The Impact of Image Descriptions on User Tagging Behavior: A Study of the Nature and Functionality of Crowdsourced Tags

Yi-ling Lin[1*], Christoph Trattner[2], Peter Brusilovsky[1], Daqing He[1]

[1] School of Information Sciences, University of Pittsburgh, USA

[2] Know-Center, Graz University of Technology, Austria

{yil54, peterb, dah44}@pitt.edu, ctrattner@know-center.at

* indicates corresponding author.

Abstract

Crowdsourcing has been emerging to harvest social wisdom from thousands of volunteers to perform series of tasks online. However, little research has been devoted to exploring the impact of various factors such as the content of a resource or crowdsourcing interface design to user tagging behavior. While images' titles and descriptions are frequently available in image digital libraries, it is not clear whether they should be displayed to crowdworkers engaged in tagging. This paper focuses on offering an insight to the curators of digital image libraries who face this dilemma by examining (i) how descriptions influence the user in his/her tagging behavior and (ii) how this relates to the (a) nature of the tags, (b) the emergent folksonomy, and (c) the findability of the images in the tagging system. We compared two different methods for collecting image tags from Amazon's Mechanical Turk's crowdworkers – with and without image descriptions. Several properties of generated tags were examined from different perspectives: diversity, specificity, reusability, quality, similarity, descriptiveness, etc. In addition, the study was carried out to examine the impact of image description on supporting users' information seeking with a tag cloud interface. The results showed that the properties of tags are affected by the crowdsourcing approach. Tags from the "with description" condition are more diverse and more specific than tags from the "without description" condition, while the latter has a higher tag re-use rate. A user study also revealed that different tag sets provided different support for search. Tags produced "with description" shortened the path to the target results, while tags produced without description increased user success in the search task.

Keywords: Crowdsourcing, image description, tagging behavior, Amazon Mechanical Turk, image search.

The Impact of Image Descriptions on User Tagging Behavior:

A Study of the Nature and Functionality of Crowdsourced Tags

1. Introduction

Crowdsourcing has emerged as a popular modern approach to perform information processing tasks that are difficult or impossible to automate. Among other applications, crowdsourcing has become a powerful mechanism to harvest collective wisdom from thousands of volunteers (Howe, 2008; Surowiecki, 2004). A good example of this kind of task is image annotation with keywords (known as image tagging). Keywords are critical for finding images, yet computers cannot tag images automatically. The need to annotate images in digital libraries and other image collections has encouraged researchers and practitioners to explore a range of crowdsourcing approaches to collect image tags. This idea was pioneered by early image sharing systems (e.g., Flickr) and was later scaled up using game-based approaches (e.g., ESP game[1]) and paid crowdsourcing marketplaces (e.g, Amazon's Mechanical Turk[2]) (Nowak & Rüger, 2010; Sorokin & Forsyth, 2008).

The importance of tag crowdsourcing, in turn, encouraged a stream of research focused on the quality and other properties of crowdsourced tags. This research was stimulated by the early work of Golder & Huberman (2006) and Kowatsch & Maass (2008), who discovered that the quality and diversity of crowdsourced tags can be affected by various components of the tagging interface, such as presentation of current and recommended tags. Following this discovery, a few other teams explored user-tagging

[1] ESP game collects image metadata by engaging users in an image tagging game which was originally conceived by Luis von Ahn of Carnegie Mellon University. Google bought a license in 2006 to increase the keywords of images for its online image search.

[2] Amazon Mechanical Turk is a crowdsourcing Internet marketplace where individuals or businesses can recruit human intelligence to perform tasks that computer are not able to do.

behavior and the impact of various parameters on the properties of produced tags. The study presented in this paper extends this research by examining the impact of *image description* on the nature and the quality of crowdsourced tags[3]. This study is important from both theoretical and practical perspectives. Image tagging context with non-textual primary content and secondary textual descriptions is considerably different from the already-explored context characterized by primary textual content and tag recommendations. Existing empirical data and models are not sufficient to reliably predict the impact of image descriptions on tag production. New empirical data has to be collected to expand and generalize known models. On the practical side, the study is important to guide managers of image collections in the process of tag crowdsourcing. While image descriptions are frequently available in image collections, it is not clear whether or not they should be displayed to crowdworkers engaged in tagging. On the one hand, the presence of image description could help the crowdworkers to generate more tags and to make them more specific. On the other hand, it could curb their creativity and harm the diversity of the resulting tags.

The goal of our study was to collect and analyze empirical data on the impact of image descriptions on tagging to advance both our understanding of the tagging process and the current practice of tag crowdsourcing in image collections. Due to the lack of relevant models that describe the process, we designed the study in an open format. That is, instead of attempting to prove or disprove outcomes predicted by the existing theory, we adopted an empirical approach that is common for examining human-computer interfaces and formulated our research questions in the following way:

[3] The quality of crowdsourced tags in this context is considered from the perspective of tag usage, which contains tag reusability (Nowak & Rüger, 2010; Sen et al., 2006), resource findability and resource discrimination (Dellschaft & Staab, 2012).

- RQ1: How descriptions influence the user in his tagging?
- RQ2: How this relates to the (a) nature of the tags, (b) the emergent folksonomy, and (c) the findability of the images in the tagging system?

According to the study of Nowak and Ruger (2010), tags crowdsourced via Amazon's Mechanical Turk are less costly, yet as reliable as expert-level tags. Therefore, this study compared two different methods for collecting image tags from Amazon's Mechanical Turk crowdworkers – with and without image descriptions. We investigated the properties of generated tags from different perspectives including generality, quality, similarity, and descriptiveness. We also conducted a user study on Amazon's Mechanical Turk to compare users' image search performance using tags sets produced with and without descriptions.

The remainder of the paper starts with the analysis of similar work (Section 2) and follows by presenting three components of our study. Section 3 explains the process of tag crowdsourcing for our study. It introduces the image collection, presents two interfaces for tag crowdsourcing (with and without descriptions) and provides some information about produced datasets. Section 4 focuses on the qualitative comparison of image tags crowdsourced with and without descriptions. It examines density, generality, consistency, and other properties of tag collections. Section 5 examines the practical differences between two collected tag sets by comparing their impact on image findability in a user study. Finally, Section 6 completes the paper with a discussion of the results and limitations of the present study.

2. Related Work

2.1 Social Tagging

Crowdsourced tags are also called “social tags” or “collaborative annotations” in the literature. They are outcomes of a distributed practice performed by internet users in organizing and indexing online digital objects, such as Web pages, video clips and images (Wu, He, Qiu, Lin, & Liu, 2012). Because they fit well with the social Web’s general principle of sharing and participating, crowdsourced tags quickly established themselves as one of the major forces for converting the static Web into a participatory information space (Ding, Jacob, Yan, George, & Guo, 2009). Consequently, there is a considerable amount of work in the literature focusing on various aspects of tags and tagging behaviors, which include tagging motivation (Ames & Naaman, 2007; Nov & Ye, 2010; Strohmaier, Körner, & Kern, 2010), navigation (Chi & Mytkowicz, 2007; Helic, Trattner, Strohmaier, & Andrews, 2010), and search (Bischoff, Firan, Nejdl, & Paiu, 2008; Heymann, Koutrika, & Garcia-Molina, 2008; Trattner, Lin, Parra, & Brusilovsky, 2012). Readers who want to know more about this topic should read Gupta’s overview article (Gupta, Li, Yin, & Han, 2011) and Smith’s book (Smith, 2007).

2.2 Tag Properties and Methods of Tag Analysis

Tagging behavior is usually modeled into a tripartite relationship. A user *u* applies *n* tags to categorize a resource *r* (Bollen & Halpin, 2009). Golder and Huberman (2006) investigated *tag-resource distribution* by discovering regularities in user activity, tag frequencies, kinds of tags used, bursts of popularity in tagging and a remarkable stability in the relative proportions of tags within a given resource. Suchanek, et al. (2008) proposed a way to quantify social tags’ *meaningfulness* and the imitating tags phenomenon by analyzing the semantic properties of tags and the tag-resource relationship. To examine the meaningfulness of tags, they applied precision and word classes as metrics. For the tag-resource relationship, they adopted the *matching rate*, the *imitation rate*, the *popularity bias* and so on. Dellschaft and Staab (2012) explored indicators of *indexing quality* by utilizing the *inter-resource consistency* (White, 1987) and the *inter-indexer consistency* including *tag reuse* (Nowak & Rüger, 2010; Sen et al., 2006) and the *size of vocabulary* (Floeck, Putzke, Steinfels, Fischbach, & Schoder, 2011; Kowatsch & Maass, 2008). Lee and Schleyer (2012) investigated to what extent social tagging can substitute for controlled indexing in the medical domain. They applied descriptive analyses of the data sets and similarity measures to compare the sets.

2.3 The Influence of the Tagging Interface on Tagging Behavior and Tag Vocabulary

With the increased popularity of social tagging, tagging systems attempted to offer better support to their users. In particular, many systems explored sophisticated approaches in recommending and auto-completing tags for users. However, researchers observed that user tagging behaviors and the tags generated from those behaviors were influenced by the presence of tag recommendation. Golder and Huberman (2006) found that the tag vocabulary of a user stabilizes over time at approximately 100 unique tags in Delicious, a social tagging system, due to tag recommendations. They attributed this to what is known as the 'rich get richer' phenomenon in the classical *Polya Urn model*. Similarly, Kowatsch and Maas (2008) demonstrated that the uncontrolled nature of collaboratively collected terms is significantly reduced if predefined terms are presented to the user. Several researchers attempted to build models of the tagging process that can represent the impact of tag recommendation. Beyond the Polya Urn model, Cattuto et al. (2007) formalized a modified *Yule-Simon model*, where the probability of an existing tag decaying along with the power-law distribution, which not only represents the addition of new tags but also the imitation of existing tags. Halpin et al. (2007)

introduced a generative model for tagging based on the preferential attachment principle that partially reproduces the frequency-rank distributions. Dellschaft and Staab (2008) introduced the epidemic model for tagging that is not only based on the users previous tag assignments but also accounts for the factor of background knowledge of the users. It is the first generative model that can also simulate well the sub-linear tag growth in tagging systems. Bollen and Halpin (2009) conducted an experiment to show that imitation is not the only reason for power-laws in tagging systems. In fact they find that power-law distributions also occur if no tag suggestion mechanism is provided in the interface.

A related stream of work explored the impact of existing tags on user tagging behavior from the prospect of implicit imitation models for tagging and the generation of folksonomies. The most notable study on the implicit imitation of tags is the work of Fu et al. (2010), who argued that imitation in tagging systems happens on the semantic level rather than on the word level. Lately, Seitlinger & Ley (2012) also introduced a multinomial model derived from Fuzzy Trace Theory for implicit and explicit tag imitation. A recent study of Lorince and Todd (2013) showed that tag imitation behavior is even simpler: the top-*n* most popular tags of a resource are a good predictor how people select their own tags. This goes in line with other related work such as Floeck et al. (2011) and Moltedo et al.(2012), who showed that existing tags have a direct influence on how tags are generated by the users for a given resource.

Much less work has been done to explore the impact of other aspects of the tagging interface on tag production. Most notable here are the studies of Heyman et al. (2008) and Lipczak and Milios (2011) who investigated the impact of features derived from *Web page content* on user tagging behavior. Investigating in-depth a number of content features such as page text, anchor text and surrounding hosts, or page title, they found that the page text features and page text title features show the highest correlation to predict the users' social tags for a given resource in a personalized manner.

In this context, our paper seeks to extend the body of research in two directions. First, we are interested to explore the impact of a different element of the tagging interface – image description – in an unexplored context of image tagging. A combination of *non-textual primary content* and *secondary textual content* makes image tagging considerably different from the explored cases. Existing empirical data and tag production models offer conflicting predictions. Imitation models focused on single user behavior predict that words from descriptions will complement the terms produced by observing images making the resulting tags *more diverse*. Global tag production models, however, predict the *decrease* of tagging diversity as a result of exposing all taggers to the same secondary textual content. A detailed analysis of differences between tags produced with and without descriptions expands our understanding of tagging process and might lead to better models.

Another contribution of this work is to extend the usual analytical approach to tag analysis with an empirical one. In the study presented below we explore both which aspects of produced tags are impacted by image descriptions and how these changes affect the functionality of tags, i.e., their ability to support search.

3. Tag Crowdsourcing Process

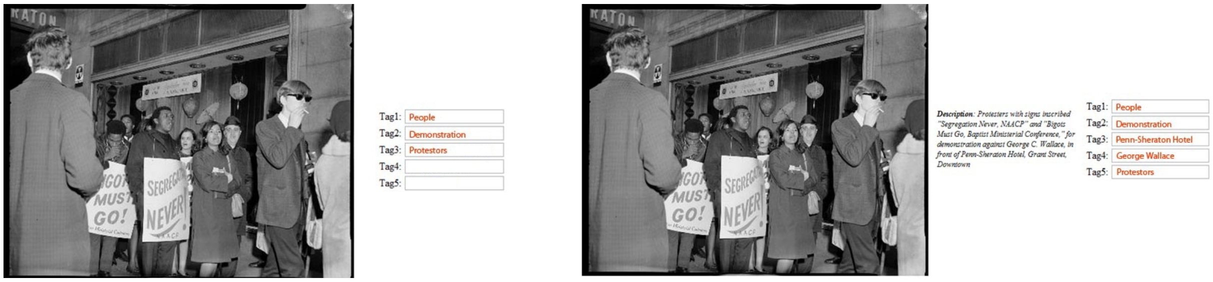
The study uses the Teenie Harris Archive, an image collection from the Carnegie Museum of Art. The collection includes more than 80,000 images and captures a 40-year period of Pittsburgh African-American life. In this study, we used 1,986 representative images, of which 986 have been selected by museum curators for a recent exhibition at the museum. The remaining 1,000 images were selected by us to provide a good overview of the entire collection.

All 1,986 selected images had curator-produced descriptions. In the process of crowdsourcing we generated two sets of tags for each image as explained below.

Amazon's Mechanical Turk (MTurk) offers a variety of pre-designed interfaces to facilitate the tag-collecting process. In order to keep our collecting methods comparable, we used the same "Image tagging" template provided by MTurk to produce both sets of tags. The only difference between the two tag-collection interfaces was the presence of image description. The interface "with descriptions" contains task requirements, an image with its original description, and five textboxes to allow a *turker* (short for MTurk crowdworkers) to input tags for the image (see Figure 1). The interface "without descriptions" included only task, image, and textboxes.

To increase the quality of produced tags while keeping the free-style nature of tagging, the tasks included a minimal number of requirements for tag collecting. First, turkers were required to apply at least two tags consisting of a maximum of three words. Second, turkers were asked to provide tags that are useful for finding the image. In order to explain to turkers how to generate meaningful tags, we suggested they imagine what kind of keywords a user on a search engine such as Google or Yahoo! would use to find the image. Beyond that, because the whole image collection is black and white, the tags "without color," "black" and "white" are not valid to our tag assignment.

We deployed each image with each interface as a separate human intelligence task on MTurk and ensured that each task was completed by three turkers (i.e., three annotators per image). Thus, each of 1,986 images was annotated by six turkers – three were able and three were not able to see image descriptions.



Tag1: People
Tag2: Demonstration
Tag3: Protestors
Tag4:
Tag5:

***Description**: Protesters with signs inscribed "Segregation Never, NAACP" and "Bigots Must Go, Baptist Ministerial Conference," for demonstration against George C. Wallace, in front of Penn-Sheraton Hotel, Grant Street, Downtown*

Tag1: People
Tag2: Demonstration
Tag3: Penn-Sheraton Hotel
Tag4: George Wallace
Tag5: Protestors

***Figure 1.** Tagging interface in Mechanical Turk: w/o descriptions (top) and w/ descriptions*

4. The Impact of Image Descriptions on User Tagging Behavior

In this section, we analyze an immediate impact of the image description on user tagging behavior by comparing the two sets of tags produced by crowdworkers with and without descriptions. To discover the nature of the difference of collected tags, we follow the literature (see Section 2.2) and focus on (1) the descriptive properties of tags (S. Golder & Huberman, 2006), (2) the imitation of image description (Dellschaft & Staab, 2008), (3) the similarity of tags assigned to the same resource (Dellschaft & Staab, 2012; Fu et al., 2010), (4) tag frequencies for popular tags in each setting (Lee & Schleyer, 2012), (5) tag diversity (unique tags in each setting), (6) tag specificity (Lee & Schleyer, 2012), tag length and dictionary matching (Suchanek et al., 2008), and (7) tag reusability (Nowak & Rüger, 2010; Sen et al., 2006) and resource discrimination (Dellschaft & Staab, 2012).

4.1 The Collected Tag Datasets

The crowdsourcing process explained above generated two tag datasets for the 1,986 selected images from the Teenie Harris Collection. **The W/O dataset** consists of 16,659 tags (4,206 unique tags) produced by 97 turkers *without* seeing image descriptions. **The W/ dataset** consists of 17,541 tags (6,418 unique tags) produced by 159 turkers who were able to see image descriptions.

Table 1. *Statistical properties of tags derived from two tagging modes*

	W/O	W/
Number of tags (unique tags)	16659 (4208)	17541 (6427)
Average working time per image (sec.)	40.42	49.49 ***
Average number of resources per tag	3.62**	2.43
Mean (median) agreement rate	.0835 (.0833)	.106 (.10)**
Average number of images per turker	61.39	37.45
Number of turkers	97	157

(**=significant at *p*<.01, ***=significant at *p*<.001)

Table 1 summarizes descriptive data of two datasets revealing some important differences. As the data shows, the taggers who were able to see the description spent significantly more time on the task (as observed by the *average working time*). It provides some preliminary evidence that image annotations were noticed and read. The data also provides some evidence about a conflicting impact of the description predicted by existing theories. On the one hand, the presence of annotations did lead to an *increased* tag production and global tag diversity as evidenced by the increase of the total *number of tags* and *unique tags* and the *average number of unique tags per image*. On the other hand, it did *reduce* inter-tagger diversity, as shown by a significantly increased *agreement rate* (agreement rate is the number of repeated tags divided by the total number of tags per image). The early analysis confirms the complex impact of descriptions on image tagging and motivates a detailed analysis provided below.

4.2 Does the Presence of Image Description Affect the Tagging Process?

The first step of our analysis is to confirm that the presence of image descriptions does, indeed, affect tagging. As mentioned in Section 2, existing models and empirical data show that the presence of textual prompts such as Web page title or proposed tags does affect tagging behavior by causing the users to use title terms or proposed tags. Following these results and models, we can hypothesize that the presence of image descriptions leads to an increased use of words found in these descriptions as tags. A proof of this hypothesis could be obtained by demonstrating an increased overlap between tags and the words from corresponding image descriptions. This would also provide evidence that the descriptions do affect tagging.

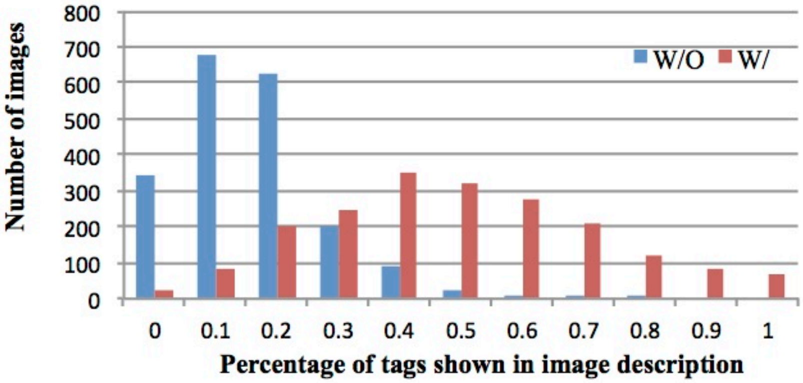


Figure 2. *Percentage of tags shown in image description*

The result of our analysis confirms the imitation hypothesis. The data shown in Figure 2 demonstrates that tags produced without image descriptions have a much smaller overlap with the words used in description than the tags produced when the description was displayed to taggers. The difference is very pronounced. In the W/O condition, 83% of images have tag sets that show 20% or less overlap with the image description. In contrast, in the W/ condition, around 70% of images have tags with greater than 30% percent overlap with image's description. Thus, the presence of description affected the taggers in the same way as the presence of suggested tags and headers, encouraging the selection of the words in the description as tags.

More reliably the impact of descriptions on the tagging process can be uncovered on the level of individual images by demonstrating a correlation between the properties of individual image descriptions and the corresponding tags. Given the observation that the presence of descriptions leads to the increased number of total and unique tags (Table 1), we might hypothesize that this dependency is continuous, i.e., *longer and richer* descriptions elicit progressively *more* tags. To prove this hypothesis we examined the correlation between common text-quality metrics of image descriptions (number of characters, number of words, entropy and number of unique words (Anderka, M., Stein, B., & Lipka, 2012; Pitler & Nenkova, 2008)) and the number of (unique) tags produced for each image in both W/ and W/O conditions. To check the assumption of normal distribution, the Kolmogorov-Smirnov test has been carried out showing that the results were highly significant (*p*<.001) for all variables. Considering the results of the KS-Test, the data apparently is not normally distributed. Although this fact may be caused by the large sample size, we cautiously applied Spearman's correlation as a non-parametric alternative to Pearson's correlation. As shown in *Table 2*, the analysis shows a positive correlation between all text-quality metrics and the number of both total and unique tags produced in W/. Although the effect size is relatively small it is statistically significant, i.e., the size/quality of the description positively impacts the number of tags produced. As expected, no correlation can be observed in the W/O: the correlation value is close to zero and the correlation is not statistically significant. This confirms that the correlation is caused by the properties of image annotations, not by the properties of images themselves.

Table 2. *Pearson correlation between description properties and tag volume*

	W/O		W/	
Text Quality Measure	Number of tags	Number of unique tags	Number of tags	Number of unique tags
Character Count	.032	.067	.121***	.205***
Entropy	.047	.077	.127***	.218***
Word Count	.044	.077	.125***	.216***
Word Unique Count	.047	.077	.127***	.218***

(***=significant at *p*<.001)

The analysis above shows clear evidence of the impact of descriptions on tagging, however, this impact might still fail to produce considerably different datasets. To proceed to a more detailed analysis, we wanted to ensure that the difference between the two tag sets is *sufficient*. To measure the scale of difference between the two tag sets, we utilized two measures, the Cosine similarity and the Jaccard similarity coefficient. According to Figure 3, most images are assigned with dissimilar tags (most images are represented with tags which are dissimilar to each other at the similarity level 0.1). This makes the two datasets sufficiently different for a detailed examination. At the same time, the descriptions do not fully define the produced tags. The rank correlation analysis demonstrates a strong positive correlation between the two conditions, r =.62, *p*<.001, confirming the lasting impact of visual content that is the only thing shared between conditions. Taken together, this data demonstrates both the images and their descriptions provide a considerable impact on produced tags.

Figure 3. *Image-level tag similarity between two conditions. The data shows the number of images for each of the similarity levels from 0.1 to 0.9 using two similarity measures*

4.3 Exploring the Nature of Differences

To get a hint of the nature of differences between the two tag sets, we compared the most popular tags in the two conditions. *Table 3* shows the data on the top 20 tags in each condition.

The tags are displayed in descending popularity order, along with the number of images annotated with the tag, and the number of turkers who used the tag. To ensure that we are looking at truly popular tags, we considered only tags used by at least 5 turkers in each condition.

Table 3. *Top 20 tags in W/O and W/ conditions used by at least 5 turkers sorted by number of tag assignments*

W/O				W/			
tag	# tag	# resources	# users	tag	# tag	# resources	# users
people	408	330	47	man	342	282	37
car	403	241	24	woman	309	243	47
men	394	354	24	men	303	277	32
man	329	269	22	car	264	182	31
women	287	256	23	people	263	220	49
street	284	247	24	women	261	240	32
tree	273	226	16	tree	140	124	11
woman	241	189	27	street	135	120	25
building	229	186	24	tree	140	124	11
table	227	179	20	cap	109	99	7
suit	165	159	11	building	106	97	23
peoples	146	138	7	suit	104	98	13
room	138	124	12	group portrait	95	81	31
cap	135	116	7	table	94	81	12
road	122	108	14	children	92	74	36
standing	117	106	10	hill district	81	71	35
children	115	101	25	girl	81	77	14
girl	110	102	15	portrait	76	71	30
smile	96	95	8	boys	65	49	22
house	89	76	21	church	63	49	27

The first look on the top tags shows an overall expected trend: general terms such as “people,” “man,” ”woman,” ”car,” “street,” “tree,” etc., are the most popular tags used in both conditions. Yet it also shows two opposite trends. First, line-by-line analysis reveals that the same top general tags are more heavily used in the W/O condition. These tags are more frequent and are used to describe more documents. Overall, taggers in W/O used considerably more tags from the top 20 tags to annotate images (28% of tags are the top 20). In contrast, users in W/ condition apply only 18% of tags from the top 20, hinting that in this condition turkers use a broader set of tags to annotate images. Second, there is a hint that the tags used in the W/ condition are not just broader and more diverse, but also more specific. This is indicated by two compound phrases “group portrait” and “hill district” that made it into the top 20 tags. In the following sections we will examine these two observations in detail.

4.4 Tag Diversity

In this section we examine in detail the hypothesis that the tags assigned by the turkers in the W/ condition are more diverse than the tags assigned in the W/O condition. The evidence in favor of this hypothesis was provided by comparing the total numbers of produced tags and was supported by the analysis of the top 20 tags. A more reliable proof of a more diverse indexing should be obtained on the image level. To demonstrate evidence that one tagging approach produces more diverse tag sets than the other, we can show that it enables a group of taggers dealing with an image to generate *more* tags (and, more importantly, *more unique* tags) and that these tags are more *diverse* as a group.

To check the tag volume hypothesis, we performed a Mann-Whitney test to compare the number of tags assigned to each image in two conditions. The analysis shown in *Table 4* and Figure 4 revealed that the W/ tag data set collected *with* descriptions has both significantly larger *total* number of tags per resource than the W/O data set (*p* <.001*)* and significantly larger number of *unique* (non-repeating) tags per resource (*p*=.023).

Table 4. *Number of tags or unique tags per resource in two conditions*

	Overall		Unique	
	W/O	W/	W/O	W/
Mean	8.392 (8)	8.836 (9) **	7.676 (8)	7.853(8)**
SD	1.703	1.864	1.702	1.741

(**=significant at *p*<.01)

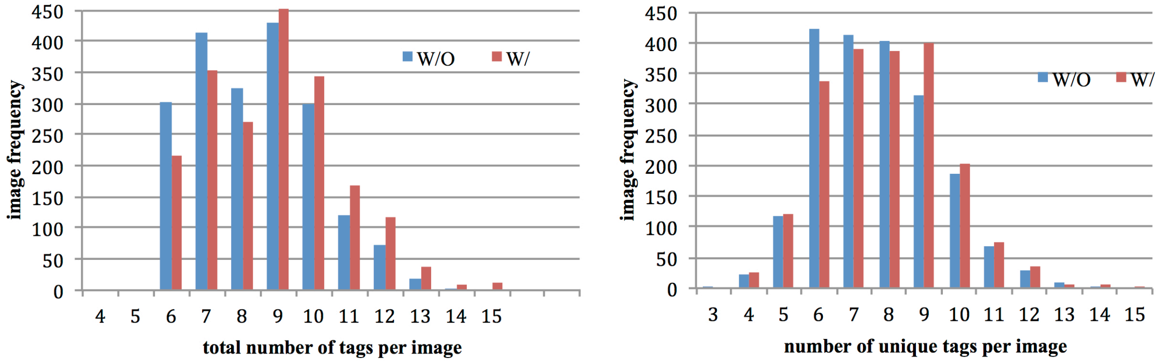


Figure 4. *The distribution of the number of tags (left) and number of unique tags per image*

While the increase of the number of tags is an important trend making tags produced with descriptions more diverse, we also discovered an opposite trend – the unifying impact of descriptions on user tagging as shown by the increase of inter-indexer agreement rate. To determine which of these trends produces larger impact of tag diversity, we compared tag diversity *within each image* using image-level inter-tag similarity. This measure tells how similar tag strings are for a given resource. We chose a well-established metric called the Levensthein (1966) distance to evaluate the difference between two given strings (tags). A *t*-test was performed to check the difference between the two data sets. The results show that tag set produced in W/ condition (M=9.68) has significantly longer Levensthein distance between image tags than W/O text description (M=7.39), *p*<.001 (F=1021.359, partial eta squared=.205). It demonstrates that tags assigned in the W/ condition to each image are more dissimilar to each other. This result provides compelling evidence that the diversity-increasing trend is stronger than the unifying trend and confirms that tags indexing produced in the W/ condition are *more diverse* than the indexing in the W/O condition.

4.5 Tag Specificity

The analysis of the top 20 tags hinted that turkers working in the W/ condition might assign *more specific* tags to images. This section attempts to explore this hypothesis more thoroughly using several measures. Since longer terms tend to carry more information and to be more specific, we started with a common way to demonstrate tag specificity by measuring the length of tags in words and characters. Figure 5 shows the distribution of number of words and number of characters per tag in two conditions. As the data shows, the W/ condition has a larger number of words (M=1.89) and a larger number of characters (M=8.91) per tag than the number of words (M=1.67) and the number of characters (M=6.81) of each tag in the W/O condition. T-test results show that both differences between these two conditions are significant (*p*<.001). The results provide evidence that the tags produced in W/ condition are more specific.

Figure 5. *The distribution of number of characters (left) and the number of words per tag (right)*

Another approach to argue for generality/specificity of a tag set is to examine which fraction of tags can be found in a common dictionary. Words in a dictionary are more general (i.e, we will not find "Hill District" in a common dictionary), so the larger the overlap, the more general the tag set. We checked whether the tags (overall tags and unique tags) in different conditions appear in the Oxford English Dictionary (Oxford University Press, 2010). As we can see from *Table 5*, for both overall and unique tags, there was a significant difference between two tagging conditions: more tags from the W/O condition appear in the Oxford English dictionary than from the W/ condition. It provides additional evidence that tags from the W/O condition are more general than those from the W/ condition.

Table 5. *Number of tags (percentage of tags) in the dictionary*

	No. of Tags in Dict (No. of unique tags in Dict)	No. of Tags (No. of unique Tags)	
W/O	10356 (1312)	16659 (4206)	62.16% (31.19%)
W/	8259(1508)	17541 (6412)	47.08% (23.52%)
			p<2.23-16 (*p*<2.23-16)

One concern about using an exact match with a dictionary is the potential use of word forms derived from general dictionary words (such as use of plurals observed in top-20 indexing). The fact that only 31% and 24% of unique tags were present in the exact form in the Oxford English dictionary provides evidence that some fraction of tags might be derived from dictionary words. To explore the fraction of these "derived" tags in both conditions, we calculated the minimum edit distance between tags and terms in the dictionary (Levenshtein, 1966). The minimum edit distance is a method to do non-word error detection. For any word not in a dictionary, we assume it is a spelling error and needs to be edited from incorrect spelling to the correct spelling in the dictionary. In the W/ condition, the average edit distance (M=2.11) is significantly higher, *p*<.001, than the average edit distance (M= 0.97) in the W/O condition. This indicates that a larger fraction of tags in the W/ condition are more specific terms, which can neither be found in the dictionary, nor easily derived from the words in the dictionary (or may be spelled incorrectly more frequently, which also points to less general terms).

4.6 Tag Reuse and Resource Discrimination

We were also interested in how often a tag was reused across the resources, which is the frequency of each tag annotated to images in the collection. According to Dellschaft and Staab (2012), this measure is an important indicator of the quality of the index induced from the tags of the given folksonomy. The more tags are reused across the resources, the more the resources are connected with each other. In addition, this measure also implies that the number of retrieved items of a tag increases when searching the specific tag.

We performed a Mann-Whitney test to discover whether there is any significant difference affected by different tag recruiting methods. A Mann-Whitney test indicated that the tag reuse rate in the W/O condition (M=.125) was significantly larger than in the W/ condition (M=.111), *p*<.001. This result shows that taggers in the W/O condition tended to reuse the same tags to annotate images more frequently that might increase more number of items retrieved back by issuing a specific tag.

In addition, we calculated inter-resource similarity in terms of tags to discover the similarity of users' behavior between two conditions as below. The inter-resource similarity is another theoretical measure (Dellschaft & Staab, 2012) to determine the quality of the tag data sets in terms of information access. If images have been annotated with more similar tags to each other, those similar tags are less powerful to distinguish a specific image. We calculated the Cosine similarity and Jaccard similarity among all images based on the tags assigned in two conditions in Figure 6. We found that W/O has significantly higher similarity rates on cosine (M=.039) and on Jaccard (M=.020) than W/'s similarity rates on cosine (M=.020) and on Jaccard

(M=.011). In other words, image tags are more distinct in the W/ conditions making it easier to discriminate resources by their tag index.

Figure 6. *Inter-resource Cosine (left) and Jaccard similarity of two conditions*

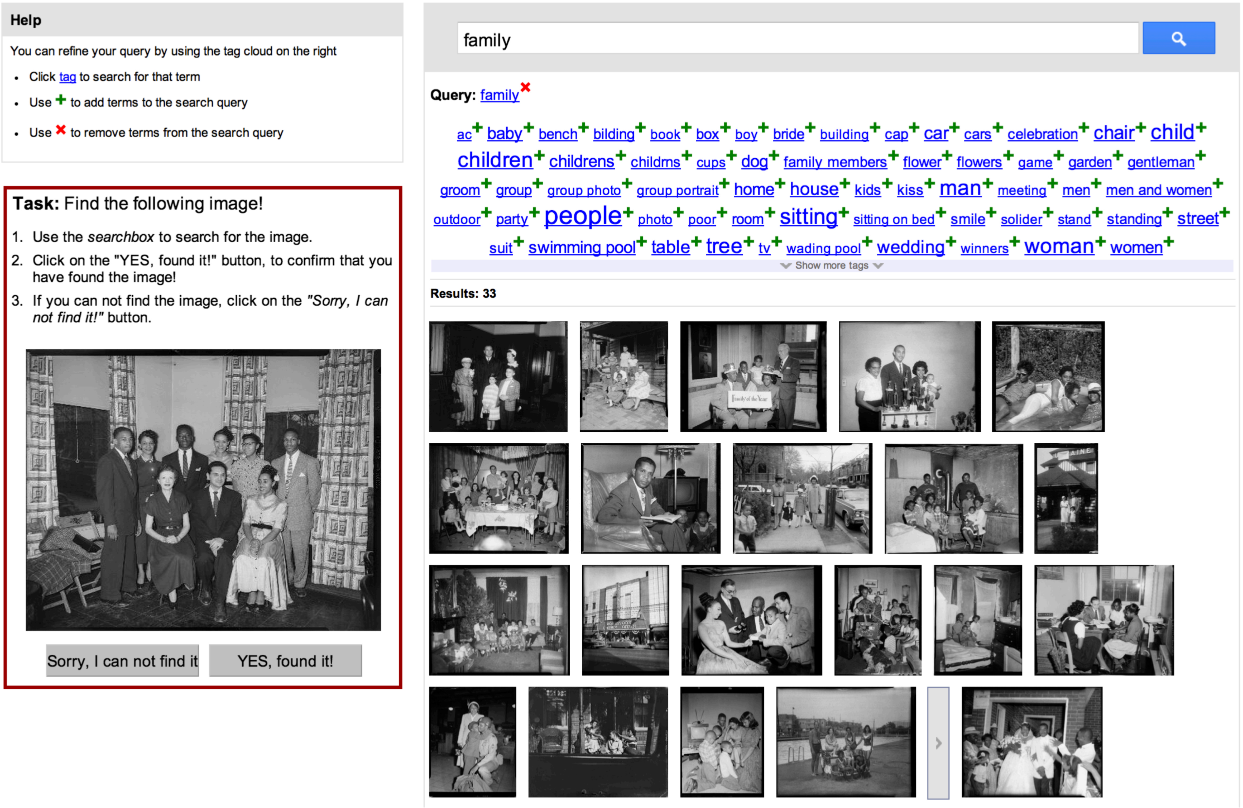
The results presented in this section are consistent with the results reported in earlier sections. Altogether these results provide an interesting insight on the nature of differences between tags obtained with and without image descriptions. In brief, the presence of image descriptions encourages taggers to consider the description text, which, in turn, results in more diverse and specific indexing that also makes images more distinct. At the same time, it results in a considerable decrease in tag reuse, making the indexing sparser. These two opposite tendencies make it hard to argue which approach is better from the perspective of image finding. On the one hand, more dense indexing with a heavier use of generic terms and more images indexed with the same term, gives a higher chance for a user in the W/O condition not to miss an image. On the other hand, a more diverse and distinct nature of indexing in W/ condition could make it easier to find a specific image among many others. Thus, a user study reported in the next section was necessary for us to find the practical differences between two approaches to tag collection.

5. The Impact of Tag Crowdsourcing Approach on Tag-Based Image Search

This section presents a study that was designed to find out which sets of tags (tags collected with or without image descriptions) provides better support for image finding. The ability to support image finding is arguably the most important value of tags, so we expected the study to reveal which tag collection approach produces better tags. The study was conducted using the MTurk platform. In the study, turkers were requested to re-find an image shown at the left hand side of the task interface (Figure 7) using the tag-based search and browsing interface described in the following section. More exactly, we provided two identically looking interfaces – one driven by the W/ set of tags and one driven by the W/O set. The turkers were randomly assigned to work with one of these conditions and were not aware of the difference. All user actions were logged and a time limit (3 minutes) was set for completing each task. If the image had not been found within the time limit, the search task was considered a failure. In total, thirty images were issued as tasks within a certain period of time on MTurk.

5.1 Tag-based Image Finding Interface

The image-finding interface used in our study offered users a combination of a search box and the classic "tag cloud" (Venetis, Koutrika, & Garcia-Molina, 2011). The "tag cloud" adopted on many popular folksonomy-based sites, such as Flickr, BibSonomy or CiteULike, is known as both an expressive representation of a set of tags associated with a collection of resources and as an efficient interface browsing-based access to these resources.



Help

You can refine your query by using the tag cloud on the right

- Click tag to search for that term
- Use + to add terms to the search query
- Use ✖ to remove terms from the search query

Task: Find the following image!

1. Use the *searchbox* to search for the image.
2. Click on the "YES, found it!" button, to confirm that you have found the image!
3. If you can not find the image, click on the *"Sorry, I can not find it!"* button.

family

Query: family ✖

Results: 33

Sorry, I can not find it
YES, found it!
ac+ baby+ bench+ bilding+ book+ box+ boy+ bride+ building+ cap+ car+ cars+ celebration+ chair+ child+
children+ childrens+ childrns+ cups+ dog+ family members+ flower+ flowers+ game+ garden+ gentleman+
groom+ group+ group photo+ group portrait+ home+ house+ kids+ kiss+ man+ meeting+ men+ men and women+
outdoor+ party+ people+ photo+ poor+ room+ sitting+ sitting on bed+ smile+ solider+ stand+ standing+ street+
suit+ swimming pool+ table+ tree+ tv+ wading pool+ wedding+ winners+ woman+ women+
Show more tags
Family of the Year

Figure 7. *Image finding interface used in Mechanical Turk*

Our interface, which was originally designed for a study of tag-based image finding (Trattner et al., 2012) uses a simple popularity-based tag cloud algorithm to generate the tag cloud, i.e., the top N most frequent co-occurring tags were displayed to the user for each query. Tags are alphabetically ordered. Hence in its functionality it works similar to the one provided by social tagging platforms such as BibSonomy where co-occurring "related" tags are shown in the form of a tag cloud to the user during the search process. The interface additionally provides functionality to increase or decrease the number of displayed tags in the tag cloud, since the number of displayed tags is an important performance factor (Sinclair & Cardew-Hall, 2007; Trattner, Körner, & Helic, 2011). The interface (Figure 7) supports a combination of search and tagging that is most typical for existing social tagging systems. At each search step the user can type a query or click on a tag to search. The user can also to expand the query by clicking "+" in the tag cloud (resulting in the addition of the tags to the current query) or to shrink the query by utilizing "x" in the query string (popular approach in BibSonomy or Yahoo! Tag Explorer[4]).

5.2 The Experiment Setup

We randomly selected 24 target images to apply with both searching conditions. We requested three different users to look for each target image with each interface. On the Mechanical Turk, we set up a restriction that each user can only perform our task once since there are powerful crowd resources in MTurk and we could reduce some potential learning effect from users' multiple attempts. However, Mechanical Turk is an open crowdsourcing market where users could apply multiple accounts. For the user with multiple accounts, we couldn't control their access to our task multiple times with different accounts. For each condition, we activated 72 tasks (cases), 3 users for each of the 24 target images.

5.3 The Results

This section examines the observed differences between two study conditions in terms of success/failure rate and efforts spent. We excluded three outliers from the W/O condition, and one outlier from the W/ condition because of their invalid input confirmation. To measure the image-finding efforts, two independent variables, search time and the number of total interface actions, were examined. Following the work on interactive search, we assumed that a more efficient interface for image finding would require a shorter search time and a lower number of actions for subjects (Kelly, 2007).

***Table 6* and**

Table 7 provide a summary of performance data for two interfaces. The first table presents an overview of the data and reports the mean performance parameters. The second table

[4] http://sandbox.yahoo.com/introducing-tagexplorer

reports medians and presents the data in two ways: first for all tasks (including failed tasks) and, second, for successfully completed search tasks only. This was done because all failed search attempts were interrupted at the same time limit and their inclusion flattens the overall differences. As we can see from *Table 6*, the use of the tag set collected with annotations produces two opposite effects. On the one hand, the users who worked with the W/O dataset were much more successful in solving their tasks. On the other hand, the users worked with the W/ condition spent much less effort (steps) and time to solve the problem.

The separation of successful cases in

Table 7 allowed us to examine the significance of these differences with and without the flattening effect of failures. Since the distributions of total actions and search time didn't meet the normality assumption, we adopted a Mann-Whitney test to discover whether there are significant differences affected by interfaces with different tag data sets. A Mann-Whitney test indicated that the number of subjects' actions with the interface with tags collected without a text description (M=8) is significantly larger than with the interface with tags collected with a text description (M=5), U=1888.5, *p*=.009, r=.198. There is no significant difference for the search time in the all cases. The analysis of successful cases reveals that the number of subjects' actions on the interface with tags collected without a text description (M=6) is also significantly larger than in the interface with tags collected with a text description (M=4.5), U=862, *p*=.035, r=.185. For search time analysis, no significant differences were discovered for successful cases as well.

Table 6. *Descriptive statistics of searching performance in two conditions*

	W/O des	**W/ des**
Success rate	79.71%	56.34%
Total actions per hit	13.94	9.58
Average search time	3 mins 18 secs	2 mins 54 secs

Table 7. *Descriptive data for success rate, total actions and search time by interface. Each statistic is calculated considering all cases and only successful search tasks*

	W/O des		**W/ des**	
Measure	All	Successful	All	Successful
Cases	69	55	71	40
Actions	8	6	5*	4.5*
Time	1 min 54 secs	1 min 20 secs	1min 45 secs	1 min 24 secs

(*=significant at $p<.05$)

To explore which kinds of actions contributed to the significant increase of total number of actions in the W/O condition, we examined the frequencies of main search and navigation actions: (1) Search (inserting a query in the search box); (2) Click Tag (issuing a query by clicking on a tag); (3) Add Tag (expanding the query with a tag by clicking the "+" sign); (4) Remove Term (removing a term from the query by clicking the "x" sign); (5) Show More Tags (clicking the show more tags button to increase the number of tags in the tag cloud); (6) Show Fewer Tags (clicking the show fewer tags button to reduce the number of tags in the tag cloud); (7) Show More Results (clicking the show more results button to increase the number of images in the result list); and (8) Click Image (clicking on an specific image). As *Table 8* shows, the significantly increased number of search actions as well as the significantly more frequent usage of the "Show More Results" link accounted for the main increase of the absolute number of actions. In terms of relative numbers, most noticeable is the several-fold increase of tag cloud manipulations (Show More/Fewer Tags).

In summary, the reported data indicated that the subjects in the W/O condition had to work harder to find the target image as evidenced by the significant increase of the number of actions. At the same time, the subjects in W/ condition experienced more failures, i.e., there was a larger fraction of images that they were not able to find within the specified time limit.

Table 8. *Median (Mean) frequencies of specific actions in two tag-based interfaces*

	W/O des		**W/ des**	
	All	Successful	All	Successful
Cases	69	55	71	40
Search	1.00(3.54)	1.00(2.22)	.00(2.06)*	.50(1.8)
Click Tag	1.00(2.17)	1.00(1.42)	1.00(1.92)	1.00(.98)
Add Tag	.00(.87)	.00(.67)	.00(.59)	.00(.28)
Remove Term	.00(.90)	.00(.69)	.00(1.10)	.00(.75)
Show More Tags	.00(.58)	.00(.45)	.00(.06)*	.00(.00)
Show Fewer Tags	.00(.12)	.00(.45)	.00(.01)*	.00(.00)
Show More Results	2.00(4.75)	1.00(3.60)	1.00(3.11)*	1.00(2.18)*
Click Image	1.00(1.01)	1.00(1.15)	1.00(.73)*	1.00(1.13)

(*=significant at *p*<.05)

6 Discussion

This study explored the impact of image descriptions in the outcome of crowdsourcing-based tagging process. We attempted to answer two questions: (1) does the provision of descriptions influences users in their tagging behavior and (2) how does it impact the nature of the tags and the fundability of the images in the tagging system. Using the Mechanical Turk crowdsourcing platform, we collected two tag data sets using two tagging interfaces: one where taggers can see only images, without any image description text (W/O), and another where users can see images *and* image descriptions (W/). We collected 16,657 tags from the W/O text description interface, 4,206 of them distinct, and 17,541 tags from the W/ text description interface, 6,418 of them distinct, for 1,986 images of the Teenie Harris Archive. We demonstrated that the presence of descriptions effected the tagging process as evidenced by a considerable reuse of words provided in the descriptions for tagging and correlation between description properties and the volume of produced tags. We also found that most images are assigned dissimilar tags in two situations. Following the insights from the analysis of the most frequent tags in both conditions, we compared the obtained tag sets in several aspects, such as diversity, specificity, tag re-use and resource discrimination.

The diversity analysis revealed two opposite trends predicted by the existing theories. On the one hand, the presence of descriptions led to the larger number of tag assignments and a much larger number of unique tags produced in the W/ condition, which also resulted in a significantly larger number of tag assignments and unique tags per image. On the other hand, descriptions provided some unified impact on tagging increasing inter-tagger agreement. Our data demonstrated that the diversity trend was stronger: the image-level inter-tag dissimilarity measured by the Levensthein distance was significantly larger for the W/ condition making tags produced in the W/ condition *more diverse*.

The specificity analysis provided evidence that tags produced in the W/ condition were also *more specific* than those produced in the W/O condition. Other evidence was obtained by examining tag length: tags from the W/ condition contain larger numbers of words and characters. Comparing two tag sets with the Oxford Dictionary produced additional evidence. We found that a larger fraction of tags in the W/ condition were more specific terms, which cannot be found in the dictionary.

These conceptual differences between the two set of tags are echoed by differences in more practical aspects such as tag re-use and resource discrimination. As we observed, a larger tag diversity in the W/ condition resulted in a significantly smaller tag re-used rate than the in the

W/O condition. At the same time, larger diversity and higher specificity of tags in the W/ condition resulted in lower inter-resource tag similarity, making resources in the W/ condition easier to discriminate by their tag index. The analysis of previous research indicated that the differences in tag re-use and resource discrimination might provide an opposite impact on the practical value of tags as a mechanism to support image finding. On the one hand, a larger fraction of generic terms and a larger number of images indexed with the same term give a higher chance for a user in the W/O condition not to miss an image. On the other hand, a more diverse and distinct nature of indexing helps users in W/ condition to find a specific image among many others.

To provide a more reliable answer about the practical value of tags in two conditions, we conducted a user study that compared users' image-finding performance using the same tag-based interface that was driven by two different tag sets. The empirical results also discovered a dual impact of image descriptions on the practical value of tags. On the one hand, we observed a higher success rate in the W/O condition. On the other hand, the data demonstrated that the users in the W/O condition performed significantly more actions in the process of their search. We believe that these results could be explained by differences between the conditions that were uncovered on the analytical stage. A higher frequency of more generic words with higher tag re-use likely gave users in the W/O condition a better chance to use the tag associated with the target image by taggers and, as a result, succeed in the search task. At the same time, they had to work harder to get to the target image by digging deeper to search results and the tag cloud and by issuing more queries. In contrast, due to the use of more specific and discriminative tags in the W/ condition, it was harder for the users to "guess" the tag that was used by the indexers, however, when there were able to do it, they got to the target image faster.

The combination of analytical and empirical data provides a rather unexpected result: neither of the conditions can be considered as superior to the other in a practical sense. More precisely, each way of indexing has its own strengths. When indexers produce image tags without text description they tend to use more generic terms that increases tag density and makes images more findable. The presence of text descriptions elicits more diverse and specific tags that make images easier to discriminate in the process of finding and can shorten user path to the images. The data suggests that a mix of tags produced with and without description could be more helpful for image search than either of these conditions alone. The generic nature of tags produced without description makes them most useful during the first steps of image search to reduce the original tag cloud and select a reasonably-sized subset of candidate images. In turn, the more specific discriminative nature of tags produced with description could be most helpful on the following steps when the appearance of more unique and more specific tags in the reduced tag cloud can guide users to the target images.

On the practical side, our data provide guidance to the managers of image collections and designers of image tagging systems. When soliciting image tags from crowdworkers, we recommend the managers ensure that part of the tags for each image are provided by indexers who can observe image text description and another part by crowdworkers who can observe only the image itself. To increase image findability at the cost of losing efficiency, the managers can increase the fraction of tags solicited without description. To shorten user paths to images, they can increase the fraction of tags solicited with description, although it might make effect ultimate image findability. Our recommendation to the designers of image tagging systems is to turn the presence of an image description into an option that can be manipulated by the system managers.

There are some limitations to this study. First, in the course of one study we were able to explore only a limited set of approaches to compare tags sets. We believe that there could be other interesting methods for analyzing the characteristics of tags from different conditions. Second, our user study was limited by the nature of the Mturk platform, which could only support our study in a between-subjects setting. A well-designed within-subjects user study might bring more insights about the practical difference between the produced tag sets.

7 References

Ames, M., & Naaman, M. (2007). Why We Tag : Motivations for Annotation in Mobile and Online Media. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 971–980). San Jose, CA, USA: ACM.

Anderka, M., Stein, B., & Lipka, N. (2012). Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of the SIGIR conference on Research and development in information retrieval* (pp. 981–990). New York, NY, USA: ACM.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can All Tags be Used for Search ? In *Proccedings of the CIKM conference of Information and Knowledge Management* (pp. 203–212). Napa Valley, California, USA: ACM.

Bollen, D., & Halpin, H. (2009). An Experimental Analysis of Suggestions in Collaborative Tagging. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (pp. 108–115). New York, NY, USA: ACM.

Cattuto, C., Loreto, V., & Pietronero, L. (2007). Semiotic Dynamics and Collaborative Tagging. *Proceedings of the National Academy of Sciences*, *104*(5), 1461–1464.

Chi, E. H., & Mytkowicz, T. (2008). Understanding the efficiency of social tagging systems using information theory. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (pp. 81-88). Pittsburgh, Pennsylvania, USA: ACM.

Dellschaft, K., & Staab, S. (2008). An Epistemic Dynamic Model for Tagging Systems. In *Proceedings of the 9th ACM conference on Hypertext and hypermedia* (pp. 71–80). Pittsburgh, Pennsylvania, USA: ACM.

Dellschaft, K., & Staab, S. (2012). Measuring the Influence of Tag Recommenders on the Indexing Quality in Tagging Systems. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 73–82), Milwaukee, Wisconsin, USA: ACM.

Ding, Y., Jacob, E. K., Yan, E., George, N. L., & Guo, L. (2009). Perspectives on Social Tagging. *Journal of the American Society for Information Science and Technology*, *60*(12), 2388–2401.

Floeck, F., Putzke, J., Steinfels, S., Fischbach, K., & Schoder, D. (2011). Imitation and Quality of Tags in Social Bookmarking Systems – Collective Intelligence Leading to Folksonomies. In *On Collective Intelligence: Advances in Intelligent and Soft Computing Volume 76* (pp. 75–91). Berlin, Germany: Springer.

Fu, W.-T., Kannampallil, T., Kang, R., & He, J. (2010). Semantic Imitation in Social Tagging. *ACM Transactions on Computer-Human Interaction*, *17*(3), 1–37.

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

White, H. D., & Griffith, B. C. (1987). Quality of indexing in online data bases. Information processing & management, 23(3), 211-224.

Gupta, M., Li, R., Yin, Z., & Han, J. (2011). An overview of social tagging and applications. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 447–497). Boston, MA, USA: Springer

Halpin, H., Robu, V., & Shepherd, H. (2007). The Complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th international conference on World Wide Web* (pp. 211–220). New York, NY, USA: ACM.

Helic, D., Trattner, C., Strohmaier, M., & Andrews, K. (2010). On the Navigability of Social Tagging Systems. In *Proceedings of 2010 IEEE International Conference on Social Computing* (pp. 161–168). Los Alamitos, CA, USA: IEEE Computer Society.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can Social Bookmarking Improve Web Search? In *Proceedings of the international conference on Web search and web data mining - WSDM ’08* (pp. 195–206). New York, NY, USA: ACM.

Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business* (p. 311). Random House LLC.

Kelly, D. (2007). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval*, *3*(1-2), 1–224.

Kowatsch, T., & Maass, W. (2008). The Impact of PreDefined Terms on the Vocabulary of Collaborative Indexing Systems. In *European Conference on Information Systems* (pp. 2136–2147).

Lee, D. H., & Schleyer, T. (2012). Social Tagging Is No Substitute For Controlled Indexing : A Comparison Of Medical Subject Headings And Citeulike Tags. *Journal of the American Society for Information Science and Technology*, *63*(9), 1747–1757.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet physics doklady*, *10*, 707.

Lipczak, M., & Milios, E. (2011). Efficient Tag Recommendation for Real-Life Data. *ACM Transactions on Intelligent Systems and Technology*, *3*(1), 1–21.

Lorince, J., & Todd, P. M. (2013). Can Simple Social Copying Heuristics Explain Tag Popularity in a Collaborative Tagging System? In *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13* (pp. 215–224). New York, NY, USA: ACM.

Moltedo, C., Informática, D. De, & Mendoza, M. (2012). Tagging Tagged Images : On the Impact of Existing Annotations on Image Tagging. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia* (pp. 3–8).

Nov, O., & Ye, C. (2010). Why Do People Tag?: Motivations for Photo Tagging. *Commun. ACM*, *53*(7), 128–131.

Nowak, S., & Rüger, S. (2010). How Reliable are Annotations via Crowdsourcing : A Study about Inter-annotator Agreement for Multi-label Image Annotation. In *Proceedings of the international conference on Multimedia information retrieval* (pp. 557–566). New York, NY, USA: ACM.

Oxford University Press, I. (2010). *New Oxford American Dictionary 3rd edition*.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *the Conference on Empirical Methods in Natural Language Processing* (pp. 186–195). Stroudsburg, PA, USA: ACM.

Seitlinger, P., & Ley, T. (2012). Implicit Imitation in Social Tagging : Familiarity and Semantic Reconstruction. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1631–1640). New York, NY, USA: ACM.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., & Riedl, J. (2006). tagging, communities, vocabulary, evolution. In *Proceedings of the 20th anniversary conference on Computer supported cooperative work CSCW 06* (Vol. 4, pp. 181–190). New York, NY, USA: ACM.

Sinclair, J., & Cardew-Hall, M. (2007). The Folksonomy Tag Cloud: When Is It Useful? *Journal of Information Science*, *34*(1), 15–29.

Smith, G. (2007). *Tagging: People-powered Metadata for the Social Web, Safari*. New Riders.

Sorokin, A., & Forsyth, D. (2008). Utility Data Annotation with Amazon Mechanical Turk. In *Porceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–8). IEEE.

Strohmaier, M., Körner, C., & Kern, R. (2010). Why Do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM2010)*. Washington, DC, USA: AAAI.

Suchanek, F. M., Vojnovi, M., & Gunawardena, D. (2008). Social Tags : Meaning and Suggestions. In *the 16th ACM International Conference on Information and Knolwedge Management* (pp. 223–232). Napa Valley, California, USA: ACM.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business. Economies, Societies and Nations* (p. 296).

Trattner, C., Körner, C., & Helic, D. (2011). Enhancing the Navigability of Social Tagging Systems with Tag Taxonomies. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11* (p. 1). Graz, Austria: ACM.

Trattner, C., Lin, Y., Parra, D., & Brusilovsky, P. (2012). Evaluating Tag-Based Information Access in Image Collections. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 113–122). Milwaukee, Wisconsin, USA: ACM.

Venetis, P., Koutrika, G., & Garcia-Molina, H. (2011). On the Selection of Tags for Tag Clouds. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11* (pp. 835–844). Hong Kong, China: ACM.

Wu, D., He, D., Qiu, J., Lin, R., & Liu, Y. (2012). Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *Journal of Information Science*, *39*(2), 169–187.