

Internet Archive and JPEG EXIF Data

Chuck Cartledge

6 February 2016

1 Introduction

The Internet Archive (IA) ¹ crawls significant portions of the Internet, stores what it finds, and makes the results of those crawls available for free. If the IA crawls a particular Internet host periodically, then the IA has snapshots of the host over time. These crawls contain both textual data and images from the host. The IA acts as a true archive because it preserves data, and makes it available for future users. There are many questions that could be asked about the veracity of these crawls, and how the crawled data could be used. This short paper focuses on two specific questions:

1. Does the IA change the metadata associated with an image?
2. Should the IA be relied upon to maintain its current image policies?

If the IA does not modify the metadata, then the metadata could be used to hold data that the originator felt was important.

2 Discussion

For our purposes, metadata is any data that is associated with an image. There are many different types of metadata that could be associated with an image. Popular image meta data include:

- EXchangeable Image File (EXIF)[1] put forward by the Japan Electronic Industries Development Association (JEIDA),
- International Press Telecommunications Council (IPTC) put forward by major news agencies, and
- Extensible Metadata Platform (XMP) International Organization for Standardization as ISO 16684-1:2012 standard originally put forward by Adobe Systems.

We will be looking at EXIF meta data because it is common to images coming from cameras in the US. A sample EXIF set is provided (see Table 1).



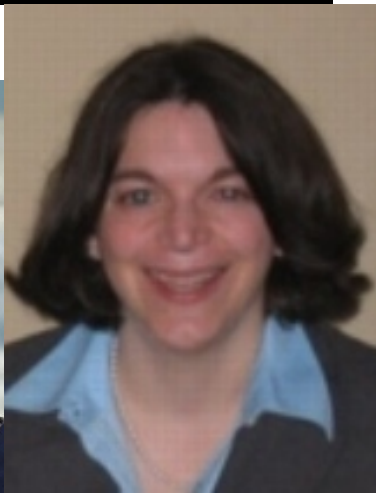
The test to see if the IA modifies the EXIF data is to compare current images with IA crawled images and see if:

1. They are the byte for byte the same as the original images, and
2. They are image from the IA and not from the live web.

Table 1: Representative EXIF data. Data taken from Wikipedia https://en.wikipedia.org/wiki/Exchangeable_image_file_format.

Tag	Value
Tag	Value
Manufacturer	CASIO
Model	QV-4000
Orientation (rotation)	top - left [8 possible values[21]]
Software	Ver1.01
Date and Time	2003:08:11 16:45:32
YCbCr	Positioning centered
Compression	JPEG compression
x-Resolution	72.00
y-Resolution	72.00
Resolution Unit	Inch
Exposure Time	1/659 sec.
FNumber	f/4.0
ExposureProgram	Normal program
Exif Version	Exif Version 2.1
Date and Time (original)	2003:08:11 16:45:32
Date and Time (digitized)	2003:08:11 16:45:32
ComponentsConfiguration	Y Cb Cr -
Compressed Bits per Pixel	4.01
Exposure Bias	0.0
MaxApertureValue	2.00
Metering Mode	Pattern
Flash	Flash did not fire.
Focal Length	20.1 mm
MakerNote	432 bytes unknown data
FlashPixVersion	FlashPix Version 1.0
Color	Space sRGB
PixelXDimension	2240
PixelYDimension	1680
File	Source DSC
InteroperabilityIndex	R98
InteroperabilityVersion	(null)

Table 2: EXIF metadata associated with IA crawled images. Data was extracted using the exiv2 program.

EXIF tag			
	http://www.cs.odu.edu/~ccartled/	http://www.cs.odu.edu/~lnguyen/	http://www.cs.odu.edu/faculty_show.shtml?p=26
File name	HPIM0235-wayback.JPG	louis1-wayback.jpg	micheleused-wayback.png.jpeg
File size	813883 Bytes	41861 Bytes	14812 Bytes
MIME type	image/jpeg	image/jpeg	image/jpeg
Image size	2848 x 2144	413 x 456	162 x 212
Camera make	Hewlett-Packard	NIKON CORPORATION	Canon
Camera model	HP PhotoSmart R727 (V01.00)	NIKON D70	Canon PowerShot S410
Image timestamp	2007:12:07 21:57:35	2009:05:22 18:37:41	2006:05:21 09:04:31
Image number			108-0859
Exposure time	1/60 s	1/500 s	1/60 s
Aperture	F5.1	F4.5	F4.9
Exposure bias	0 EV	0 EV	0 EV
Flash	Yes, auto, return light detected	Yes, auto, return light detected	Yes, auto, red-eye reduction
Flash bias			0 EV
Focal length	6.5 mm (35 mm equivalent: 41.0 mm)	56.0 mm	22.2 mm
Subject distance			0
ISO speed	100	200	3.125
Exposure mode	Auto	Aperture priority	Easy shooting (Auto)
Metering mode	Center weighted average	Multi-segment	Multi-segment
Macro mode			(25964)
Image quality			(0)
Exif Resolution	2848 x 2144	413 x 456	2272 x 1704
White balance	Auto		Auto
Thumbnail	image/jpeg, 6596 Bytes	None	None
Copyright			
Exif comment			256 (null)

The program `exiv2` was used to extract the EXIF data from each of the IA crawled images (see Table 2). Of particular interest is the row (see Table 2) labeled “Exif comment” because in most of the samples it is empty on the images, but in one case it has 256 NULL characters.

`exiv2` is a command line program that can read, write, set, and modify EXIF, IPTC, and XMP meta data. The program can be used to set/write EXIF metadata into the comment section like this (line broken for readability):

```
exiv2 -M"set Exif.Photo.UserComment charset=Ascii This is a test of the silly stuff"
      HPIM0235-orig.JPG
```

Which would cause the Exif comment field to change to:

```
Exif comment      : This is a test of the silly stuff
```

If the format of information in the comment field is important, then base 64 encoding the original data may be advisable, using these commands:

```
base64 -w0 DataToAdd
exiv2 -M"set Exif.Photo.UserComment charset=Ascii base64 encoded RGF0YVRvQWRk" image.jpg
```

According to the EXIF specification, any type of data, of any length could be added as a `UserComment` ([1], pg 46). My testing was limited to exercising `exiv2` from the command line, which has a limit on the number of characters that can be on the command line. I was able to insert and retrieve 3115 characters in the `UserComment` field.

I was able to find python, C++, and C# program examples of updating EXIF data, so the limits imposed by using the command line may not be an issue.

3 Conclusion

Using a very limited dataset, I was able to confirm that the Internet Archive (IA) does not appear to modify the EXchangeable Image File (EXIF) meta data associated with an image during the crawling process. I was able to put arbitrary information into the EXIF `UserComment` field using `exiv2`. The time spent investigating this question did not allow for full testing using an IA crawl to discover the new data, but I am confident the new data would remain intact.

Returning to the original questions:

1. Does the IA change the metadata associated with an image? No. Based on limited data sampling, the original EXIF data is unchanged by an IA crawl.
2. Should the IA be relied upon to maintain its current image policies? No. I think that depending long term on a policy that uses a backdoor approach is not a good idea.

References

- [1] Camera & Imaging Products Association et al., *Exchangeable image file format for digital still cameras: Exif version 2.3*, Tech. report, CIPA DC-008-2010 & JEITA CP-3451B Standard, 2012.

¹<https://archive.org/>