No Bull, No Spin: A comparison of tags with other forms of user metadata

Catherine C. Marshall
Microsoft Research, Silicon Valley
1065 La Avenida
Mountain View, CA 94043
+1 650 693 1308
cathymar@microsoft.com

ABSTRACT

User-contributed tags have shown promise as a means of indexing multimedia collections by harnessing the combined efforts and enthusiasm of online communities. But tags are only one way of describing multimedia items. In this study, I compare the characteristics of public tags with other forms of descriptive metadata—titles and narrative captions—that users have assigned to a collection of very similar images gathered from the photo-sharing service Flickr. The study shows that tags converge on different descriptions than the other forms of metadata do, and that narrative metadata may be more effective than tags for capturing certain aspects of images that may influence their subsequent retrieval and use. The study also examines how photographers use peoples' names to personalize the different types of metadata and how they tell stories across short sequences of images. The study results are then brought to bear on design recommendations for user tagging tools and automated tagging algorithms and on using photo sharing sites as de facto art and architecture resources.

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *system issues*, *user issues*

General Terms

Documentation, Design, Human Factors.

Keywords

Metadata, tags, study, collaborative information management.

1. INTRODUCTION

User-contributed tags have been embraced as a means of describing, organizing, and otherwise enhancing the value of collections [14,19], especially multimedia collections with little or no textual content [3]. Tags play a central role in making large, loosely-organized collections searchable, especially those that are proving difficult to index using conventional automated analysis methods. While local context is a great help in indexing multimedia material embedded in web pages—the surrounding text and inbound link anchors give substantial clues about video or image content—there are also collections that do not have this sort of contextual description available, and hence must rely on user-contributed metadata to make them tractable.

Image resources—including stock photography, picture collections, and personal photo albums—are notoriously difficult to index [10]. A number of different standards have been proposed and implemented to help trained catalogers manage large image collections (e.g [16]), but even with these standards in place, there is still the larger problem of marshaling the labor necessary to perform the actual cataloging. Increasingly this problem has been approached by harnessing user enthusiasm, with the idea that even naïve labeling efforts are better than no cataloging at all, and may in fact improve on rigid classification schemes that must anticipate content in advance [19].

Most efforts to assign metadata to public picture collections have focused on tagging as a primary method of classifying images or describing their content [12]. People may tag their own images, as they do in Flickr, Smugmug, or Photobucket, or they may tag images as part of a growing wave of purposeful games, such as von Ahn and Dabbish's ESP Game [18]. In the first case, self-tagging, participants are motivated not only by an impulse to communicate, but also by a desire to make their own contributions easy to find and prominent [2]; in the second case, game-based tagging, participants are motivated by competition, a clever means of making the activity interesting and coordinating the tags.

Although user-contributed tags are not as plentiful as their proponents had originally envisioned, they are beginning to play a vital role in organizing and indexing extensive information resources, and results have shown that even if there is great variability in tags, there is also significant consensus [6]. To date, most studies have settled on characterizing the tags that have been assigned to describe dissimilar objects in a large collection [15] or in randomly chosen samples (e.g. [9]). As such, they have given us an early profile of tag use and have served as the impetus for tag recommenders based on word co-occurrence or on spatial, temporal, and social proximity [12,13].

Although research to date has made a compelling case for tagging, questions still remain about the relationship between tags and other types of metadata. To take a closer look at this relationship, it is helpful to compare how a wide range of people tag and otherwise describe the same or similar images. If the images are very similar, and they have been shared beyond the photographer's immediate sphere of friends and family, both congruencies and divergences will tell us something about the

JCDL '09, June 15–19, 2009, Austin, Texas, USA.

nature of image description and the relationship between tags and other types of descriptive metadata.

Fortunately, large image databases on the web provide us with the ability to assemble the study collections necessary to answer questions of this sort. Flickr, for example, contains well over three billion images, a substantial portion of which are public. Because there are so many images, and because people snap photos in similar circumstances and places (as noted in [12]), Flickr and its ilk provide us with the capability of assembling test collections that we never could have gathered before.

To perform this qualitative study, I have gathered such an image collection. The subject of these photographs is a mosaic in Milan; the photographers were largely motivated by a folktale found in guidebooks and familiar to locals. Using the metadata many different people have assigned to these very similar images, I have pursued three related research questions:

- How do tags compare with other forms of user-supplied image metadata such as titles and captions?
- Are there content patterns that distinguish tags from other image metadata forms?
- How can we take advantage of other metadata creation practices to improve contributors' ability to tag public content beyond existing tag suggestion mechanisms?

Because much of this analysis hinges on the quality of the image collection, I first discuss the image itself and how the dataset was compiled from Flickr. I go on to profile the dataset—its characteristics and how much of each type of metadata (titles, narrative captions, and tags) is available for the individual images. This profile is followed by a detailed analysis of the metadata content for the English-language portion of the collection. I then summarize the findings and discuss their implications for tag-based functionality, building on related work in this area.

2. GATHERING THE DATASET

To answer the study's research questions, I gathered metadata from as many nearly identical photos as I could find on the popular photo-sharing site, Flickr. The subject of the photos is a mosaic of a bull embedded in the floor of a public place, the Galleria Vittorio Emanuele II in Milan, Italy.[1]

This subject has four specific attributes that make it well suited to answer the study's research questions. First, the subject is located in an accessible public place, a shopping area that attracts both tourists and local residents, and the photos themselves have been made public on Flickr; this way, sufficient photos are available to compare. Second, the relevance of any photo is easy to ascertain, given a set of basic criteria (discussed later). Third, the subject is visually recognizable; it is relatively easy to scan a large number of photos and pick out the relevant ones. Finally, the subject (the bull mosaic) has an associated story, one that appears in guidebooks and on travel websites; even without the story, the mosaic is sufficiently interesting as a photo subject that people would take pictures of it, but the story gives us a common point of reference, one that suggests possible descriptive metadata (although photo selection was based on visual characteristics, not metadata).

Needless to say, the folklore about the mosaic varies to some extent, but the stories are more alike than they are different. Fodor's online guidebook [5] offers the following description:

> Like its suburban American cousins, the Galleria Vittorio Emanuele fulfills numerous social functions… The floor mosaics are a vastly underrated source of pleasure, even if they are not to be taken too seriously… Be sure to follow tradition and spin your heels once or twice on the more-"delicate" parts of the bull beneath your feet in the northern apse; the Milanese believe it brings good luck.

By choosing a specific photo subject, a physical artifact that is in a particular place and has a limited, well-defined bit of folklore associated with it, we can make a detailed comparison of the tagging and description practices of many different people. Furthermore, the mosaic is probably incidental to the larger reason the photos' contributors are in the Galleria (most of the Galleria photos are of other sights, especially the Galleria's domed roof); thus the reason for taking and sharing a picture is likely to be similar: the mosaic is something of a curiosity, and by participating in the local ritual, the photographers will have a tale to tell, or will have good luck, or both.

A photo's relevance was assessed according to these criteria:

1) The photo must unambiguously show the proper bull mosaic. There are several related mosaics in the immediate area; enough of the mosaic must be visible in the photo so that it can be positively identified.
2) The mosaic must consume at least 20% of the photo's height and at least 75% of its width. For example, in Figure 2a, because of the shot's angle, the mosaic takes up around one third of the photo's height, but fills its width. In photos without human subjects, the mosaic should generally fill the frame.
3) The photo must be centered horizontally on the mosaic.
4) Metadata is insufficient to establish relevance; a photo that does not meet the other criteria is not considered relevant even if its metadata establishes it to be of the correct mosaic beyond the shadow of a doubt.

In principle, one way to maximize recall without relying on user metadata would be to linearly scan Flickr's three billion-plus photos for the target photo subject. Or, if all Flickr photos were geotagged, location information could be used to retrieve a smaller set of potentially relevant photos to evaluate. In ideal circumstances, image similarity algorithms could be applied to identify candidate photos.

Instead I resorted to the three complementary strategies listed below to gather instances of the target photo from Flickr's public database; through these means, I evaluated over 40,000 candidate photos over the course of three weeks. The numbers in parentheses indicate how many candidate photos were evaluated for relevance and how many new relevant photos were identified as a result. The techniques were applied in the order listed; hence using geotags yielded the fewest new photos.

[1] An earlier, lighthearted (unreviewed) version of this study based on a smaller collection has appeared in the humanities computing magazine *TEKKA* [11].

- Using queries (34119 candidates/540 new relevant);
- Browsing photos proximate to relevant photos in photostreams (~4250 candidates/62 new relevant); and
- Browsing geotagged photos (2555 candidates/1 new relevant).

Naturally, the photos with the least metadata are the hardest to find; yet I wanted to gather as many relevant photos as possible. Hence I used a combination of place names, travel guide terms, and naïve terms that describe the photo's subject without referring to story-specific elements (in the event that the photographer was either unaware of the story, or was telling it in a different way). Some queries avoided slang terms that referred to the bull testicles; others included colloquialisms. Because Flickr stems the terms, but does not correct the spelling, I included some common misspellings (heal for heel, e.g.) to ensure I did not systematically exclude poor spellers (which may account for up to 40% of the tags [7]).

Because non-native speakers sometimes attempt to use the language of the place they are visiting to describe the photos, I included some queries that used simple Italian words (e.g. *toro* and *fortuna*). Relevant photos included items with Italian, Portuguese, Spanish, Dutch, and Chinese metadata. These examples are included in the initial profile of the dataset, since some characteristics, such as number of tags, are evident regardless of language. However, items clearly in another language are omitted from the detailed analysis in Section 4.

To expand the recall of the dataset, I browsed 425 likely photosets (photosets that contained at least one photo of the target subject) to find additional photos. This technique yielded 62 new photos. I also scanned geotagged photos in the Galleria area; because this technique was applied last, I only found one new photo of the mosaic (with Italian metadata) this way.

I applied each strategy until it reached a logical conclusion or it stopped revealing any new photos. Of course, many photos are added to Flickr daily; for pragmatic reasons, it was necessary to freeze the dataset as of a particular date (December 19, 2008). While further gleaning might extend the dataset, generally it seemed that these outliers would either have very little metadata, non-English metadata, or misleading metadata and would thus be difficult for an English-speaking user to find. Hypothetically this relevance judgment task could be delegated to Amazon's Mechanical Turk using the criteria listed earlier if a complete metadata-independent dataset were desired and one's budget could support extensive human effort [1].

Table 1 summarizes the kinds and number of queries used to assemble the dataset, and the aggregate precision given the different types of terms. In all, I submitted 73 two- and three-word queries in English and 8 additional queries in naïve Italian to find photos of the bull mosaic; the naïve Italian queries turned up 76 new photos. Because Flickr does not display results beyond the first 4008, three of the broad queries were capped at 4008 results.

Table 1. A summary of the query portion of the data-gathering process. Aggregate precision was calculated before duplicates were eliminated.

Query type	# of queries	# photos examined	aggregate precision	representative query
3 word narrow	20	835	.90	Galleria Italy bull
3 word broad	9	8087	.02	Galleria Vittorio Emanuele
2 word narrow	9	1135	.64	Milan bull
2 word broad	35	23158	.05	bull ball
Italian narrow	5	447	.62	Milano toro
Italian broad	3	457	.16	Milano fortuna
Total	81	34119		

After duplicate relevant photos were eliminated, 603 photos of the mosaic remained. Using an Excel spreadsheet to build a dataset, for each picture, I recorded the photo's URL; the narrative caption; the photo's title; the tag set; the contributor's identity; and the predominant language used for the metadata. Figure 1 shows a typical collection item as it appears in Flickr. I also added a unique identifier (PID) and some derived metadata: the number of tags; the caption length; and title length.

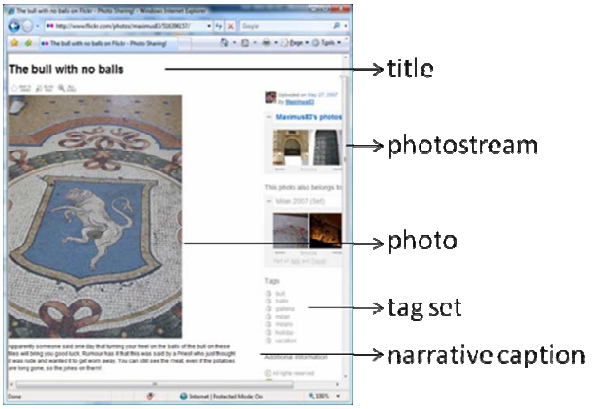

Figure 1. Collection item as it appears in Flickr.

The photos are categorized as one of three types: mosaic, person, or action. *Mosaic* means that the photo is of the mosaic itself, a nearly identical close-up of the Galleria's floor. *Person* has a human subject; because the mosaic must be clearly visible in the photo to meet the selection criteria, these are usually full-length pictures taken from close by. *Action* is a hybrid of the two; these photos are focused on the subjects' feet as they start their spin and do not show their faces; generally the mosaic occupies most of the photo. Figure 2 shows the number of photos in the study collection by type and an example of each.

Subject (#)	Example
(a) person (306)	
(b) mosaic (186)	
(c) action (111)	

Figure 2. The three types of photos in the collection

The 603 photos were taken by 427 unique photographers; 105 photographers took more than one photo of the mosaic. The most frequent sequences involve either a mosaic shot coupled with a person shot or a series of person shots. In one instance, a photographer took 13 pictures of the mosaic, each with a different person standing on it, poised to spin. These series are helpful for exploring how metadata crosses photographic boundaries (see Section 3.2). Table 2 shows how many photos of the bull mosaic each photographer took.

Table 2. Number of independent photographers

number of photos	# of instances
1	322
2	71
3	19
4	6
5	4
6	3
7	1
13	1
Total	**427**

3. DATASET PROFILE

Before we further cull the photos for specific analytic purposes, it is worthwhile to look at the characteristics of the entire dataset, especially since general studies have characterized tagging using very large heterogeneous collections [6,7,9,13,15]. By comparing the study collection's characteristics with those of the larger datasets, we can determine whether the study collection represents the whole resource.

First, let's examine metadata omission. Table 3 summarizes the number of photos lacking each descriptive element.

Table 3. Photos lacking metadata elements

type	#	# w/o tags (%)	# w/o title (%)	# w/o caption (%)
action	111	31 (28%)	2 (2%)	53 (48%)
mosaic	186	47 (25%)	4 (1%)	72 (39%)
person	306	113 (37%)	2 (2%)	117 (38%)
total	*603*	*191 (32%)*	*10 (2%)*	*242 (40%)*

It is relatively uncommon for this type of public photo to be submitted to Flickr without a title, although this can be in part attributed to Flickr's assignment of default titles. Nonetheless, only 11 photos in the dataset lack titles (which implies that the default title, the filename of the uploaded file, has actually been removed), coupled with an additional 65 photos that appear to have titles corresponding to filenames assigned by the user's camera. Even with these minimal titles factored in, the number of photos with non-descriptive titles is still only 12.6%, well below the number of photos that lack tags or captions; furthermore, the vast majority of these titles have been assigned individually, since the names meaningfully describe the photos' subjects.

Tags are next most frequently supplied metadata element; fewer than one-third of the photos—191 photos in all—lack tags. This seems to confirm that people expect tags to be helpful. More 'person' photos lack tags than the other two types and 'mosaic' photos are most apt to be tagged, perhaps because the subject is architectural and not self-explanatory.

By contrast, the narrative caption is missing from slightly over 40% of the photos. Consider that this photo subject represents a best-case scenario for eliciting a caption: folklore about the mosaic forms a natural basis for explaining the photos, especially when they don't have a recognizable human subject to motivate them. It seems odd then that action photos are less apt to have a caption than the other two categories; the absence of a narrative explanation may mean that the photographers think the photos themselves tell the story, or possibly that the story is spread over more than one photo (i.e. a companion photo's caption tells the story). We explore this possibility in Section 3.2.

Naturally the method used to select these photos has emphasized their findability in a social setting; we might expect to encounter different (and possibly less) metadata if the photos were private and described for one's own use. An individual might be able to remember that she went to Italy in August, 2005, and snapped a picture of the mosaic, but a stranger would not be able to recover that picture given only the three tags *Italy*, *August*, *2005*. Thus, although the dataset may not fully represent the set of photos with *no* metadata, there are sufficient photos with minimal metadata to

give us a picture of which metadata is offered, and which is withheld, all other things being equal.

Table 4. Lengths of captions, titles, and tags assigned to the public photos of the bull mosaic

	mean words per narrative (stdev)	mean words per title (stdev)	mean tags per photo (stdev)
action	21.9 (26.7)	3.5 (3.0)	9.6 (9.3)
mosaic	22.9 (21.7)	4.1 (3.3)	4.6 (3.5)
person	19.9 (20.1)	4.0 (2.7)	4.5 (3.5)
total	21.2 (21.7)	4.0 (2.9)	5.5 (5.5)

Table 4 further breaks down the study collection's per-photo metadata. This table omits the missing metadata accounted for by Table 3 with an eye to characterizing how much descriptive metadata people contribute when they do add metadata to their photos; there is substantial (but not unexpected) variation metadata quantity. Each kind of descriptive metadata includes a few noteworthy outliers on the high end, but tags seem particularly susceptible to variation; for example, one 26 element tag set was used to tag 13 action photos. If these outliers are removed from the action category, the remaining 396 tag sets have an average length of 4.7 tags, which makes the mean more consistent with the other types of photos.

Titles are the most common type of user-assigned metadata. The four word mean length is not surprising, given the text input box's size, and that four words may be sufficient to summarize what is in the photo and/or where it is. For example, two typical four word titles for the bull mosaic are *The Bull in Milan* (PID 484) and *Galleria Vittorio Emanuele II* (PIDs 266, 336, 361, 411, 536, 537, 569, and 602). A seven word title allows the photographer to be more expansive: *Rotate on the Bull's Balls for luck* (PID 491) or *Bull Mosaic in Galleria Vittorio Emanuele II* (PID 470). Although there are variations, more surprising are the similarities. Figure 3 lists the 22 PIDs and English-language titles that begin with the word *spinning*.

Each of these titles was contributed by a different photographer. It is unlikely any of the photographers were aware of the other photos of the bull mosaic: although some of the photos were part of Milan-related photo sets, none of them were members of the same one. This apparent regularity suggests not only a need for a more systematic analysis of word frequency in the titles, but also that we examine the other metadata types for similar patterns.

1. Spinning (PID 576)
2. Spinning (PID 234)
3. Spinning 3 times on the Bulls Balls. (PID 582)
4. Spinning after a wish (PID 589)
5. Spinning at the balls (PID 351)
6. Spinning for luck (PID 375)
7. Spinning on bull testicles in Milan (PID 324)
8. Spinning on bulls balls, Milan (PID 271)
9. Spinning on Taurus the Bull (PID 462)
10. spinning on the bull balls (PID 209)
11. Spinning on the Bull for Good Luck in Milan (PID 546)
12. spinning on the bull for luck (PID 150)
13. Spinning on the bull in Milan (PID 373)
14. Spinning on the Bull in Milan (PID 500)
15. Spinning on the bull in the Galleria Vittorio Emanuele II (PID 230)
16. spinning on the bull's balls (PID 301)
17. Spinning on the bull's balls in Milan (PID 99)
18. spinning on the bulls balls...its supposed to give me luck (PID 557)
19. spinning on the little Bull (PID 43)
20. Spinning on the Taurus (PID 109)
21. Spinning on the toretta (PID 100)
22. Spinning your heel on the bull's testicles is apparently good luck in Milan (PID 533)

Figure 3. Titles that begin with the word *spinning*.

Although the lengths of the captions vary, the content conforms roughly to the same story structure. Longer captions expand on the details. The following are examples from the study collection:

> *step and spin on the bull's testicles for good luck. really*. (11 words; PID 483)
>
> *So in Milan they say that if you spin on your heel three times on the Bull's balls, you get good luck*. (22 words; PID 581)
>
> *At the center of Galleria Vittorio Emanuele II, it is said that you will have good luck if you step on the Taurus' testicles (not a real taurus, just a mosaic) and turn twice!* (34 words; PID 322)
>
> *Planting your heel and twisting with a flourish on the 'private parts' of the mosaic picture of the bull in Galleria Vittorio Emanuele, Milano, is a tradition for Italians and tourists alike. The bull offers good luck. A nun just did it, the other can't miss doing it but is hesitant.* (51 words; PID 286)

Are tags—which are by definition classifiers and therefore intended to be at least somewhat regular—consistent as well? By all rights they should be, since regularities in titles and narrative imply that the photographers are referring to the same folklore.

But at first blush, the tags seem less regular, rather than more, possibly due to important variations in tagging strategies, a number of which seem to be at work. Some focus on the individual photo, where it was taken and what is in it; others seem to unite a whole set of photos and document, for example, other stops on a travel itinerary; still others refer to what else the photographer was doing in Milan (e.g. shopping or attending a conference). By design, tags are said to be relative to the tagger [9], so this variation is unsurprising. That the tags diverge from the other forms of description—and their regularities—bears further investigation.

Given the role of tags in personal information management [8], individual variation has an obvious purpose. While {*Arriving Home*, *Dinner With Lia*, *Milan or Bust*} (PID 120) is unhelpful if one is attempting to use Flickr as a database of stock photography, it may be very effective if the photographer is trying to find the photo he took on his vacation to Milan the day he had dinner with Lia. But are the bulk of tags this personalized and quirky, or are there other reasons for their heterogeneity?

As other studies have noted, short tag sets (3 tags or fewer) are much more common than long tag sets (6 tags or more) [13]. Recall that 191 of the photos have no tags, 273 of the photos have 1-4 tags; 106 of the photos have 5-12 tags; 17 have 13-19 tags;

and 16 outliers have very large tag sets consisting of 26 or 30 tags. Figure 4 shows the frequencies of tag set lengths; this small scale frequency distribution mirrors the larger-scale results found in Flickr-wide studies [15]; the average of 3.8 tags per photo is slightly higher than the ZoneTag results reported in [2] (2.2 tags per photo), which may reflect the fact that those photos were uploaded directly from cameraphones.

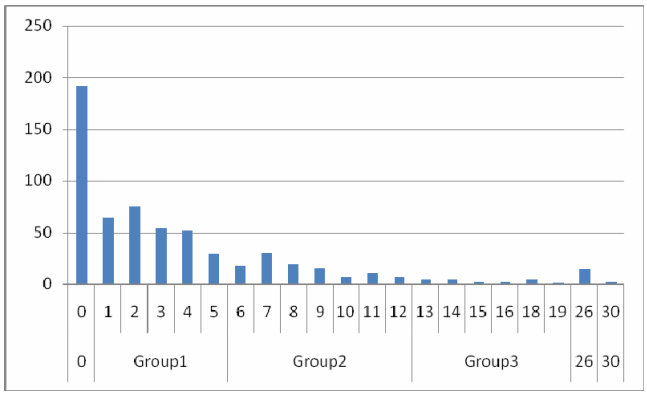

Figure 4. Frequency of n-length tag sets

3.1 Personal Names / Public Photos

Names are a useful form of personalization in many Flickr use cases: they are a vital element of social interaction and communication among people who know each other, and they may be one of the most useful terms to help contributors retrieve their own photos from Flickr, or to sift through friends' and family members' photo sets (dates also help serve this purpose); as Van House points out, image content and use are often social [17].

Names are often a way to give agency to the story about spinning on the bull—who is spinning, who the photographer is, or who the photographer's travel companions are. The interesting thing to examine is where names are used in the metadata, and whether they are used uniformly by those contributors who use them.

Most of the metadata does not include peoples' names, possibly because the photos are public. Forty-five names are included in the titles; forty-two in the narrative captions; and 100 in the tags. That the tags include more than twice as many names is not surprising, since tags are a good means of capturing contextual information. If we look at each tag set as a unit (since they may use more than one name each), 70 use peoples' names. Hence we might decide that tags are the metadata element that is the most apt to be personalized. Unsurprisingly, when names are used in more than one form of metadata, they are the same names.

Are names more commonly used when people are in the picture or when an action is suggested by the composition? It is telling that names are most commonly used in titles when people are in the picture; there are no photos of type 'mosaic' with names in the title, and there are only three of type 'action' with names in the title. The same pattern holds for the narrative captions: mostly captions describing photos with people in them (apart from three action photos) have names in them.

Tags are different. Six of the action photos use a total of 15 name-based tags; 14 of the mosaic photos use a total of 20 name-based tags; and 38 of the person photos use a total of 65 name-based tags. Thus while name-based tags are still most commonly used when there is a person in the photo, they are also clearly used for other purposes. This finding confirms the notion that tags assignment is often done as a form of personalization [2]; it also suggests that tags may serve a function that is not duplicated by the other metadata forms.

3.2 Photographic Series

As we saw earlier, some of the photos are members of short series. These series consist of 2 to 13 photos of the mosaic with and without a person standing on the bull, either poised to spin or in mid-spin. The sequence of images often suggests the story; for example, there is a shot of the mosaic, followed by one or more shots of people spinning on the bull. There are 105 such series that can serve as examples to investigate across-photo metadata.

What happens to the metadata in these photo series? Is the metadata the same across the individual photos in the series, or does the photographer create distinct per-photo metadata to tell his or her story in a way that spans the entire series? Or does the photographer implicitly designate a lead photo and assign all of the metadata to it? Do all forms of metadata—the titles, narrative captions, and tags—follow the same pattern? Are these series more likely to have metadata (because clearly the photographer was sufficiently invested in the story to post and share a series of pictures) or less (because the story is told visually)?

Table 5 summarizes how metadata is assigned across the individual series. The coding scheme captures whether the metadata is either (a) the *same* for each item in the series; (b) *different* for each item in the series; (c) assigned to a single *lead* item (i.e. the others are left blank); or (d) missing (*none*).

Table 5. Metadata patterns for photographic series

	titles	% of total	captions	% of total	tags	% of total
same	31	30%	12	11%	46	44%
different	61	58%	37	35%	24	23%
lead	1	1%	15	14%	2	2%
none	12	11%	41	39%	33	31%

Let's examine the titles first. Sixty-one out of 105 photo series have titles that suggest a story; that is, each title is different, and taken together, they describe the sequence of events in the folktale. For example, one of the two-photo series consists of a picture of the mosaic itself titled *Bull Mosaic in Galleria Vittorio Emanuele II* (PID 470) and a photo of a woman spinning on the bull titled, *Beth Spinning on Bull's Balls* (PID 469). On the other hand, in 31 of the series, the photos all share the same title (or the same title modified by a counter, e.g. *Italy! 091*); these titles appear to have either been copy-pasted or taken directly from the pictures' file names. Interestingly, a greater proportion of the photos that are in series remain unnamed—12 of the series, including the longest series (13 photos of people captured mid-spin) and several other of the larger groups of photos (a 6-photo series, a 5-photo series, and a 4-photo series) have non-descriptive names such as *IMG_9712* (PID 106); perhaps in these cases, the contributor feels that the action captured by the series is adequate to tell the story without further narration or that the photos are too similar to warrant different titles. The one remaining series (the

type *lead*) couples a titled photo, *Recharging her luck* (PID 158) with an untitled photo.

If in the majority of the cases, the titles in a photo series are different from one another (and tell the story when they are taken together), does this finding hold for the rest of the metadata? As is true in the data overall, close to 40% of the photos in series have no captions. As is the case with the titles, the captions in over a third of the series span multiple photos (coded as *different*). For example, in a three photo series consisting of a 'mosaic' and two photos of a woman standing on the bull and subsequently captured mid-spin, the captions read:

1) *This is Turin's Bull in the Galleria Vittorio Emanuele. Spinning on his privates is supposed to be good luck. The poor thing has none left...* (PID 116)

2) *Me posing on the Bull.* (PID 115) and

3) *Me, spinning on Turin's Bull in the Galleria Vittorio Emanuele. I look mildly retarded.* (PID 114)

The use of ellipses at the beginnings and ends of titles and captions appears to signal the continuation of a story from one photo to the next. But there seems to be a difference between titles and captions as well: in 14% of the photo series, a single photo has the complete caption and the others in the series have no captions at all. In other words, the story that's associated with multiple photos is consolidated as metadata for a single photo. An alternative strategy, copying the entire story into each of the captions, occurs in 11% of the cases.

The tags differ from the other forms of metadata in one important way: tag sets are more apt to be copied among all of the photos in a series; this phenomenon has been noted in previous studies as well [2]. Furthermore, when they differ, they often differ by one or two tags, unlike the narrative captions, which tell different parts of the story. Consistent with titles, tags rarely are assigned to a single photo in the series. Thus, the general tendency with tags is to treat all of the pictures in the series as standing alone, perhaps because tags have a greater anticipated role in browsing and retrieval.

4. METADATA CONTENT PATTERNS

By examining relative word frequencies in different metadata types, we can begin to see patterns that distinguish the metadata types from each other. While in principle, the same words can be used in any of the three roles (as part of a caption, in a title, or as a tag), we have already seen that in practice, each type of metadata is used in a different way.

To compare relative word frequencies in the various types of metadata, it is necessary to subset the data by language, especially since the non-English portions of the collection are incomplete owing to the method used to compile the collection. The collection includes 460 photos that are described with predominantly English metadata, and 16 mixed-language items that each include some English-language metadata.

To compare the metadata types, I have created four categories of words, *place*, *artifact*, *context*, and *story*.[2] *Place*-related words reflect the photo's geographic location. This set of words includes: {*Galleria Vittorio Emanuele II*, *Milan*, *Italy*, *Europe*}. It may also include nearby locations, for example *Duomo* or other regional designations (e.g. *Lombardy*), but not unrelated geographical entities such as *Turin* or *Switzerland*, which belong to other categories by the criteria I am setting out. (*Turin* is story-related and *Switzerland* is related to the photographer's personal context—it is probably another destination on the itinerary.)

The second category, *artifact*, includes words that describe the visual or physical characteristics of the artifact in the photo; they may be assigned without knowing the story. In fact, these are the kinds of words that are usually assigned in tagging games like von Ahn's and Dabbish's ESP game [18]. In this case, the set of artifact-related words includes {*mosaic*, *bull*, *tile*, *floor*}.

The third category of words are *story*-derived; they include {*good*, *luck*, *spin*, *around*, *three*, *times*, *tradition*}. Although there may be a slight overlap with artifact-related words, the guideline of 'visually apparent from the photo' helps sort things out: *bull* is a tag that a casual tagger may assign; *luck* is not.

The final category is *context*. These are words that are derived from the relationship between the contributor and photo: {*travel*, *2008*, *Natalie*, *Switzerland*} are examples of contextual terms. They may be useful for the contributor's own retrieval purposes, but not for public retrieval.

Table 6 shows the ten most common tags or tag-parts; words have been stemmed and multiple-word tags have been broken apart (if need be) for uniformity. Thus *luck* represents the tags *luck*, *good luck*, *lucky bull*. *Good luck* and *lucky bull* are also counted as matching the tags *good* and *bull*, respectively. Similarly, sometimes *Galleria Vittorio Emanuele* was used as a single tag and sometimes the words were spread among three tags; in either case, the words were counted individually. *200x* refers to any of the years *2003*-*2008*.

The top three tags refer to the photo's location. Indeed, the vast majority of tags are the place-related terms *Milan* and *Italy*, with the most specific place name, *Galleria*, coming in a distant third; this aligns with Sigurbjörnsson and van Zwol's findings [15]. In the English language subcollection, 26 of the 66 photos with two tags have the tag set {*Milan*, *Italy*}. One-quarter of the tagged photos have a tag that refers to their main visual attribute (in this case, *bull*, an artifact tag). Next most common are contextual tags, in this case, the word *travel* or the date.

[2] These categories are different than Sigurbjörnsson and van Zwol's WordNet-based categories [15] in that they attempt to distinguish the way the metadata is being used, rather than relying on the word alone.

Table 6. Word frequency in tag sets

tag word	count	% of all items	%items w/tags	word category
Milan	272	57%	85%	place
Italy	213	45%	66%	place
Galleria (& var)	83	17%	26%	place
bull	79	17%	25%	artifact
Emanuele	46	10%	14%	place
Vittorio	42	9%	13%	place
Europe	38	8%	12%	place
200x	36	8%	11%	context
travel	30	6%	9%	context
luck	29	6%	9%	story

Story-related tags were the least common; in addition to the 6% that used the tag *luck*, a smaller percentage used the story-related tags *foot/feet* (5%), *funny* (3%), or *spin* (3%). Like *luck, spin* is a useful tag to set apart the photos of the bull mosaic from the Galleria's other popular photographic subjects. When it is run over full text, the Flickr query *Milan spin*, for example, returns 160 photos, 101 of which are photos of the mosaic.

Does this simply mean that most of the photographers were unfamiliar with the story, and they were taking photos of an eye-catching mosaic? Table 7 shows the ten most common content words in the titles.

Table 7. Word frequency in titles

title word	count	% of all items	%items w/titles	word category
bull	166	35%	42%	artifact
Milan	107	22%	27%	place
luck	101	21%	26%	story
Galleria (& var)	88	18%	22%	place
balls	67	14%	17%	artifact
Vittorio	58	12%	15%	place
spin/spun	57	12%	15%	story
Emanuele	55	12%	14%	place
good	52	11%	13%	story
Il	35	7%	9%	place

An artifact-related word, *bull*, is by far the most prominent word in the titles; this is unsurprising, since there is always a bull (i.e. the mosaic) in the photo. But in examining the photos' titles, it is also relatively more common to encounter story-related words such as *luck*, *spin*, and *good* than it was in the tags. While *Milan* is still common, at 27% for all items with titles, it is by no means as dominant as it is in the tags (at 85% of all items with tags).

Finally, let's look at the narrative captions: are they more similar to titles—since they're freeform—or to tags? Table 8 shows the comparable list of the ten most common words in the captions. For the purpose of these counts, multiple occurrences of a word in the same caption are treated the same as a single appearance. In other words, the count of 174 uses of the word *bull* means that *bull* appeared one or more times in the caption field of 174 out of the 603 items in the collection.

Table 8. Word frequency in captions

caption word	count	% of all items	%items w/captions	word category
bull	174	37%	65%	artifact
luck	136	29%	51%	story
good	121	25%	45%	story
spin	119	25%	45%	story
Milan	82	17%	31%	place
heel/heal	80	17%	30%	story
balls	67	14%	25%	artifact
Galleria (& var)	64	13%	24%	place
around (round)	54	11%	20%	story
testicle	49	10%	18%	artifact

What is most notable is that story-related words are much more common in the captions than they are in the tags, and more prevalent even than they are in the titles. Although many of the words overlap with words used in the titles, the shift in emphasis is striking: the caption is often used to tell the story.

If we look at the high frequency words in each of the three categories (those appearing in over ten percent of each type of metadata), further evidence of this phenomenon arises. For tags these five words are (in descending order of frequency): *Milan*, *Italy*, *Galleria*, *bull*, and *Emanuele*. For titles, consensus is higher—there are nine words shared among ten percent of the titles—and the words include (again, in descending order of frequency): *bull*, *Milan*, *luck*, *Galleria*, *balls*, *Vittorio*, *spin*, *Emanuele*, and *good*. Interestingly, the words used in the narrative captions are similar to those used in the titles. Not only that, but also ten of these words were shared among more than ten percent of the items: *bull*, *luck*, *good*, *spin*, *Milan*, *heel* (or *heal*), *balls*, *Galleria*, *around* (or *round*) and *testicle*.

Which words were common among all three kinds of metadata? *Milan*, *Galleria*, and *bull*. This is not altogether surprising: taken together, the three words are an accurate description of what is in the photo and where it was taken. What is more interesting are the words used frequently in the titles and narrative, but omitted from the tags: *luck*, *balls*, *spin*, and *good*. They refer to the story rather than to what's visible in the photo or to personal context (as the tag *travel* does).

And here we come to the crux of the problem with naïve tagging strategies, such as those employed by players of the ESP game: the tags often describe what's in the photo (in very general terms), rather than referring to aspects of the photo—the story, for example—that distinguish it from other photos of bulls and other photos taken in the Galleria; naïve tags do not provide information that allows the viewer to better understand the photo. Tags are seldom unique verbs, such as *spin*, that can serve as useful descriptors.

The high-frequency words that distinguish captions from titles also bear looking into. *Around*—an adverb—is used in 20% of the items with captions (and in 11% of the items overall). In this case, it shouldn't be a stop word, but it is hard to elicit as part of a title (it occurs in only 1% of the titles) or as a tag (it is not used as a tag). *Heel* is a story detail (you are instructed to put your heel in the hole and spin), and as such often appears in captions, but rarely in titles (ten times), and almost never in tags (twice).

Thus a close look at words used in each type of metadata reveals that tags may be less effective descriptors for image retrieval, classification, and description than titles and captions are. In fact, for the case we explored in this study, if the titles and captions are enhanced by time stamps, geotagging, and gazetteer-like functions, much of the utility of tags for non-personal use is subsumed by other forms of description.

5. DISCUSSION

People interpret the invitation to contribute each form of image metadata differently: partly according to convention (because they've seen how others assign metadata to their own photos); partly according to how they perceive the metadata as being used; and partly according to clues that the user interface offers them (after all, the title is allotted a single line input box, while the caption is given a multi-line rectangle; tags are typed in one at a time, while the other forms are input at a single go). Ideally these distinctions would be useful; each form of metadata would complement the others.

But the study collection tells us that this is only partially true. First, as we might expect, not all forms of metadata accompany every photo. Furthermore, the quality of one form of metadata may not compensate for deficits in the others; just because a photo has no tags doesn't mean that it will have a good caption.

More importantly, this study shows that tags may converge on different types of description than other kinds of metadata do; tags converge on more general place-related metadata, while titles and captions may include a greater number of story-specific elements. Yet the expectations for tags are higher than those for titles and captions. Certain words classes—verbs, for example—may concisely describe some aspect of a public photo; yet they are rarely assigned as tags. Others—names, for example—may be personally or socially useful, but photographers may not think to assign them as tags if they have been used in captions or titles.

Given this high degree of descriptive variability and the number of very similar photos, how can we make these forms of metadata work together better, inform and complement each other, and describe the image to fulfill metadata's larger mission? After all, without good metadata, the power of such an extensive resource—an image collection that dwarfs all of the print picture collections that have preceded it—is reduced.[3]

Much of the research directed at solving the tagging problem (that, getting people to assign tags at all) has focused on making recommendations [12,15], or on improving the tag elicitation process [12], in some cases by normalizing existing tags [7]. Recommendations arise from many techniques, for example, from word co-occurrence in existing tag sets [15], from the social setting (i.e. the tags other people have assigned when they are taking pictures at the same event), or from the larger community (possibly from more experienced taggers) [13]. Improving the elicitation process may involve allowing photographers to tag at the time the photo is taken, or by making it easier to type in tags (either tags that have been recommended or one's own).

Yet there are reasons to proceed with caution; previous studies have found that inexplicable tag suggestions alarm users or discourage more accurate tagging; they may even disrupt tagging strategies that improve retrieval (especially recall) [2].

What then are we to make of the results of this study? There are two aspects of the results to consider when we talk about their implications: (1) the specific properties of the study collection and (2) the potentially wider-ranging applicability of the differences between the tags and the other forms of metadata.

Using Flickr as a de facto art and architecture resource. A combination of image similarity algorithms (e.g. [4]) and geotagging applied to social sharing sites like Flickr may help identify photos of the same place or artifact; that there are so many photos of the bull mosaic may help establish the image's significance. Then word co-occurrence techniques (across all metadata sources) may help identify consensus base tag sets for common art and architecture images. Thus instead of using the photos in their intended social role, this technique proposes to use them as a well-described art and architecture resource.

Instead of propagating tags among these very similar images, a query can instead return alternative forms of the image. That is, a well-tagged photo can act as a surrogate for any number of poorly tagged photos; that way, photo quality can be assessed separately from tag quality.

Improving existing tag suggestion mechanisms. Four results from this study may form the basis of methods to improve others' strong starts on tag suggestion mechanisms: (1) the apparent differences among metadata types; (2) the use of metadata to bind together short photo sequences; (3) the relative prominence and varying scope of location-based tags; and (4) the role of names in metadata personalization.

In at least some cases, people appear to be better at titling and telling stories than they are at coming up with tags. Using a combination of part-of-speech detection and other heuristics, potential tags may be extracted from titles and captions; depending on situational factors (including the user's desire to tag and the strength of the suggestions), tags can either be extracted from the narrative or recommended based on the evidence it offers. Furthermore, because we have a sense of the kinds of tags that might best describe certain types of photos, tags may be elicited via stories or story templates instead of by simply asking for tags.

Most photo sharing sites provide a means for contributors to create larger groupings of images (in Flickr, photostreams or photosets fulfill this function), but there is no way to create lightweight cohesive sequences of images based on a narrative thread.

This study confirms previous results about location-based tags [15]: they are common and useful, yet users find it difficult to specify them at a proper scope. For example, common tags for the

[3] While it is true that there are larger image collections—for example, as of last spring, Facebook had over five billion photos—in Flickr, the community is formed around the images, rather than the other way around.

mosaic include *Galleria*, *Milan*, *Lombardy*, *Italy*, and *Europe*. A gazetteer function can be used along with a sensible means for tidying up tags [7] to improve their utility.

Names (i.e. proper nouns representing peoples' names) can find their way from narrative captions and titles into tags. Again, heuristics along with part-of-speech detection would make this possible. Although peoples' names are seldom consequential for the resource's public use, they are decidedly useful for the social use of the individual photos.

Finally, the study confirms the problematic nature of actual tag suggestions. Take, for example, a base set of tags that might be appropriate for the bull mosaic based on word frequency: {*Milan*, *Galleria*, *bull*, *balls*, *luck*, *spin*}. A problem is immediately evident: many of the contributors are unwilling to use colloquialisms for bull testicles (or even to refer to bull testicles at all, preferring instead to say, *the bulls... er... parts* (PID 172) or *check out where you have to put your heel* (PID 374). Others chose more colorful and unpredictable euphemisms, including *plums* or *knackers* (PID 61) or the more obscure *stugats* (PID 539). It would be inadvisable to breach taboos simply in the name of assigning metadata. Furthermore, what is appropriate for a photo collection in its social function [2,17] might be less so for an art and architecture image collection.

While it has been duly noted that most tagging is done for one's own personal benefit rather than for social good, tags and other metadata still play an important role in public retrieval from non-textual resources like Flickr or YouTube. Good metadata should thus include elements that are personally useful, socially meaningful, and mutually intelligible.

6. REFERENCES

[1] Alonso, O., Rose, D., and Stewart, B. 2008. Crowdsourcing for Relevance Evaluation. *ACM SIGIR Forum* 42 (2), 11-15.

[2] Ames, M. and Naaman, M. 2007. Why We Tag: Motivations for annotation in mobile and online media. *Proc. CHI 2007*. ACM Press, New York, NY, 971-980.

[3] Besser, H. 1997. Image Databases: The First Decade, the Present, and the Future. in *Digital Image Access & Retrieval,* P. B. Heydorn and B. Sandore (eds.), U. Illinois Press, Urbana, IL, 11-28.

[4] Chum, O., Philbin, J., Isard, M., Zisserman, A. 2007. Scalable Near Identical Image and Shot Detection. Proc. ACM CIVR 2007. ACM Press, New York, NY, 549-556.

[5] Fodor's Travel Guide for Milan, Lombardy, and the Lakes. http://www.fodors.com/world/europe/italy/milan-lombardy-and-the-lakes/review-98689.html.

[6] Golder S. and Huberman, B. 2006. Usage Patterns of Collaborative Tagging Systems. *J. Information Science 32*, 2, 198-208.

[7] Guy, M. and Tonkin, E. 2006. Folksonomies: Tidying up Tags? *D-Lib Magazine* 12, 1.

[8] Jones, W., Phuwanartnurak, A., Gill, R., Bruce, H. 2005. Don't take my folders away! *Proc. CHI'05*, ACM Press, New York, NY, 1505-1508.

[9] Marlow, C., Naaman, M., boyd, d., Davis, M. 2006. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. *Proc HT'06*, ACM Press, New York, NY, 31-40.

[10] Marshall, C.C. 1998. Making Metadata: a study of metadata creation for a mixed physical-digital collection. 1998. *Proc. DL '98*, ACM Press, New York, NY, 162-171.

[11] Marshall, C.C. 2009. Do Tags Work? *TEKKA 4,* 1.

[12] Naaman, M. and Nair, R. 2008. ZoneTag's Collaborative Tag Suggestions. *IEEE Multimedia 15*, 3, 34-40.

[13] Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J. 2006. tagging, communities, vocabulary, evolution. *Proc. CSCW'06*, ACM Press, New York, NY, 181-190.

[14] Shirky, C. 2005. Ontology is Overrated: Categories, Links, and Tags. (retrieved 1/7/2009) http://www.shirky.com/writings/ontology_overrated.html.

[15] Sigurbjörnsson, B. and van Zwol, R. 2008. Flickr Tag Recommendation Based on Collective Knowledge. *Proc. WWW 2008*, ACM Press, New York, NY, 327-335.

[16] Simons, W. and Tansey, L.1970. *A Slide Classification System for the Organization and Automatic Indexing of Interdisciplinary Collections of Slides and Pictures*. University of California, Santa Cruz, August, 1970.

[17] Van House, N. 2007. Flickr and Public Image-Sharing: Distant Closeness and Photo Exhibition. *Proc. CHI'07*, ACM Press, New York, NY, 2717 – 2722.

[18] Von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. *Proc. CHI'04*, ACM Press, New York, 319-326.

[19] Weinberger, D. 2007. *Everything is Miscellaneous*. Times Books, New York.