# LOOKING FOR HEALTH INFORMATION ON THE INTERNET: CAN SOCIAL BOOKMARKING SYSTEMS REPLACE EXPERT GATEWAYS?

V. Durieux[a1], G. Kerdelhué[b]
[a] *ReSIC, Université Libre de Bruxelles, 1050 Brussels, Belgium*
[b] *CISMeF, Rouen University Hospital, 76031 Rouen, France*
[1] *vdurieux@ulb.ac.be*

## Abstract

Social bookmarking systems allow Web users to store and organize their bookmarks of online content by assigning them metadata in the form of tags (natural language terms). The process of adding and sharing tags is called collaborative tagging and leads to the construction of folksonomies. Recently, collaborative tagging has been described as an alternative method for creating classification systems.

In this paper, we conduct a quantitative and qualitative analysis of online resources and their associated tags assigned by Web users compared with descriptors provided by librarians for the same content. Our study is based on a data set of 113 online health resources listed in both the social bookmarking system Delicious (formerly Del.icio.us) and the expert gateway CISMeF. The aim of our study is to measure the overlap between healthcare resources listed in CISMeF and Delicious, and especially between the metadata that both have assigned to those resources. In other words, the study tries to determine the extent to which tools such as CISMeF and Delicious are redundant.

## Introduction

Over the last few years, the World Wide Web has become the first and for many the sole source of scientific and health information. Numerous studies have shown that Google is frequently used by scientists for information retrieval [1]. However, there is neither validation, nor organisation process on the Web.

In order to overcome these limitations, several systems have been proposed as alternative tools to search engines to organize and retrieve online content. Some of the oldest ones are Web sites designed as gateways wherein online resources are selected, described, organized, and made available by experts and librarians. At the start of the World Wide Web, experts' gateways were an efficient way to find online documents. Nevertheless, over the years, the content of the Internet kept increasing exponentially. The amount of online resources exceeds the librarians' capacities to classify.

Since the appearance of social bookmarking, its use has increased dramatically. It allows Web users to store and organize their bookmarks of Web content by assigning them metadata in the form of tags. The assigned tags become immediately available for others to see and use as a mean of information retrieval. The process by which ordinary users add keywords to online content such as website bookmarks, photographs, computer games or music is called collaborative tagging [2]. The list of terms made of all the tags freely assigned to particular resources is commonly referred to as a 'folksonomy' (short for "folk taxonomy"), meaning a user-generated taxonomy [3]. Collaborative tagging systems were rapidly seen as an alternative method for creating classification systems [4]. Collaborative tagging and its resulting folksonomies are thus generally compared to professional indexing and taxonomies.

In this paper, we analyse and compare the social bookmarking service Delicious with the experts' gateway CISMeF, especially the resources they both list and their associated metadata for a sample of healthcare documents from Belgium.

*CISMeF*

CISMeF (http://www.chu-rouen.fr/cismef) is the French acronym for Catalogue and Index of French Language Health Resources on the Internet [5]. This experts' gateway was initiated at Rouen University hospital which hosts its website since 1995.

It offers a selection of websites and online documents to Health professionals, students and consumers. Survey, selection and description of Web resources is based on human expertise and is currently realized by a team of four librarians.

CISMeF uses two standard tools to organize information: the MeSH thesaurus from the US National Library of Medicine in its French translation and several metadata element sets, in particular the Dublin Core metadata format. There are currently 67,000 indexed resources.

*Delicious*

Delicious (http://delicious.com/) is a social bookmarking service launched in 2003. Bookmarks are saved and shared on the servers of Delicious and are thus accessible from any computer connected to the Internet. The assigned tags immediately become available for others to see and use as a mean of information retrieval. Through a search by tag, users may retrieve resources from their own collection of bookmarks, as well as from other users' libraries. Searches can also be performed by user and URL.

Users may share sources they found interesting with others who have similar information needs or personal interests by sending each other bookmarks. Moreover, Delicious provides subscription services that allow users to keep track of tags and users they find interesting [6].

Though there are no official statistics about Delicious, one source mentioned 3 million registered users and 100 million unique URLs bookmarked in September 2007 [7].

Since Delicious tags and CISMeF descriptors are assigned to online content, it is possible to collect tags and descriptors associated with resources listed in both systems. A comparison can thus be made between users' tags and professionals' descriptors assigned to a particular online health resource. This study addressed the following research questions:

- To what extent do healthcare resources listed in CISMeF overlap those available in Delicious?
- To what extent do tags in Delicious differ from descriptors in CISMeF for the same content?
- To what extent do users' tags provide added value to traditional descriptors in the information retrieval process?

## Materials and Methods

This study identifies the Web resources common to CISMeF and Delicious and compares tags assigned by Delicious users with descriptors provided by CISMeF librarians for the same online healthcare resources. The identification of resources listed in both systems was

performed by a Mash-up (http://www.cismef.org/blog/?p=170) between Delicious and CISMeF using Yahoo! Pipes[tm]. The Mash-up retrieved simultaneously data from CISMeF and Delicious, and aggregated results (see Table I). It permitted to directly focus on the URLs common to the two information systems and provided a direct link to the concerned bookmarks in Delicious.

| **Steps** | **Example** |
|---|---|
| A given query | "SOGC"<br>acronym for The Society of Obstetricians and Gynaecologists of Canada |
| Get the results from the CISMeF search engine | http://doccismef.chu-rouen.fr/servlets/Simple?Mot=sogc (2 results) |
| Fetch URLs from these results | http://www.sogc.org/ and 2 others |
| Get the reciprocal URLs from Delicious (feeds in RSS format) | http://feeds.delicious.com//rss/url/data?url=http://www.sogc.org/ and 2 others |
| Count number of bookmarks and get the URLs from Delicious (html format)* | count = 22 and URL is http://delicious.com/url/d80afdea147f6c1f9e301d1c5ce99994 |
| Filter out all URLs where bookmarks count < 1 (ie not bookmarked) | 2 URLs are excluded |
| Display results according to the bookmark counts in descending order | See http://pipes.yahoo.com/pipes/pipe.run?_id=2d14daad0582e616875d498393f4b070&textinput1=sog |

** Note: Delicious provide a feed even though an URL is not bookmarked yet*
Table I: Main steps of the data processing in the Mash-up.

The query used for this study was "Belgique.pa" which returns all Delicious bookmarks of the resources from Belgium listed in the CISMeF catalogue. The query was executed on October 9, 2008. Using this method, a total of 113 resources were identified.

Data concerning resources listed in both systems were then collected manually; from Delicious, i.e. the URL of the bookmarks, their associated tags and taggers, and from CISMeF, i.e. the URL of the bibliographic records, their associated MeSH descriptors, resource types and location indications (see Table II).

| **CISMeF** | **Delicious** |
|---|---|
| URL | URL |
| Title | Title |
| Description | Annotations |
| Publication date (if any) | Bookmarking dates |
| MeSH descriptors | Tags |
| Resource Types | |
| Location | |
| Authors and Publisher | |
| | Delicious users |

Table II: Correspondence of main metadata between Delicious and CISMeF

Once collected, the data set was used to measure the overlap between the content of Delicious and CISMeF. To do so, we needed 3 different data: the number of healthcare resources from Belgium listed in Delicious, the number of healthcare resources from Belgium coded in CISMeF, and the number of these resources listed in both systems. While the last two data are provided by the Mash-up, the number of healthcare resources from Belgium listed in Delicious has to be determined. Due to the unsystematic description of resources by Delicious users, the anarchic nature of folksonomies and, the weaknesses of Delicious search functions, this number can neither be precisely determined, nor even estimated. To identify a sample of these resources, we performed the following search in Delicious: (belgium OR belgique OR brussels) AND (health OR santé OR medicine OR médecine OR medical). The presence of these resources within the CISMeF catalogue was checked manually.

Two forms of analysis were then performed on our data set: descriptive statistics and term comparison. For the comparison of tags with descriptors, we used the seven-point scale proposed by Kipp to compare tags from CiteUlike users, author's keywords and librarian's descriptors assigned to journal articles in the field of Information science [4]. Each tag of our dataset was manually compared to descriptors assigned to the concerned resource and placed in one of the following categories:

1. Same – the tag and one of the descriptors are the same or almost the same (e.g. plurals, spelling variations, acronyms and multiword terms);
2. Synonym – the tag and one of the descriptors are synonyms according to the MeSH thesaurus;
3. Broader Term – the tag is a broader term of one of the descriptors;
4. Narrower Term – the tag is a narrower term of one of the descriptors;
5. Related Term – the tag is a related term of one of the descriptors according to the MeSH thesaurus (e.g. the "See also" relation);
6. Related – there is a relationship (conceptual, etc) between the tag and a descriptor but it is not formally expressed in the thesaurus;
7. Not Related – the tag has no apparent relationship to one of the descriptors.

## Results

### *Quantitative analysis of resources overlap*

According to the Mash-up, 1747 healthcare resources from Belgium were listed in CISMeF, while 113 of those resources were bookmarked in Delicious as well. We also identified a sample of 415 resources in Delicious, 13 of which were present in CISMeF (see Figure 1). The overlap between CISMeF and Delicious is thus very poor as only 6.5% [113/1747] of the CISMeF resources are listed in Delicious.

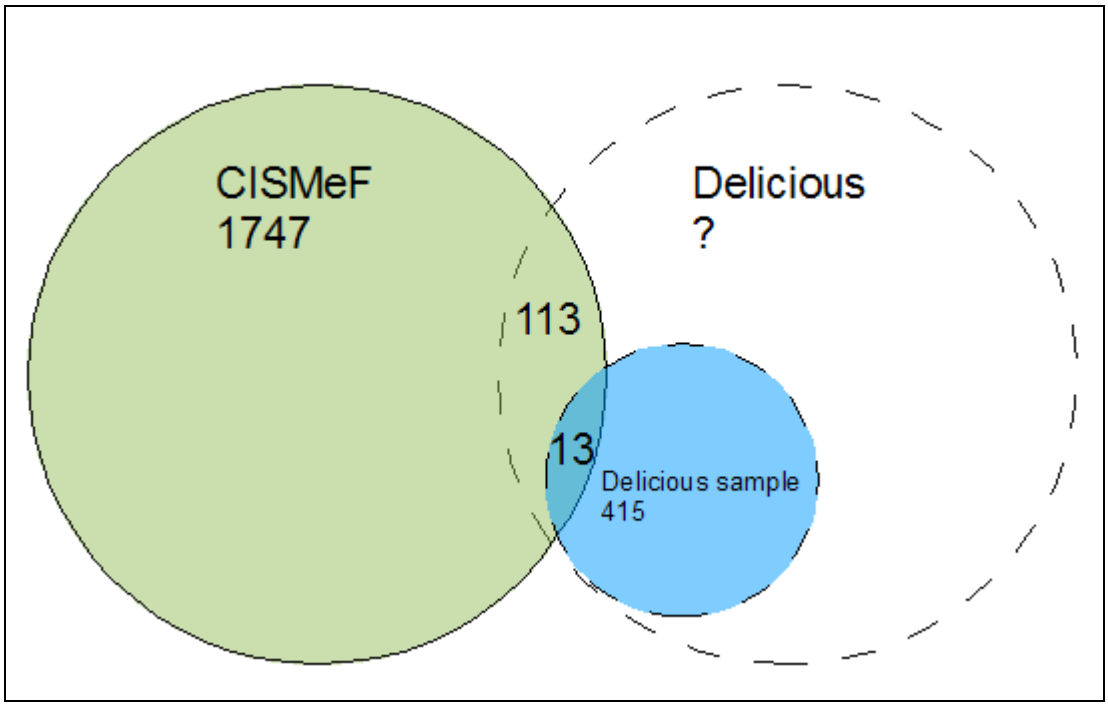


Figure 1: Overlap between CISMeF and Delicious

The resources listed in Delicious are very different in nature from the resources in CISMeF. Indeed, 85% of the URLs listed in both CISMeF and Delicious were websites homepages. The proportion is only 18.5% for the whole CISMeF URLs showing that CISMeF tend to index more deeper links and thus more specific documents than the Delicious users of our study. There's only 4 Belgian journal articles bookmarked in Delicious out of the 1228 indexed in CISMeF. It shows that Delicious users tend to focus on more general and popularised resources, while CISMeF tend to focus on scholar content.

Given the low overlap between Delicious and CISMeF, we can conclude that these systems are not redundant. The comparison of the resources listed in only one system shows that Delicous and CISMeF are actually complementary as providers of healthcare resources from Belgium.

*Qualitative analysis of resources metadata*

Descriptive statistics

Metadata for a total of 113 online resources was collected from Delicious and CISMeF. From this set, we discarded 7 resources which had not received any tag. (The potential absence of tag is due to the fact that Delicious allows users to bookmark resources without any obligation to assign them tags.) This data set included thus all Belgian resources listed in CISMeF and tagged by at least one Delicious user. Given that numerous resources were tagged by more than one user, the data set contained a total of 288 posts.

There were a total of 230 unique users in the data set. Each user name was associated with at least one post. The highest number of tagged resources per user is 8. A total of 197 users (86%) have tagged only one resource in the data set.

Similarly, when we examined the number of users who have tagged a specific resource, the maximum number of users per resource was 17, the minimum 1, and the median 2. A total

of 51 resources (49%) were tagged by only one user in the data set. Only 8 resources were tagged by 10 or more users.

The total number of descriptors in the sample was 561 whereas the total number of tags was 747. We considered that two tags are unique when similarly written, whatever their case and accentuation; Delicious does not detect those variations. For example, “Liège” and “liege” are considered as one and unique tag. The huge difference between the number of unique descriptors and unique tags shows that the indexation performed by taggers is more diversified than the librarians’ indexation. This can be partially explained by the fact that different languages were used by taggers.

| | Descriptors | Tags |
|---|---|---|
| Unique | 231 | 431 |
| Total | 561 | 747 |

Table III: Number of indexing terms

Term comparison

Using the seven-point scale proposed by Kipp [4], the term comparison has shown that the most frequent relationship between users’ tags and librarians’ descriptors was Not related (see Table IV), meaning that the most commonly used tags had no apparent relationship to one of the descriptors. Nevertheless, it does not mean that those tags are meaningless or useless (as showed below). The second most common relationship was Same, where the tag and one of the descriptors were identical or almost the same (e.g. plurals, spelling variations, acronyms and multiword terms).

| Same | Synonym | Broader Term | Narrower Term | Related Term | Related | Not Related |
|---|---|---|---|---|---|---|
| 201 (27%) | 20 (3%) | 118 (16%) | 16 (2%) | 21 (3%) | 114 (15%) | 257 (34%) |

Table IV: Frequency of relationship between users’ tags and librarians’ descriptors

Thesaural relations

Tags having a thesaural relation with descriptors according to the MeSH thesaurus fell into one of the following categories: Same, Synonym, Broader Term, Narrower Term and Related Term. Thesaural relations were as frequent as non thesaural ones. Approximately half of all the assigned tags (51%) did fall into a thesaural category. Users’ tags could thus compete with librarians’ descriptors.

The majority of the thesaural relations (85%) in our study were Same and Broader Term. The taggers tended to assign tags that were identical to descriptors or more general. A similar trend has also been observed above; Delicious users tended to bookmark more general and popularized resources, while CISMeF librarians tend to focus on scholar content. The most common Broader Term tags were “health” and “medicine”, and their translation in French and Dutch. Very few tags fell into Synonym, Narrower Term and Related Term categories.

Related tags

Tags falling into the Related category provide additional access points to a resource compared with librarian’s descriptors because their relation with descriptors is not formally expressed in the thesaurus. For example, the governmental web site of the Social Security

(http://inami.fgov.be) received the tag “government”, which did not have any thesaural relation with one of the assigned descriptors.

Unrelated tags

Tags that were not related to any descriptors assigned to the same resource tended to fall into two main categories, i.e. Tagger-related tags and Resource-related tags, which are themselves divided in several subcategories. Our categories and subcategories were inspired by the typologies of tags proposed by Golder [2] and Kipp [4, 8].

The Tagger-related category included 87 tags out of the 257 unrelated ones. These 87 tags tended to fall into two subcategories, i.e. Time & Task Management tags and Self Signification. The majority of the Tagger-related tags (77%) fell into the Self Signification category. This type of tags is meaningless for the other users of Delicious (ex.: “question_21_incitervisites”, “OutilsMG”, “B*”). The most frequent Time & Task Management tags were “thesis”, “work”, “stage” (French translation of “work placement” or “intership”), which are completely useless for any users other than the tagger himself.

The Resource-related category included the majority of the unrelated tags (68%). This category is divided in subcategories, i.e. Subject-related and Qualities & characteristics. The most common subcategory was Subject-related (90%) where tags described the content of the tagged resources. The most frequent Subject-related tags were acronyms and proper nouns (ex.: “institute_of_tropical_disease”, “Redcross”). There were few tags that described qualities or characteristics of the resource. This type of tags provides information on resources concerning their type (ex.: “guide”, “base.de.données” (French translation of “database”)), their language or their characteristic (ex.: “interesting”). By describing the subject and the characteristics, the Resources-related tags create additional access points to the resource compared with descriptors, as tags falling in the Related category do.

Usefulness of tags

A tag is considered as useful if it provides a relevant access point to the tagged resources. Tags falling in the following categories were thus considered as useful: Thesaural categories (i.e. Same, Synonym, Broader Term, Narrower Term and Related Term), Related and Resource-related (the subcategory of the Not Related category) tags. On the basis of this assumption, the majority of the tags (88%) from our dataset can be considered as useful in the information retrieval process; only 87 tags are meaningless and useless for Delicious users other than the tagger himself. Moreover, nearly half of those useful tags (284 tags) constitute access points that are not provided by descriptors. Such tags were those falling in Related and Resource-related categories.

Nearly 25% of the descriptors were not related to any of the tags by a Thesaural or Related relation. It means that numerous resources are incompletely described by their assigned tags but still findable in Delicious. Only 6 resources of our dataset (0.5%) are impossible to retrieve in Delicious due to a lack of useful tags. For example, the resource titled “Ligue Nationale Belge de la Sclérose en Plaques” (http://www.ms-sep.be) only received the tag « mam » only. Therefore, it cannot be found by other Delicious users other than the tagger himself.

## Conclusions

This study demonstrated that there is a low overlap between healthcare resources from Belgium listed in CISMeF and Delicious, which showed that those two information systems are not redundant. They are actually complementary by reaching out to different audiences. While Delicious tends to provide popularized resources, CISMeF lists more scholar documents.

The comparison of users' tags with librarians' descriptors showed a similar trend; Delicious users tend to assign tags that are identical to descriptors or more general. Moreover, numerous tags provide additional access points to the tagged resources compared with descriptors. The majority of tags are thus relevant and useful for the information retrieval process. Nevertheless a quarter of the assigned descriptors were not represented at all by any of the users' tags. This study thus demonstrated that users' tags complement and even compete with librarian's descriptors but in no way could substitute for them.

In further research, we plan to compare tags assigned by CiteULike users with descriptors provided by librarians to a set of scholar papers in the Healthcare field. The librarian's descriptors will be collected in bibliographic databases, such as PubMed. In this way, we assume that users' tags will be more specialized than those in Delicious given that CiteULike is expressly made for academics.

## References


[1] Doyle T, Hammond JL. Net cred: evaluating the internet as a research source. Reference Services Review. 2006;34(1):56-70.

[2] Golder SA, Huberman BA. Usage Patterns of Collaborative tagging systems. Journal of Information Science. 2006 Apr;32(2):198-208.

[3] Vander Wal T. Explaining and Showing Broad and Narrow Folksonomies. 2005 Feb 21 [cited 2009 Apr 10]. Available from: http://www.vanderwal.net/random/entrysel.php?blog=1635.

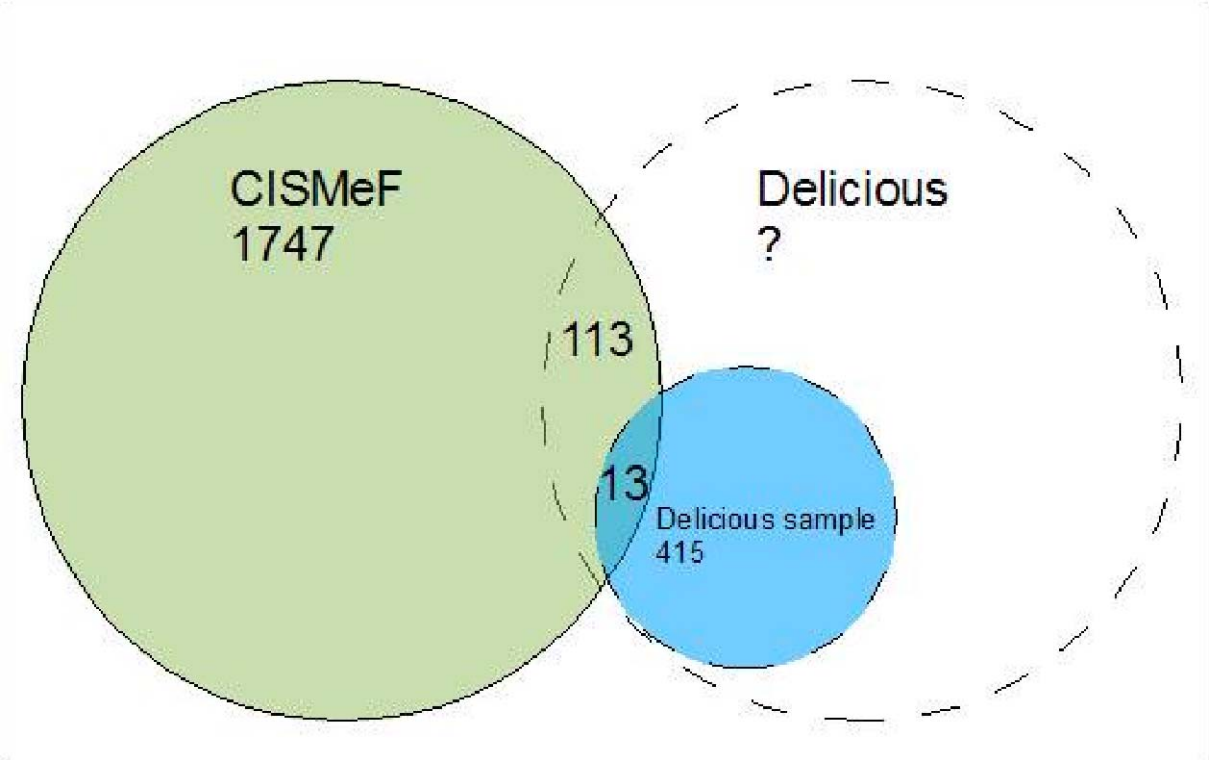
[4] Kipp M. Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords. In Proceedings Canadian Association for Information Science. 2006; York University, Toronto, Ontario, Canada.

[5] Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B. CISMeF: a structured health resource guide. Methods Inf Med. 2000 Mar;39(1):30-5.

[6] Gordon-Murnane L. Social bookmarking, folksonomies, and web 2.0 tools. Searcher. 2006; 14(6):26-38.

[7] Arrington M. Exclusive: Screen Shots And Feature Overview of Delicious 2.0 Preview. 2006 Sept 6 [cited 2009 Apr 10]. Available from: http://www.techcrunch.com/2007/09/06/exclusive-screen-shots-and-feature-overview-of-delicious-20-preview/.

[8] Kipp M. @toread and Cool: Tagging for Time, Task and Emotion. poster presented at the Annual Meeting of the American Society for Information Science and Technology (ASIST). 2006 Nov 4; Austin, Texas.


CISMeF
1747
Delicious
?
113
13
Delicious sample
415