

Reverse Image Lookup of a Small Academic Library Digital Collection

Elizabeth Kelly, Loyola University, New Orleans

Abstract

This article details a case study using Reverse Image Lookup (RIL) tools to analyze the reuse value of the digitized University Photographs collection at Loyola University New Orleans. A sample of 63 images from the digitized collection was searched using three different RIL tools. The results were then analyzed to see which RIL tool was the most effective, whether RIL is an effective and time-efficient method for determining reuse, and whether reuse is an effective measure of success for this type of collection. Existing literature on using RIL tools on digital libraries focuses on popular, well-known collections. This study is the first known published example of using RIL tools on a smaller, lesser-known collection and can help enable other institutions to determine if reuse is an effective measurement of success for their digital collections.

Keywords: Image Reuse, Digitization, Digital Library, Assessment, Impact

Introduction

Reverse Image Lookup (RIL) or Reverse Image Search technologies allow for the identification of reused images online and have historically been used primarily to locate commercial images without properly licensing. These same technologies, however, can also be useful for cultural heritage institutions trying to evaluate the impact of their digitized content online. Existing studies in the digital library literature have focused on how these technologies can be used to evaluate the reuse of large popular image collections. This paper instead seeks to analyze the effectiveness of RIL technologies on a smaller, lesser-known collection to see if reuse can be accurately determined using these tools and whether reuse is as an effective measurement of success.

Knowing how, and how often, images are being reused provides a method for establishing the value of digitized collections as well as for determining digitization priorities, outreach channels, and more. Several free tools exist for searching for copies of images online, and each has varying degrees of success. In this article, a case study of image reuse of the Loyola University New Orleans Photographs Collection will be conducted to analyze the effectiveness of performing this type of analysis on a lesser-known collection and whether or not reuse is an indicator of success.

Literature Review

The professional literature about the use of RIL with digital library collections is limited, but two previous publications have analyzed the reuse of fairly large and well-known collections. A 2010 case study investigated images from NASA to see how frequently the images were reused in academic publications (Kousha, Thelwall, & Rezaie, 2010). The researchers used TinEye (<https://www.tineye.com>), a free web image search tool, to find online reuse of 260

images from the NASA astronomy digital collection. The images had all been published online between 2000 and 2006. In the process of uploading the images to TinEye, the researchers found that results were most accurate when using high-resolution images; since they did not have high-resolution copies of all of the photos, however, they used a medium-resolution copy for each upload. After all the images had been uploaded, the researchers did a content analysis on the results to see where images were being reused. The results showed that the selected images were reused a median of 11 times online, but that only 1.4% of the results were in academic publications. Some limitations of TinEye may account for this: TinEye doesn't index PDF, Doc, or PS files, and it only searches the open web where many research publications are not available. Consequently, the results of the study reflected primarily non-academic research. The study also attempted to determine if images that had been available online for longer periods of time were more likely to be reused; while this was not found to be the case (images uploaded in 2006 had a median reuse of 11.5, while images uploaded in 2000 had a median reuse of 11), the study did find by re-testing 50 images that reuse increased over time, with a 12.4% increase in reuse after a five month period (Kousha et al., 2010).

The same article also did smaller, secondary studies of fine art and scientific image reuse online and found that the amount and type of reuse varied greatly by discipline (Kousha et al., 2010). The fine art study used 300 randomly selected paintings by ten historically important artists taken from a DVD of collected artworks and online galleries. The mean for reuse was 38.4 while the median was 8, with just under three quarters of the reuse results coming from Leonardo da Vinci (58%) and Vincent van Gogh (13%). For the scientific image reuse study, 190 and 96 images were tested from WebPath and the Protein Data Bank (PDB), respectively. The reuse for these images was much lower than the NASA and fine arts images; the median for

both databases was 0, and the means 0.44 and 0.87. The researchers felt that this was because the scientific images have a very small niche audience. However, in contrast with the NASA images, 3% of the scientific images were reused in scholarly output (Kousha et al., 2010).

A more recent study searched for reuse of images from the National Gallery in London. This study used both TinEye and Google Image Search (<https://images.google.com/>), another RIL tool, to find reuse of 32 paintings from two different collections (Kirton & Terras, 2013). The researchers used the URL search function on TinEye rather than uploading images directly to the search engine. The first collection analyzed had a mean reuse of 65 per painting and a median of 27. They then performed content analysis on the results to provide context for the reuse and performed the search again using the top three used paintings, but this time with Google Image Search. For the three paintings examined, Google Image Search returned many more results, and significantly fewer out-of-date results, than TinEye did. TinEye was found to be more transparent in its search results as Google Image Search omits similar results, but Google's results were more current (Kirton & Terras, 2013).

Finally, web statistics from the National Gallery websites were utilized to triangulate findings and establish patterns of access and reuse (Kirton & Terras, 2013). The web statistics showed that the most accessed images on the National Gallery's website from the collection examined were also the most reused online, but that the most reused images also fed users directly back to the National Gallery website. This has implications for the licensing of images, as the freer the license, the more likely that it will be reused, and the more the original institution will benefit (Kirton & Terras, 2013).

Research Questions

Kirton and Terras (2013) note that the paintings in the National Gallery are some of the most famous in the world and that RIL of lesser-known collections may not yield enough results to justify the time-consuming nature of performing RIL. Since there is not yet an example in the professional literature of RIL and smaller, lesser-known collections, a case-study of a digitized academic library collection is used here to attempt to answer the following questions:

- Which RIL tool returns the most accurate results?
- Is RIL an effective and time-efficient method for determining reuse of a University Archives image collection?
- Is reuse an effective measure of success for this type of collection?

Methodology

Selection of Images

It is essential that a study like this use truly original images that would not be found elsewhere so as not to skew the results. Images for this study were chosen from the Loyola University New Orleans Photographs collection. The collection has tens of thousands of images which are being systematically digitized and added to the university's CONTENTdm instance in the Louisiana Digital Library (<http://www.louisianadigitallibrary.com/cdm/search/collection/p16313coll28>). The photographs in the collection were taken by both official and unofficial university photographers and show campus life, buildings, student activities, athletic competitions and practices, important people, and other relevant scenes to the university's history. Because these photos were taken mostly by university employees, they are largely unique and not duplicated in collections elsewhere. They are therefore ideal for this project because any reuse found should be due to the images' origin in

the digital collection.

The first batch of images was uploaded to the digital collection in November of 2012. There are currently over 6,000 images in the digital collection. 63 were selected for inclusion in this study. First, the most viewed images from July 2013 to June 2014 were compiled using ContentDM's Webalizer Usage Reports. Of the 4,669 images then available in the collection, only 868 had been viewed during the selected time period. Of those images, 34 had been viewed ten or more times; the mean number of views for these images was 15.38, and the median was 12. These 34 images were all included in the RIL search. Ten additional images which had been viewed at least once were also selected at random for inclusion to determine if there was a correlation between number of views and likelihood of reuse. Finally, images promoted in publicly available blog posts written by Special Collections & Archives (SCA) staff at Loyola University New Orleans were investigated to see if promotion of certain images led to greater reuse. Of these images, seven had already been selected for inclusion in the study as they had ten or more views from the selected time period, but 19 new images had not been included yet and were added to the set.

Best practices Based on Previous Studies

The findings of the previously referenced studies were used to develop the methodology for this case. The highest resolution images available were used with the RIL tools to get the best results set, as Kousha et al. (2010) discovered higher resolution images resulted in greater findability; in this case, the 300 dpi jpg original scans were used. Google Image Search and TinEye were selected as tools as they index different sites and file types and therefore might have different results. TinEye has been in existence longer than Google Image Search and looks only for direct reuse of an image, as opposed to Google Image Search which also searches for

similar images. In addition, the use of Google Image Search and TinEye for the previous studies referenced allows for a more direct comparison to existing scholarship. The searches were all completed within a small time span (two days) for consistency.

Google Image Search and TinEye both allow users to either upload images directly or search for an image using the URL at which it is found. Google and TinEye also have web extensions that can be installed so that users can right click on an image on a website and automatically search for that image. After searching using a sample of ten images, it was found that uploading high-resolution images directly to the RIL tools produced more results than URL or browser extension searches did, so this method was used throughout. In exploring available RIL tools, the author came across a website called ImgOps (<http://imgops.com/>). ImgOps allows users to upload images and apply different image utilities at once, including RIL tools. Rather than separately uploading an image to Google Image Search and TinEye separately, ImgOps was used so that each image only had to be uploaded once but could then be searched using each different tool. This resulted in slightly less time (approximately 3-5 seconds per image) being spent on uploading the images.

Results

Of the 63 images uploaded, only ten had results, and these results all came from Google Image Search—TinEye did not find a single reused image among the set. All of the images which were reused were from the top 36 most viewed images except for one which was featured on a blog. None of the randomly selected images were reused. Of the 10 images which were found to have reuse there were 20 resulting URLs. The reuse results were analyzed and false positives (images that were not actually matches) and duplicates (two or more URLs that went to the same webpage) were removed; only 11 of the 20 reuse results were true instances of reuse, so the

search was only 55% accurate. While the previous two studies detailed did mention removal of duplicate or false results, neither provides an exact count so it is unknown how this accuracy compares to the previous attempts.

Because the results are minimal, it is easy to analyze the context for the images reuse. Four links were from university history webpages on the Loyola University New Orleans website. Two were from blogs on the Loyola library's website. Two were from a PDF of a presentation given by Loyola librarians at a local conference found on the conference's website, proving that Google, unlike TinEye, can find images in PDFs. One was a link to the actual CONTENTdm digital collection and not actually an example of reuse. One was from the New Orleans news website for the Times Picayune newspaper (<http://www.nola.com/>), and one was from a historical article on the Louisiana Public Broadcasting website (<http://www.lpb.org/>). If the CONTENTdm link is removed from the results, then there are ten actual instances of reuse, and of these, only two were created by users not affiliated with Loyola.

Image Title	Reuse Link	Context for reuse
UP000007	http://www.lpb.org/programs/radio/002.htm	Historical education
UP000007	http://www.loyno.edu/2012/gallery/1920-1929	Historical education (Loyola)
UP000089	http://www.loyno.edu/2012/gallery/1930-1939	Historical education (Loyola)
EW000001D	http://www.loyno.edu/2012/gallery/1930-1939	Historical education (Loyola)
EW000001D	http://www.nola.com/education/index.ssf/2012/04/loyola_university_celebrates_1.html	News
EW000016B	https://sites01.lsu.edu/wp/louis/files/2013/10/gallowaykellysullivanwille_embedlib.pdf	Illustrative--presentation
EW000047B	http://www.loyno.edu/2012/gallery/1930-1939	Historical education (Loyola)
EW000047B	http://library.loyno.edu/blog/?p=1870	Blog (Loyola)
UP001941	https://sites01.lsu.edu/wp/louis/files/2013/10/gallowaykellysullivanwille_embedlib.pdf	Illustrative--presentation
UP004358	http://library.loyno.edu/blog/?p=3647	Blog (Loyola)

Table 1: Results from Google Image Search

Discussion

Perhaps the most surprising result of the RIL search was that many known instances of reuse (or even of original display) were not found by the RIL tools. The original image uploads in the CONTENTdm collection were not returned as results except in one instance. The RIL tools also did not find the images in Loyola’s own blog posts or social media except in two instances. Five of the faulty results were duplicates, where the same page on a site was returned as a result multiple times; four of the faulty results were simply incorrect matches. Google Image Search did find results not only in HTML pages (which is all TinEye searches) but also in PDFs

and even in an animated gif. It is alarming that TinEye returned no results, especially as one of the reuse results came from a popular news site. There does not appear to be a correlation between blogging by SCA staff and image reuse.

In answer to this study's research questions: Google Image Search was found to be the most accurate tool for RIL in this instance as TinEye did not find any instances of reuse for any of the images tested. RIL was not a particularly time-efficient method for determining reuse of this image collection, although occasional application to find reuse of specific images may be beneficial for collections like this. Unlike the images searched in the previously mentioned studies which were largely found to be reused for informal scholarly communication and commercial purposes, it is assumed that images from the University Photographs collection would likely be reused on personal websites, genealogy blogs and sites, and by other projects of the Loyola community—and if the Loyola library's blog is to be taken as an example, these sites may not be indexed by the RIL tools used.

As for whether reuse is an effective measure of success for this type of collection, the question remains unanswered. The scarcity with which the images in the collection are reused is not out of proportion with how often they are viewed in the digital collection (see Methodology, "Selection of Images"). Looking at reuse can provide some context for what users want out of digitized images which can be triangulated with other assessment methods. Other assessment methods currently employed for analyzing the success of the Loyola library's digital collections include usability testing and web statistics. In addition, the library utilizes a "Librarian Transaction Form" created in Google Docs so that librarians can log interactions of all types, and SCA staff also use this form to log patron interactions whether they be face-to-face, on the phone, or through email. Using the data recorded on this form for analysis, 5% of archival

reference and known item requests to SCA staff from July 2013-June 2014 were able to be answered using the digitized University Photographs. Additionally, 55% of requests to SCA for university photos during this time period were answered using the University Photos online (sometimes in conjunction with digitized university periodicals). Measuring the success of digitized collections is a holistic process, and for this particular collection, reuse seems less a reflection of a useful collection than some other metrics reflect. Because of the small results set for this study, measuring the number of reused images is not as useful as measuring image views.

Limitations

The results of this study may be affected by a combination of indexing policies by the RIL tools and lack of Search Engine Optimization (SEO) expertise among the types of sites likely to reuse this particular collection. In addition, the very small sample set limits both the ability to represent this collection as a whole and to extrapolate the results to other collections. To determine the effectiveness of RIL on lesser-known collections, other digital collections will need to be tested and compared. The age of a digital collection may also be a factor in its reuse; the case study involving NASA images used a collection that had been online for 5-10 years at the time of research, so there had been 2-5 times as many years for users to find and reuse the images as there were for the University Photographs (though that study did not find a correlation between time available online and likelihood of reuse (Kousha et al., 2010)). To complete this study, the same images will need to be analyzed for reuse in several years to see if reuse has increased.

Conclusion

While reuse of digital images from the Loyola University New Orleans Photographs collection is limited, RIL is still a helpful exercise in determining what users are doing with

digital images. The RIL tools used in this study, Google Image Search and TinEye, are free and easy to learn, so experimentation is possible for smaller institutions. Even if reuse is limited, content analysis of even a small sample of reused images can be used to determine digitization priorities, and the addition of better outreach channels can be prescribed if reuse is not what the researcher was hoping. Performing RIL on lesser-known digital collections also contributes to the literature on image reuse which currently only focuses on popular, well-known collections. Continuing additions to the scholarly canon on image reuse will enable other institutions to determine if reuse is an effective measurement of success for their digital collections.

References

Kirton, I., & Terras, M. (2013). Where do images of art go once they go online? A reverse image lookup study to assess the dissemination of digitized cultural heritage. *MW2013: Museums and the Web 2013*. Retrieved from <http://mw2013.museumsandtheweb.com/paper/where-do-images-of-art-go-once-they-go-online-a-reverse-image-lookup-study-to-assess-the-dissemination-of-digitized-cultural-heritage/>

Kousha, K., Thelwall, M., & Rezaie, S. (2010). Can the impact of scholarly images be assessed online? An exploratory study using image identification technology. *Journal of the American Society for Information Science & Technology*, *61*(9), 1734–1744. doi:10.1002/asi.21370

Louisiana Digital Library (n.d.). Retrieved from <http://www.louisianadigitallibrary.com/>

Louisiana Public Broadcasting (n.d.). Retrieved from <http://www.lpb.org/>

The Times Picayune (n.d.). Retrieved from <http://www.nola.com/>