# Folksonomy Indexing From the Assignment of Free Tags to Setup Subject: A Search Analysis into the Domain of Legal History

José-Antonio Moreiro-González* and Carmen Bolaños-Mejías**

* Universidad Carlos III de Madrid, Facultad de Humanidades, Comunicación y Documentación, Departamento de Biblioteconomía y Documentación, calle Madrid, 126-128. 28903-Getafe-Madrid-Spain, <jamore@bib.uc3m.es>

** Universidad Nacional de Educacion a Distancia, Facultad de Derecho, Departamento de Historia del Derecho y de las Instituciones, calle Obispo Trejo, s/n. 28049-Madrid-Spain, <cbolanos@der.uned.es>

José A. Moreiro-González holds a PhD in history from the Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain. He got his BSc in education from the Complutense University and in arts and humanities from the UNED. He is Professor at the School of Humanities, Communication and Library Science in Carlos III University of Madrid. He has supervised thirty PhD theses in library and information science. Since 1994, he has been part of the Knowledge Reuse Group and is involved in projects and research and publication on thesauri, ontologies, indexing and retrieval systems and natural language processing techniques.

Carmen Bolaños-Mejías holds a PhD (1999) as well a bachelor's degree in law from the Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain. She has been a lecturer in the legal history and institutions department since 2000. Her preferred line of research has dealt with the legal institutions of the nineteenth century: history of legal-political thought, family discrimination or political leading figure. She belongs to the interdisciplinary research group on law, history and social intervention, from which she has made several collaborations with Information science researchers.

Moreiro-González, José-Antonio and Carmen Bolaños-Mejías. 2018. "Folksonomy Indexing From the Assignment of Free Tags to Setup Subject: A Search Analysis into the Domain of Legal History." *Knowledge Organization* 45(7): 574-585. 48 references. DOI:10.5771/0943-7444-2018-7-574.

**Abstract:** The behaviour and lexical quality of the folksonomies is examined by comparing two online social networks: LibraryThing (for books) and Flickr (for photos). We presented a case study that combines quantitative and qualitative elements, singularized by the lexical and functional framework. Our query was made by "Legal History" and by the synonyms "Law History" and "History of Law." We then examined the relevance, consistency and precision of the tags attached to the retrieved documents, in addition to their lexical composition. We identified the difficulties caused by free tagging and some of the folksonomy solutions that have been found to solve them. The results are presented in comparative tables, giving special attention to related tags within each retrieved document. Although the number of ambiguous or inconsistent tags is not very large, these do nevertheless represent the most obvious problem to search and retrieval in folksonomies. Relevance is high when the terms are assigned by especially competent taggers. Even with less expert taggers, ambiguity is often successfully corrected by contextualizing the concepts within related tags. A propinquity to associative and taxonomic lexical semantic knowledge is reached via contextual relationships.

Received: 11 May 2018; Revised: 4 August 2018; Accepted: 21 September 2018

Keywords: history, tags, law, legal, folksonomy

## 1.0 Introduction

Our most constant professional activity has always been that of ordering the universe of knowledge. Today, information technology and communication ("ICT") are widely used by the public, and cooperative activities on the web have become popular, to such an extent that the initial notion of order on the web and the abilities of professional specialists to create order, are becoming destabilized. Recently, the former authority of established systems is being affected by the empowering of other information organization systems (Mai 2011), and controlled vocabularies are proving to be insufficient to satisfy the requirements of retrieval, in the context of the enormous availability of re-

sources on the web. This is the context in which we find that folksonomies operate: for example, the practice of co-operative indexing, which is done by tagging with free keywords, in an attempt to represent both implicit and explicit concepts. It is also possible to index by using word commands, but it is expensive, time-consuming, and work intensive to carry out the semantic checking.

"Folksonomies" is the term we apply if documents have individual URL links attached to them, which have been created by the users themselves. Users choose the words that they introduce in these tags from their pragmatic knowledge of the subject (Hjørland 2010). That is to say: the labels that are assigned or juxtaposed in the same document are not created either by the system itself, or by professionals; the labelling or tagging, is created by the users themselves. What we find in folksonomies is a flat indexing system, where all words have the same weight and the same level of importance with respect to the indexed resources. Already, in the earliest databases, the keywords were combined with descriptor terms, to assist the search (Salton and McGill 1983). More recently, however, folksonomies prefer to accept the fact that each chosen word remains independent of its relevance and origin. A bottom-up indexing system provides everyone with user-oriented searchable computing.

In folksonomies, anybody can take the opportunity to make their own individual contribution and share their tagging with other users (Spiteri 2007). Any item of information on the web may be indexed repeatedly by many different users. The result is a decentralized, collaborative vision, in contrast with the traditional information systems that were hierarchical and directed (Fox 2016). Now, each new tag for an item is aggregated to that item in the manner of an "intersubjective description," creating something that is very different from the keywords that authors traditionally assigned to their own articles, and the set terms of a professional information indexing system (Kipp 2011). For this reason, it is so important to understand the behaviour and motivation of modern users of the web, because the way they choose to proceed in their tagging will be a significant conditioning factor for successful retrieval of information subsequently.

## 2.0 Background, purpose and scope

Folksonomies can be useful languages for information retrieval in poorly controlled environments that require user participation—where we frequently find that the criterion of the user is to employ some "popular language" tags, together with other more "specialist" tags. Each user benefits from the work of all those who have contributed tags to describe the content mass of books and images. The task of divide and rule, applying collaborative web-based tagging (Olson 2007), favours the connecting analysis of huge volumes of information. Without web-based tagging things would be chaotic; but there are still some problems with free tagging that have not been solved. Users' tags do not follow any rules, and this means that folksonomies can finally fall into "messiness" and inconsistency (Thomas et al. 2010).

The use of folksonomies has become widespread, undoubtedly because it is possible for users to choose the exact terms and vocabulary that they want, while the system immediately incorporates the changes, and adapts to the requirements of the users. It is easier and more flexible to use and update than the controlled vocabulary system. Looking into how folksonomies operate, it becomes clear that they are extremely useful (Moreiro 2006). This is plainly demonstrated, for example, not so much in indexing, but by the fact that the tags applied in folksonomies can be used to extract all the elements of a specific vocabulary, and can also be employed as a tool to enable us to identify the tendencies followed by groups of users when describing certain concepts (Al-Khalifa et al. 2007; Van Damme et al. 2007). That something has begun to change is confirmed by the never-ending search for solutions to problems like the following, for example: how to achieve an automatic differentiation between connotative tags and denotative tags (Basile et al. 2015); how to generate links between related tags (Mathes 2004); how to carry out tag gardening (Peters and Weller, 2008); and how to recomend appropriate tags for indexing (Cantador et al. 2011; Font et al. 2015). Another aspect, very relevant to our stated objectives, has its basis in the fact that, over a period, folksonomies tend to self-regulate, and thus achieve a certain consistency. This does away with the need for any external control (Mai 2011). Our literature review has revealed the existence of previous analytical studies that have investigated some of the social networks—such as, for example, Delicious, Flickr and LibraryThing—which started using uncontrolled tags (Golder and Huberman 2006; Bartley 2009; Pera et al. 2009). These earlier studies focused on the possibility of categorizing the different tags found in these social networks. One of the studies addressed the linguistic and formal construction of tags in Delicious, Furl, and Technorati (Spiteri 2007).

The goal of our study was to ascertain how free collective tagging functions in two online books, also in photo management, and in shared social networks, which permit new ways of organising texts and photos. The subject is located within the context of legal history in LibraryThing (https://www.librarything.com/), a cataloguing and social web portal that assists in generating tags for users searching library resources. And, at the same time, in Flickr (https://www.flickr.com/), which applies collaborative knowledge to the organization of images. LibraryThing

and Flickr both associate tag information objects in any way they wish. We have chosen the legal history field because of its scope of social problems and because of the power relations that govern society, its groups, interests, and ideologies. So, it offers both academic study and general appeal.

The stated purpose of LibraryThing folksonomy is "to tag your own book collections." LibraryThing has added over 136 million tags to over 113 million books and nineteen million works (films, music, etc) in the system: "socially generated access points assist in the task of searching and browsing library resources" (https://www.librarything.com/zeitgeist 2017). This type of democratization of the processes of information analysis is also the idea behind Flickr, although the ultimate purpose of Flickr is to manage, organize, and share photos online, in a new way, that makes the photos readily accessible to users (Rafferty and Hidderley 2007).

Our main aim in this study was to analyse and compare results of human behaviour in indexing sets in the legal history context of academic knowledge, with two specific objectives. First, to evaluate the quality of the resulting representation by examining the level of consistency, pertinence and recall, using concrete examples. Then, to verify the behaviour of both social networks in relation to the amendatory corrections that have been introduced by the folksonomies in order to avoid the drawbacks of free tag allocation. We wanted to evaluate the results of our retrieval by conducting an analysis of the indexing tags chosen by the folksonomy taggers. During the process of analysis we were interested in observing, not only whether folksonomy tagging behaviour is different in the case of a printed text than in the case of an image; but also in taking our observations a step further, namely, to trace how folksonomy tagging in both types of format (text and image) has evolved over time as a result of the search for solutions to the problems of retrieval. We were interested in identifying the specific reasons for this evolution.

As a first hypothesis, we can consider that LibraryThing folksonomy is partly elaborated from a specialist user approach, in contrast with Flickr, where a more open and generic approach is used. Of the two, we might assert that Flickr reflects the "community world-view" better: by aggregating greater individualistic relevance it achieves universal relevance. In general, folksonomies address the set of subjects contained within a contextual situation (Morato et al. 2003), without neglecting to attend to the multiple facets into which the context may be subdivided or to which it is related (Van Dijk 1977). We may assume, perhaps, that, over time, folksonomies will manage to self-regulate and improve their consistency without any external control (Trant 2009).

## 3.0 Methods and tools

Methodological coherence was achieved by combining quantitative and qualitative procedures. We used the case study method as the basis for designing our strategy, having realized that the particular and complex nature of the case in hand would require an exploratory and also a descriptive approach that would demand quantitative data. In order to explain the results obtained we realized that we would need to make a qualitative appreciation (Eisenhart 1989). Our methodology would necessarily be affected by the representative aspect that pertains to any linguistic question and also by the interpretative-evaluative intention of the folksonomy taggers. Added to which, our aim was to present one single object of interest, although we have made the assessment in two different models of folksonomy. Our methodological approach proceeded step by step, as follows:

1. Obtain information that could be used to demonstrate how the folksonomy tagging evolved, and explore the consequences for subsequent retrieval, both good and bad.
2. Analysis generalisation, to evaluate the information retrieved (Yin 1994). Include an exploration of the behaviour of the folksonomy user-indexers, from classifications of the data obtained.
3. Establish the credibility of the conclusions reached on the basis of our descriptive and analytical work.

To obtain the information, our initial search in LibraryThing was carried out in "legal history," a concept that is also expressed in the synonyms "law history" and "history of law." A search by history of law in *Burton's Legal Thesaurus* redirects to legal history as a preferred term. *The UNESCO Thesaurus* also prefers legal history. The *New Oxford American Dictionary* accepts the use of law history. We conducted our query in 11.10.2017, using the following three links: https://www.librarything.com/tag/history+of+law; https://www.librarything.com/tag/Law-History; https://www.librarything.com/tag/legal+history

The evaluation of the collected information was verified from the relevance of the tags. The documents retrieved from LibraryThing were examined individually and their precision ratio calculated (Cleverdon 1972). Noise ratio was also calculated, with respect to tags that were not relevant. The operation was carried out on the tags that were considered to be valid with respect to the desired retrieval, but only after deducting those of a subjective or organizational nature. Likewise, the information capability and consistency of the tags used was evaluated, and their effectivity in discriminating between the content of the various different documents retrieved (Bartley 2009). With

regard to the in-depth representation of the textual content, we analysed whether the tags referred to the entire text; whether the tags were extracted from possible reviews; or whether they could have been taken from a description of the book sent out for publicity purposes by the publishers or the distributers, which could then be confirmed in the pages of each of the books retrieved.

As our work of analysis revealed the existence of many very diverse rationales for tagging in the folksonomies, it seemed all the more probable that retrieval would have to be made from tags that went from being either far too generic to being far too specific. It was impossible to achieve a sufficiently exhaustive and exact recall that would allow us to identify the relevant documents and exclude the non-relevant ones (Rolling 1981). In our case, the retrieval of the documents that would prove to be useful for a search could only be achieved by following the meaning transmitted by the labels. We, therefore, proceeded to check the pertinence and exactness of the tags and conducted our search by observing the qualitative limits of linguistics and of usage.

In trying to evaluate the pragmatic approach to tagging in folksonomies, and the relevance of the documentary content, we experienced the feeling of moving over a permanently shifting ground. Any attempt to analyse free tagging, using tools that have in fact been designed for the formal purpose of analysing controlled indexes, creates considerable uncertainty in the mind of the analyst. Consequently, we decided that if we did not want to confine our search simply to how particular words are used in tagging, we would have to find an alternative approach. We concluded that we would have to establish the nature of the contextual fields of the tags that had been used in the retrieved documents. In order to achieve this, we would have to go beyond establishing the relationship between pairs of words, as prescribed in the technique of co-word analysis (Callon et al. 1983). We would take this technique of analysis further, until it was possible to detect a coincidental occurrence that could serve as a reliable confirmation of the existence of certain incipient semantic relationships between the words employed in tagging—and until we eventually reached a point where we could confirm that certain hierarchical and associative relationships were emerging.

We repeated the search in Flickr (https://www.flickr.com/), by the same subjects, to evaluate the tags assigned to each photo and to set them in context. Initially, using the category "All;" and subsequently, the category "Everyone photos." To locate photographs in Flickr that would be suitable for the purposes of our work of analysis, we made use of the function "Advanced search," in order to search for photographs filed under "Creative Commons." By using "Advanced search" it is also possible to extend or limit a search according to particular tags and get access to the texts that describe the content of the photographs. It should be borne in mind that photographs are like a frozen analogy of reality, with no intermediary code. Our ability to conduct any analysis is necessarily affected by the connotative contribution made by the person who attaches the tag. Inaccuracies and alterations to meaning inevitably appear in the tagging (Barthes 2000).

First, we planned to approach the task of evaluation by observing the correspondence between the tags themselves and our own analysis of their content (Rose 2016). However, we then realized that it was incorrect to expect to be able to establish an objective procedure that would identify categories of representation, because any attempt to set up categories to include the various objects visible in the images would necessarily represent a very different approach from that of a system of free, and individualized, tagging. The steps generally prescribed in any formal analysis of content—namely, denotation-quantification-inclusion in categories—had to be reduced here, we realized, to a purely denotative observation of visual representation. This was due to the completely free origin of each representation—all attempts at classification in this free context originated from the common sense of the taggers. Our work of analysis in Flickr was therefore limited to carrying out a lexical and tagging usage evaluation (as occurred also in our analysis of the selected books). In analysing images, it is not possible to evaluate to what extent each folkindexer may be socially or historically competent, as this goes completely against the nature of folksonomies. The symbolic interpretation of images requires that formal image-indexers demonstrate an even greater knowledge of context than the indexers of written texts (Rui and Huang 2001). However, any image-indexing folksonomy will certainly provide evidence of many superficial tags and many other tags that are excessively refined.

In both LibraryThing and Flickr, our search focused particularly on the lexical composition of tags and the semantic conditions of the indexation context. The quality of the results with respect to the search subject was compared. First, the lexical composition factors involved in the retrieval were considered (Copestake 1992). Then the number and relevance of the documents retrieved for each search term was assessed. And, finally, tables were prepared, listing the related tags assigned to the books that had received each tag in question, to determine the indexation context. The difficulties caused by tagging were then identified with respect to synonymy, ambiguity and denotative forms, as well as the different solutions that each folksonomy provides to solve them.

### 4.0 Results

In the following two sections we show the results obtained. First, we analyse the tags assigned to the retrieved documents. We then go on to present a lexical and functional evaluation of the tagging.

### 4.1 Analysis of retrieved information

The outcome was different when we entered using the search engine option "in search LibraryThing," instead of making the retrieval by searching for tags. History of law and law history are less relevant, because they offer a higher retrieval in the general search than by tags, while legal history acts in in the opposite way (Schamber 1994). Legal history is, in all three groups of hits, the term most frequently used to describe the concept searched. During recall, the greatest difficulty arises when the syntagmatic tags are broken down into their lexical units and these are then used as separate tags ("legal"/ "history"). If we consider the case of the tag "medieval," we see that it appears as a simple entry on 307 occasions, one hundred times more than in the phrase "medieval history," which allows us to deduce that only one out of every three retrievals exactly matches the subject sought. We note that the use of lexical units increases the number of books retrieved very considerably, but noise also increases. Legal history occurrences are distributed among those of the other two tags that do not relate to the desired meaning, and the location of relevant books is, therefore, encumbered. Although the use of lexical units such as history (6,744 retrievals), law (5,075 retrievals), or legal (363 retrievals) increases the number of books retrieved, noise also increases in parallel, because the crossing of concepts is made possible. Law and history, as lexical units, are the related terms that have the greatest presence throughout the entire query. But their composition can only be calculated by the established intersections between tags that relate to the same content. The relationships between the tags history of law and law history is closer, because they are really syntagmatic true synonyms, the former by composition, the latter by prepositional complement. We obtained results of very analogous relationships when we searched by history of law and law history, always finding a somewhat greater term weighting in favour of the latter. Hence, their related tags show a high similarity. Legal history content is found in the combination of lexical units law, history, legal, and so on, although these units had to be selected in pairs, because if we took them one by one, they proceeded to combine in a way that corresponded to subjects other than legal history.

Looking at the range of tags that appear simultaneously in the retrievals made using the three terms, the maximum value always comes from the tags related to legal history. This syntagm and its variations was used 3,677 times by 446 members. Each of the books recalled has been indexed with an average of 28.58 tags. Legal history offers a few more related tags that are a wider application, which establishes it finally as the broadest term. Retrieval by legal history includes fifty (37.03%) of the 135 books that are retrieved by history of law, and all the books retrieved by law history and its variations. The semantic primacy of legal history is also determined by the number of relevant books retrieved: a value of 50%, compared with 43.75% by history of law and 12.5% obtained by law history. If we analyse the tags assigned by members of more formal indexing groups, we perceive a greater pertinence and preference for syntagmatic terms, compared with independent users, who prefer lexical units, and more generic and ambiguous tags.

We also conducted a search in Flickr, for photos, searching by legal history. The total retrieval result was of 28,425 photos; and 787 when these were searched by legal history, which were mostly related to the world of law. Some were retrieved, because legal history appeared in the same document as the one in which the photo was published. Searching by history of law, nine photos were retrieved. Every document was found to contain a syntagmatic tag. Some of them are assigned connotatively, according to some kind of judicial process related to the environment, although what is seen in the photo is, for example, a snowy forest and a stone plinth in memory of a nineteenth century jurist. The rest are photos of a law book launch at the WTO Public Forum of 2015, whose

| Search tags | Times tags used | Number of Users | Works in LibraryThing search | Books retrieved | Unrelated works | % of noise in retrieval |
|---|---|---|---|---|---|---|
| **History of law** | 135 | 45 | 863 | 617 | 246 | 56.20% |
| **Law history** | 64 | 8 | 1,208 | 868 | 340 | 87,5% |
| **Legal history** | 3,677 | 446 | 691 | 552 | 139 | 50% |

*Table 1.* Number of documents retrieved using the three tags.

reference to history of law is non-contextual and thematic, as published by the organizers of the event. None of the photos relate directly to history of law, although phrases such as "Commerce works" appear, without any reference to the picture retrieved. Searching by law history (in "View all") we retrieved 106,005 photos, but when we searched only by law history, no results at all were retrieved.

### 4.2 Tagging review

In our retrieval of books on legal history we encountered a variety of situations. In five cases, decimal classification codes are shown in the tags, which are only comprehensible if they are correctly allocated by experts and are very far from an all-kind indexation. This confirms Rolla's findings (2011). At the same time, very generic terminological classifications appear, which are applicable to all specialties: non-fiction, fantasy, trademark … or the existence of auxiliary descriptors such as perspective, which function as contextualisation qualifiers. However, when used independently they generate undesirable noise (Thomas et al. 2010).

In some cases, the tagging collects official catalogue numbers (which is redundant) or makes other references to the documents as objects: 950 printed copies; eBook; $5.50; 217 p.; from Amazon; second ed.; diskette included; hardcover; University of Chicago Press; ftc; ip; copyright; borrowed from library; … We also retrieved non-relevant tags that relate to the publishers or libraries in which a book was either published or deposited, and reflect the intervention of experts, who indicate the specific libraries where they accessed the text. We even found the comment: Digest (which is already in the title); A comprehensive overview of the law; classical; classics.

In addition, we encountered purposive tags regarding the contents of the books. And then we should also note the subjective tags, which are important from a personal point of view, or that express personal actions or the opinion of the tagger, but which, perhaps, offer little benefit to other information seekers (Xu et al. 2006). Evaluative expressions are not infrequent either, such as, for example: *true story*. The frequent use of expressions such as: needs label or to-read should be avoided, mainly because this increases synonymy and ambiguity. These last two examples indicate personal tasks that are still pending in respect of the book analysed. Similar terms appear in retrievals resulting from our search by history of law and by law history: non-fiction (827-4,346 respectively); read (66-333); reference (74-501); to read (250-1,426). Or, in the case of legal history: crime (400); fiction (393) and true crime (383). Subjective and appreciative tags take up 19.1% of related tags. The same thing happens with True crime—in an attempt to evaluate the content—or with to read; read or unread; already read; read in … in order to make manifest a personal interest in exploring the tagged book. Knowledge of indexed books can be assessed by comparing the number of occurrences of the read and unread tags in the three search options followed. Most books are "to read," so the main function of the search is thus fulfilled.

In history of law and in legal history, we find that the retrieval is less relevant, if we look at the tags assigned to specific books. We note, in "People's welfare" (Novak 1996) that there are tags like: rivers; roads; safety; individualism; Box 10. The lack of relevance of many terms (Green 1995) is well represented by the use of the homograph "mine," which is employed with ambiguous effect—either as a pronoun; or as a verb, to indicate that the book's information should be exploited, as in the meaning of the noun "mine," which refers to the places where minerals are extracted. Among the retrievals that do not include books retrieved both by legal history and history of law, some show more specific key-words such as legal by artist; legal issues; legal; legal theory; legal interpretation; legal philosophy or legal system.

In the three retrievals that were made, we encountered many inconsistent tags. Some of them are purpose tags (Gupta et al. 2010): interpretation; owned; missing; were to donate books; recycling books or not-available-yet. There are even some nonsensical tags, that are very subjective in character: !cc; !cnd; #z-date?; #z-no date; $z-new_tags_done; %z-not_read; … All of these nonsensical tags have a weight in the tag count.

Flickr accepts key words that are expressed by syntagms, although most of these are formed by lexical units, among which some are adjectives, such as legal or red, which produce miscellaneous, scattered, and incomprehensible retrieval sets. It is also common to find that the entries are repeated for the same photo: a concept is expressed with synonyms, both grammatically and then also in different languages. We even find the same word repeated, or in both singular and plural (Freixa 2011). Some tags are not very objective, along with others that are not relevant, and have been extracted by the search, because the term "legal" has been used in the sense of cool, incredible, or beautiful. There are many tags in Spanish or Portuguese about cars, animals, or natural landscapes. Both terms appear by lexical units—that is to say: only legal or only history, and the meaning is maintained in the syntagm legal history. We find other tags in Flickr that are much more connotative of legal institutions. Such as the Magna Carta, for example; or the album that illustrates the Dred Scott Decision at BAM (the contradiction of the US that was founded on democracy, but also has slavery woven into its founding document); or emblems of social activism, such as the defense of marital equality, Lambda legal, or the Gay pride parade. The context is so distant at times that even some baseball softballs or photos without tagging may appear.

In Flickr, when the search was limited in the "Advanced" mode in tags, the result obtained was seventy-eight photos. In general, the tags are more relevant in relation to law books and libraries, codex, legal documents, book talks, courts of justice, parliaments, museums (Medieval Torture Museum in Vienna), law schools, or law conferences and conventions. All tags were in the English language, with some exceptions, assigned in Spanish, such as: arquitectura or antigüedad. In addition, unassigned tags will retrieve, from personal comments: theatrical performance. Occasionally, the terms used are so generic that massive retrieval results were obtained, such as in the case of the term "people," for example. We have also found, not infrequently, that the formal elements of photographs are tagged: size, type of camera used, and even document type and technical aspects, as in the case of photo or Canon EOS. These should appear in Exchangeable Image File Format (EXIF), a standard that specifies the technical details of the photos, and the ancillary tags used and recorded by digital cameras, which Flickr posts on each of the photos. The assignment of physical descriptive EXIF tags is frequent in the case of technical elements, but is not frequent in the case of the actual content. Flickr also links photos that are geotagged in a map.

## 5.0 Improvements to retrieval performance

The results we obtained are completed here. In this section we present the different proposals that are applied in both folksonomies to improve observed performance. LibraryThing-related tags allow us to navigate and search for subjects that are closely related to the initial concept search, both because they contain their lexical compositional units and for other, derivative reasons, such as the fact that a word can have various meanings. In this case, they facilitate links to homonymic terms, and also to other terms with which the concept is associated. This is a way of associating concepts with hierarchical and contextual implications, as will be seen later in the discussion about "tag clouds" (Van Damme et al. 2007). Related tags distinguish two levels: subject, controlled by the system, and individual, that has greater freedom. In the set determined by the selected related tags, the contents of the searched concept are recalled, in addition to others that are closely related to it, but some that are not so close are also recalled. The lexical unit "law" occurs 5,075 times: of these, 1,197 occurrences are more retrieval-related, and are due to the appearance of law in the law history list of related tags. "History" follows, with 930 occurrences as a related tag of law history; and appears 6,744 times, in total. So, the occurrences of both may be considered frequent in general, although occurrences are relatively infrequent when either is related to law science. The following tables compare the occurrence of tags related to legal history, history of law, and law history. They are displayed in descending order of frequency.

| **Number of related tags** |
|---|
| History (6,744) |
| Law (5,075) |
| Philosophy (4,044) |
| American history (1,964)<br>US history (695) |
| Politics (1,603)<br>Political science (429) |
| Supreme Court (1,248) |
| Biography (1,157) |
| Legal history (952) |
| USA (757) |
| Constitution (732)<br>Constitutional law (296)<br>U.S. Constitution (316) |
| Government (429) |
| Legal (363) |
| Religion (339) |
| 19th century (218) |
| Sociology (214) |
| Medieval history (206) |

*Table 2.* Related tags in all three terms.

The fact that law history does not appear in our related terms table is due to the fact that its components are taken as lexical units, so it is not taken to be a synonym of history of law. During the search by history of law, the system automatically includes History of Law, law-history of, History of Law, History of law. Most combinations of history of law are established with other syntagm tags.

| **Number of related tags** |
|---|
| Mythology (970) + myth (226) + myths (46) = 1,242 |
| Anthropology (139) |
| Democracy (94) |
| Roman law (94) |
| Spirituality (90) |
| Psychology (88) |
| History of law (64) |
| Culture (59) |
| Jurisprudence (55) |
| Archetypes (52) |
| Comparative religion (46) |

*Table 3.* Related tags in history of law and in law history.

A slant that favours interdisciplinary intersection is evidenced by the presence of anthropology, culture, philosophy, politics, religion... that reveal the cognitive associations, particularly noteworthy in Table 3. Without abandoning

hierarchical approximations, an analysis of related tags makes it possible to descend to specific topics in legal history, that are so extensive that in many universities they are given as differentiated disciplines: roman law, constitutional law, constitutional history, comparative law. Alternatively, it is possible to particularize towards national legal history, especially in respect of the national legal history of the United States, England, or France. It is also possible to search by large sectors within national legal history such as the Supreme Court, jurisprudence, or even judges' biographies.

| Number of related tags |
|---|
| Medieval (307) |
| 18th century (297) |
| Holocaust (280) |
| France (275) |
| 20th century (257) |
| Witchcraft (254) |
| England (246) |
| America (201) |
| American (193) |
| American Revolution (198) |
| Theory (187) |
| Political theory (243) |

*Table 4*. Tags related only with legal history.

In these lists, the lexical units occur four and even five times more frequently than those of the syntagms. Here there is an inverse relationship between accuracy and exhaustiveness. Tags that are related to legal history show a greater hierarchical status, proving that it is a greater conceptual hypernym.

| Search tags | Relevant | No relevant | Total |
|---|---|---|---|
| **Legal history** | 644 | 1790 | 2434 |
| **Law history** | 38 | 51 | 89 |
| **History of law** | 113 | 108 | 221 |

*Table 5*. Number of related tagmashes.

Legal history and history of law function as syntagms, but law history functions as two intersected tags in the same document, that then become searchable for everyone. This is called a "tagmash," and this function is what makes it possible to search a book by two or three tag combinations when you have a general idea of the theme or can describe it in "fairly broad terms" (Rolla 2009). The link shows common tagmashes related to a given tag. If you are the first user to try a tagmash, the system calculates the results. The tagmashes that we may consider relevant to users' needs, in respect of the subject of the texts, are set out in Table 6, where the frequencies of the occurrences are also indicated:

The highest number of related tagmashes retrieved also corresponds to searches by legal history. It is observed that all tags corresponding to legal history also appear in the other two terms. The only case where the contrary effect occurs is that of history of law, in which there are more relevant tags than irrelevant ones, because, as we have seen previously, the syntagm history of law is itself formed by a tagmash.

| Search tags | Related tagmashes (Number) | % of related tagmashes retrieved |
|---|---|---|
| **Legal history** | history, law, legal (174)<br>history, law (170)<br>constitution, pre-law, law, legal (102)<br>USA, history, law (101)<br>Supreme Court, history, law (97) | 26.4% |
| **Law history** | history, law (10)<br>history, law, legal (8)<br>Supreme Court, history, law (8)<br>constitution, pre-law, law, legal (5)<br>Supreme Court, biography (5)<br>USA, history, law (4)<br>history, law, medieval (3) | 2.3% |
| **History of law** | history, law (23)<br>history, law, legal (22)<br>Supreme Court, history, law (19)<br>USA, history, law (15)<br>constitution, pre-law, law, legal (12)<br>England, history, legal (8)<br>England, history, legal (7)<br>law, medieval history (7) | 45.2% |

*Table 6*. Relevant tagmashes to the users' needs.

### 5.1 Use of synonym processing

LibraryThing introduces all the synonyms that name a concept or fact, together with their aliases and variations. Where members have combined a tag with other tags that have the same meaning, these are automatically included in the same search. The difficulties generated by the rise in unlimited, formal, and synonymous spelling expressions, or a variation in terms, causes this platform to specify the equivalent tags. For example, in the case of "to-read," the platform includes all graphical variations: to read, #toread; To Read, TO READ!, To Read, ToBeRead, toberead, … and also includes their inter-idiomatic equivalences: noch zu lesen, à lire, da leggere, DA LEGGERE… A very representative example is the search by legal history, which includes possible synonyms and all the variant spellings: legal history, legal history, Legal History, Legal history, legal History, LEGAL HISTORY … In addition, consider the case of law history and its variations: LAW History, law history, Law History, Law-History … In this way, a conceptual and formal synonym ring is established, "which connects a set of terms as being equivalent for search purposes" (Garshol 2004) and then proceeds to control the vocabulary. Expressive disambiguation, as with synonym rings, only affects the retrieval phase, not the assignment of tags.

### 5.2 Lemmatization

The tags assigned reproduce, logically, the grammatical difficulties of any type of free indexing with key words. Each of the denotative forms of the same word is shown as a different entry, without preference for the canonical form (myth; myths). The idiosyncrasy of the user-indexer is shown in the use of derivative morphemes (mythology) in preference to its lexeme (myth). Also, derivative adjectives (legal; american) appear at the same time as their nominal base (law; america). The reiterative presence of adjectives as tags that can only be assigned to related nouns in the context of indexing, but which sometimes give rise to the most imaginative relations (urban, rural, ...) because the compound terms have been sectioned in the process of assigning them to tags. LibraryThing offers the asset that all these term variations are included when a search is carried out. In this way, all the possible retrievals obtained by searching each of the words derived from the same lexeme (nominal and also adjective forms) and by searching their different flexible morphemes (feminine-masculine; singular-plural) are retrieved together.

Finally, LibraryThing makes it possible to establish a comparison between user-created tags and a professional classification (such as, for example, the Library of Congress Subject Headings). Thus, a specific tag shows all related subjects given to the tagged books and vice versa. Any possible inconsistencies produced in the tagging are contrasted with the related subjects of a taxonomic vocabulary (Zubiaga et al. 2011).

### 5.3 Corporate collections and controlled indexing

Flickr does not limit its service to the contributions of folksonomies. While one part of Flickr's collection contains free indexing, another part is controlled at source. Many companies and organizations tag their photos themselves and broadcast them using Flickr, which functions like a public archive. Only the administrators can modify these tags and delete or upload images (Bolaños and Moreiro 2014). The advantage is that the photos can be placed in a connotative institutional context that contributes to the interpretation of the message by imposing meanings that might otherwise be debatable (Weller 2010). Subsequently, they are organized by a method of classification that usually reflects the same criteria that are used to organize the institution in question (Noruzi 2006).

Another response, still in development, are the "beta tags," which show up with a white outline and are generated from an image-recognition algorithm at the testing phase. The Flickr algorithm determines appropriate tags for image-finding (https://help.yahoo.com/kb/flickr/tag-keywords-flickr-sln7455.html 2015). Beta tags are not, therefore, "pure" folksonomy tags that users have decided to attach. In our retrieval, the beta tags are denotative of visible objects, and of the specific physical context: techniques, black and white, monochromatic... None of these details can be properly said to be elements relating to the content of the photographs, and so they should not appear in the index tags, because they are either in EXIF: outdoors, indoors, for example; or because they create a type of indexation that is too generic: people, group photo…

| Search tags | Read | to-read | un-read | |
|---|---|---|---|---|
| **Legal history** | 73 | 222 | 58 | |
| **History of law** | 66 | 256 | 67 | |
| **Law-history** | 19 | 40 | - | |
| **TOTAL** | 158 | 518 | 125 | |

Table 7. Number of occurrences of the "read" and "unread" tags. (https://www.librarything.com/tag/Law-History).

In the search carried out by "law-history," the related subjects that were retrieved present a higher level of specificity, and mainly concern institutions, biographies of jurists, and supra-national cases (Europe), national cases (United States), regional cases (Castilla, Wales) as well as the various different approaches to law history from the philo-

sophical, political, sociological, and religious perspective (history of ideas, natural law, jewish law…). Also retrieved were other cases related to "law" as judicial power, constitutionalism, constitutional history or criminal law.

### 6.0 Conclusions and recommendations

It has been demonstrated that folksonomies are useful languages for information retrieval in scantily controlled environments that show very considerable levels of growth and functionality in respect of user participation. Folksonomies provide a superficial and easy type of indexing that can comprise all the subjects covered in the document, using generic terms for specific names. It appears that an implicit classificatory will exists, generated by the contextual situation, with which each user approaches the document to be indexed. This means that classification is necessarily initiated from some kind of general organizational scheme that reflects cultural, social, and pragmatic knowledge.

It is certainly the case that LibraryThing and Flickr pursue the task of working without any language restrictions. The way they operate shows that the use of free language offers an undeniable advantage, due to the speed and simplicity of word identification. There is, however, a concomitant disadvantage: the greater level of distortion in communication, a distortion which folksonomies do certainly attempt to overcome. The main problems in folksonomies have been observed to arise in relation to inconsistencies between plural and singular nouns; and also in relation to ambiguous tags, either in the form of homonyms, or as unqualified abbreviations or acronyms. However, our study has proved to be determinant, in that it has allowed us to estimate that these cases, in which the tags were ambiguous or inconsistent, represent less than a quarter of the total quantity of assigned tags in each of the platforms analysed. We have found that more than a third of this "messiness" is in the form of tag variations containing non-alphabetic characters. The other types of messiness measured were less significant, so we can conclude that tag variations are the most prominent hindrance to search and retrieval.

We have found that another weakness in the current usage of folksonomies is the inconsistency caused by the lack of control over linguistic or usage synonyms, along with imprecise naming, that has its origin in the preference for lexical units. At the same time, the presence of polysemy is inevitable, since folksonomies are lacking in definitions and context classification, except in the particular examples mentioned in our study.

Folksonomies can certainly be said to represent, nonetheless, a major improvement over traditional subject indexing. They have the singularity of using a type of indexing that is both free and specialized at the same time. Successful performance depends, however, on the domains to which the tags are applied, because of the greater or lesser terminological stability in their scientific, cultural, or general characterization. Because of the cognitive-field investigations carried out in this study, we have verified the pertinence that is achieved when the terms are assigned by specialists. Our study has also shown that affectivity is greater when lexical units are used, than when syntagmatic terms are used. We have also observed the existence of an inverse relationship between accuracy and completeness.

In LibraryThing, the hierarchical relation between non-classifying tags is not absolutely dislocated. It is true that the LibraryThing folksonomy does not follow the common practices of taxonomic trees, but most of the taggers are especially competent, as we have seen in the case of the tag legal history, and the indexing procedure in the mental categorization of this legal frame. The hierarchical relationship between the concepts handled is properly represented in the tags. Nevertheless, the related terms obtained do not always have an identical meaning, since they are often overloaded with subjective evaluations. But LibraryThing has been validated as a vocabulary that uses equivalences to find and use effective terms. To this extent, the starting hypothesis of this paper has been verified. Further progress will have to be made in differentiating tags by their context.

In the case of Flickr, we have seen that the "open part" does not have instruments that serve to counteract the drawbacks of free language, so irrelevant documents are frequently retrieved. Even when very specialized searches are carried out, the ambiguity one can encounter is considerable, because there are no user guidelines, nor are there any limitations or restrictions about how to tag the contents. For one person, a term may express a reality that is quite different from that of another.

Connotation is another cause of difference in the tagging of photos in Flickr, in addition to the fact that so many people want to view them. In our study, we have perceived the presence of various tags, attached to the same photo, that have been assigned by specialists, alongside other tags that we could describe as being of unexpected and even impulsive origin. One reason for the fact that relevance of the retrieved photos is very considerable is that the taggers are indeed users, but they are specialist users. In folksonomies, hierarchy is created by the individual's commitment to the content that is tagged; each created category is both personal and social. The number of documents retrieved is significantly greater in LibraryThing, and this is due to the fact that the content of books attracts more connoisseurs of academic subjects. Flickr, dealing in images, organizes, for public benefit, the personal interests that can be represented through images.

The folksonomies that we have analysed in this study have been created by means of an increasingly hybrid system of indexing, that employs a mix of free and controlled language. We perceive that the initial problems of folksonomies are gradually being overcome, starting from a certain approximation to the usual semantic relations of the controlled vocabularies. From an initial situation of unlimited tags being posted, a certain formalization is being achieved by creating inter-conceptual associations, to better reflect the subject complexity of photos and books, by means of tagmashes and related terms, and by assigning automatic identifiers in Flickr: EXIF. This is due, in part, to the creation of co-word deductions, but also to the finding of solutions that offer a hierarchy of concepts, and contextual dependence, with proximity to associative and taxonomic lexical semantic knowledge. Thus, folksonomies are overcoming some of the disadvantages caused by retrievals made from free tags. The system has learned to group tags automatically by semantic similarity. This is perhaps facilitated by the fact that the terminology of legal history homogenizes the logical-semantic similarities. Folksonomies also uses morphological analysis programmes to channel the inflected forms into their corresponding headword and reduce tag variations. Lastly, folksonomies have opened lines of business, operating as warehouses and centres for the dissemination of documents produced in corporative and institutional activity, the organisation of which is restricted.

## References

Al-Khalifa, Hend S. and Hugh C. Davis. 2007. "Exploring the Value of Folksonomies for Creating Semantic Metadata." *International Journal on Semantic Web and Information Systems* 3, no. 1: 12-38.

Barthes, Roland. 2000. "The Photographic Message," trans. Stephen Heath. In *Theorizing Communication: Readings across Traditions*, ed. Robert Craig and Heidi Muller. London: Sage, 191-9. Translation of "Le message photographique" 1961.

Bartley, Peishan. 2009. "Book Tagging on LibraryThing: How, Why, and What Are in the Tags?" *Proceedings of the American Society for Information Science and Technology* 46: 1-22. doi:10.1002/meet.2009.1450460228

Basile, Valerio, Silvio Peroni, Fabio Tamburini and Fabio Vitali. 2015. "Topical Tags vs Non-Topical Tags: Towards a Bipartite Classification?" *Journal of Information Science* 41, no 4: 486-505. doi:10.1177/0165551515585283

Bates, Jo and Jennifer Rowley. 2011. "Social Reproduction and Exclusion in Subject Indexing: A Comparison of Public Library OPACs and LibraryThing Folksonomy." *Journal of Documentation* 67: 431-48.

Bolaños-Mejías, Carmen and José A. Moreiro-González. 2014. "Indización fotográfica en Flickr: Palabras-clave frente a organización del conocimiento corporativo." *Revista temas em educação* 23, no 2: 70-81.

Callon, Michel, Jean P. Courtial, Bill Turner and Serge Bauin. 1983. "From Translations to Problematic Networks: An Introduction to Co-word Analysis." *Social Science Information* 22: 191-235.

Cantador, Iván, Ioannis Konstas and Joemon M. Jose. 2011. "Categorising Social Tags to Improve Folksonomy-Based Recommendations." *Web Semantics: Science, Services and Agents on the World Wide Web* 9, no.1: 1-15.

Cleverdon, Ciryl W. 1972. "On the Inverse Relationship of Recall and Precision." *Journal of Documentation* 28: 195-201. doi:10.1108/eb026538

Copestake, Ann. 1992. "The Representation of Lexical Semantic Information" PhD diss., University of Sussex. https://www.cl.cam.ac.uk/~aac10/papers/thesis.pdf

Eisenhart, Kathleen M. 1989. Building Theories from Case Study Research. *Academy of Management Review* 14: 532-50.

Font, Frederic, Joan Serra and Xavier Serra. 2015. "Folksonomy-Based Tag Recommendation for Collaborative Tagging Systems." *International Journal on Semantic Web and Information Systems* 9, no 2: 1-30.

Fox, Melodie J. 2016. "Priorities of Arrangement" or a "Hierarchy of Oppressions?": Perspectives on Intersectionality in Knowledge Organization." *Knowledge Organization* 43: 373-83.

Freixa-Font, Pere. 2011. "Patrimonio fotográfico y web 2.0: la experiencia Flickr The Commons." *El Profesional de la Información* 20: 432-8.

Garshol, Lars Marius. 2004. "Metadata? Thesauri? Taxonomies? Topic maps! Making Sense of It All." *Journal of Information Science* 30: 378-91.

Green, Rebecca. 1995. "Topical Relevance Relationships. I. Why Topic Matching Fails." *Journal of the American Society for Information Science* 46: 646-53.

Gupta, Manish, Rui Li, Zhijun Yin and Jiawei Han. 2010. "Survey on Social Tagging Techniques." *ACM SIGKDD Explorations Newsletter* 12: 58-72.

Hjørland, Birger. 2010. "The Foundation of the Concept of Relevance." *Journal of the American Society for Information Science and Technology* 61: 217-37.

Kipp, Margaret. 2011."Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing." *Knowledge Organization* 35: 17-48.

Lu, Caimei, Jung-ran Park and Xiaohua Hu. 2010. "User Tags versus Expert-Assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings." *Journal of Information Science 36:* 763-79.

Mai, Jens-Erik. 2011. "Folksonomies and the New Order: Authority in the Digital Disorder." *Knowledge Organization* 38: 114-22.

Mathes, Adam. 2004. "Folksonomies-Cooperative Classification and Communication through Shared Metadata." http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Morato, Jorge, Juan Llorens, Gonzalo Génova and José A. Moreiro. 2003. Experiments in Discourse Analysis Impact on Information Classification and Retrieval Algorithms. *Information Processing & Management* 39: 825-51.

Moreiro-González, José A. 2006. "La representación y recuperación de los contenidos digitales de los tesauros conceptuales a las folksonomías." In *Tendencias en documentación digital*, ed. Jesús Tramullas. Biblioteconomía y administración cultural 155. Gijón: Trea: 81-109.

Noruzi, Alireza. 2006. "Folksonomies: (Un)Controlled Vocabulary?" *Knowledge Organization* 33: 199-203.

Novak, William J. 1996. *The People's Welfare: Law and Regulation in Nineteenth-Century America*. Studies in Legal History. Chapel Hill, NC: University of North Carolina Press.

Olson, Hope A. 2007. "How We Construct Subjects: A Feminist Analysis." *Library Trends* 56, no 2: 509-41.

Pera, María Soledad, William Lund and Yiu-Kai Ng. 2009. "A sophisticated Library Search Strategy Using Folksonomies and Similarity Matching." *Journal of the American Society for Information Science and Technology* 60: 1392-406.

Peters, Isabella and Katrin Weller. 2008. "Tag Gardening for Folksonomy Enrichment and Maintenance." *Webology* 5, no 3: 1-18.

Rafferty, Pauline and Rob Hidderley. 2007. "Flickr and Democratic Indexing: Dialogic Approaches to Indexing." In *Aslib Proceedings* 59, no. 4/5: 397-410.

Rolla, Peter J. 2011. "User Tags versus Subject Headings." *Library Resources & Technical Services* 53: 174-84.

Rolling, L. 1981. Indexing Consistency, Quality and Efficiency. *Information Processing and Management* 17, no. 2: 69-76.

Rose, Gillian. 2016. *Visual Methodologies: An Introduction to Researching with Visual Materials*. 4th ed. London: Sage.

Rui, Yong and Thomas S. Huang. 2001. "Relevance Feedback Techniques in Image Retrieval." In *Principles of Visual Information Retrieval*, ed. Michael S. Advances in Pattern Recognition. Lew. London: Springer, 219-58.

Salton, Gerard and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Computer Science Series. New York: McGraw-Hill.

Schamber, Linda. 1994. "Relevance and Information Behaviour." *Annual Review of Information Science and Technology* 29: 3-48.

Spiteri, Louise F. 2007. "The Structure and Form of Folksonomy Tags: The Road to the Public Library Catalogue." *Webology* 4, no. 2: 13-25. http://www.webology.org/2007/v4n2/a41.html

Sturges, Paul. 2005. "Understanding Cultures and IFLA's Freedom of Access to Information and Freedom of Expression (FAIFE) Core Activity." *Journal of Documentation* 61: 296-305. doi:10.1108/00220410510585232

Thomas, Marliese, Dana Caudle and Cecilia Schmitz. 2010. "Trashy Tags: Problematic Tags in LibraryThing." *New Library World* 111, no. 5/6: 223-35. doi:10.1108/03074801011044098

Trant, Jennifer. 2009. "Studying Social Tagging and Folksonomy: A Review and Framework." *Journal of Digital Information* 10, no. 1: 1-44.

Van Damme, Céline, Martin Hepp and Katharina Siorpaes. 2007. "Folksontology: An Integrated Approach for Turning Folksonomies into Ontologies." Paper presented at Bridging the Gap between Semantic Web and Web 2.0 International Workshop located at the European Semantic Web Conference ESWC 2007 June 7, 2007, Innsbruck, Austria.

Van Dijk, Teun A. 1977. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Longman Linguistics Library 21. London: Longman.

Weller, Katrin. 2010. *Knowledge Representation in the Social Semantic Web*. Knowledge & Information. Berlin: De Gruyter Saur.

Xu, Zhichen, Yun Fu, Jianchang Mao and Difu Su. 2006. "Towards the Semantic Web: Collaborative Tag Suggestions." In *Proceedings of the 1st Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, ed. Sheila McIlraith et al. Springer. http://ra.ethz.ch/CDstore/www2006/www.rawsugar.com/www2006/13.pdf

Yin, Robert K. 2018. *Case Study Research and Applications: Design and Methods*. 6th ed. Thousand Oaks, CA: Sage.

Zubiaga, Arkaitz, Christian Körner and Markus Strohmaier. 2011. "Tags vs Shelves: from Social Tagging to Social Classification." In *HT'11:Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia; June 6-9, 2011, Eindhoven, the Netherlands*. New York, NY: Association for Computing Machinery, 93-102. doi:10.1145/1995966.1995981