An Experimental Study of Social Tagging Behavior and Image Content

Jennifer Golbeck, Jes Koepfler, and Beth Emmerling
College of Information Studies, University of Maryland, College Park, Maryland 20742.
E-mail: {jgolbeck, koepfler, bemm14}@umd.edu

Social tags have become an important tool for improving access to online resources, particularly non-text media. With the dramatic growth of user-generated content, the importance of tags is likely to grow. However, while tagging behavior is well studied, the relationship between tagging behavior and features of the media being tagged is not well understood. In this paper, we examine the relationship between tagging behavior and image type. Through a lab-based study with 51 subjects and an analysis of an online dataset of image tags, we show that there are significant differences in the number, order, and type of tags that users assign based on their past experience with an image, the type of image being tagged, and other image features. We present these results and discuss the significant implications this work has for tag-based search algorithms, tag recommendation systems, and other interface issues.

Introduction

Social tags are words or short phrases that users add as annotations to Web-content, be it documents, Web pages, images, videos, or other media. They help with personal organization and labeling, and can also be helpful for search and retrieval. Social tags have become a popular approach for facilitating search and browsing as well as for making recommendations (e.g., Sen, Vig, & Riedl, 2009). User-generated tags are sometimes created for the internal use of a system, as with Google Image Labeler/the ESP game (von Ahn, Ginosar, Kedia, Liu, & Blum, 2006a) or, more often, in a social tagging environment where the tags are shared with the public. The types of tags that users create have the potential to strongly impact the functionality of the systems that use them.

To date, the studies of social tagging systems have not looked much at how tags relate to the content of the item being tagged. The content of an image includes the people, places, and things depicted in it. This specifically excludes metadata about the image (e.g., the artist, title, date it was created, etc.) and personal tags that users may add. We choose to focus specifically on content for several reasons. First, metadata is usually much cleaner and more available. It is often reflected in the text about an image, and it is managed by object owners, curators, and catalogers as well as social taggers. Descriptions of content, on the other hand, are not as widely available in the text form or from authoritative sources, and this is a place where social taggers have an opportunity to add great value to our understanding of objects and our ability to search for objects such as images or videos based on their contents.

The features of an item may affect how users tag its content, and this will have implications on how the tags are used in systems. For example, consider social tags that describe paintings. An abstract painting that shows a small green blob on a white background may frequently be tagged “green.” A painting of a jungle, on the other hand, may be more green and have more green in it, but because there are other things to tag, it may receive the “green” tag less frequently. This would bias searches for “green” images toward more abstract paintings, not because they are greener, but because taggers tend to use tags about color more frequently on that type of image. This is just one example of many ways features of the object’s content can affect the tags being used to describe it.

This study is designed to better understand the relationship between social tags about the content of an image and the image features and type. We focus on images as our tagged objects since they are non-textual, and thus the content is not easily computer-processable, and because they are quick to process (as opposed to, say, watching a video).

Specifically, we will study how the number, order, and type of tags that users create differ when the type of image or features of an image change. While there have been studies that have investigated tag types (e.g., Bischoff, Firan, Nejdl, & Pau, 2008), and patterns of tagging (e.g., Golder & Huberman, 2005), none have delved into the content of the tagged object in a systematic way nor its relationship to the tags that users assign.
The relationship between image content and user-generated tags has implications for many applications that use tags. For example, if abstract (or non-representational) images receive more “color” tags than representational images do, then tag-based search algorithms may return biased results. The same is true for browsing interfaces that use tags, and such differences may also confound recommender systems. Furthermore, tag recommenders, which suggest tags to users (e.g., Heymann, Ramage, & Garcia-Molina, 2008b; Zhang, Zhang, & Tang, 2009), may use our results to suggest underutilized types of tags for a particular image.

We have chosen to work with images selected from online art museum collections. Although the content and purpose of these images are certainly different than one might find in the average Flickr or Facebook image, the range of content is actually quite broad. They include landscapes, portraits, scenes of people involved in activities, events, still life images, and abstract images. That allows us to perform analyses that would not be possible on an image corpus that features mostly snapshots (as is common in many social media sites).

Using two-dimensional paintings with a range of different styles and content, we performed an analysis of tags provided in the Steve.museum dataset (Trank & Project, 2006), a large collection of images and social tags, and ran an experimental study with 51 subjects.

**Related Work**

Social tagging has become increasingly important as the amount of user-generated content on the Web has expanded. Much work has gone into understanding tagging behavior (e.g., Farooq et al., 2007), its implications for search (Begelman, Keller, & Smadja, 2006; Heymann, Koutrika, & Garcia-Molina, 2008a) and navigation (Kammerer, Nairn, Pirolli, & Chi, 2009), the effect on and motivation of users (Ames & Naaman, 2007; Cherubini, Gutierrez, de Oliveira, & Oliver, 2010; Nov, Naaman, & Ye, 2008), tag generation (von Ahn et al., 2006a; von Ahn, Liu, & Blum, 2006b; Law & von Ahn, 2009), and personalization (Bateman, Muller, & Freyne, 2009).

Some of the basic questions about social tagging address whether or not the tags are useful. Traditionally, systems have relied on professional cataloging and indexing. Many studies have shown that, although professionals cannot be replaced, social tags enhance the description of images for users and provide data useful for retrieval.

Bar-Ilan, Shoham, Idan, Miller, and Shachak (2008) found that for image retrieval, structured tagging done by professionals remains the best method for successful retrieval. However, credence should be paid to social tagging (Unstructured tagging) as it allows for a wider range of tags that may prove more useful in completing successful queries.

The results of Matusiak (2006) showed that while indexing done by professionals is necessary to create the framework that allows consistency and accuracy in tagging, social tagging can enhance the professional indexes as some of this terminology feels more accessible to the average user.

Mathes (2004) also found that in creating an image retrieval categorization, metadata created by professionals in isolation is not as helpful in creating a wide and more user-centered vocabulary resulting from a communal tagging experience. The usefulness of tagging to information retrieval has also been widely studied and the results have shown that tags have the potential to improve retrieval. Heckner, Mühlbacher, and Wolff (2008) researched tagging in a scientific bibliography management system to find a way to index the vocabulary with a variety of names for easier retrieval.

Social tagging enhanced the retrieval process by including users’ tags in addition to authors and experts.

Jørgensen (2007) found that there remains a semantic gap in the tagging of professionals and the search queries of users. Some of these gaps are based on the dichotomy between programmers and the users. While some objects have single meanings, other words, such as “peach,” exist in a variety of contexts. This makes it more difficult to create a clear index. The authors argue that this needs to be taken into account as work continues on the tagging/query issue and the paradigm in which the index is created needs to be considered for its appropriateness for reaching users. The results of Rorissa and Hastings (2004) also showed that the best way to meet user query on indexing categories and groupings of images is with interpretive descriptions rather than the perceptual ones. Social tagging is one way to obtain these descriptions and improve search. In a later work, Rorissa (2008) found that when users label individual images, they tend to use more basic tags whereas when users tag groups of images, they use more abstract terms. This has implications on search and retrieval. Our analysis is similar in spirit, as we look at the types of tags users assign under different conditions and consider the implications this has on search, indexing, and retrieval.

Chung and Yoon (2009) found that with user-oriented search interfaces, specific categories could prove useful in helping users in a successful search. This is based on research utilizing the Shatford Matrix (Shatford, 1986), which revealed that a generic category covered the majority of tags and queries, followed by specific and abstract. Colors were the category least identified.

Our work addresses the relationship between tagging behavior and image content. Understanding what and how people tag is an important question that has been addressed before. Often, even when the functionality is available, people simply do not tag (Miller & Edwards, 2007). When they do, they do it for self-promotion (Cherubini et al., 2010), describing images to their social circles (Nov et al., 2008), and/or personal organization (La Cascia, Sethi, & Sciaroff, 1998).

Different types of tagging behavior were defined in Farooq et al. (2007). They are tag reuse, non-obviousness, discrimination, frequency, and patterns. These categories were designed to describe behaviors in social bookmarking
systems. We are interested in the relationship between tagging behavior and the content of an image, rather than the way tags are used in the system as a whole. However, we use the frequency measure in our work as well as a tag-type measure that borrows some of the features of non-obviousness and patterns.

Tagging behavior was also addressed in Golder and Huberman (2005). They found that majority consensus tags emerge for objects. This has interesting implications for our work. Although we study smaller tag collections, the common tags—and the features of them that we address here—are likely to be these consensus tags. While these papers look at individual behavior or high-level descriptions of tagging behavior, we are interested in a finer-grained analysis of tag types and image content in this paper.

Analyzing Tags and Images

The core interest we have in this research is tagging behavior, particularly relationships between tagging behavior and the attributes of the items being tagged as well as interrelationships between types of tagging behavior. Before presenting our research questions, we introduce two background concepts that support them: the Panofsky–Shatford Matrix and Areas of Interest (AOIs).

Tag Categorization: Panofsky–Shatford

Social tags may describe an image’s content, stylistic features (color, style, etc.), metadata (title, author, holding museum, etc.), or may be personal tags (e.g., “like,” “favorite,” etc.). In this study, we are interested in the tags that describe an image’s content. To categorize tags precisely, we leverage two frameworks for describing the content of art. Panofsky (1972) presents three major categories, which were later refined by Shatford (1986). We describe them here as they apply to tags.

- **Pre-iconographic (General—G)**—This category includes general types of items, people, and events. Examples of pre-iconographic tags would include “dog,” “blizzard,” “basement,” or “spring.”
- **Iconographic (Specific—S)**—Iconographic understanding requires some background or cultural knowledge on the part of the viewer, and includes specifically named items, people, or events in the content. For example, “Jesus” would be an iconographic tag. It is a specifically named person (rather than, say, “man”) and it requires the user to understand the context to know who or what is depicted. Other iconographic tags would include “World War II,” “New England,” or “Battle of Gettysburg.”
- **Iconologic (Abstract—A)**—The less concrete descriptions of an image’s content fall under the iconologic category. These will include things such as mythical creatures (“unicorn”), symbolic representations (“youth”), and emotions (“happiness”).

Shatford (1986) added four categories perpendicular to those described above.

- **Who**—This describes people or objects in the image. Examples include “Americans,” “tree,” and “mermaid.”
- **What**—Since the “who” category describes both people and items, the “what” category does not describe items but rather events or actions. Example tags could include “death” or “Birth of Christ.”
- **Where**—Locations are described in this category. Examples include “beach,” “heaven,” and “Chicago.”
- **When**—Dates, times, and periods are covered by this category, including tags such as “daytime” or “Battle of Gettysburg.”

When these are taken together, the result is a matrix of 12 categories describing image content. This matrix has been used in many studies analyzing user interaction with images (e.g., Enser & Sandom, 2003; Rorissa, 2008; Westman & Oittinen, 2006; Yang et al., 2003).

We added two categories: one for visual elements (V) which includes colors and shapes, and the other “unknown” (U) to catch all other tags. Since the matrix contains only codes for describing people, places, things, and events in an image, these two additional categories allowed coders to classify all tags. Visual elements, which do describe the content, can be coded as “V” and anything else can go into the “U” category. That will include personal tags, metadata, tags that the coder does not understand (perhaps a foreign word or a word they do not know), etc. The full matrix is shown in Table 1.

### Areas of Interest

For each image we used, we identified AOIs. AOIs are areas of an image, Web site, or other visual stimulus that

<table>
<thead>
<tr>
<th>Areas of Interest</th>
<th>Pre-Iconographic (General)</th>
<th>Iconographic (Specific)</th>
<th>Iconologic (Abstract)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Who</strong></td>
<td>Types of people, items: Dog, tree, boy</td>
<td>Named people or items: Jesus, Americans</td>
<td>Mythical beings: Unicorn, mermaid</td>
</tr>
<tr>
<td><strong>What</strong></td>
<td>General events or state of being: Death, blizzard, game</td>
<td>Specific events: World War II, Birth of Christ</td>
<td>Emotions or abstraction: Happiness, mysterious</td>
</tr>
<tr>
<td><strong>Where</strong></td>
<td>Type of location: Beach, basement</td>
<td>Specific location: New England, Chicago</td>
<td>Place symbolized: Heaven</td>
</tr>
<tr>
<td><strong>When</strong></td>
<td>Cyclical time: Spring, daytime</td>
<td>Specific time period: Battle of Gettysburg</td>
<td>Time symbolized: Youth</td>
</tr>
<tr>
<td><strong>Visual Elements</strong></td>
<td>Red, circle, heavy brush strokes</td>
<td>(Tags that do not fit into the categories above or that could not be classified)</td>
<td></td>
</tr>
<tr>
<td><strong>Unknown</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 1. The Panofsky/Shatford matrix with our additions of Visual Elements and Unknown. These 14 options were used to categorize the type of tags we collected.
Research Questions

1. Does tag order relate to tag types?
2. Does tagging behavior change based on the type of image being tagged?
3. Does tagging behavior relate to the number of AOI in an image?
4. Does a user’s previous experience with an image affect tagging behavior?

Question 1 looks at tag order. For example, do people assign tags that describe the objects in an image (e.g., dog, vase, ocean) before assigning more abstract tags (e.g., mysterious, sad, fertility)? Alternately, there may be no apparent order based on the type of tags.

With question 2, we are interested in the relationship between the type of tags and number of tags for the types of images being tagged. Some images may naturally lead to tags that fall into different areas of the Panofsky–Shatford matrix; a religious painting of the Madonna and child has more specific tags since the people are namable for someone with background knowledge, whereas an anonymous painting of nudes has no clearly identifiable people. Similarly, we may expect to see more tags that describe the visual elements of an abstract image (e.g., colors) because there are fewer objects to tag than would be seen in a representational image. This has implications for the use of tags in search.

Question 3 involves more directly into the issue of tags and image content. AOIs in an image may be objects, features, background areas, etc. For example, Figure 1 shows the AOIs on one of the images in our study. Some images, like the one shown, have many AOIs while other images will have few. How does the number of AOIs affect tagging? We might assume that objects or features within an AOI are likely to be tagged. However, will a user tag all of them? If there are many AOIs, will they come up with more general tags that capture the idea of the image rather than tagging each of many featured objects? How tagging behavior relates to AOIs has many interesting dimensions, which we will address through this question.

Finally, through question 4, we are interested in whether users’ experiences with an image affect their tagging behavior. If users see an image for the first time, they may tag differently than if they have seen and studied the image before. For example, when a user sees an image for the first time, he/she may tag objects first and then move on to more abstract concepts. On the other hand, someone who has studied the image may already have impressions about it and they may start by tagging those rather than tagging things as they see them.

We had three main dependent variables in this study: tag count, tag type (from the Panofsky–Shatford matrix), and tag order. Three different independent variables were used: image type for research questions 1 and 2, number of AOIs in the image research question 3, and the subject’s experience with the image for research question 4.

Experiment

Participants

Fifty-one subjects participated in this study. They ranged in age from 20 to 52 years of age, with an average age of 29 (standard deviation of 8.6 years). Thirty-two participants were female and 19 were male. Subjects came from a variety of academic concentrations and backgrounds, including social sciences ($n = 19$), math/engineering/computer sciences ($n = 16$), humanities ($n = 9$), and physical science ($n = 1$). Subjects were asked to report their interest and/or experience with art to confirm that results in this study were not biased for the selected art images. Art experience levels were fairly evenly distributed among subjects who reported it: 17 had little to no interest or experience with art, 11 cited passing interest in art, such as visiting art museums, and 15 reported high interest in art with experience in creating it themselves and/or studying it as an academic discipline. Finally, participants were asked to describe their familiarity and experience with tagging. Once again experience levels were evenly distributed for subjects responding to this question: 14 reported no experience or prior knowledge, 16
reported awareness of the concept but little to no experience engaging in the activity, and 14 reported that they tagged frequently on various sites such as Facebook, Flickr, and Del.ici.ous, or for the purposes of providing metadata in a database context (e.g., archives, archaeological indexing).

**Procedures**

Subjects were recruited and brought into a lab where they completed the experiment. The experiment itself was browser-based and all subjects used Internet Explorer on a Windows machine.

All subjects were given a short training session to begin. This explained social tagging and familiarized them with the study interface. Subjects were then divided into two groups.

The first group was shown each of the six images, one at a time, in a window with a box for tagging underneath the image. An example screen is shown in Figure 2. Each image was shown for 1 min and 45 sec. This time frame was selected by considering the experience of the Steve Museum team and through pilot studies. The Steve team found that most subjects who were tagging art images usually completed their tagging in about 1 min (Trant & Project, 2006). We wanted to give at least that much time so that more prolific taggers would have time to complete their work. To decide on this time, we ran a pilot study with seven subjects and asked them to take a series of images. We found all but one user completed their tagging of each image in 1:45, and the one user who did not finish was experienced with art and took a lot of time in analyzing each image. Since art experts were not our target subject pool, we set 1:45 as the maximum time allowed for subject to tag.

Subjects were permitted, but not required, to tag that entire period. They could not advance the screen early or move on to the next image until the time expired, but they were free to stop tagging whenever they liked. Test instructions also made clear that they were not expected to tag for the full time and that they could stop whenever they wanted and wait for the screen to advance.

The second group had the same interface except that they began by simply looking at three images (image (b), (c), and (e) from Figure 3) of the study without the tagging option. The image was shown for 30 sec and then the screen automatically advanced to the next image. After studying all three, they moved on to the interface used for the first group where they tagged the same three images they had just studied plus the remaining three images in the image set. This group was used to compare the tagging behavior of users who had never seen an image before to those who had some experience with the image.

The first group had 34 members and the second had 17.

**Images and Data Sources**

We used two data sources in our study. First, we ran a lab-based experiment with users. For this experiment, we chose six images covering a range of subjects and periods. They also included representational paintings (pictures of recognizable things) and non-representational paintings (more abstract work that does not depict objects) (Gardner, De la Croix, Tansey, & Kirkpatrick, 1980; Janson, 1995). These are shown in Figure 3. They include (a) nudes, (b) surreal, (c) pastoral landscape, (d) abstract, (e) religious painting, and (f) war scene. This variety of content allows us to analyze the relationship between image type and tags. For example, we might expect to see different types of tags or frequency of tags used on an abstract painting compared with a representational painting.

The lab study allowed us to capture tag order, frequency, and other specific behaviors, as well as control when and how users saw the images. However, for some factors of interest, we required more tags and more images. In such instances, we supplemented the lab study with data from the steve.museum project (Trant & Project, 2006). The creators of this project built and ran a tagger for museum images on the Web. They collected data over several years and released that dataset for public use. That data include over 1,750 images and over 90,000 tags. As part of a separate project, trained experts classified tags for 100 of these images into the Panofsky–Shatford matrix described above. We used the tags and tag-type classifications from these images for additional analysis. Researchers also coded AOIs on this set of images following the procedure described above.

The tags that we used from the Steve museum dataset were already classified into the Panofsky–Shatford matrix by trained experts in another experiment (Golbeck & LaPlante, 2010). We also classified the tags that subjects provided in our experiment. Each tag was classified by two coders. We achieved an inter-coder reliability of 93.1% (Cohen’s Kappa) and used only tags on which there was a coding agreement.

As mentioned above, art images are certainly different from images found in many photo sharing and social media sites on the Web. However, we believe that the art images are
FIG. 3. Images used in our study. They represent a variety of types and periods: (a) Untitled (three nudes in a landscape), N.D., artist unknown. (b) Les valeurs personnelles (personal values), 1952, Ren Magritte. (c) The water mill (the Trevor landscape), 1667, Meindert Hobbema. (d) Sunny beach life, 1974, Karel Appel. (e) Virgin and child, ca. 1520, attributed to Simon Bening. (f) The defense of Champigny, 1879, Edouard Detaille. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
actually far more generalizable than if we were to have sampled images from Flickr or Facebook. This study is focused on the content of images, and we chose our paintings because, as a set, they represent a wide variety of image of content types. Other than being painted, they have content that one would expect to find in any image—actions, events, people (some that a user can specifically name and others they cannot), and locations. The tags that may be art specific (e.g., art periods, artistic techniques, artist names, etc.) are ignored in this study; all such tags would be classified into the “U” category for tags that do not describe the content of the image, and these are not included in our analysis.

While there will certainly be differences between paintings and what one might find on Flickr and Facebook, that is more in terms of distribution of content types. Indeed, paintings are more representative of different types of image content than, say, the majority of Facebook photos, which are snapshots of people.

**Results**

We found many significant results relating tagging behavior and image features, including ordering, number of tags, and tag types. Interestingly, while we found many results with respect to the tag type being General, Specific, Abstract, or a Visual Element, we found no significant result on the who/what/where/when axis of the matrix. Thus, when we discuss tag types in the Panofsky–Shatford matrix below, we will be referring only to the General, Specific, Abstract, or Visual Elements axis.

We begin by describing the overall behavior we observed from our experimental results and then move on to an analysis of how behaviors relate to each of our independent variables.

**Overall Tagging Behavior**

Subjects described the surreal image (b) with the most tags providing an average of 10.63 tags ($SD = 6.31$). Subjects described the war scene (f) with an average of 8.92 tags ($SD = 5.02$) followed by 8.47 tags each for the landscape (c) and abstract (d) images ($SD = 4.45$ and $SD = 5.01$, respectively). The religious image (e) received an average of 7.72 tags ($SD = 4.42$) and the nude image (a) received the fewest tags with an average of 7.54 ($SD = 4.42$).

In general, the proportion of types of tags was similar across the two treatment groups. The vast majority of tags fell into the General category (65.6 and 74.0%, respectively), followed next by Visual Elements (11.8 and 10.6%). Fewer tags were coded into the Abstract (6.8 and 3.9%) and Specific (4.8 and 2.6%) categories. The remaining tags were unrelated to the content of the image and fell into our “U” category (11.8 and 9.0%).

**Tag Type and Tag Order**

To look at the relationship between tag type and tag order, we assigned each tag a number indicating the order in which the user assigned it. For each image, the first tag was 1, the second tag was 2, and so on. We then grouped the tag ranks by categories in the Panofsky–Shatford matrix, and compared the average rank for each category. Note that for this analysis, we only considered images that subjects had tagged without seeing first. We excluded tags from the group of subjects who had seen and studied three images before tagging them.

Specific tags had a significantly lower rank (i.e. they appeared earlier) than tags in any other category. An ANOVA showed significant differences among the tag categories, and pairwise t-tests showed Specific tags ranking lower than each other category. All values were significant for $P < 0.01$.

We saw this effect most clearly in the religious painting (e). The average rank of the Specific tags was 4.12 compared with 6.27 for General tags, 7.19 for Abstract tags, and 9.73 for Visual Elements tags.

This image (e) also offers a possible explanation into why we see this behavior. Among all of our images, this one has the most obvious iconographic elements. Looking at the tags, 21 of the 49 first tags identified the image as containing “Madonna,” “Virgin,” or “Mary” (sometimes “with child”). This suggests that when taggers can identify a specific person in the image, they are likely to record that among their first tags rather than waiting until later.

**Tagging Behavior and Image Type**

We found several relationships between tagging behavior and image type. Table 2 shows the percentage of tags that come from each category in the Panofsky–Shatford matrix for each image. The most striking result there, shown in bold, is that 43.3% of tags for the abstract image (d) are Visual Element tags—a far higher percentage than for any of the other images. We hypothesized that this was because this

<table>
<thead>
<tr>
<th>Image</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>64.7%</td>
<td>80.5%</td>
<td>80.3%</td>
<td>29.0%</td>
<td>68.5%</td>
<td>81.8%</td>
</tr>
<tr>
<td>S</td>
<td>2.2%</td>
<td>0.4%</td>
<td>0.9%</td>
<td>1.0%</td>
<td>14.4%</td>
<td>5.5%</td>
</tr>
<tr>
<td>A</td>
<td>4.1%</td>
<td>3.6%</td>
<td>7.0%</td>
<td>6.5%</td>
<td>6.5%</td>
<td>6.6%</td>
</tr>
<tr>
<td>V</td>
<td>14.6%</td>
<td>5.1%</td>
<td>4.4%</td>
<td>43.3%</td>
<td>3.6%</td>
<td>2.3%</td>
</tr>
<tr>
<td>U</td>
<td>14.3%</td>
<td>10.4%</td>
<td>7.3%</td>
<td>20.1%</td>
<td>7.0%</td>
<td>3.9%</td>
</tr>
</tbody>
</table>
image is non-representational; it is abstract and not a depiction of any specific thing, and thus users may be more inclined to tag colors and shapes.

To verify this, we turned to the steve.museum dataset so we could verify the hypothesis with data from a real-world application. Our set included 77 representational images and 8 non-representational images. We found nearly identical results; 41.9% of all tags applied to non-representational images were Visual Elements tags, and only 26.1% were General, compared with 12.7 and 65.3%, respectively, for representational images. These differences were statistically significant for $P<0.05$, and are shown in Figure 4.

**Tagging Behavior and AOIs**

In our lab study, the surreal image (b) had the highest number of AOIs at 9. The abstract image (d) had 3, the war (f), nudes (a), and religious (e) images had 5, and pastoral image (c) had 7. We grouped the average number of tags together for images with the same number of AOIs. We found that the surreal image (b), with 9 AOIs, had significantly more tags on average (10.36) than images with 7 (8.47), 5 (8.31), or 3 AOIs (7.72). However, with such a small sample of images, we could not draw any generalizable conclusions on these images alone, so we moved on to analyze the steve.museum dataset.

We identified the number of AOIs for 78 images in the dataset. We then calculated the average number of tags assigned to each image, grouped by the number of AOIs. While the lab study hinted that we may see more tags for images with more AOIs, this was not the case. Instead, as shown in Figure 5, images with five AOIs had the highest number of tags, and images with more or fewer AOIs all had fewer tags. A statistical analysis showed that images with five AOIs had significantly more tags than images with seven and two AOIs ($P<0.05$).

**Further investigating AOIs.** Why would we see the number of tags increasing up through 5 AOIs and then start to decline? We hypothesized that it may be that users are comfortable tagging up to 4 or 5 AOIs, but around that point the image is so visually complex that it may become overwhelming and instead of tagging the objects they may assign more general thematic tags resulting in fewer tags overall. To test this, we showed several subjects, an image with seven AOIs and an image with three AOIs, asked them to tag each, and then interviewed them. Some of their comments supported our hypothesis.

When describing the first more complex image, users said:

- “There is way too much going on for me to figure out what to tag.”
- “Too much information for me to figure it out.”
- “I didn’t know where to begin, and once I began I didn’t know where to stop.”
- “This seems easier to identify from a thematic perspective than individualistic.”

For the simpler image, users were much more comfortable:

- “This picture is much easier to tag because it just makes sense.”
- “This is more in my comfort zone.”
- “The images are clear and describing this picture doesn’t make my brain hurt.”

This is a preliminary evidence, and there is space for future work to study this relationship. We believe these results justify further investigation.

**Tagging Behavior and Experience With the Image**

We chose three of our images for comparing tagging behavior when users have experience with the image or when they are tagging when seeing it for the first time: Image 4, Image 8, and Image 9. We then compared the number and types of tags...
FIG. 5. Average number of tags used on images with a given number of AOIs. The peak, five AOIs, has significantly more tags than the extreme endpoints of two and seven AOIs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

from the first group of subjects, who saw the image for the first time when they began to take it (the “No Experience” group) and the second group who looked at the images first before seeing the screen where they began to tag (the “Experienced” group).

We found significant differences between the groups. First, the Experienced group added significantly more tags in the allotted time than the No Experience group. Overall, on the three images, the Experienced group added an average of 10.6 tags versus 8.1 for the No Experience group. A student’s two-tailed t-test showed that this was significant for \( P < 0.01 \).

There were also differences in the types of tags. For each group, we looked at the percentage of each tag type that they used. For the Experienced group, 10.2% of their tags were Visual Elements tags versus only 5.7% for the No Experience group. This difference was significant for \( P < 0.05 \).

Discussion

The most significant implication of our results is for tag-based search. We found that abstract images receive significantly more tags describing visual elements than other images. Representational images may be stronger examples of a color or shape, but since users tag objects rather than colors and shapes in these images, they will not be properly represented in a search for a color. Indeed, the more abstract images are likely to be over-represented when a user searches for a Visual Element tag.

While we studied this in the context of art images, it is possible (and we believe likely) that the same effects will apply to other images and to other media. This should prompt a reconsideration of search algorithms that rely on tags. In some cases, it makes sense to return the items to which the searched tag has been applied most often. However, in other cases, if a tag is of a type that may be over- or under-represented for certain kinds of items, this should be taken into consideration. There is future work to be done to investigate the manifestation of this result in other domains and to integrate the insight into search algorithms.

Our results also may impact the design of tag recommender systems and tag collection systems. Tag recommender systems predict tags and suggest them to users. Tag collection systems, like the Google Image Labeler, prompt users to enter new tags. Our results have shown that users may be more inclined to add one type of tag over another based on the image. This will affect the base data that is used for suggesting tags in recommender systems. Tagging collection systems have mechanisms for rewarding tags that are less “common.” However, with the insights we have developed about types of tags that are more common on different types of images, the definition of common and uncommon tags may change. In both tag recommender and tag collection systems, our results about which types of tags are assigned most frequently and which type are assigned earliest may be used to influence interaction between these users and systems to affect the tag distribution in the system.

Finally, our results about users experience with an image may also have implications on the use of tag collection systems and when users are prompted to tag. We found significant differences in tagging behavior when users had seen and briefly studied an image compared with when they were tagging an image as they saw it for the first time. When users are prompted to tag varies between Web sites, and tagging on first look is common in tag collection systems. Designers should consider what types of tags they are interested in and the implications the tags will have on their eventual use in their system. Our results may affect their decision about when to prompt users for tags and if and how to use data from tag collection systems.
Conclusions

In this paper, we have presented an analysis of the relationship between tagging behavior and image content. Using a lab-based study and analysis of data from a large online tagging site, we identified several important connections.

Most significantly, and with the greatest potential impact for the existing systems, we found that abstract images receive disproportionally more tags describing visual elements than representational images do. For search systems that use tags, this finding indicates that search results may be biased toward abstract images for this type of search term; even if a representational image better typifies a visual element, the tag is more likely to be applied to the abstract image.

We also found that users are more likely to assign tags describing identifiable people, dates, or locations (Iconographic tags) earlier in their list of tags. Furthermore, up to a point, users assign more tags to images that are more visually complex. However, once the number of AOs increase beyond a certain point, the number of tags decreases.

We have begun this work in the domain of art images. Similar techniques have been used on other types of images and in other domains (e.g., image search for journalists in Westman and Oittinen, 2006). We believe that an important area of future work will be to extend our analysis to other types of media. The general types of content categorized in the Panofsky–Shatford matrix can be applied to many domains—video, text, MP3s, and mostly other types of media—and thus the experiments could be directly repeated. If our results are duplicated, this will lead to powerful general suggestions about how to cultivate and use tags. If the results are not duplicated on other media, then this will highlight important differences in how tagging behavior varies among types of media.

Acknowledgments

This work was supported by Grant No. LG-30-08-0117-08 from the Institute of Museum and Library Services. Any views, findings, conclusions, or recommendations expressed in this paper do not necessarily represent those of the Institute of Museum and Library Services. Thanks to Judith Klavans, Irene Eleta, Raul Guerra, and Rebecca LaPlante for their help in the design of this work, and the HCIL for their comments on earlier drafts.

References


Jorgensen, C. (2007, October). Image access, the semantic gap, and social tagging as a paradigm shift. Paper presented at the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, WI.


