

User-Contributed Descriptive Metadata for Libraries and Cultural Institutions

Michael A. Zarro and Robert B. Allen

*College of Information Science and Technology, Drexel University
michael.a.zarro@drexel.edu, rba@drexel.edu*

Abstract:

The Library of Congress and other cultural institutions are collecting highly informative user-contributed metadata as comments and notes expressing historical and factual information not previously identified with a resource. In this observational study we find a number of valuable annotations added to sets of images posted by the Library of Congress on the Flickr Commons. We propose a classification scheme to manage contributions and mitigate information overload issues. Implications for information retrieval and search are discussed. Additionally, the limits of a “collection” are becoming blurred as connections are being built via hyperlinks to related resources outside of the library collection, such as Wikipedia and locally relevant websites. Ideas are suggested for future projects, including interface design and institutional use of user-contributed information.

Keywords: Annotation, Descriptors, Metadata, Social Media

1 INTRODUCTION

Contributions of metadata by users to online library holdings is an emerging phenomenon where the control of metadata generation is shared between librarians and curators, the traditional gatekeepers, and patrons. Notably, the Flickr Commons (flickr.com/commons) on the photo-sharing website Flickr.com allows cultural institutions to post digitized resources for public tagging, comment, annotation, and discussion. The implications of this phenomenon are significant in that patrons will have increased information and description of a resource, an enhanced ability to find resources through search and browsing, and use of the resource as a connector to additional materials on the Web.

The public’s response to the efforts of the United States Library of Congress, the first participating institution, has been “overwhelming” [8] and the resulting set of user-contributed metadata is a valuable source of descriptive information that may be utilized for information retrieval, resource identification, and outreach. The public has also taken it upon themselves to create linked, or “extended collections” connecting an institution’s resources to other material on the web; such as Wikipedia articles, commercial web pages and personal Flickr images. The use of Web 2.0 technologies and “social-sharing” websites has allowed libraries and other institutions a way to capture the collective intelligence of their patrons.

Zarro, M., & Allen, R. (2010). User-Contributed Descriptive Metadata for Libraries and Cultural Institutions. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science* (Vol. 6273, pp. 46-54). Springer Berlin / Heidelberg. http://dx.doi.org/10.1007/978-3-642-15464-5_7

The original publication is available at: <http://www.springerlink.com/>

Comments and notes are types of “descriptive metadata” that describe, identify, and add context to a resource. User-contributed metadata enhances the identification of resources, connects resources to similar holdings in a private collection, and supports the retrieval of a resource through search and browsing. The ability to annotate a cultural institution’s resources is a significantly different experience from tagging, and one that has not yet been fully explored.

Given the lack of resources for cataloging and indexing historically available in libraries, patrons may be able to give a level of attention to rapidly growing digital collections that library staff cannot be expected to provide. With high processing speeds and bandwidth, users are equipped with systems fully capable of working with virtual image collections [10]. It is important to note that library patrons here are considered those who access library holdings wherever they exist. Thus, a visitor to the photo sharing website Flickr viewing a digitized photograph posted by a library is considered a patron of that library. Also, included in this definition are visitors to the library website where user contributions may be collected and/or displayed.

The Library of Congress in January of 2008 first posted photos to The Commons on Flickr (www.flickr.com/commons). The Commons is an area of the site where museums, libraries and other cultural institutions may post digital images for interaction with the community. The ongoing Library of Congress pilot project has three main objectives in sharing their content on the Flickr Commons; to increase awareness of the photographs in the Library’s holdings with people not likely to visit the Library’s website, gain experience with Web communities, and to understand how user-contributions may enhance the Library’s holding [8]. The Library invited users to annotate the images and was rewarded with a wealth of historical information and individual contributions.

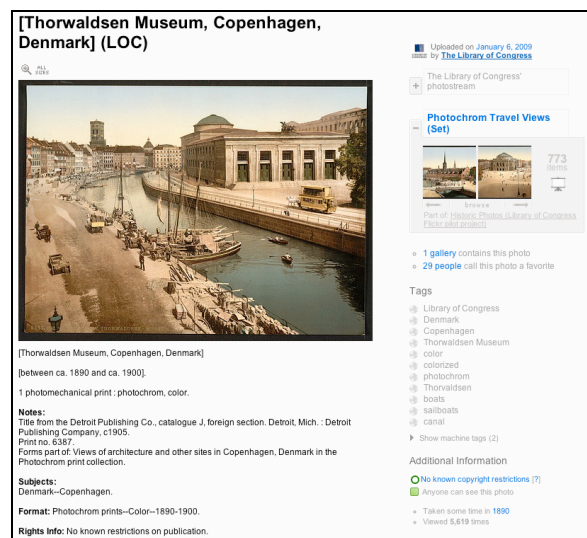


Figure 1. Typical resource from the Photochrom travel set posted by the Library of Congress on the Flickr Commons.

Much previous research in the area of user-contributed metadata has studied users tagging their own resources, rather than those held by others or institutions. Marshall [7] described the tagging and annotation contributions for several hundred images of a popular mosaic in Venice and suggested that people may be better at story telling through titles and narrative metadata than they are at assigning one word labels and tags. An analysis of user contributions and user queries in the National Archives of the Netherlands found that the most popular query type is for geographic locations, while the most popular comment type is

for named persons, places, or objects [10]. Thus, “stories” about named entities may prove to be most useful.

We aim to expand understanding of this phenomenon in two areas. First, we explore the annotation of a library’s holdings in the Flickr Commons by the public at large using three sets of digitized images that elicited significant user interaction. Second, we analyze comments and notes contributed by the user for the purposes of:

1. Expansion of historical or factual knowledge about the holding through user-contributed statements of fact or personal recollection.
2. Linking Out [2], or linking the holding to other resources such as personal Flickr photos or Wikipedia articles, through hyperlinks or textual pointers. In some cases, multiple links are contributed for a single library resource, building an “extended collection” related to a subject in the original, or “seed” image.
3. Corrections to existing descriptions and translations, as supplied by the Library or the public.
4. Linking In [2], or adding the Library’s images to groups of similar images.

2 DATA COLLECTION

We collected user-contributed metadata in the form of comments and notes for three images sets containing 1043 total images hosted by the Library of Congress on Flickr Commons. We downloaded as a tab delineated file the Flickr identifier, image title, and digest of tags for each image in three unique sets. First, we analyzed the Photocrom Travel Views set held by the Library of Congress. This set contains digitizations of color photochroms for various locations in Scandinavia, the British Isles, and Canada created between 1890-1900 [3]. There were a total of 657 images in the set at the time of data collection in October and November, 2009. Next, we looked at a digitized set of 364 black and white illustrated newspaper supplements published beginning in the 1880’s. The final set we examined was labeled “Mystery Pictures.” It is a somewhat smaller group of 22 color images, depicting content similar to the Photochrom set. In this case, the Library posted a call to action on the set homepage,

“HELP!! Please let us know if you recognize any of these images. France and the Mediterranean coast are likely locations for these travel views, called photochroms, from ca. 1900. They are the only pictures in a collection of 6,000 that arrived without titles.”¹

We manually examined each resource’s notes and comment stream to determine the types of metadata present. The categorization of comments was non-exclusive. Because these can be paragraph length pieces of text, a single comment could be judged to belong to more than category (fig. 2). Multiple comments in a category were not tabulated, only the presence of at least one comment of the type. The intent was to determine at a high level, the presence of valuable contributions across the set.

¹ http://www.flickr.com/photos/library_of_congress/sets/72157623063035332/

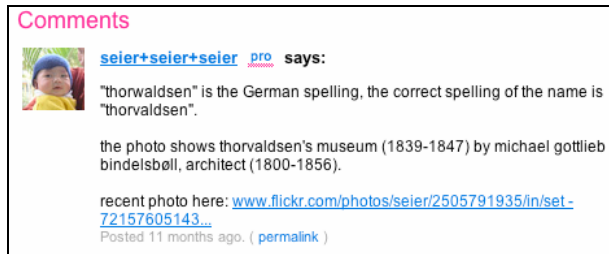


Figure 2. Comment for resource shown in Figure 1. This single comment includes a correction, historical fact, and connector to a more recent photo of the subject.

Only English language comments were used in this study. Given the international flavor of many social sharing sites and the content, with users and images from around the globe, this may prove to be a shortcoming to be addressed in future studies.

3 DATA SET

We found 37% of resources in the Photochrom set had at least one comment judged to fall within at least one of our categorizations listed above. As shown in figure 2, a comment may contain multiple categories of comments in just a few sentences. For example, a comment indicating a building was renovated in the 1950's that also contained a link to a photo of the reconstructed building counted as both a link out comment and a personal/historical comment. The counts shown in Table 1 indicate images where at least one contribution in the category was judged to appear.

The sets with geographic focus (Photochrom and Mystery) have a higher percentage of user-contribution than found for the Illustrated Newspaper Supplements, which received a contribution on 25% of its 364 images. Additionally, the set where the Library specifically asked for help – Mystery Photos - 100% of the images received a contribution. The small size of this set, 22 images, versus 657 for the Photochroms and 364 in the Newspaper Supplements, may be a factor.

Table 1. Metadata categorization and results

| Set | Metadata Category | Images | Percent |
|--|----------------------------|--------|---------|
| Photochrom Views | Personal and Historical | 165 | 25% |
| | Link Out | 105 | 16% |
| | Correction and Translation | 83 | 13% |
| | Link In | 86 | 13% |
| Illustrated Newspaper Supplements | Personal and Historical | 62 | 17% |
| | Link Out | 48 | 13% |
| | Correction and Translation | 5 | 1% |
| | Link In | 16 | 4% |
| Mystery Pictures | Personal and Historical | 21 | 95% |
| | Link Out | 16 | 73% |
| | Correction and Translation | 4 | 1% |
| | Link In | 0 | 0% |

These user-contributions are indexed by commercial search engines today and are being used to return results in web searches. For example, the phrase

“Consul F.G. Gade” appears only in the comments of an image of Fantoft Church, in Bergen, Norway². A search in early December, 2009 for the term “Consul F.G. Gade” returned results in both Google and Bing. This term appears only in a user-contributed comment, it was not a tag or other form of metadata at the time of the search. Using Bing, the Flickr resource was the first result with and without the use of quotes designating a search for the exact phrase. On Google, using quotes to search for the exact phrase, the Flickr page is the first result. It is noteworthy that the search term does return a match in the local Flickr site search.

Corrections of cataloging mistakes in the library record by a user have the potential to be a great benefit of a commenting system. For example, the comment by chistenm.hielberg, “The correct spelling is Tyssestrengene (meaning the Tysse Strings, where Tysse refers to the catchment area). These beautiful “strings” (about 300 meters free fall) are part of a large hydroelectric power scheme and are sadly no longer to be seen” on a page titled “Tåyssestrengene, Hardanger Fjord, Norway.”³ A search on the translation of the corrected term, “Tysse Strings” is the first result in both Google and Bing web searches.

In these two examples, we see user-contributions being used to enable the finding of library resources. It is unlikely that cataloging or indexing staff in a typical library in the United States would have access to this level of local knowledge for a foreign resource. While they may certainly find this knowledge through research, the access to a crowd-based solution may prove to be more effective and more efficient.

4 DISCUSSION

The “folksonomic flaw” as described in [4], suggests that tags may not be very useful for public information retrieval. Tags attached to a resource are generally for the use of the submitter and have a personal meaning. There is little control over the submission of tags, and meanings can be ambiguous or hidden to the reader. For example, the tag “spike” may have many meanings as a noun or verb. The user searching for an image likely has little or no concept of the context in which the tag was submitted, and therefore may be presented many results in a search that are in fact not relevant in context. In contrast, the comments and notes we studied appear to be meant almost entirely intended to add context to an image.

Today, there are no widespread accepted practices for identifying the nature of user-contributions. Several categorizations have been created for metadata, particularly tags, in a social-sharing environment. [6] [7] [8] [10]. To make sense of the contributions we examined, they were grouped into four categories; Personal and Historical, Link Out, Corrections and Translations, and Link In. Contributions classified as Personal and Historical are related to a known historical fact or a user’s personal knowledge of the subject of an image. An example of a historical fact might be an architect’s name for a building in the photo. An example of a personal fact is shown in Figure 3.

² http://www.flickr.com/photos/library_of_congress/3175010584/

³ http://www.flickr.com/photos/library_of_congress/3174183473/

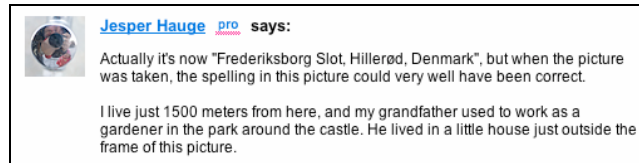


Figure 3. A personal recollection submitted as a comment.

While it may seem trivial to include personal facts, from our example we would learn that the castle mentioned had a garden of sufficient size and/or complexity as to require the employment of at least one groundskeeper. We hope the inclusion of personal remembrances give the researcher access to “hidden” facts about a resource.

The Link Out comments are connections from the current resource to additional resources on the web. A number of commenters linked to their personal photos of the scene depicted in the Library of Congress resource. Additional links went to Wikipedia pages and commercial websites. Better visualization of these links may facilitate resource discovery

Correction and Translations are a correction to a factual error or a translation of one language to another. For example, the place names in one of our sets have changed over time, with the original title now being out of date. The modern place name was supplied by a user, as demonstrated in Figure 3, allowing for search crawlers to index the image under its modern label and presumably make it more likely to be found in a web search. Users occasionally engaged in threaded discussions in the comment stream, correcting or verifying each other’s comments.

The Link In category contains comments that allow for the image to be found via browsing groups of like images. This is another, unofficial, access point to the resource. These groups may be seen as “self-curated” collections. Unlike analog resources, the digital may reside in multiple sets or collections simultaneously.

Specific user-types, such as those searching for images with certain aesthetic may be served by an expanded categorization system. An expanded classification scheme might be produced to include ideas such as: transcription of text in the image, aesthetic judgments, or technical details. Using a classification as outlined above, specific user groups may be given the ability to filter out comments that do not fit their information needs.

Currently, there is no way for users to identify the type of comment they submit to the system. It may be useful to provide users with a user interface that allows them to identify their comments as belonging to one of the categories we propose. These user-supplied classifications could then be used as filters, allowing patrons to view only those comments which are of interest in a particular search. As the comment stream grows ever longer, it might prove beneficial, for example, for the historian to view only historical comments while the artist views only comments categorized as aesthetic.

User-contributed comments and notes are a new phenomenon in the library world. We are looking at a source of rich and valuable content for library resources. Tools and processes should be developed to encourage continued interaction on a deep level between interested and willing patrons and the library.

5 FUTURE RESEARCH

A fuller categorization than the four we use in our study of user-contributed metadata is a likely next step. Comparing this new categorization to previous

tag-based efforts might prove to expose additional insight into users motivations and practices while annotating images. Differences of annotations for user's own images, versus their annotating of library resources might be useful in showing how ownership and motivation affect metadata creation.

Determining patron and librarian reaction to the use of user-contributed metadata may help guide libraries through the early years of social networking. For example, are librarians and library staff accepting of "crowd-sourced" information? Libraries and museums [9] are experimenting with user-contributions, but there is no widely accepted practice of incorporating this information into catalogs or other data stores. More research is needed to determine the best ways to use this new type of metadata.

Librarians may also be given better tools to act in a gatekeeper role. Notably, the National Archives of the Netherlands instituted a professional review of user contributions [10]. The *steve.museum* initiative has implemented a thumbs-up, thumbs-down voting protocol for moderators in a museum [9]. It may prove beneficial to expand on this sort of voting system. Or, as is found in commenting systems across the Web, patrons may find it useful to self-moderate comments, with some threshold of "thumbs-up" points needed for a particular comment to be considered trusted. However, presumably users are accessing library materials primarily to gather information thus the solicitation of too much additional effort might prove to interfere with users' information seeking behaviors, undermining the very benefits libraries hope to see.

A visualization of the structure of links and connections between resources would be a valuable next step in creating new uses for user-contributions. The user interface may greatly influence how a user both contributes a comment, and how a patron uses it. There are several interesting possibilities in this domain. A timeline view [1], or "places in time" interface may be built to visualize the same location at varying points in its history. Links to these resources could be detected programmatically and an interface built to allow comparison of two or more images. As the user moves away from the original, or seed, image in the library's holdings they may wind up multiple steps removed from the library. An interface allowing the user to remain within the realm of the library while viewing these external resources might foster additional exploration.

A goal of The Smithsonian's activity with Flickr is to "enhance the documentation and interpretation of our collections using the knowledge, perspectives, and experiences of these audiences" [5]. They also anticipate collecting and storing user contributions, possibly in a catalog record with attribution to the source of the information. While this is an exciting possibility, there are many technical and organizational hurdles to overcome before such a system may be implemented. We reach a point of information overload, where the patron may not have the ability or patience to parse multiple comments, regardless of their value. An analysis of queries and information retrieval techniques used on a set of images may prove to identify the most valuable annotation types [10], allowing a library to solicit comments targeted towards information retrieval.

6 CONCLUSION

Patrons are no longer passive consumers; they have the desire and ability to enhance library collections for the use of other patrons and professionals. In response to a request for help at the launch of the Flickr Commons, the public has shown they are willing and able to provide detailed and valuable annotations, corrections, and translations for the Library. This community-based effort may provide expertise where a library has none, and increase the number of "catalogers and indexers" identifying and annotating images. Trust and

authority of the work is an issue. However, an open exchange of comments, as the discussions have shown, may mitigate the effects of many erroneous postings.

Libraries may be able to leverage “crowd-curated” collections, which take an authoritative resource as a “seed” and build sets and groups without central control. Much like the open-source software movement, we may see open-source collection building, some of these efforts closely affiliated with institutions, others less so. Additional efforts on the development of the user-interface, institutional policies, and privacy/authority assurance are needed before the full use of user-contributed metadata may be realized.

ACKNOWLEDGEMENTS

We would like to thank the Library of Congress and Flickr teams for initiating the Commons and making their data available via an API.

REFERENCES

- [1] Allen, R. B. 1995. Timelines as Information System Interfaces. In Proceedings International Symposium on Digital Libraries (Tsukuba, Japan, Aug. 1995). <http://www.-cis.drexel.edu/faculty/ballen/PAPERS/TL/isdl.pdf>.
- [2] Anderson, S., and Allen, R.B. 2009. Envisioning the Archival Commons. *The American Archivist*, 72, 2 (Fall/Winter 2009), 383-400.
- [3] Flickr. Photochrom Travel Views (Set). http://www.flickr.com/photos/library_of_congress/sets/72157612249760312/
- [4] Guy, M., and Tonkin, E. 2006. Folksonomies. Tidying Up Tags? *D-Lib Magazine*. 12, 1. <http://www.dlib.org/dlib/january06/guy/01guy.html>.
- [5] Kalfatovic, M., Kapsalis, E., Spiess, K., Van Camp, A., and Edson, M. 2009. Smithsonian team Flickr: A library, archives, and museums collaboration in web 2.0 space. *Archival Science*. 8, 4 (Dec. 2008), 267-277. DOI=10.1007/s10502-009-9089-y
- [6] Marlow, C., Naaman, M., Boyd, D., and Davis, M. 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (Odense, Denmark, August 22 - 25, 2006). HYPertext '06. ACM, New York, NY, 31-40. DOI= <http://doi.acm.org/10.1145/1149941.1149949>
- [7] Marshall, C. C. 2009. No bull, no spin: a comparison of tags with other forms of user metadata. In Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (Austin, TX, USA, June 15 - 19, 2009). JCDL '09. ACM, New York, NY, 241-250. DOI= <http://doi.acm.org/10.1145/1555400.1555438>
- [8] Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D., and Zinkham, H. 2008. For the Common Good: The Library of Congress Flickr Pilot Project. Technical Report. Library of Congress.
- [9] Trant, J. 2009. Tagging, folksonomy and art museums: Early experiments and ongoing research. *Journal of Digital Information*, 12, 1.
- [10] van Hooland, S. 2006. Spectator Becomes Annotator: Possibilities Offered by User-Generated Metadata for Image Databases. In Proceedings of the CILIP Conference (University of East Anglia, UK).