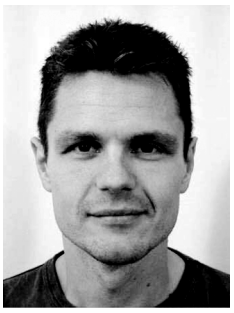
# Folksonomies, Tagging Communities, and Tagging Strategies– An Empirical Study †

Timme Bisgaard Munk* and Kristian Mørk**

*Faculty of Humanities, University of Copenhagen, Njalsgade 80, DK-2300, Copenhagen S, Denmark, <timme@kforum.dk>

**Danish Broadcasting Corporation, Emil Holms Kanal 20, 0999 Copenhagen C, Denmark,<krmr@dr.dk>

Timme Bisgaard Munk is a Ph.D. student at the University of Copenhagen in Denmark. He is currently working on a thesis about how Web 2.0 technologies can be used to improve knowledge sharing. He has written several articles on information architecture, the semantic web and a book about Human-Computer Interaction together with Kristian Mørk (*Brugervenlighed påinternettet: en introduction*. Frederiksberg: Samfundslitteratur, 2002.). Finally, he is the editor of the Danish online journal *Kommunikationsforum* (http://www.kommunikationsforum.dk/) on new media and technology.



Kristian Mørk has a Master of Comparative Literature from the University of Aarhus in Denmark. He works as a team leader in the web development department of the Danish Broadcasting Corporation (DR). He has written several articles about the Internet and the semantic web and has published a book on human-computer interaction together with Timme Bisgaard Munk (*Brugervenlighed påinternettet: en introduction*. Frederiksberg: Samfundslitteratur, 2002.). Finally, he is the editor of Denmark's largest film guide on the Internet *Scope* (http://www.scope.dk/).

† Together, they have written an earlier article: "Folksonomy, the power law & the significance of the least effort" (*Knowledge organization* 34(1), 2007: 16-33) on folksonomies and the computer program *del.icio.us*.This article is the empirical part of their joint study of folksonomies and the basis of their theoretical contemplations about the phenomenon.

Munk, Timme Bisgaard and Mørk, Kristian. **Folksonomies, Tagging Communities, and Tagging Strategies—An Empirical Study**. *Knowledge Organization, 34(3)*, 115-127. 23 references.

**ABSTRACT:** The subject of this article is folksonomies on the Internet. One of the largest folksonomies on the Internet in terms of number of users and tagged websites is the computer program *del.icio.us*, where more than 100,000 people have tagged the websites that they and others find using their own keywords. How this is done in practice and the patterns to be found are the focus of this article. The empirical basis is the collection of 76,601 different keywords with a total frequency of 178,215 from 500 randomly chosen taggers on *del.icio.us* at the end of 2005. The keywords collected were then analyzed quantitatively statistically by uncovering their frequency and percentage distribution and through a statistical correspondence analysis in order to uncover possible patterns in the users' tags. Subsequently, a qualitative textual analysis of the tags was made in order to find out by analysis which tagging strategies are represented in the data material. This led to four conclusions. 1) the distribution of keywords follows classic power law; 2) distinct tagging communities are identifiable; 3) the most frequently used tags are situated on a general-specific axis; and 4) nine distinct tagging strategies are observed. These four conclusions are put into perspective collectively in respect of a number of more general and theoretical considerations concerning folksonomies and the classification systems of the future.

## 1. Introduction

The subject of this article is folksonomies on the Internet. Folksonomies are user-created taxonomies where the users themselves are free to create descriptive meta-data for the tagging of data. One of the largest folksonomies on the Internet in terms of number of users and tagged websites is the computer program *del.icio.us*. Here, more than 100,000 people have tagged the websites that they and others find on the Internet using their own keywords. How this is done in practice and the patterns to be found are the focus of this article. The empirical basis is the collection of 76,601 different keywords with a total frequency of 178,215 from 500 randomly chosen taggers on *del.icio.us* at the end of 2005. The keywords collected were then analyzed quantitatively statistically by uncovering their frequency and percentage distribution and through a statistical correspondence analysis in order to uncover possible patterns in the users' tags. Subsequently, a qualitative textual analysis of the tags was made in order to find out by analysis which tagging strategies are represented in the data material. The tagging strategies found by analysis were then sorted into eight groups and processed statistically to see how frequently they appear among the 245 most popular keywords and in the complete data material with 178,215 tags. This leads to four conclusions:

**Firstly**, the distribution of keywords in *del.icio.us*, as in many other complex systems, follows the classic power law where very few keywords are dominant. These keywords are primarily the so-called cognitive basic categories and essentially consist of a number of very broad and general content categories that are common to all people or common to the people working professionally in the IT field. Categories which everyone may use directly and which in the cognitive sense represent the fastest, most basic and most economical creation of a category.

**Secondly**, statistically, there are three distinct tagging communities in *del.icio.us* which may subsequently be described qualitatively as three different types of taggers having different interests and tagging strategies: The well-informed and curious citizen, who tags in very broad common cultural categories, the professional IT worker, who tags in a number of very specific IT-related technical categories, and the professional IT designer, who tags with a number of specific design-related terms.

**Thirdly**, the 245 most frequently used tags are situated along an axis from general societal subjects to specific IT concepts. What is significant for what and how the websites are categorized in respect of keywords.

**Fourthly**, there are nine distinct tagging strategies that constitute a repertoire of tagging strategies for all taggers, where broad content categorization is dominant followed by formalistic media categorization, process categorization and meta-categorization.

Finally, these four conclusions are put into perspective collectively in respect of a number of more general and theoretical considerations concerning folksonomies and the classification systems of the future in continuation of the authors' previously published theoretical article on folksonomies.

## 2. What, When and How have the Empirical Data been Collected?

For this empirical study, 76,601 tags were collected from 500 random users in the *del.icio.us* system in the period from 30 November to 1 December 2005. The total number of tags is 178,215. Technically, the tags were collected by having a computer program record every time a tagger among the more than 100,000 participants tagged a website in the computer program during the above period. Subsequently, the computer program copied the tagger's entire list of tags and recorded it as individual tags. For this reason, the collection of tags does not comprise sequences of tags in respect of different websites, just isolated keywords. The technical collection process was then repeated until tags from 500 taggers were collected. Thus, the 500 taggers have only been chosen because they were tagging a website in the *del.icio.us* program during the collection period. 500 taggers is a limited amount of people relative to the total number of more then 100,000 taggers in *del.icio.us*. For this reason, the result of the study may not be generalized uncritically as applying to all taggers in the system. The objective is only to find tendencies in the material, not statistical interrelationships for the entire *del.icio.us* database in the strict sense. The reason for selecting 500 taggers is to

have a manageable amount of tags which is still sufficient to be able to come to any conclusions regarding possible tendencies in the use of keywords.

The material collected was subsequently entered in a database and processed statistically in order to chart frequency, percentage distribution and to make a correspondence analysis. In addition to the statistical material we have collected ourselves, a number of publicly available statistics of taggers, keywords used and websites tagged in *del.icio.us* have been used in the article. These statistics are either available from the *del.icio.us* website or from separate statistics sites for the program. This statistical material is, of course, used subject to reservations since the collection method used for these official lists is not scientifically qualified and since it is not always clear on which terms and how these figures have been collected. This material should also be regarded as material indicating the tendency in the database.

## 3. The general pattern in the tags

It is clear in the *del.icio.us* system that the user-generated tagging does not result in anarchy or coincidence, but tends towards a consensus on specific words related to a specific website over time. This will be obvious to any user of the system. You would think that the opportunity of choosing tags freely would create a chaotic and idiosyncratic mass of tags without a clear pattern. However, figures show that this is not the case. On the contrary, a number of patterns relating to frequency, interpretation communities and relations between the different tags can be seen in the *del.icio.us* database and among over 100,000 users. There is regularity in the way the tags are used, their frequency and the relative distribution of frequency between the different tags. The overall pattern that can be demonstrated is that tagging follows the so-called power law, which has also been pointed out in other studies (Shirky 2003; Cozy 2005; Shen and Wu 2005). As mentioned, the power law results in some tags being very dominant and frequently used. It is the broad and general content-categorized keywords that dominate, while the majority of tags are peripheral and used very rarely. The statistical evidence that the randomized 178,215 tags follow the mathematical power-law distribution is provided in the graph below (Figure 1).

The graph shows a power law with an exponent (i.e. slope) of 1. This is interesting, because it shows that the frequency of the chosen tags does not fall as quickly as in natural language (an exponent of 1.5) (Vogt 2000). In other words, tags are more distrib-

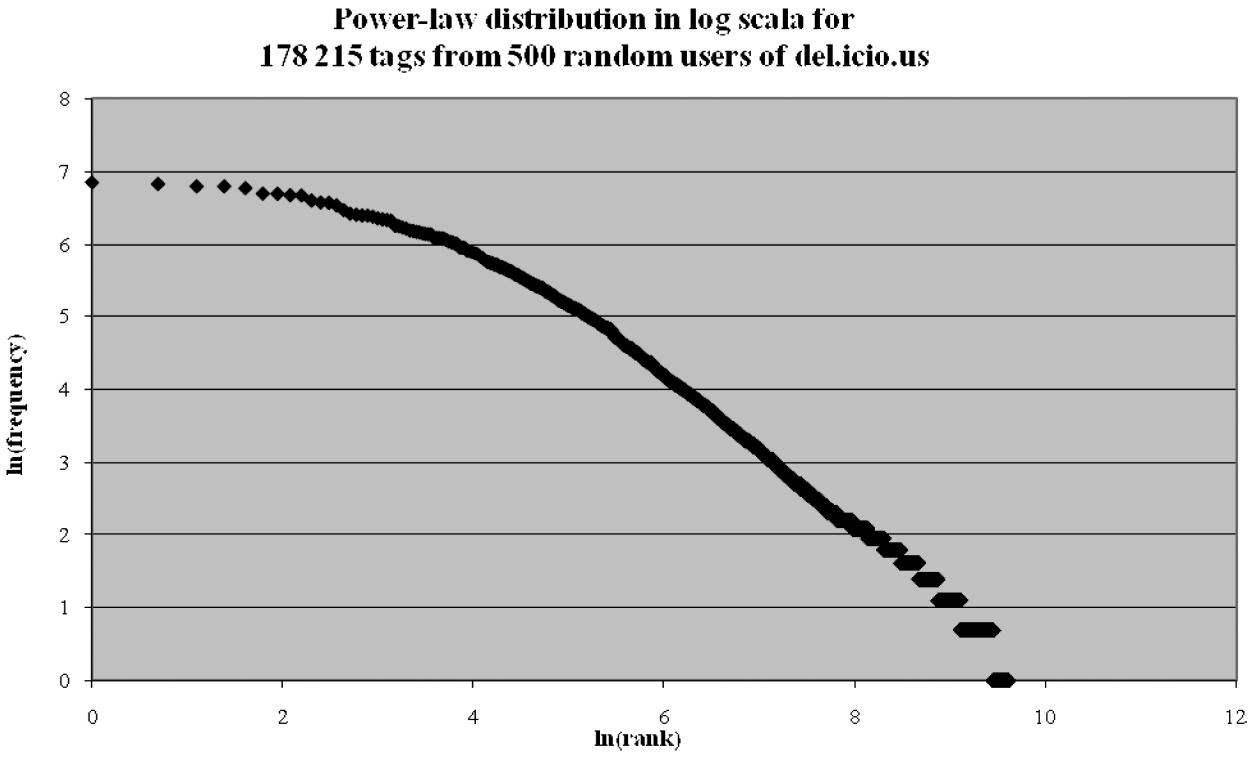



*Figure 1.*

uted than they would be if a person were to write a text (Zipf 1949). However, this is not strange, because the tags have a certain encyclopaedic character as expressions and language action reflecting the many different interests of the users. An encyclopaedic list or a dictionary has an exponent of 0. Consequently, this graph shows that the *del.icio.us* users are relatively heterogeneous and more heterogeneous than the typical language users in the written language (exponent of 1.5). On the other hand, if you take a closer look at some of the 178,215 tags, it is clear to see that they are "strange", i.e. that they refer to certain acronym traditions, IT-related catchphrases and IT slang. They are certainly not everyday words. This means that there are groups of people who have a specific tag language and specific communities.

In order to find out whether such tagging communities exist, tags from the 500 users been processed through a statistical correspondence analysis of the tags (Hill 1974; Tabachnik 1989; Greenacre 1993). This method has been chosen because it is especially suited for uncovering possible patterns in data material with many variables and with many variables relative to many users. The correspondence analysis was developed by the French mathematician Jean-Paul Benzécri as a tool to analyze systematic relations between data (Benzécri 1980; Benzécri 1992). The analysis does not require any assumptions on distribution or scale properties, as all variables are considered to be nominal or categorical variables. The correspondence principle is, in short, that it generates a profile for each row and column in a cross table, and the analysis results in a map where similar profiles end up close to each other and different profiles far away from each other. The method constructs a profile for each of the subjects of the analysis (in this case the 500 tagging users and the 245 most frequent tags). As a part of this method, the data collected and patterns found are subsequently used to create a qualitative profile description of the types of taggers found among the 500 taggers (Benzécri 1992; Blasius and Greenacre 1998). The 245 have been chosen because the most frequent are also the most interesting in light of the hypothesis and strategy of the analysis of uncovering a number of interpretation communities. On the basis of a table of the profiles, the analysis constructs a map, where users/tags with similar profiles are situated close together, while users with more diversified profiles are situated far away from each other. The closer to the corners they are situated, the more significant they are statistically (Hill 1974; Tabachnik 1989). The map may in this sense be regarded as a visualization of statistic over-representation for both users and tags (for a detailed account of the underlying mathematics, please refer to (Hill 1974; Greenacre 1993)). A mathematical map has been created of the 245 most frequent tags among the 500 users (Figure 2).

The conclusion is that the predominant tagging strategy is a content categorization. The patterns of statistical visualizations show that there is a left-right axis from general use to IT-specific use. On the left side, very general common basic cultural categories which all people share, such as travel, politics and culture, are used. On the right side, there are a number of very specific IT-related keywords connected to certain IT-related fields and work functions such as Apache (database technology/language), Perl (database language) and SQL (database system). At the top of the map is a cluster of keywords describing a number of graphical and design-related words and concepts. All words connected with the work of a graphical designer and the graphical layout of websites. The statistical analysis thus indicates that there are three different significant interpretation communities in the system. They have been described qualitatively below on the basis of the static clusters of keywords, where the individual keywords are used to describe a fictive person behind the behaviour.

### *3.1. The Well-Informed Citizen—"The Curious Citizen"*

(See Figure 2, on the left, in a red circle.) The purpose is to obtain general information within a number of wide social subjects. The categories activated are the general human categories existing across social boundaries and interests. It is the dedicated citizen who tags about his or her hobbies and to fulfil a general need to "keep up with developments/the Internet." The motivation is curiosity regarding what is going on in society, and the contextual framework is a combination of spare time and work.

### *3.2. The Professional IT Worker—"The Professional IT Expert"*

(See Figure 2, on the right, in a green circle.) The purpose here is primarily work-related. The objective is to keep up-to-date with programs and technologies within narrowly defined IT fields. In this case, database technology. This is a professional IT expert


*Figure 2.*

working with very heavy and complicated database solutions. In all probability, a man who is not interested in the "softer" part of the IT world relating to graphics and user experience. Here, the focus is on the basic structure of the system and the possibilities provided by technology. He is not interested in the system as presentation, rather as representation and structure, focusing on the underlying structures—the so-called back-end. His basic categories are at a completely different level and much more specifically related to a professional community. The motivation is to have an updated toolbox containing knowledge on specific technologies, and the framework is specifically work-related.

### *3.3. The Professional IT Designer—"The Soft and General IT Expert"*

(See Figure 2, in the centre, in a blue circle.) Here, the purpose is to stay up-to-date with developments within graphical design and the design of websites. Here, pictures, icons and fonts for the graphical work in programs such as *Photoshop* are collected. This is probably a woman who is interested in the meeting between technology, people and graphical design. The tendency is to use broader and more diffuse categories which are not only related to specific narrowly defined subjects (e.g., freeware, Webdev article). Here, the focus is on the IT system as presentation—the so-called front-end. Her basic categories are related to the work with graphical design and the most commonly used keywords in that world. Her motivation is to have an updated and shared important knowledge of graphical design and web design.

## 4. Which Websites are Tagged in *del.icio.us*?

Information on the tendency in the database may be drawn from looking at the fifty websites since the launch of *del.icio.us*. Even though it must be noted that the more than 100,000 users have different reasons for sharing their bookmarks with others, there is still a tendency that *del.icio.us* is essentially a community for people with an interest in or who work with information technology. Bookmarks to IT-related pages and the use of IT-related tags are thus over-represented. Considering that folksonomies is a relatively new concept on the Internet, this characteristic is not surprising, since the professional users of the medium will typically be the frontrunners in connection with new technology, software and opportunities.

The content may be divided into the following categories on the basis of a qualitative analysis of the fifty most popular websites since the launch of *del.icio.us* (*Populicio* http://populicio.us/fulltotal.html). The different websites fall into four groups, where the first group is the largest one by volume according to number and popularity. These are:

**How-to** websites/tutorials on a number of popular computer programs and technologies. All websites that can help the user solve concrete IT-related jobs.
**Websites** that can help the users make better use of the *del.icio.us* system.
**Websites** which in one way or another live up to one or more of the classic news criteria in respect of the IT field as a profession. This is information which may broadly be perceived as being news within the IT field. Here, there is an over-representation of websites that may be considered to be quaint news, ideas or concepts.
**Websites** which constitute a digital public such as digital debate media and may be perceived as joint forums for users with an interest in IT.

The most important function of the system is therefore to support knowledge sharing and help the users share tools and knowledge about how to solve different IT jobs. This also includes sharing new knowledge (news) in respect of specific IT-related jobs.

## 5. Which Tags are Dominant in *del.icio.us*?

If you look at the 87 most popular keywords for all more than 100, 000 users as listed on the *del.icio.us* website, and thus the strongest tags in the system at any given time, the image of a database for people interested in IT created by people interested in IT is reinforced. The users here tend to have a far more varied and broad repertoire of keywords on IT than on other categories in the system. In addition, the creation of hierarchies through sub-categorizations and synonyms appear more often with IT-related bookmarks, while other categories outside the IT area appear as detached and without sub-categories. The following list, which was retrieved from *del.icio.us* in mid-October 2005, shows the most frequently used tags sorted by frequency (http://*Del.icio.us*/tag/?sort=freq):

> Blog programming software web design reference music news ajax tols linux javascript css howto web2.0 firefox art blogs politics games webdesign shopping technology humor business google science photography tutorial mac search tips fun java tech windows opensource video development free internet daily toread flash diy security mp3 osx ruby php hardware books funny cool media travel article productivity language webdev *del.icio.us* research comics microsoft photo education downloads maps computer work tv email blogging apple culture electronics rss podcast audio hacks food wiki movies religion freeware geek rails community graphics entertainment writing architecture game download inspiration shop extensions html tutorials history library photos philosophy radio python code .net tool hosting lifehacks.

Generally, the most frequently used keywords may be divided into two groups. Firstly, very broad and very general keywords on cultural and societal, generally human categories (music, politics, science). Secondly, a number of keywords which may all be regarded as a sub-category under terms such as computers or the Internet. From this list, it may be seen that tags which are practically related to the solution of concrete IT tasks are predominant (e.g., tools, tips, howto, tutorial, tutorials), while terms and programs related thereto are also very predominant, including a number of IT-related abbreviations (html, osx). There are also a number of job-based tags which are very active (such as: toread, 2read, read later, to do) and a number of value judgments (funny, cool). Another interesting observation is that the most popular keywords are very general and may be said to belong to a number of cognitive basic categories which many people working professionally in the IT field will consider if asked to mention some IT terms off the top of their heads (Rosch and B.Caroly 1975; Rosch 1976; Mervis and Crisafi 1982; Jolicocur and Kosslyn 1984; Tanaka and Taylor 1991; Vogt 2000).

The following is the list of the 245 most frequently used tags sorted by frequency among the 500 randomly chosen taggers in our data material for comparison:

> Blog software web music design CSS news programming tools google Linux reference javascript ajax flash RSS system:unfiled art search books php video tutorial security Firefox internet webdesign java blogs science howto wiki XML photography opensource windows free mp3 HTML games technology FlickR tips *del.icio.us* fun business web2.0 ruby graphics photo TV email development media mac ipod history hardware politics travel photos maps tech Microsoft food audio radio mobile writing language cms social shopping python community research apple hacks database humor marketing fonts photoshop health p2p funny network tutorials game images podcast osx movies book wordpress perl productivity gtd architecture computer blogging library webdev cool culture download dvd bittorrent unix hosting education magazine forum usability freeware English code gmail apache management japan mysql religion tags browser hack rails geek diy map math comics work article plugin Wikipedia bookmarks backup tool delicious itunes links learning jobs privacy api seo editor folksonomy online pdf typography information sql networking daily copyright calendar animation spam movie tagging Yahoo collaboration podcasting color illustration reviews Mozilla xhtml literature mail shop os framework portal voip dhtml philosophy patternsUK server amazon gui standards gadgets wirelessimage lifehacks utilities film toread gallery wifi hacking humour psychology finance money life people statistics advertising greasemonkey computers extension phoneservice torrent interview skype testing environment eclipse layout London China 3d dictionary im xp debian documentation guide resources electronics inspiration icons ui desktop accessibility cooking physics reading toys storage Generator performance recipes archive faq law organization subversion xtensions

Our study showed that the 245 most frequently used tags were used 66,398 times, corresponding to 37.2 per cent of the total number of keywords. This is a substantial percentage, which only proves statistically that a few broad keywords are dominant. In a cognitive sense, these keywords may all be described as the shortest possible and most economical encoding of a relevant category (Chater 1999; Chater and Pothos 2002; Chater and Hurley 2005). These are categories that everyone could think of without giving it much thought (Tanaka and Taylor 1991). Again, we see the same patterns as in *del.icio.us*' own official list of the 87 most frequent tags. Broad basic cultural content categories are predominant (such as: music, news, design), while broad basic categories for everyone working professionally with IT are also frequent

(such as: programming, Linux, javascript, flash). The many tags referring to how to solve tasks (such as: todo, learning, tutorial) are also pronounced. This yet again confirms that for many of the users, *del.icio.us* is a learning community and a knowledge resource where education and learning are important, especially in respect of IT technology. The list also reveals that the formalistic media technology categorization is a popular way of categorizing websites by means of tagging (such as: photo, video, mp3, maps, or books). On the lists of the most frequently used tags, there are also several examples of media genre categorizations where it is not the media technology but the genre that determines the choice of categorization (such as: games, blogs, articles, Faq, or reviews). There are also many tags describing a process in connection with the tagging (toread, howto) or expressing a positive assessment (funny, fun, humour). It is surprising that meta-references to tagging as a process (tag, bookmarks, folksonomy, links) are so prominent since they do not say anything essential about what is tagged, but only refer to the context of meaning which should be a given. In both cases, no information is added which is not part of the implicit context. A possible explanation is that the use of meta-tags refers to the many websites providing advice and services regarding how to make better use of *del.icio.us*. That is, meta-knowledge and meta-services in respect of *del.icio.us*.

## 6. The Different Tagging Strategies used in *del.icio.us*

Through a quantitative analysis of the 178,215 tags collected and the list of the 245 most frequently used tags in the database, it is possible to prove the existence of nine distinct tagging strategies in the database which constitute the entire repertoire of possible tagging strategies. This conclusion is confirmed by other statistical studies. However, these nine distinct strategies are not identical to the seven strategies defined in Golder and Huberman's study (Golder and Huberman 2005). The difference is that media, genre, time and egocentric categorization have been added, while subcategorization is not perceived as a separate strategy and an essential quality of the keyword but as a relationship between keywords used in content categorization. For this reason, it has been removed from our list of tagging strategies. Our studies of the tags used shows that the different tags and tagging strategies are either based on content, medium, genre, copyright, value judgments, meta-reference, process description or egocentric. The different strategies may be described as follows:

1. **Tagging is content categorization.** Categorization through content with different degrees of specification relative to different dimensions. This is by far the most popular categorization method. Here, tagging marks how the content of the bookmark may be described as belonging to a number of content categories such as: business, marketing and web design. Either societal categories or IT categories. The subject categories are often very broad. Here, the horizontal structure of scattered keywords is sometimes replaced by hierarchizations and relations between the individual tags through tagging in the content categorization, where a combination of tags provides a more precise categorisation such as: business, marketing, branding.

2. **Tagging is media categorization.** This takes the form of a formalistic reference to the medium which the bookmark concerns or in which it appears. Here, it is not the content that is described but the form of provision of information. This may, for example, be a reference to media such as book, TV or newspaper or the media technology used such as photo, film.

3. **Tagging is categorization subject to copyright.** This takes the form of a reference to ownership or legal rights. In most cases as a categorization referring to information, the program or the resource is free of charge such as *free*, *freeware*, *open source*.

4. **Tagging is a type of categorization.** This is a categorization where the tagged website is put in relation to different textual genres or text types. This may, for example, take place by tagging the website with different keywords connected with different textual genres such as *Wishlist*, *academic paper*, *application* or *reference list*. This is a formalistic categorization which says nothing about the content of the tagged resource.

5. **Tagging is categorization through a value judgment.** Here, a value judgment on the bookmark is given. There are, for example, often positive value judgments: cute, funny, nice or cool. Here, the

bookmark is related to personal references and values.

6. **Tagging is categorization referring to the categorization.** This meta-reflexive categorization takes place through a meta-reference to tagging as an action or structure or to the social bookmarking system as such. This could, for example, take the form of a reference to tagging as an individual action: Mytags, worktags, tagging or *del.icio.us.*

7. **Tagging is categorization on the basis of jobs and processes.** There is often a reference to the bookmark in relation to a future job. Examples include: Toread later, checkout later.

8. **Tagging is a mark in time.** Here, the objective is tagging where categorization identifies the bookmark's content in time in relation to the tagger's time abstract such as: news, old, history, future, or personally relating to process, such as: remember.

9. **Tagging is an exclusive personal categorization.** Here, tags which only make sense to the taggers are often used. We often see abbreviations or combinations of figures and letters. This strategy often conceals one of the other strategies, but the motive and the strategy stay hidden because the meaning is personal. The material collected shows that it is often used to a very limited extent in the database. For this reason, it has been left out of the following analysis because it is so infrequently used and is thus not particularly prominent in the material.

These different tagging strategies must be perceived as ideal-typical descriptions, where it may be hard in practice to set clear boundaries between the different strategies. They may, for example, appear in a combination as several tags for the same website or in the same words as hybrid tagging strategies, where a content categorization is also a job description, and in some cases, the different strategies may be subsets of each other. Furthermore, it must be noted that the individual tags may often cover different motives pointing further than their immediate meaning. In the strict sense, tagging of the word RSS is a media categorization of a bookmark on the RSS technology; however, the purpose may be to collect everything new and relevant on RSS in respect of a certain job. Thus, the meaning is really: "Exciting news (→ value judgment) on RSS which must be read (→ job categorization)". The meaning of the word may thus be far more complex in the tagger's consciousness than what may immediately be seen from the presentation form as a tagging strategy. In the same way categories are essential ambiguous, and finding their actual meaning would require that you ask all the 500 people personally about all their tags.

## 7. Frequency of the Different Tagging Strategies in *del.icio.us*

For these reasons it is difficult to quantify the frequency of the nine tagging strategies, as most keywords may cover more than one strategy and different motives. The strategy used is thus impossible to demonstrate statistically and with any certainty. In our study design, we have manually sorted 76,601 tags used into eight categories on the basis of our perception of which category they belonged to. This could be a potential source of error and a tendency which impairs the validity of the result of the study. It is far from being the ideal methodical solution, but it is the most expedient considering the volume of data material and the mass of possible methodical approaches to categorizing the 76,601 keywords. Sorting the keywords collected on the basis of the nine strategies thus only expresses a tendency in the material. In the following, the 245 most frequently used tags and the entire mass of tags were sorted according to the eight strategies. The number of strategies in this sorting has been reduced to eight although there are nine strategies. The reason for this is that the personal and egocentric strategy is, of course, not represented among the 245 most frequently used tags for all users. However, in the graph of all tags, the personal categorization has been removed because it is a strategy implicitly covered by one of the other strategies, although it is impossible to determine which one. Furthermore, it is so rarely represented in the entire database that it is not appreciable.

With regard to the data material, the reader should also note that the tendency among the 245 most frequently used keywords and the tendency in the entire database do not necessarily show how the nine strategies are distributed in respect of the frequency of a randomly chosen tagger's use of the different tagging strategies. The result is only an expression of the tendencies applying to the entire group of all taggers or more precisely to around a third of all the tags. With these reservations, the graph of the distribution of the eight tagging strategies relative to the

245 most frequently used tags would look as shown in Figure 3.

Then, with a sorting of the eight strategies according to frequency (all 178,215) in the entire database, the resulting graph looks as shown in Figure 4.

Subject to the mentioned reservations, the conclusion is that the tendency is that the classic content categorization is dominant, followed by formalistic media categorization with genre categorization and process categorization in third place. In the top 245, process categorization is the third most used strategy, while genre categorization is in third place in the entire database. This difference is possibly due to the fact that in the top 245, the ranking is only an expression of the relational difference between tags with a high frequency, while all the possible genres are in-

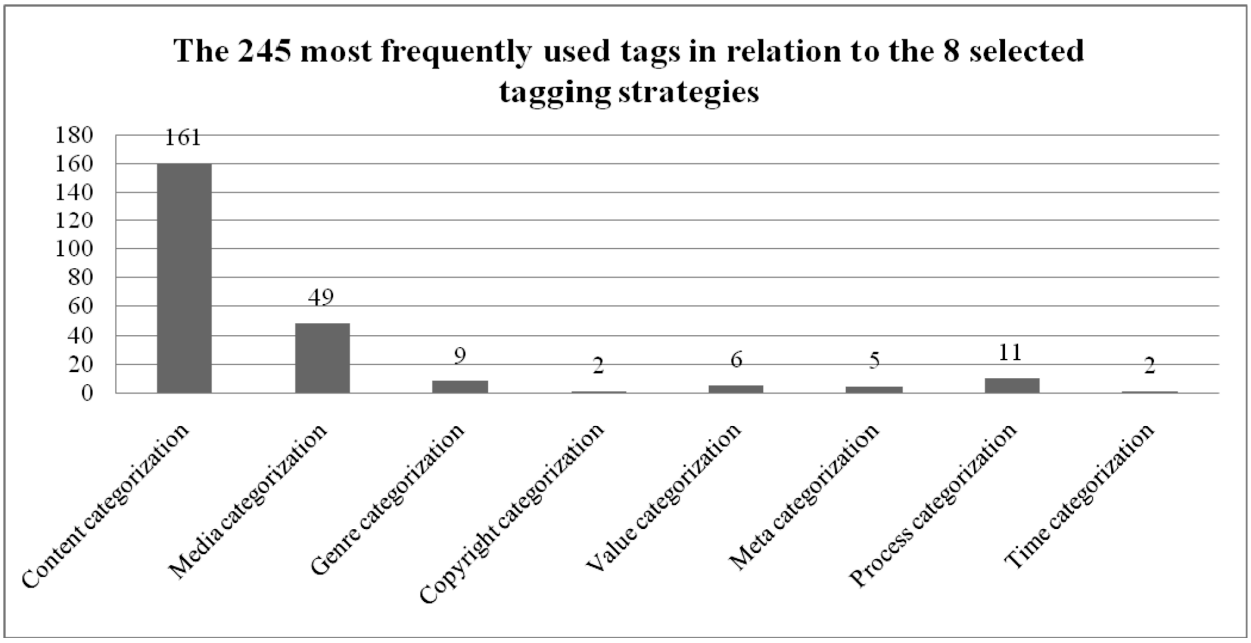



*Figure 3.*

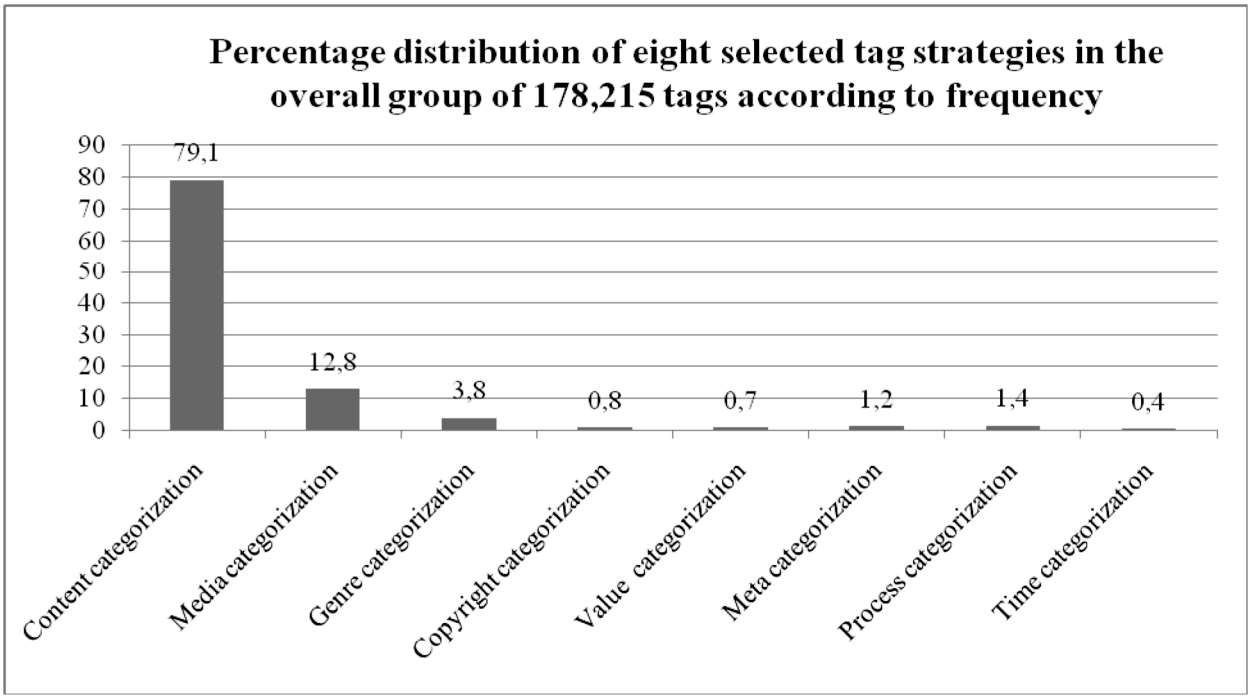



*Figure 4.*

cluded in the entire database, also the ones with a low frequency. However, process categorization has a more limited and defined set of possible conceptualizations, which results in a lower frequency in the entire database. Meanwhile, this difference does not change the fact that they are among the preferred tagging strategies. Considering these differences, it is possible to come to a conclusion on the underlying tendencies and cognitive questions which the tagger attempts to answer through the use of different keywords. This may also be perceived as the total repertoire of motives behind the different strategies.

The overall conclusion in respect of the results that may be read from both graphs is that tagging essentially tends to be about the keyword answering the following questions ranked by importance, where questions 3 and 4 are on the same ranking:

1. What is the content of the tagged information in general, described in very broad cognitive basic categories?

2. What is the media format, media type and media technology used to store the tagged information?

3. Which media genre may describe the content?

4. What do I as a tagger wish to do with the information, and in which contextual processes is the tagging included?

5. What is my immediate, often positive, emotional perception of the tagged information?

6. How is the tagging connected to the process of tagging information through a number of meta-references?

7. How may the tagged information be categorized in time, not least in relation to the taggers' own personal time?

## 8. The Overall Conclusion from the Empirical Study of the *del.icio.us* Folksonomy

On the basis of our quantitative and qualitative empirical study of the *del.icio.us* folksonomy, we have come to the following five conclusions:

**Firstly**, it is possible to demonstrate statistically that the distribution of keywords in *del.icio.us*, as in many other complex systems, follows the classic power law where very few keywords are dominant. These keywords are primarily the so-called cognitive basic categories and essentially consist of a number of very broad and general content categories that are common to all people or common to the people working professionally in the IT field. This creates a number of structural limits for what may be conceptualized and the depth of categorization. The power law, the comprehensive and influential profiling of the many keywords which, in the cognitive sense, is the shortest possible and most economical encoding of a category, suggests that cognitive economizing and information cascades influence the choice of tagging strategy (Bikchandani, Hirshleifer et al. 1992; Bikhchandani and Welch 1998; Chater 1999; Chater and Pothos 2002; Chater and Hurley 2005) We have previously dealt with this issue in the theoretical discussion of folksonomies and *del.icio.us* (Munk and Mørk 2007).

**Secondly**, statistically, there are three distinct tagging communities in *del.icio.us* which may be described qualitatively as three different types of taggers having different interests and tagging strategies: The well-informed and curious citizen, who tags in very broad common cultural categories, the professional IT worker, who tags in a number of very specific IT-related technical categories, and the professional IT designer, who tags with a number of specific design-related terms. If you belong to one of these three types of interpretation communities, you benefit more from folksonomy than if you do not belong to these three interpretation communities. This is due to the fact that categorization is more varied in respect of these three interpretation communities, and the knowledge resources collected support the interests and learning needs of these three professional communities. Which interpretation communities are dominant, and whether you belong to these communities and will accept the specific contextual purpose of folksonomy (in *del.icio.us*, primarily IT-related updating and IT-related self-help), thus, probably influence how the individual user benefits from folksonomies.

**Thirdly**, the 245 most frequently used tags are situated along an axis from general societal subjects to specific IT concepts. This also means that your benefit from folksonomy depends on

whether your own way of categorizing the world and information is identical to the categories which are dominant in folksonomy and the immediate relation between the specific IT concepts and the general societal subjects. In other words, it depends on whether the societal categories are too broad for your societal commitment and the IT-specific concepts too technical and incomprehensible relative to your need for updating and professional IT help.

**Fourthly**, there are nine distinct tagging strategies that constitute a repertoire of tagging strategies for all taggers, where the broad content categorization is dominant followed by formalistic media categorization, genre categorization, process categorization and meta-categorization. This tendency also influences the possibility of finding information in folksonomy. If you use the dominant tagging strategies as a search strategy, your benefit will, everything else being equal, be greater. The frequency of the different tagging strategies and the overrepresentation of five strategies thus also influence the benefit retrieved from the system.

**Fifthly**, our study demonstrates profiling of value, process and time tagging strategies, although they are not the dominant strategies. This generally indicates that taggers bring a new time-related, personal and procedural dimension to categorization as a process. The energy and popularity characterizing folksonomies such as del.icio.us and user-created meta-data in general suggest that there is potential in more dynamic and multi-dimensional classification systems, where time, value and process are included.

The overall conclusion is that folksonomies are not necessarily the alternative classification system of the future; however, it is an important inspiration for a future where taggers wish to and must be involved in order to create improved classification systems.

## References

Benzécri, Jean-Paul. 1980. *L'analyse des données tome 2:l'analyse des correspondances*. Paris. Bordas.

Benzécri, Jean-Paul. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Bikchandani, S., D. Hirshleifer, et al. 1992. A theory of fads, fashion and cultural change as informational cascades. *Journal of political economy* 100:992-1026.

Bikhchandani, S., Hirshleifer, D. and Welch, I. 1998. Learning from the behavior of others:conformity, fads and informational cascades. *The journal of economic perspectives* 12: 151-70.

Blasius, Jörg and Greenacre, Michael. 1998. *Visualization of categorical data*. San Diego: Academic Press.

Chater, Nick. 1999. The search for simplicity: a fundamental cognitive principle? *The quarterly journal of experimental psychology* 52: 273-302.

Chater, Nick and Hurley, Susan. 2005. *Perspectives on imitation - from neuroscience to social science*. Cambridge: MIT Press.

Chater, Nick and Pothos, Emmanuel M. 2002. A simplicity principle in unsupervised human categorization. *Cognitive science* 26: 303-43.

Cozy, M. 2005. *Power laws in single websites in del.icio.us* from http://www.cozy.org/d/.

Golder, Scott and Huberman, Bernardo. 2005. *The structure of collaborative tagging systems.* Retrieved 10-08, 2007, from http://arxiv.org/abs/cs.DL/0508082.

Greenacre, Michael. 1993. *Correspondence analysis in practice*. San Diego: Academic Press.

Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. *Applied statistics* 3: 340-54.

Jolicocur, P. G., M.A. and S. M. Kosslyn. 1984. Picture and names: making the connection. *Psychology* 16: 243-75.

Mervis, Carolyn B. and Crisafi, Maria A. 1982. Order of acquisition of subordinate-, basic-, and superordinate-level categories *Child development* 53: 258-66.

Munk, Timme Bisgaard and Mørk, Kristian. 2007. Folksonomy, the power law & the significance of the least effort. *Knowledge organization* 34: 16-33.

Rosch, Eleanor and Caroly, Mervis B. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive psychology* 7: 573-605.

Rosch, Eleanor et al. 1976. Basic objects in natural categories. *Cognitive psychology* 8: 382-439.

Shen, Kaikai and Wu, Lide. 2005. *Folksonomy as a complex network.* Retrieved 10-08, 2007, from http://arxiv.org/PS_cache/cs/pdf/0509/0509072.pdf.

Shirky, Clay. 2003. *Power laws, weblogs, and inequality networks, economics, and culture.* Retrieved 24-08, 2007, from http://www.shirky.com/writings/powerlaw_weblog.html.

Tabachnik, Barbara G. 1989. *Using multivariate statistics*. New York. HarperCollins.

Tanaka, J. and Taylor M. 1991. Object categories and expertise: is the basic level in the eye of the beholder? *Cognitive psychology* 23: 457-82.

Vogt, Paul. 2000. *Minimum cost and the emergence of the Zipf-Mandelbrot law. introduction to linguistic knowledge/computational linguistics.* Retrieved 10-08, 2007, from http://www.ling.ed.ac.uk/~paulv/publications/alife9.pdf.

Zipf, George Kingsley. 1949. *Human behaviour and the principle of least effort: an introduction to human ecology.* Cambridge, Mass.: Addison-Wesley.