IJDC | *General Article*

Embedded Metadata Patterns Across Web Sharing Environments

Santi Thompson
University of Houston

Michele Reilly
University of Arkansas

Abstract

This research project tried to determine how or if embedded metadata followed the digital object as it was shared on social media platforms by using EXIFTool, a variety of social media platforms and user profiles, the embedded metadata extracted from selected New York Public Library (NYPL) and Europeana images, PDFs from open access science journals, and captured mobile phone images. The goal of the project was to clarify which embedded metadata fields, if any, migrated with the object as it was shared across social media.

Received 21 January 2018 ~ Accepted 20 February 2018

Correspondence should be addressed to Santi Thompson, 4333 University Drive, Houston, TX 77204-2000, USA. Email: sathompson3@uh.edu

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v13i1.607

Introduction

As digital objects are downloaded, copied, or shared from cultural heritage digital repositories and Open Access science journals to social media sites such as Pinterest, Facebook, Instagram, Twitter, and others, the ability to follow the provenance or determine any associated rights of a shared object is virtually impossible for cultural heritage professionals and data curators. The continual sharing over social media also presents authenticity issues, as Jessica Bushey (2013) wrote:

> 'the recent convergence of digital cameras into mobile phones, laptops and tablets with Internet connectivity to cloud based services has provided the tools and means for anyone to quickly create and store digital images; but, without the awareness or concern of professional photographers and information professionals for capturing metadata that contributes to record identity and integrity.'

The researchers of this study have conducted several previous studies, using logged usage data (Reilly and Thompson, 2014) and Reverse Image Lookup (RIL) technology (Reilly and Thompson, 2017; Thompson and Reilly, 2017), in an attempt to understand the reuse of digital images over the web. While they have found that these approaches yielded interesting results about users and their reuse, these methods have not been able to ascertain the exact provenance of reused images. While RIL finds similar images across the web, it is not developed to identify discrete instances of image reuse, particularly within sharing environments. Additionally, RIL is unable to query objects in PDF format. The researchers contend that an object's embedded metadata, which could be unique to the object, may be one potential strategy for following this sharing activity. According to Banerjee and Anderson (2013), the Exchangeable Image File Format (Exif) metadata (one type of embedded technical metadata), which includes rights management and provenance fields, follows the object as it travels through the web.

This research project tried to determine how or if embedded metadata followed the digital object as it was shared on social media platforms by using EXIFTool, a variety of social media platforms and user profiles, the embedded metadata extracted from selected New York Public Library (NYPL) and Europeana images, PDFs from open access science journals, and captured mobile phone images. The goal of the project was to clarify which embedded metadata fields, if any, migrated with the object as it was shared across social media.

Background

Human written descriptive, administrative, and technical metadata are useful tools for discoverability and access, but additional metadata is created at the point of capture by the capture device itself, i.e. camera, cell phone camera, scanner, etc. This research study focused on a variety of embedded metadata schema, specifications, profiles, and tags, including the Exchangeable Image File Format (Exif), Composite tags, the International Press Telecommunications Council (IPTC) Photo Metadata Standard, the

International Color Consortium (ICC) Profile, the JPEG File Interchange Format (JFIF), Adobe's Extensible Metadata Platform (XMP), and APP14 (an Adobe JPEG Tag).

Table 1. Descriptions of image metadata used in study.

Image Metadata	Description
Exif	"Exchangeable Information File Format (EXIF) is a standard used by camera manufacturers to store camera-created information in the file. This includes camera settings like aperture and shutter speed as well as information like the white balance selected for the image. EXIF also describes the characteristics of the image data itself so programs can know how to open the file properly" (Krough, 2018).
Composite	"The values of the composite tags are Derived From the values of other tags. These are convenience tags which are calculated after all other information is extracted. Only a few of these tags are writable directly, the others are changed by writing the corresponding Derived From tags. User-defined Composite tags, also useful for custom-formatting of tag values, may be created via the ExifTool configuration file" (Harvey, 2016).
IPTC	"IPTC Core and IPTC Extension define metadata properties with comprehensive sets of fields that allow users to add precise and reliable data about people, locations, and products shown in an image. It also supports dates, names and identifiers regarding the creation of the photo, and a flexible way to express rights information" (IPTC, 2017).
ICC Profile	"The ICC profile which describe the color attributes of a particular device or viewing requirement by defining a mapping between the source or target color space and a profile connection space (PCS)" (Wikipedia, 2016).
JFIF	"The JPEG File Interchange Format (JFIF) is an image file format standard. It is a format for exchanging JPEG encoded files compliant with the JPEG Interchange Format (JIF) standard" (Wikipedia, 2018).
XMP	"Adobe's Extensible Metadata Platform (XMP) is a file labeling technology that lets you embed metadata into files themselves during the content creation process" (Adobe, n.d.). XMP "makes a file self describing so that the file can be identified and described outside of its home system" (Christensen and Dunlop, 2011).
APP14	"The 'Adobe' APP14 segment stores image encoding information for DCT filters. This segment may be copied or deleted as a block using the Extra 'Adobe' tag, but note that it is not deleted by default when deleting all metadata because it may affect the appearance of the image" (Harvey, 2014).

Information professionals can employ the ExifTool potentially to 'reveal' and/or 'manipulate' this hidden and embedded metadata. Developed by Phil Harvey (2003), the tool is "a platform-independent Perl library plus a command-line application for reading, writing and editing meta information in a wide variety of files." As Shala and Shala (2016) wrote, EXIFTool "is mainly designed for extracting and modifying

 doi:10.2218/ijdc.v13i1.607

metadata from EXIF (Exchangeable Image File Format) file format which is specialized to store metadata of digital camera and scanners output."

Literature Review

Recent years have seen an increase in the attention paid to embedded metadata by the information profession. Foundational research has explored the advantages of embedding metadata into digital images and objects. Smith, Saunders, and Kejser (2014) discussed how embedded metadata can include technical, descriptive, and administrative elements. They wrote: "properly applied, embedded descriptive metadata can be as easily understood and used as technical metadata. Knowing who created the object(s) shown in a digital image can be as easy as knowing when that image file was created." Fuhrig (2012) and Smith, Saunders, and Kejser (2014) also noted that while technical metadata is automatically recorded by the capture device, descriptive and administrative metadata can be manually added and manipulated using software designed for this purpose.

Embedded metadata also comes with limitations, including: (a) it is not always persistent (Smith, Saunders, and Kejser, 2014), (b) it can be removed "during actions of uploading and downloading digital files into and out of social media platforms" (Bushey, 2015), and (c) "embedded descriptive metadata... can be incorrect, incomplete, or missing entirely" (Corrado and Jaffe, 2017).

Previous groups have completed studies on embedded metadata. Some are focused on developing standards for capturing and populating embedded metadata elements. A team at the Smithsonian Institution identified core minimal embedded metadata fields for their digital image production studio (Christensen and Dunlop, 2011). They wrote that "using existing standards for embedded metadata, whether in the form of descriptive, technical, structural or administrative can aid in searchability, provenance, rights management, interoperability, and data repurposing" (Christensen and Dunlop, 2011). Another project, funded by The Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and led by the American Society of Media Photographers (ASMP), designed and published "guidelines for refined production workflows, archiving methods, and best practices for digital photography based on a variety of capture methods and intended image use" (Krough, 2015). These guidelines contained recommendations and commentary on embedded metadata, including IPTC, Exif, XMP, and Global Positioning System (GPS).

Closely linked to the authors' own research project, the IPTC Photo Metadata Working Group study investigated how embedded metadata is shared across social media. As Bushey (2013) noted, the working group's findings:

> 'reveal image metadata is inconsistently supported across social media sites and that the two most popular sites for sharing digital images, Flickr and Facebook, remove embedded metadata from the image file header during procedures for uploading a digital image to the social media platform and downloading a digital image onto the desktop from the social media platform.'

The authors' own work further engages the conversation about embedded metadata, how it persists, and how it is shared across social media.

Methodology

Data Preparation

The researchers initially selected ten images to use in this study. They downloaded four random images from the Public Domain Collection at the New York Public Library, two in JPEG format and two in TIFF format; two images from The Europeana Collections in JPEG format; two open access journal articles from the *Journal of Librarianship and Scholarly Communication* in PDF format; one image captured by an iPhone in JPEG format; and one image captured by an Android mobile phone in JPEG format.

After selecting the images, the researchers created test accounts on multiple social media platforms, including: Pinterest (two accounts), Facebook (two accounts), Twitter (two accounts). Later they also determined that they needed data from additional platforms, including Flickr and Instagram, for a valid comparison. These accounts would be the mechanism used to transfer the selected images across social media platforms. They originally developed multiple accounts for Pinterest, Facebook, and Twitter because the researchers intended to test images shared from one like social media platform account to another. More information on these accounts will be discussed in the data collection portion of the methodology.

Before starting data collection, the researchers decreased the number of images used in the study from ten to four. There were three primary reasons for this decrease: (a) most social media platforms (including Pinterest, Facebook, and Twitter) did not support the sharing of files in PDF or TIFF format; (b) the researchers elected to test only one JPEG image from NYPL and Europeana because testing any additional JPEG images would have yielded similar results; and (c) PDF and TIFF formats in Flickr were not attempted because the other social media platforms in this study did not support these file types. Once the file selection was completed, the researchers stored the images on a local hard drive while conducting analysis on the images.

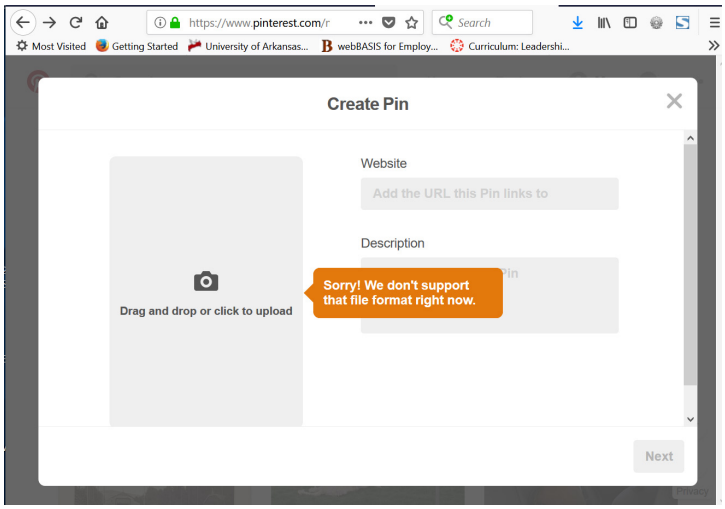

Figure 1. Screenshot of unsupported file error in Pinterest.

To record the results of the study, the researchers created a spreadsheet using Google Sheets. Each image had a sheet in the spreadsheet. Each column aligned with a sharing

 doi:10.2218/ijdc.v13i1.607

activity in the study (for example, sharing to Pinterest Account 1, Facebook Account 1, etc.). Each row recorded the embedded metadata field values, with the first row containing field labels.

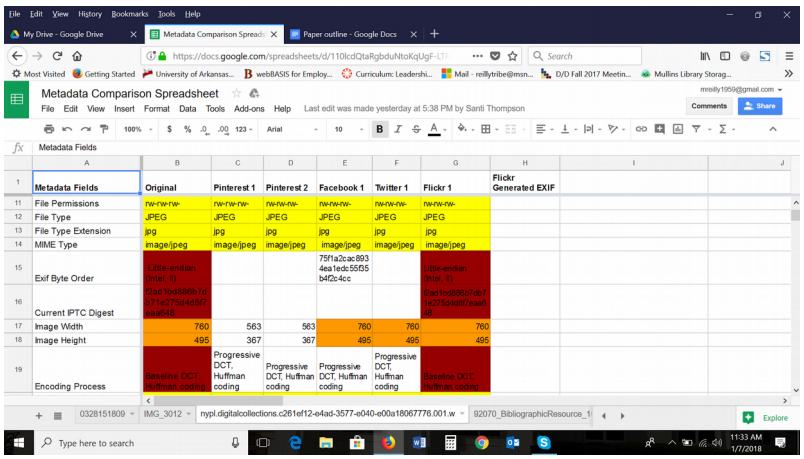

Figure 2. Google Sheets screenshot.

While conducting the study, the researchers observed that different image capture devices and institutions populated embedded metadata fields in varying degrees of comprehensiveness and arrangement. They developed a metadata template that accounted all metadata fields contained in any image used for this study, whether original or produced through sharing across social media accounts. They applied the template to each image. By the end of the data collection process, the template contained 215 metadata fields.

Before collecting any data, the researchers ran an experiment to identify the most efficient way to download images without altering the original embedded metadata for the test images. This experiment showed that third party software image viewers, such as Photoshop and Microsoft Image Viewer, changed the embedded metadata upon being loaded into the software. This confirmed observations made by Smith, Saunders, and Kejser (2014), who wrote, "if a file is copied or edited, its technical metadata may be updated automatically by the software being used." As a result, the researchers avoided the use of any third party image viewing or editing software as part of this study. Instead, they elected to take advantage of either 'Save As' feature in browsers, download features in NYPL and Europeana image repositories, and the 'Download Original' feature in Flickr.

Data Collection

The researchers ran EXIFTool on the original four images to determine the baseline embedded metadata. They recorded all metadata that the EXIFTool retrieved. Exported data was saved to the spreadsheet for later comparison.

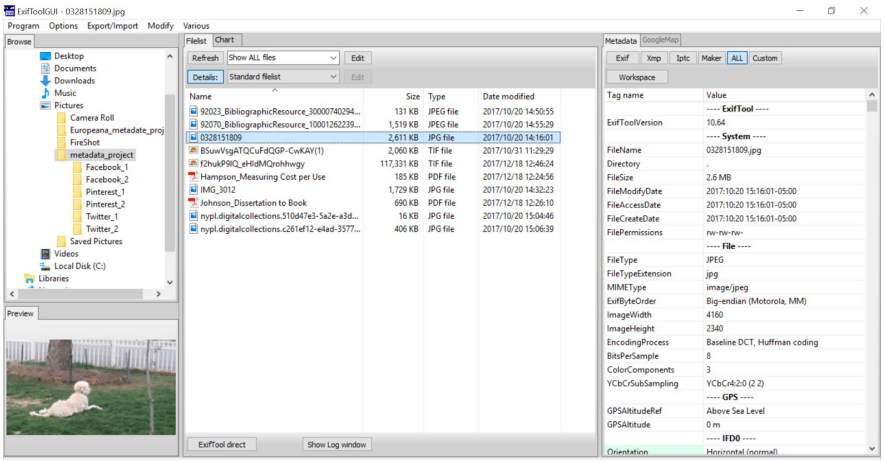

Figure 3. EXIFToolGUI screenshot.

After recording baseline values, the researchers uploaded images to the first Pinterest, Facebook, Twitter, and Flickr accounts. The researchers abandoned the use of Instagram at this point because it did not support desktop upload. Transferring images to a mobile device and then uploading had the potential to change the embedded metadata of the original image.

After the four image files were transferred to each of the social media accounts, they downloaded the files to the local desktop using the 'Save Images As' operation, extracted the embedded metadata using EXIFTool, and recorded results in the spreadsheet.

Next, the researchers attempted to share images from the first Pinterest, Facebook, and Twitter accounts to the second accounts for each of the platforms. They had limited success with this portion of the research project. Sharing from the first to second Pinterest accounts was possible. The researchers downloaded the images from the second Pinterest account to the local desktop, extracted embedded metadata using EXIFTool, and recorded results in the spreadsheet. However, the researchers discovered that they could not complete similar actions for Facebook or Twitter. While both platforms offer the ability to 'share' images from one like-account to another, the researchers noticed that the platforms produced links from the first account to the second account instead of actually transferring images from one to another. As a result, they could not collect data for the second Facebook or Twitter accounts. Consequently, they eliminated these accounts from the spreadsheet.

Finally, the researchers attempted to share images across differing platforms. When 'sharing' images from Pinterest to Facebook, they noticed that the images did not transfer. Instead, Facebook links back to the original Pinterest image. They noticed similar linking activities when working from Facebook and Twitter. As a result, they could not collect data for these actions.

Data Analysis

For each image, the researchers compared the embedded metadata of the original image against the metadata collected from the same images that were shared in Pinterest, Facebook, Twitter, and Flickr. They color-coded fields that had matches across two, three, and four platforms. They recorded the highest number of metadata matches per

 doi:10.2218/ijdc.v13i1.607

platform and logged them into an additional worksheet to visualize results (see Figure 4 below).

Results

The goal of the research project was to clarify which embedded metadata fields, if any, migrated with the object as it was shared across social media. The researchers found no meaningful, manipulatable metadata field that travelled with the image across all social media platforms. Given this result, the researchers analysed which metadata types contained fields that were more frequently shared across social media platforms.

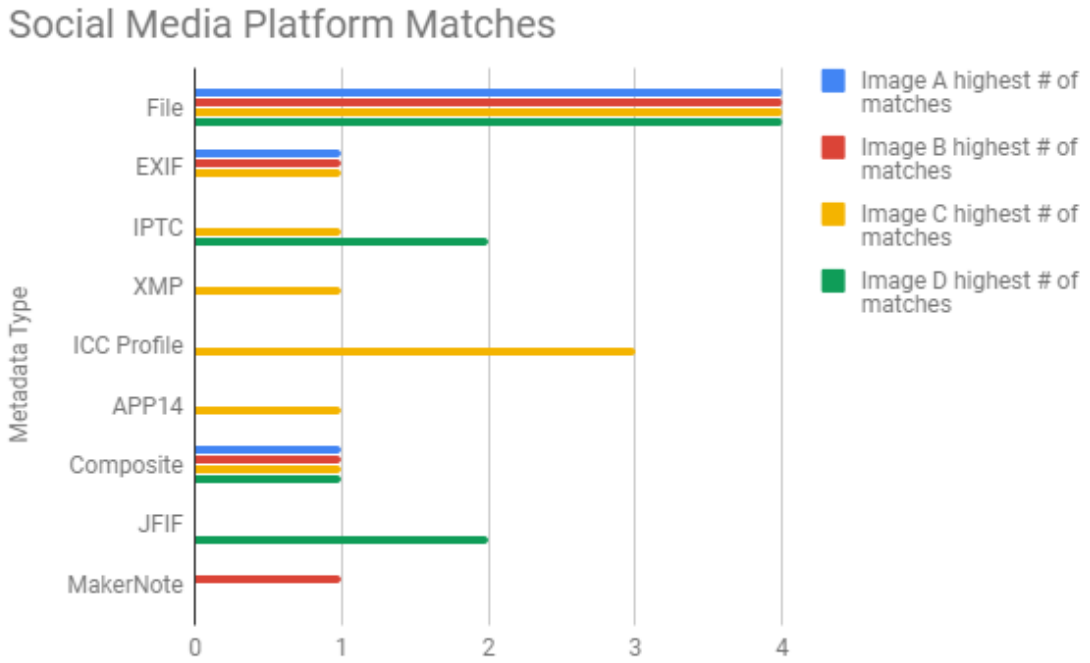

Figure 4. Social media platform matches by metadata type.

Discussion

The researchers drew upon Figure 4 to identify optimal metadata schema and fields that could potentially trace reuse across social media platforms. The researchers considered the optimal metadata type one that encompasses: (a) fields that are shared across the most platforms; and (b) fields that can be easily manipulatable in order to embed provenance or rights management information.

The most promising type, upon first glance, was File, as it shared the most values. Unfortunately, these values represented general, non-manipulatable fields, like File Type (JPEG), File Extensions (jpg), and MIME Type (image/jpeg). Fields like these, however, were not ideal candidates for tracing reuse over social media because of several factors, including (a) these fields did not contain distinct-enough values to

differentiate one JPEG image from another or (b) the values that were distinct (like File Name) were altered by some social media platforms (for example, see Table 2 below).

Table 2. File names altered by social media platforms.

Metadata Fields	Original	Pinterest 1 account	Pinterest 2 account	Facebook 1 account	Twitter 1 account	Flickr 1 account
File Name:	0328151809.jpg	b4325efd176b4b96f2845046fbb8dde9.jpg	b4325efd176b4b96f2845046fbb8dde9.jpg	03281809_25396300_154271818671737_656573911975886754_n .jpg	03281809_DRW6HUMUEAAjseX.jpg	25540521548_4d1ef6bb59_o.jpg

Two additional embedded metadata types, ICC Profiles and JFIF, demonstrated multiple instances of sharing across social media. ICC Profiles, as discussed in the background section, document color properties and characteristics of an image. According to Wikipedia (2017), "the ICC defines the format precisely but does not define algorithms or processing details." For the purposes of this study, the researchers observed that JFIF data captured the X/Y resolutions of the JPEG image. According to Wikipedia (2018), JFIF "defines the number of details left unspecified in the JPEG part 1 standard." The researchers found that ICC Profile and JFIF metadata were not ideal candidates for tracing reuse over social media because there was no way to differentiate an original image and an exact copy using data from these metadata types. ICC metadata focuses on the color output of the capture device and JFIF acts only as an extension for JPEG properties. Furthermore, both metadata types are not intended to be manipulated.

The researchers hypothesize that IPTC metadata shows the most promise for tracing reuse over social media. IPTC is designed to contain unique, manipulatable data about an object – "descriptive information, including photographer name, subject and copyright/licensing terms" (Bushey, 2015) – that could be theoretically traced back to the original object. The researchers' preliminary analysis found that IPTC metadata fields can be changed easily within the desktop environment. Several metadata fields have free text properties that can be edited in whichever image viewer available to a user. Additionally, IPTC metadata not only traveled to two of the four social media platforms (Flickr and Facebook) but also transferred the kinds of fields that were manipulatable (see Figure 6 below for example of editing using Windows Properties interface).

 doi:10.2218/ijdc.v13i1.607

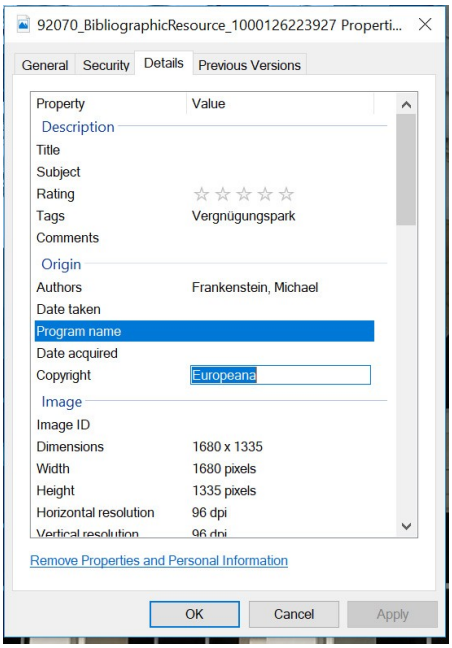

Figure 5. Screenshot of editing IPTC metadata field in Windows Explorer.

The remaining metadata types were not analyzed by the researchers because they only had one social media match. Additionally, several of these, including Exif, were not manipulatable after the point of capture.

Conclusion

This is a very early and small study on tracking embedded metadata in social media platforms and is part of a larger research agenda focused on understanding the reuse of digital images. As such, the researchers have more to learn about the various kinds of software and applications available to view and edit embedded metadata and the intricacies of specific embedded metadata types and fields.

Based on preliminary reading, the researchers presumed that an object's embedded metadata, which could be unique to the object, may be one potential strategy for tracking shared images across social media. After completing this study, they found no reliable metadata field that extended to all platforms studied. This complicates and contradicts previous research by others.

The researchers identified one metadata type, IPTC, that holds promise towards their larger research agenda. Future research is still needed to verify this hypothesis. It should address several questions: (a) what are the sharing and manipulation possibilities of IPTC metadata? (b) What flexibility exists within the IPTC standard to allow for metadata manipulation? (c) What tools are needed to effectively manipulate data that will transfer? (d) What implications arise when metadata types are supported or not supported by social media platforms? This final question is particularly important given that "existing software and file formats don't support locking, and there's no magical way to make them do that" (Krough, 2018).

While this research has developed more questions than answers it has determined that some embedded metadata is shared across social media platforms, giving hope to the possibility of tracing digital image reuse.

References

Adobe. (n.d.). Extensible metadata platform (XMP). Retrieved from http://www.adobe.com/products/xmp.html

Banerjee, K., & Anderson, M. (2013). Batch metadata assignment to archival photograph collections using facial recognition software. *Code4Lib Journal*, 21. https://journal.code4lib.org/articles/8486

Bushey, J. (2015, January). Trustworthy citizen-generated images and video on social media platforms. In *System Sciences (HICSS) 2015 – 48th Hawaii International Conference* (pp. 1553-1564). IEEE.

Bushey, J. (2013, September). Trustworthy digital images and the cloud: Early findings of the records in the cloud project. In *International Symposium on Information Management in a Changing World* (pp. 43-53). Springer, Berlin, Heidelberg.

Christensen, S.O., & Dunlop, D. (2011, January). The case for implementing core descriptive embedded metadata at the Smithsonian. In *Archiving Conference*, 2011(1), pp. 116-120. Society for Imaging Science and Technology.

Corrado, E.M., & Jaffe, R. (2017). Access's unsung hero: The [impending] rise of embedded metadata. *International Information and Library Review*, *49*(2), pp. 124-130. doi:10.1080/10572317.2017.1314142

Fuhrig, L.S. (2012). Three cheers for embedded metadata. Smithsonian Institution Archives blog. Retrieved from https://siarchives.si.edu/blog/three-cheers-embedded-metadata

Harvey, P. (2016). Composite tags. Retrieved from https://sno.phy.queensu.ca/~phil/exiftool/TagNames/Composite.html

Harvey, P. (2014). JPEG tags. Retrieved from https://sno.phy.queensu.ca/~phil/exiftool/TagNames/JPEG.html

Harvey, P. (2003). Exiftool application documentation. Retrieved from https://sno.phy.queensu.ca/~phil/exiftool/exiftool_pod.html#SYNOPSIS

International Press Telecommunications Council. (2017). IPTC photo metadata standard. Retrieved from https://iptc.org/standards/photo-metadata/iptc-standard

Krough, P. (2018) Metadata overview. *Digital photography best practices and workflow*. American Society of Media Photographers. Retrieved from http://www.dpbestflow.org/metadata/metadata-overview#handle

Krough, P. (2015) Project overview. *Digital photography best practices and workflow*. American Society of Media Photographers. Retrieved from http://www.dpbestflow.org/project-overview

Reilly, M. & Thompson, S. (2017). Reverse image lookup: Assessing digital library users and reuses. Journal of Web Librarianship, 11(1), 56-68. doi:10.1080/19322909.2016.1223573

Reilly, M. & Thompson, S. (2014). Understanding ultimate use data and its implication for digital library management: A case study. *Journal of Web Librarianship*, *8*(2), 196-213. doi:10.1080/19322909.2014.901211

Shala, L. & Shala, A. (2016). File formats-characterization and validation. *IFAC-PapersOnLine*, *49*(29), 253-258. doi:10.1016/j.ifacol.2016.11.062

Smith, K.R., Saunders, S., & Kejser, U.B. (2014, June). Making the case for embedded metadata in digital images. In *Archiving Conference*, 2014 (1), pp. 52-57. Society for Imaging Science and Technology.

Thompson, S. & Reilly, M. (2017). "A picture is worth a thousand words": Reverse image lookup and digital library assessment. *Journal of the Association for Information Science and Technology*, *68*(9), 2264-2266. doi:10.1002/asi.23847

Wikipedia. (2018). JPEG File Interchange Format. Retrieved from https://en.wikipedia.org/wiki/JPEG_File_Interchange_Format#JFIF_extension_APP0_marker_segment

Wikipedia. (2017). ICC Profile. Retrieved from https://en.wikipedia.org/wiki/ICC_profile

Wikipedia. (2016). International Color Consortium. Retrieved from https://en.wikipedia.org/wiki/International_Color_Consortium