschaft* [Information industry], Oldenbourg, München, Wien, 2000.

[33] W.G. Stock, On relevance distributions, *Journal of the American Society for Information Science and Technology* **57**(8) (2006), 1126–1129.

[34] W.G. Stock, *Information Retrieval. Informationen suchen und finden* [Information Retrieval. Searching and Finding Information], Oldenbourg, München, Wien, 2007.

[35] A. Toffler, *The Third Wave*, Morrow, New York, 1980.

[36] T. Vander Wal, Explaining and showing broad and narrow folksonomies. [Blog post 2005-02-21.] Online: http://www.vanderwal.net/random/category.php?cat=153.

[37] A. Weiss, The power of collective intelligence, *netWorker* **9**(3) (2005), 16–23.

[38] H. Wu, M. Zubair and K. Maly, Harvesting Social Knowledge from Folksonomies, in: *Proceedings of the 17th Conference on Hypertext and Hypermedia*, ACM, New York, 2006, pp. 111–114.