# Digital Forensics Tool Testing – Image Metadata in the Cloud

Philip Clark

Master's Thesis
Master of Science in Information Security
30 ECTS
Department of Computer Science and Media Technology
Gjøvik University College, 2011

Avdeling for
informatikk og medieteknikk
Høgskolen i Gjøvik
Postboks 191
2802 Gjøvik

Department of Computer Science
and Media Technology
Gjøvik University College
Box 191
N-2802 Gjøvik
Norway

# Abstract

As cloud based services are becoming a common way for users to store and share images on the internet, this adds a new layer to the traditional digital forensics examination, which could cause additional potential errors in the investigation. Courtroom forensics evidence has historically been criticised for lacking a scientific basis. This thesis aims to present an approach for testing to what extent cloud based services alter or remove metadata in the images stored through such services.

To exemplify what information which could potentially reveal sensitive information through image metadata, an overview of what information is publically shared will be presented, by looking at a selective section of images published on the internet through image sharing services in the cloud.

The main contributions to be made through this thesis will be to provide an overview of what information regular users give away while publishing images through sharing services on the internet, either willingly or unwittingly, as well as provide an overview of how cloud based services handle Exif metadata today, along with how a forensic practitioner can verify to what extent information through a given cloud based service is reliable. Further, a methodology for testing cloud based storage services will be presented, and followed through an example using a selection of cloud based storage services to see what Exif information is altered after images are stored in the cloud.

The results presented in this thesis could be practically useful for forensic practitioners as a means to demonstrate the need to carry out experiments before using information gathered through cloud services, as a business case in the scoping of a forensics assignment. Further, the results can be used by cloud service providers to help develop their cloud service to be able to aid forensics practitioners in tracing the origin of an image. The suggested approach could be used not only to test cloud based storage services in regards to image metadata, but could also be used to look at metadata in documents and other media files stored in the cloud.

# Acknowledgement

I would like to thank my fiancée Linda for her enormous support and understanding throughout the process of writing this thesis. Without her patience, I could not have completed this work in the timeframe available.

The work done by Peter E. Berg, as student opponent was greatly appreciated, and a big thank you is extended to Peter for his thorough second reading and productive input to the process of writing this thesis.

I would further like to thank my thesis supervisor, Professor Katrin Franke, for her input and mentoring in my work. Through her guidance, she has helped bring quality to the process. Without her help, this thesis would have been more difficult to produce.

Finally, I would like to thank Marit Gjerde, who willingly set out to be my second thesis supervisor, and helped me greatly in the initial scoping of this thesis. She has been a huge motivator throughout the entire process of writing this thesis, by giving me inspiration to keep going no matter what.

# Table of Contents

# Table of figures

# Table of tables

# 1. Introduction

## 1.1. Background

The use of digital forensics techniques are often used as a means to support an argument presented in a courtroom. Thus the work of the forensics examiner will be subject to scrutiny from both defence lawyers and the prosecution, and both parties are equally dependent on how the expert's integrity holds up through the interrogation process. A large part of this depends on the forensic practitioners' choice of both tools and methods in coming up with his or her statement. Historically courtroom forensic testimony has often been criticized by defence lawyers as lacking scientific basis [1-3]. This, in part, has been a result of forensic practitioners using (semi) automated tools to extract and analyse the evidence, without necessarily knowing in-depth how the tools they use actually work on a lower level – or at least not knowing with a sufficient amount of certainty, to be able to present a convincing explanation in the courtroom. Another reoccurring issue is the lack of proper scientific methods of testing the performance of the tools used prior to use in the courtroom case. Such testing could be either to make sure the tools in use produce accurate results based on different data sets, or that they produce reliable results when tested again later, and reproducible results if tested with a separate set of tools. This is often referred to as dual tool verification, and is a leading practice in digital forensics.

The court rulings of Frye [4] and Daubert [5] are a common reference point in the forensics community, as these rulings urgently address the issue of scientifically proven methods to be incorporated in working with forensic evidence. This is not unique to the field of digital forensics, however, but is pressing also in the field of computational forensics, as described by Katrin Franke in "Trends and challenges in applying artificial intelligence methodologies to digital forensics" [1] and in the handling of physical evidences. There is quite a lot of work done in the field of physical evidence, but there is still a need to address the domain of digital evidence in this regard [1]. Physical evidence is, however, outside the scope of this thesis.

When forensic evidence is brought before a court of law, there are certain principles that need to be adhered to in regards to the validity and admissibility of expert opinion testimony. The main relevant court rulings are the ones of Frye [6] and Daubert [7], which will be introduced below.

### 1.1.1. Daubert criteria of admissibility

The article Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner by Josh Brunty [8] summarises the Daubert Standard quite well:

> *In the legal community, the Daubert Standard can be used for guidance when drafting software/tool validations. The Daubert Standard allows novel tests to be admitted in court, as long as certain criteria are met. According to the ruling in Daubert v. Merrell Dow Pharmaceuticals Inc. [5] the following criteria were identified to determine the reliability of a particular scientific technique:*

1. Has the method in question undergone empirical testing?
2. Has the method been subjected to peer review?
3. Does the method have any known or potential error rate?
4. Do standards exist for the control of the technique's operation?
5. Has the method received general acceptance in the relevant scientific community?

This means that if either party calls for a Daubert motion, the judge has to decide, based upon the aforementioned criteria, if the presented evidence or expert witness testimony should be admitted or excluded from the case data.

### 1.1.2. Frye standard

The previous US Supreme Court ruling of Frye v. United States [4] from 1923 mainly concluded that expert scientific testimony was to be admitted only when it had received "general acceptance" in the relevant scientific community [9, 10]. To meet the Frye standard, scientific evidence presented to the court must be interpreted by the court as *"generally accepted" by a meaningful segment of the associated scientific community. This applies to procedures, principles or techniques that may be presented in the proceedings of a court case* [9].

The Frye acceptance test preceded the Daubert ruling, and nowadays Daubert is most commonly referred to in regards to acceptance tests in the courtroom. The main difference aside from the acceptance criteria between Frye and Daubert is that in the case of Frye it is the jury or a panel of experts who decide if the evidence is admissible or not, whereas in the case of Daubert, the judge makes the decision [7].

### 1.1.3. Need for standardisation

The attention especially through the court, but also driven forward from the fact that new tools are released every day to perform digital forensics tasks, create the need for standardisation.

There are currently several different initiatives that have been made from the academic community to help provide a solid foundation for the expert testimony. There has been done some research in the field of tool testing for use in digital forensics, and a couple of guidelines have been made available. Chapter *2.1 Review of the previous research in the field* intends to give an overview of the current state regarding publically available methodologies for tool testing. Examples of relevant initiatives mentioned are:

- NIST Computer Forensics Tool Testing Project [11].
- ISO/IEC 17025: General requirements for the competence of testing and calibration laboratories [12].
- Scientific Working Group on Digital Evidence (SWGDE) – "Recommendations for Validation Testing" [13].

These initiatives will be described in chapter *2.1 Review of the previous research in the field*.

## 1.2. Motivation

As digital evidence is becoming more and more relevant in the ever increasingly technology driven world, new challenges arise in handling such evidence. Considering that new technology and new software is released continuously, keeping up with all developments is a daunting task. In being able to present digital evidence in a court of law, the scientific principles of validity and admissibility have to be met. To be able to do so, there has to be some generally accepted guidelines for testing of the forensics tools used in the handling of digital evidence, to be able to say with a reasonable amount of certainty that the results are a cause of one thing, rather than being introduced through the use of a tool or function that the practitioner is unaware of.

As new services on the internet makes it easier to collaborate and share information, the forensic artefacts are more and more being moved from the local storage on a person's computer over to storage spaces in the cloud. With the addition of an extra link in the process, there is a need to verify that the process of cloud storage itself does not introduce extra unknown variables. This makes it interesting to see whether such cloud services retain the original information uploaded, or if they are altered somehow in the process. If altered, the artefacts stored in the cloud could give a forensic examiner incorrect data, which could have severe implications if used in a court of law – either for the implicated parties or the integrity of the forensics examiner.

With the emerging use of different cloud based services for storing and sharing information, there could also be a possible privacy issue. Digital cameras today have the ability to store additional information, or metadata, along with the photograph itself. This information could include information such as time and date when the picture was taken, GPS location of where the picture was taken, and an original thumbnail of the photograph (before any potential editing has taken place). To get an overview of the extent of how widespread such information is available, a mapping of what information is freely available on the internet is interesting.

## 1.3. Research problem

This thesis will mainly focus on two aspects of image metadata and cloud storage, namely metadata contained in images, and to what extent this information is retained when the image is stored or shared using a cloud service.

This is especially interesting for cloud service providers, who seek a guidance to what type of information a user willingly (or unwittingly) shares with the cloud provider, and what measures they can take to at least keep this information on file to be able to aid in a forensic investigation, and what is shared freely to the world.

Further, an approach for testing clod based storage services will be presented.

The main questions this thesis will aim to answer are:

1. What information do people share freely (or unwittingly) through their photographs on the Internet, contained in the image metadata?
    - To what extent is information which could help identify a person or location shared with the images themselves?
2. How is this information affected by storage in the cloud?
    - What information is lost or altered when uploading to these types of cloud based storage services in regards to image metadata?

## 1.4. Limitations

There are many standards for storing images digitally, and also many standards for storing image metadata within these images. This thesis will focus mainly on the most commonly used image file formats for sharing images today. The selection process is discussed in chapter *2.3 The most commonly used image file formats today*. This thesis will therefore be limited to the following image file formats:

- **JPEG**
- GIF
- PNG
- TIFF

However, out of the four, it was found that JPEG images outnumbered the other formats significantly, and as such, JPEG images will be the main focus in this thesis.

Due to these findings, and the additional observation that most digital cameras today store image metadata using the Exif format, and that most other formats were mainly added while editing, the thesis will mainly focus on finding relevant Exif information.

This thesis will additionally direct its focus on information which could be used to aid in identifying persons or locations, and thus the Exif fields of main interest are the ones containing information about:

- Location information (GPS-data)
- Camera make, model and serial number
- Distance setting for camera focus
- Date and time when the photo was taken
- Thumbnail image of the original picture

There is an increasingly growing amount of cloud based storage services available, and only a sub-section of these services could be tested within the given time limitation of this thesis. The cloud based storage services tested in this thesis are:

- Windows Live SkyDrive
- Windows Azure Platform Storage service
- Flickr

# 2. Theoretical background – the state of the art

## 2.1. Review of the previous research in the field

To be able to present digital evidence in a court of law, the aforementioned criteria of Daubert/Frye have to be met. To make sure that a subject matter expert is able to present the evidence in a satisfactory way, several different initiatives have been made from the academic community to help provide a solid foundation for the expert testimony. There has been done some research in the field of tool testing for use in digital forensics, and a couple of guidelines have been made available. This section intends to give an overview of the current state regarding publically available methodologies for tool testing.

### 2.1.1. NIST Computer Forensics Tool Testing Project

The National Institute of Standards and Technology (NIST)[14] have an on-going project called the Computer Forensics Tool Testing (CFTT) project [11]. The goal of the CFTT project at NIST is to: *"establish a methodology for testing computer forensic software tools by development of general tool specifications, test procedures, test criteria, test sets, and test hardware."* [11]. The project has produced a general test methodology for computer forensics tools [15], which is heavily based upon the guidelines provided in ISO/IEC 17025, *General Requirements for the Competence of Testing and Calibration Laboratories*. [12]

The general approach defined in the NIST methodology is to:

1. establish categories of forensic requirements,
2. identify requirements for a specific category,
3. develop test assertions based on requirements,
4. develop test code for assertions,
5. identify relevant test cases,
6. develop testing procedures and method,
7. report test results.

Based on this test methodology, the CFTT project has produced a large amount of test reports for different digital forensics tools, mainly divided into the following categories:

- Data acquisition
- Software write block
- Hardware write block
- Mobile device (cell phone)
- Drive wipe

General observations the NIST CFTT project has identified is that error rates are hard to define and quantify; an error rate may have a theoretical error rate, an implementation of an algorithm may have errors, and the execution of a procedure may have a blunder that affects the results [16].

### 2.1.2. Scientific Working Group on Digital Evidence (SWGDE) – "Recommendations for Validation Testing"

The work done by the Scientific Working Group on Digital Evidence (SWGDE) [17] is in the broader field of validation testing. SWGDE defines validation testing as: *"An evaluation to determine if a tool, technique or procedure functions correctly and as intended."* [13]. SWGDE have published a paper called: "*2009-01-15 SWGDE Recommendations for Validation Testing Version v1.1*" [13]. The SWGDE mission statement states: *"SWGDE brings together organizations actively engaged in the field of digital and multimedia evidence to foster communication and cooperation as well as ensuring quality and consistency within the forensic community"* [17], and their Forensics Committee states: *"It is the mission of the Forensic Committee to promote the use of forensically sound techniques in the collection and analysis of digital and multimedia evidence. The Forensic Committee will endeavour to accomplish our mission through the direction of subject matter experts and the publication of technical notes and papers."* [17].

The guidelines presented in: "*2009-01-15 SWGDE Recommendations for Validation Testing Version v1.1*" [13] are a straight to the point list of actions to follow to complete a tool testing validation process. They do not provide a specific form to fill out, but provide guidelines to follow step by step for the tool testing process and a suggested list of topics to include when typing up the report for each tool tested.

The process presented consists of the following steps [13]:

1. Develop and document test plan before testing begins
   a. Purpose and scope
   b. Requirements to be tested – what does the tool have to do?
   c. Methodology – how to test? (Identify support tools required to assist in evaluation of results when applicable)
   d. Test scenarios
      i. Condition or environment required for test scenario
      ii. Actions to perform during utilization of the tool, technique or procedure
      iii. Expected results - determine pass/fail criteria
      iv. One test may be sufficient depending on the tool, technique or procedure being tested. The number of test scenarios should be sufficient to cover the various environments encountered – for example, different file systems, media sizes, platforms, device types, etc.
      v. Different options may need to be tested such as user configurable option settings, switch settings, etc., in accordance with purpose and scope

   e. Test data to fulfil conditions of test scenarios – can the existing reference data set be used? (Identify support tools required to assist in the development of test data when applicable)
   f. Document test data used
2. Perform test scenario(s) and document results in test report
   a. Use media and/or other sample materials that are in a known state or condition
   b. Use test equipment with known configuration which corresponds to your examination environment
   c. If anomaly occurs then:
      i. Attempt to identify conditions causing anomaly
      ii. Attempt to independently verify conditions causing anomaly
      iii. If feasible, implement alternative procedure and re-test
   d. If re-tests are performed, results of all tests must be documented
   e. Be sure pass/fail status for each requirement is annotated in test report
   f. Ensure to annotate all testers and dates assigned to test scenario
   g. Individual test scenario(s) must be documented separately, but a summary report should be written which states the overall pass/fail status of the tool, technique or procedure, along with any recommendations, concerns, etc.
   h. Validation of results: comparison between actual and expected results must be performed and discrepancies between the two must be documented.

### 2.1.3. Lynn M. Batten and Lei Pan – "Testing Digital Forensic Software Tools Used in Expert Testimony"

The work presented in the paper by Lynn M. Batten and Lei Pan from Deakin University, Australia, proposes an experimental framework that aims to help digital forensic experts to compare sets of digital forensic tools of similar functionality based on specific outcomes. The results can be used by an expert witness to justify the choice of tools and experimental settings, calculate the testing cost in advance, and to be assured of obtaining results of good quality [18, 19]. The framework combines the advantages of the Jain [20] and Taguchi [21] models, and claims to avoid their drawbacks.

The structure in the model proposed in the paper is summarised in Figure 1.

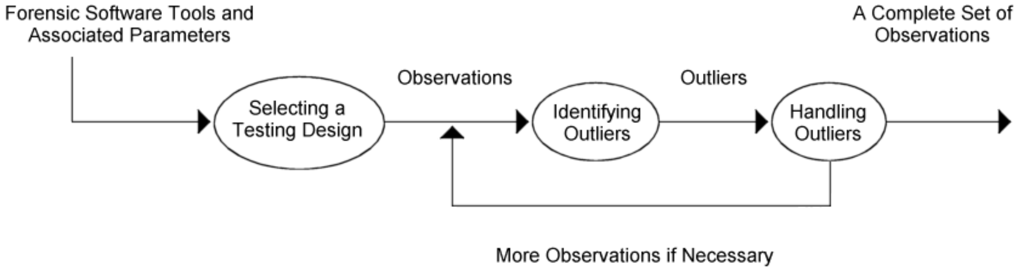

Forensic Software Tools and Associated Parameters
A Complete Set of Observations
Observations
Outliers
Selecting a Testing Design
Identifying Outliers
Handling Outliers
More Observations if Necessary

*Figure 1: 3-component testing framework as presented in [19].*

The overall process in Batten and Pan's framework starts with the tester deciding on what tools need testing, the important parameters and the settings of those parameters on which to test. This is done by:

1. Selecting a testing design based on these tools and parameters.
2. Identifying outliers in the observations, since errors can occur in experiments due to inaccurate measurement tools and human errors.
3. Reducing the negative impact of the identified outliers to a safe level so that the tester can correctly draw the conclusion regarding the test.

At the last stage of the test, the tester obtains a complete set of observations which should be applied to appropriate statistical models, for example the Analysis of Variance (ANOVA) method, as presented by Taguchi et al. [22].

### 2.1.4. Further research

The topic of digital forensics tool testing is being further researched by other organisations and individuals as well, which shows that this field of research is highly relevant now and in the future. As an example, a paper is in the process of being published by Mario Hildebrandt, Stefan Kiltz and Jana Dittmann from Universitaet Magdeburg in Germany with the title: "A common Scheme for Evaluation of Forensic Software", and was presented at the 6th International Conference on IT Security Incident Management and IT Forensics (IMF2011) in Stuttgart in May 2011 [23].

At the time of writing of this thesis, the paper was only circulated for use at the conference; however the relevancy of the subject is apparent, and shows that the field of research has several yet unsolved problems which need to be addressed.

## 2.2. Introduction to graphics file formats

There are a large number of different image file types to choose from. No two formats are the same, and each type stores graphics data in a different way. When choosing a file type one has to take into account for what purpose the image shall be used. They all have different characteristics and parameters. Some have a more realistic colour representation, can be enlarged and printed on a billboard and can be easily manipulated without significant visible data loss. Other image file types are better adapted to viewing on the internet with a compressed file size, but often with a loss in quality as a result, rendering the image grainy if tried to enlarge. There are also a variety of other differences.

This thesis will not go into great detail in describing all different image file formats, but will rather present an overview of the most common image file types used for photography storage today. Furthermore, as the most relevant part of the image file format in this thesis is where the image metadata is stored, theory about the image data itself will not be widely described.

To try and characterize all these different image file formats, they can mainly be put in groups as either raster graphics (bitmaps) or vector graphics. In addition to these format types, The *Encyclopedia of Graphics File Formats* [24] specifies some additional format types of lesser importance, which are mentioned for completeness; these format types include metafile formats, scene formats, animation and multimedia. These format types will not be discussed further in this thesis.

### 2.2.1. Raster images

Raster images are popularly referred to as bitmap images or pixel images, and are used for storing bitmap data. They consist of a rectangular set of pixels, or bits, which are each given a colour, which represent the image. An illustration of this is given in Figure 2.

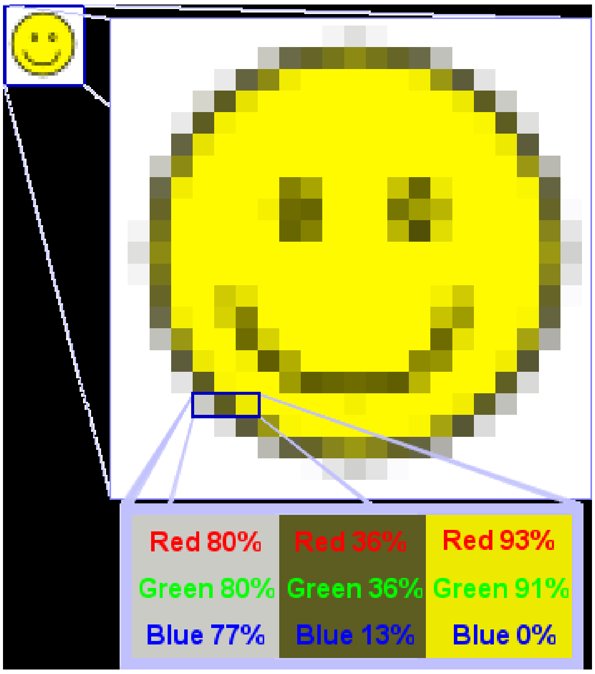


*Figure 2: Illustration of a raster image [25].*

Raster images consist of a header, bitmap data and other information, which may include a colour palette and other data. The components in the file format can be arranged in different ways depending on the file format, and can consist of a large number of different sections. However, a common example consists of a header, bitmap data and footer, as shown in Figure 3.

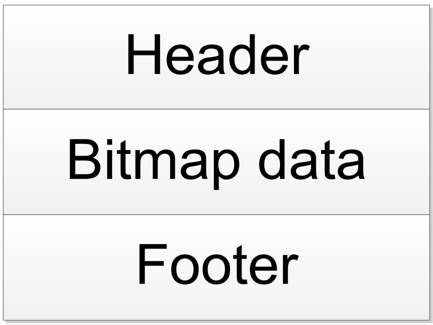


*Figure 3: A common composition of a raster image.*

### Header

The header section consists of binary or ASCII-format data containing information about the bitmap data found elsewhere in the file. The header is usually found at the beginning of the file format, and often contains an identification field or magic number which enables software to differentiate the header from other fields in the file format. The header usually consists of fixed fields. All bitmap files have some kind of header, but the format of the header and the information stored in it varies considerably from format to format.

Figure 4 gives an example of fields that could be part of the header structure for raster images, as portrayed by the *Encyclopedia of Graphics File Formats [24]*.

Each header field will not be studied in detail in this thesis, however the subject of what information is stored about the raster image will be explored in chapter *2.4 Image metadata*.

### Bitmap data

This section contains the actual bitmap data which makes up the visible part of the image. Naturally this is the largest section of the file format. How the bitmap is physically stored in this section varies from image format to image format, and is amongst other factors dependent on the compression level and colour palette used, to mention a few possibilities. As this thesis focuses mainly on image metadata, this section will not be explored further.

### Footer

The *Encyclopedia of Graphics File Formats* [24] defines the footer, sometimes referred to as the trailer, as a data structure similar to a header which is often an addition to the original header, but appended to the end of a file. The existence of a footer is mainly a result of a desire to maintain backwards compatibility with previous versions of a file format, and is appended when it is no longer convenient to add or change information in the header.

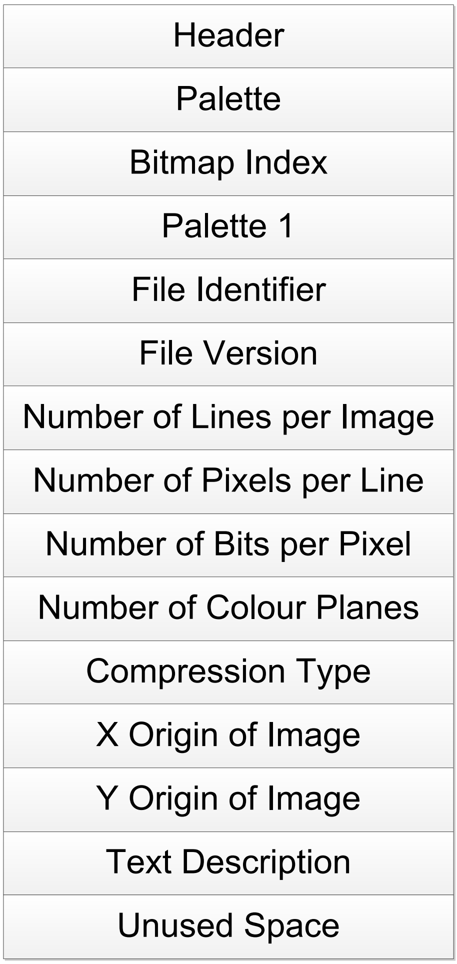


*Figure 4: Example of header fields [24].*

### 2.2.2. Vector images

Vector images are designed to store vector data, such as lines and geometric data. Although vectors can be more complex, a vector is minimally defined as an element containing a starting point, a direction and a length. As opposed to raster images which store a pixel by pixel bitmap of the image, vector images store mathematical descriptions of one or more image elements. These elements are then interpreted by the rendering engine to construct the final image. Because of the mathematical interpretation within the rendering engine, rather than having a pixel by pixel mapping, vector images are easily scalable, and thus provide a more visually appealing appearance when for instance magnified. Figure 5 aims to illustrate this.

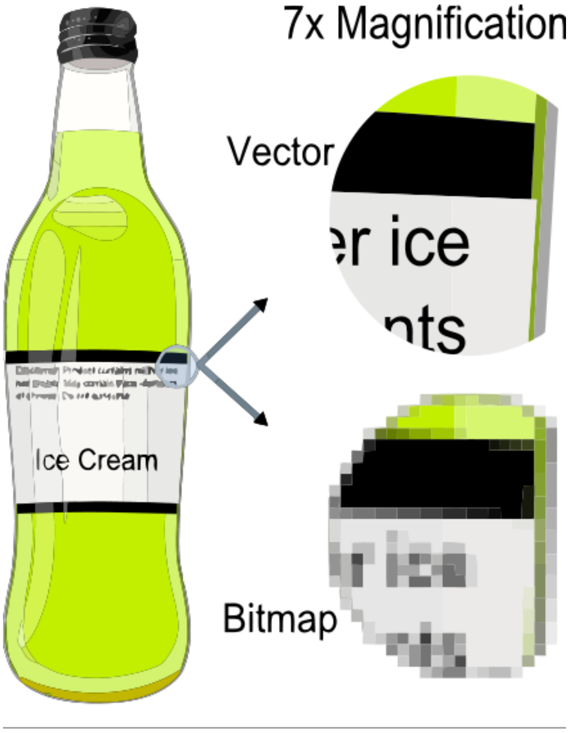


*Figure 5: Illustration of difference between vector image and raster image (bitmap) [26].*

Vector images are organised in a similar fashion as raster images, with the difference that most vector formats consist of fewer structure types than what raster can accumulate to. Generally, vector images consist mainly of a header, image data and an End of File (EOF) marker, as illustrated in Figure 6, but can also be scaled to include a palette and footer.

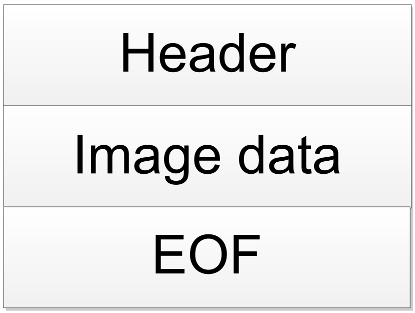


*Figure 6: A common composition of a vector image.*

The file size of a vector image is proportionate to the number of elements it contains, and can so grow to be quite large in file size in proportion to its visible image size. Raster images, on the other hand, maintains its file size regardless of complexity, but varies based on the amount of pixels, and the compression available to the file creator.

For images of photographs, vector images are at a disadvantage to raster images in that they cannot be used to store extremely complex images, where colour information is paramount and may vary on a pixel by pixel basis.

## 2.3. The most commonly used image file formats today

This thesis will focus on the most common image file formats found on the internet and on personal computers today. The starting point for the selection of file formats was a review of how the different web based services which give users a platform to share their images with others, chose their supported file types. The web services included in the initial selection were Flickr [27], Picasa [28], PhotoBucket [29] and Facebook [30]. All four services allow uploading of images stored in JPEG, GIF, PNG or TIFF file formats. Some allow a few other formats such as BMP, PSD (Photoshop), TGA and selected RAW formats as well. For the most part it was observed that BMPs are converted to JPEG on upload, and Facebook converts all images uploaded to JPEG format.

The images stored in a large amount of publically shared accounts were enumerated, and there was a clear tendency towards users sharing their images stored in JPEG format. There were also some occurrences of images stored in GIF, PNG and TIFF, but in a much lesser extent than JPEG.

Based on these findings, the image formats in focus in this thesis are JPEG, GIF, PNG and TIFF. This is because these four file formats combined, based on the findings, represent the largest group of image file types used for personal and business storage today. The image file types in focus are all categorized as raster file formats, or bitmaps. Common properties of raster images are presented in chapter *2.2.1 Raster images*. However, JPEG files will be given extra attention, due to the wide use of this particular format in regards to the other top four file formats.

Whether or not to include PDFs with pixel graphics was considered, but was chosen to omit in this thesis due to practical limitations in the tools chosen for testing, and because there is already work being done on this field in parallel to the writing of this thesis.

As this thesis will focus mainly on the image metadata rather than the physical image itself, the details of the differences in how the image is stored will just briefly be discussed, while the main focus lies in how such metadata is handled and stored.

### 2.3.1. JPEG

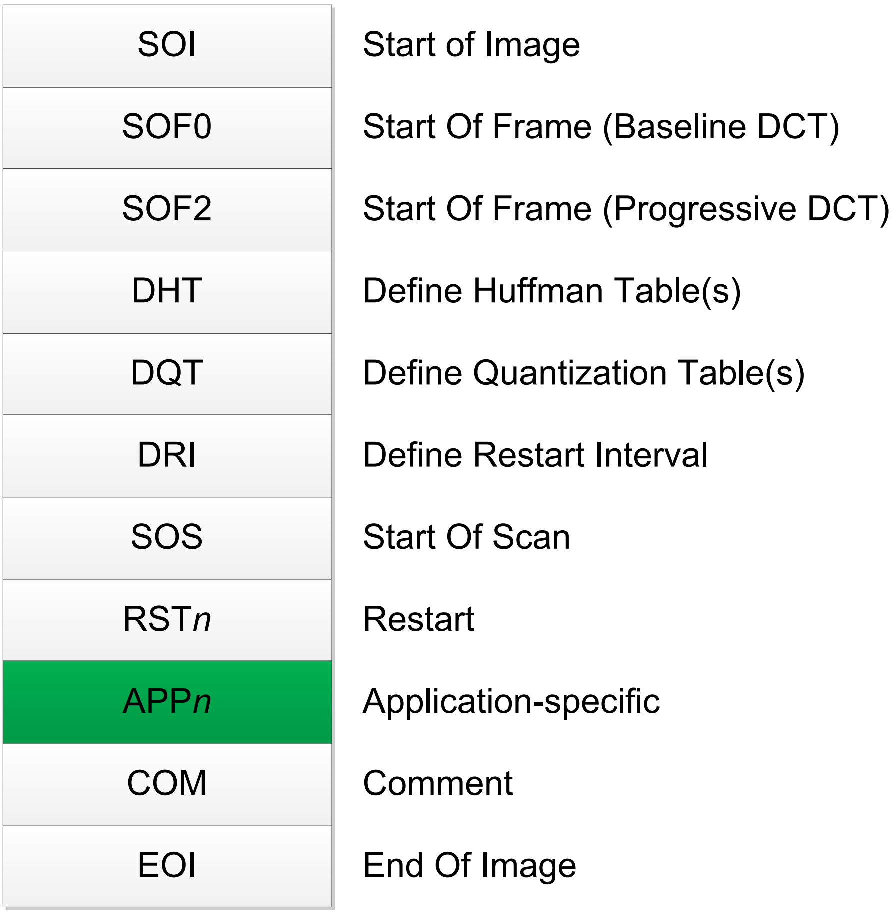
The acronym JPEG stands for the Joint Photographic Experts Group, which is a standards committee that had its origins within the International Standard Organisation (ISO). JPEG images supports up to 24 bits colour depth. The JPEG format uses a lossy compression method which can greatly reduce the file size needed to store the image. However, JPEGs are known to generate some visible distorted artefacts if compressed too much. The degree of compression can be adjusted, to allow a trade-off between storage size and image quality. What is today commonly known simply as JPEG, consists of a subset of different formats, where the two most common are JPEG/Exif and JPEG/JFIF.

*The JPEG standard specifies the codec, which defines how an image is compressed into a stream of bytes and decompressed back into an image, but not the file format used to contain that stream. The Exif and JFIF standards define the commonly used formats for interchange of JPEG-compressed images* [31].

The JPEG compression algorithm produce best results on photographs with smooth transitions in tone and colour, and produce the most visible distorted artefacts when used on line drawings and textual graphics where sharp contrasts in adjacent pixels have a great visual influence on the perceived image. Because of its great compression algorithm, an image can generally be compressed to about a tenth of its uncompressed size without showing signs of visible distorted artefacts. This makes the file format ideal for sharing on the internet in comparison to other image formats, and is one of the reasons why it is so widely used today. The fact that most digital cameras today store images in the JPEG/Exif format further improves its standing as the most commonly used file format for sharing images.

The JPEG/JFIF format is specified in Annex B of the JPEG standard [32], but was not as widespread in use due to some shortcomings in the standard, which result in some challenges with implementation of encoders and decoders. In an attempt to remedy this, several other standards have emerged, whereas JPEG/Exif is the most widely used today. The Exif (Exchangeable Image File) format is described in more detail in chapter *2.4.1 Exif metadata description*, but the JPEG/Exif format is based on the actual byte layout of the JFIF format, but where JFIF mainly uses application marker APP0, Exif mainly uses APP1, and thus the formats are used today in a close to compliant way to the original standard. Extensible Metadata Platform (XMP) information, which is another standard for storing metadata, can also be embedded within the APP1 segment in place of Exif metadata.

A JPEG image consists of a sequence of segments, each beginning with a marker. Figure 7 shows common JPEG markers. The most interesting marker for this thesis is the APP*n* marker (APP1), which is where the Exif metadata is stored.



| | |
|---|---|
| SOI | Start of Image |
| SOF0 | Start Of Frame (Baseline DCT) |
| SOF2 | Start Of Frame (Progressive DCT) |
| DHT | Define Huffman Table(s) |
| DQT | Define Quantization Table(s) |
| DRI | Define Restart Interval |
| SOS | Start Of Scan |
| RST*n* | Restart |
| APP*n* | Application-specific |
| COM | Comment |
| EOI | End Of Image |

*Figure 7: Common JPEG markers [32].*

### 2.3.2. GIF

The acronym GIF stands for Graphics Interchange Format and was originally created by CompuServe Inc. GIF images supports 1 bit up to 8 bits colour depth. The GIF format is used to store multiple bitmap images in a single file for exchange between platforms and systems. Due to its wide support of a multitude of systems and its portability, the GIF format was one of the most common image file formats in the internet's early days. The first version of the GIF format was called 87a. Today, GIFs are mainly used for small animated image sequences using the GIF 89a version, and PNG has to a large extent taken over the other main areas of usage for GIFs.

GIF images are compressed using the Lempel-Ziv-Welch (LZW) [33] lossless data compression technique to reduce the file size without degrading the visual quality. The compression technique was patented before CompuServe released the GIF format, but later issues with controversy over the license agreement were what generated the development of the PNG format, as an open alternative.

GIF images in 87a revision do not support additional metadata, as the file format has no segments in which to store such additional information. The feature to store application specific metadata arrived in revision 89a, along with features such as animation delays and transparent background colours. Although GIF 89a has some extension information, Exif metadata is not supported. In the application extension field, application specific information can be stored. Extensible Metadata Platform (XMP) information can be embedded within this field.

Due to the colour space of the GIF format, it is not as well suited for photographs as JPEG, PNG or TIFF, but it is more than sufficient to store logos or low resolution animations, as is seen by its use today. Figure 8 shows the layout of a GIF 89a file.

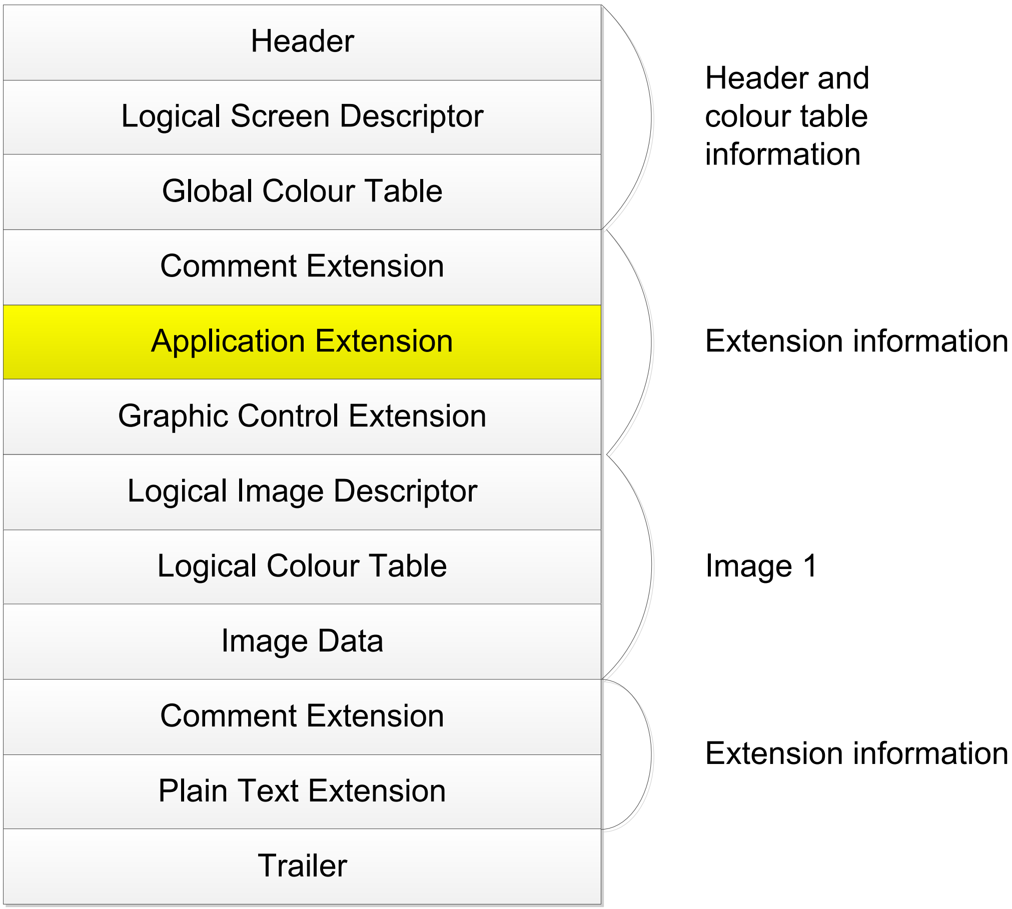


*Figure 8: Layout of a GIF 89a file [24].*

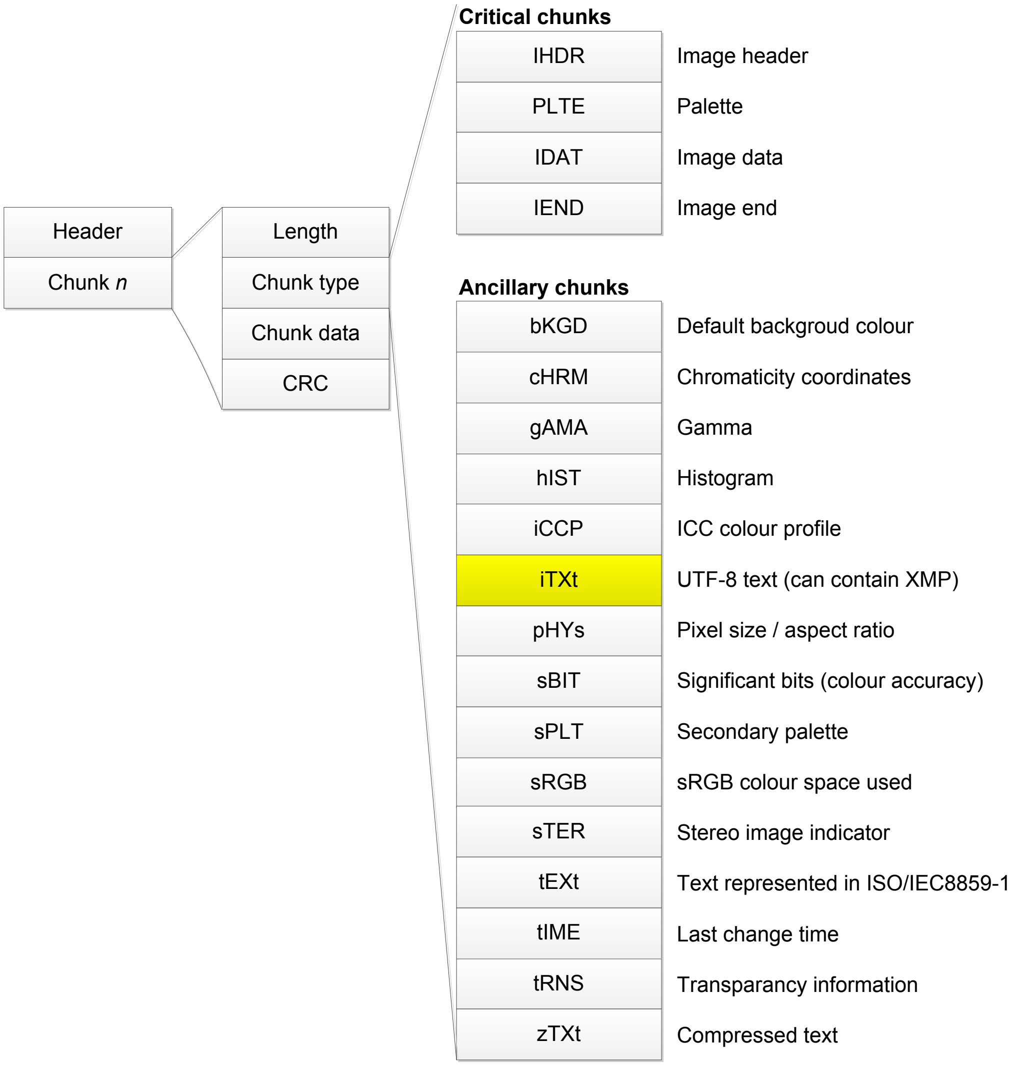
### 2.3.3. PNG

The acronym PNG stands for Portable Network Graphics. PNG images supports up to 48 bits colour depth. PNG was mainly developed to improve upon and replace the GIF format, due to patent issues associated with the compression method used in GIFs. As such, PNGs are designed for transferring images on the internet, rather than professional grade photographs for use in printed material. The PNG format is an improvement on all of the original qualities of GIFs, except that PNGs do not support animation. A later unofficial extension to the format called MNG (Multiple-Image Network Graphics) makes animation possible, but contains such vast extensions that most regular PNG decoders are not able to render them.

PNG uses a non-patented lossless data compression method known as DEFLATE [34], which is the same algorithm used in the zlib compression library. Although PNGs can be significantly larger in file size than GIFs, this is mainly due to the increase in colour depth, however if an image is stored as a GIF and a PNG from the same source with the same colour depth, PNGs are usually smaller in file size.

Some of the main strengths of the PNG format are that it can accommodate more transparency options than GIF, and that PNGs for the most part have a more efficient compression algorithm than GIFs and is without the patent issues. Although efficient compression is obtained on lines, gradients and curves, it is not as efficient on photographs as JPEG. This leads to a much larger file size when used on photographs.

The PNG format consists of a header and a series of chunks containing information. The chunks are divided into critical chunks and ancillary chunks. The critical chunks specify sections which are obligatory for the successful rendering of the image, however the ancillary chunks may be omitted, and the image will still render. PNG images do not support a standard means of embedding Exif metadata; however Extensible Metadata Platform (XMP) information can be embedded within the iTXt chunk. Figure 9 shows the main components of a PNG image.



Critical chunks

| | |
|---|---|
| IHDR | Image header |
| PLTE | Palette |
| IDAT | Image data |
| IEND | Image end |

Ancillary chunks

| | |
|---|---|
| bKGD | Default backgroud colour |
| cHRM | Chromaticity coordinates |
| gAMA | Gamma |
| hIST | Histogram |
| iCCP | ICC colour profile |
| iTXt | UTF-8 text (can contain XMP) |
| pHYs | Pixel size / aspect ratio |
| sBIT | Significant bits (colour accuracy) |
| sPLT | Secondary palette |
| sRGB | sRGB colour space used |
| sTER | Stereo image indicator |
| tEXt | Text represented in ISO/IEC8859-1 |
| tIME | Last change time |
| tRNS | Transparancy information |
| zTXt | Compressed text |

*Figure 9: Main components of a PNG image.*

### 2.3.4. TIFF

The acronym TIFF stands for Tagged Image File Format. TIFF images supports up to 24 bits colour depth. The file format was originally released as a standard method of storing black and white images created by scanners and desktop publishing applications. TIFF is a very flexible file format, and can be used to store either lossy compressed images, or lossless images.

TIFF files are organised into three main sections; the image file header (IFH), the image file directory (IFD) and the bitmap data. Within the TIFF image, the sections can vary in position, and only the header has a fixed place at the very beginning of the file. There are also a multitude of different fields or tags, and very few existing rendering software have implemented support for all of the tags. For the user this can cause a lot of confusion when handling images stored as TIFFs. The format can incorporate a variety of different compression algorithms as well. This makes implementing support for all aspects of the format a huge task. Figure 10 shows three possible arrangements of the components of a TIFF file.

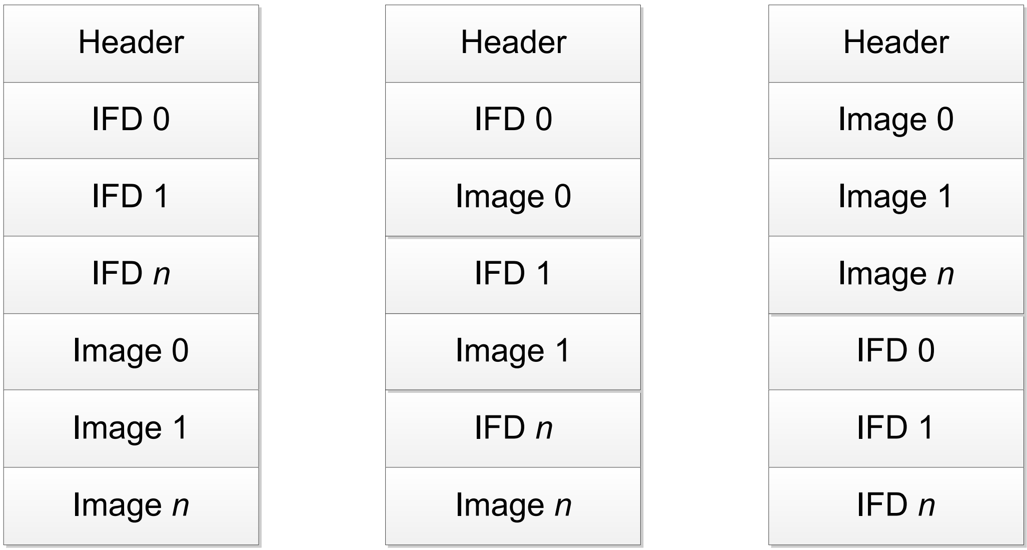


*Figure 10: Three possible physical arrangements of data in a TIFF file [24].*

Each IFD (Image File Directory) can contain a number of data records called tags, containing information. There is a long list of supported tag types, however for the purpose of this thesis, it is sufficient to mention the following specific tags:

- Tag id 700 (0x02BC), which is the tag that contains XMP metadata
- Tag id 34665 (0x8769) which contains the Exif specific TIFF tag
- Tag id 34853 (0x8825) which contains the Global Positioning System (GPS) Exif information, if applicable
- Tag id 40965 (0xA005) which contains Exif related interoperability IFD.

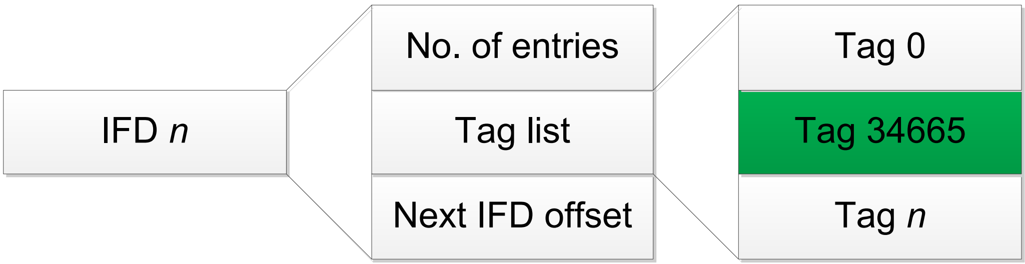


*Figure 11: Structure of TIFF Image File Directory (IFD).*

## 2.4. Image metadata

The National Information Standards Organization (NISO) [35] describes metadata as *structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information* [36]. Through the metadata contained within digital images, information about when the image was taken, the photographer who took the picture, the equipment used to take the picture and the settings this equipment was set up with including camera serial number and lens type used, the location of where the picture was taken, whether or not the flash was used, and a variety of other information can be stored.

Before the image metadata standards were common, a lot of the different image software kept metadata records in their own proprietary format which could only be read by the software that wrote the metadata in the first place. This often lead to loss of the metadata when sharing files and using different software to work with the images.

Image metadata can be stored in a variety of different ways within the file itself, and there are several standards for storing and structuring the metadata. The most common metadata formats are Exif (Exchangeable Image File Format), XMP (Extensible Metadata Platform), and IPTC (International Press Telecommunications Council).

Although all of the image file formats described in chapter *2.3 The most commonly used image file formats today* can accommodate XMP information, whilst only JPEGs and TIFFs can accommodate Exif information, this thesis will focus mainly on Exif information. This is because most digital cameras today store Exif information when a picture is being taken [37], and this format is readable by most image software today. XMP is not supported by most camera models on the market today, and as such, XMP information is only added manually by the user during processing of the image, if added at all. This makes Exif information more widely available, and thus more interesting to study further.

### 2.4.1. Exif metadata description

Exchangeable image file format (Exif) was created by the Japan Electronic Industries Development Association (JEIDA) [38], and is a format to describe metadata. Images can have more or less metadata attached to them. Due to the fact that Exif information stores information about the image, this information can pose a privacy issue, especially concerning information describing time and date a picture was taken, the location the picture was taken and the original unedited thumbnail of the picture.

The Exif file structure is tag based in the same way as TIFF files. This makes the format very versatile. This also means that the Exif information can be stored sequentially in physically different places in the image file itself using offset pointers, inevitably risking being overwritten and lost by image processing software.

When Exif information is stored in a TIFF image, the basic structure is as described in chapter *2.3.4 TIFF* and illustrated in the Exif 2.2 standard, presented in Figure 12.

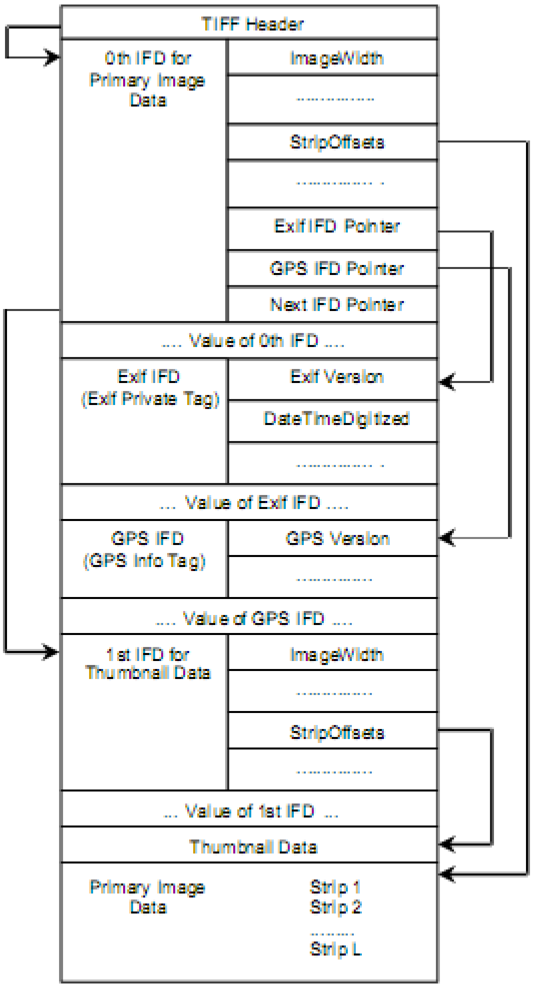


*Figure 12: Basic structure of uncompressed data files[39].*

When Exif information is stored in a JPEG image, the basic structure is still contained within a TIFF structure as described in chapter *2.3.4 TIFF*, however the TIFF structure is located within the APP1 section of the JPEG as described in chapter *2.3.1 JPEG*, and illustrated in the Exif 2.2 standard, presented in Figure 13. The Exif information attached to JPEGs are limited by the size of the APP1 segment, which has a size limit of 64 Kbytes. However, extended Exif information for Flashpix extension data is possible within multiple APP2 segments, but this is not as widespread in use as the basic Exif information.

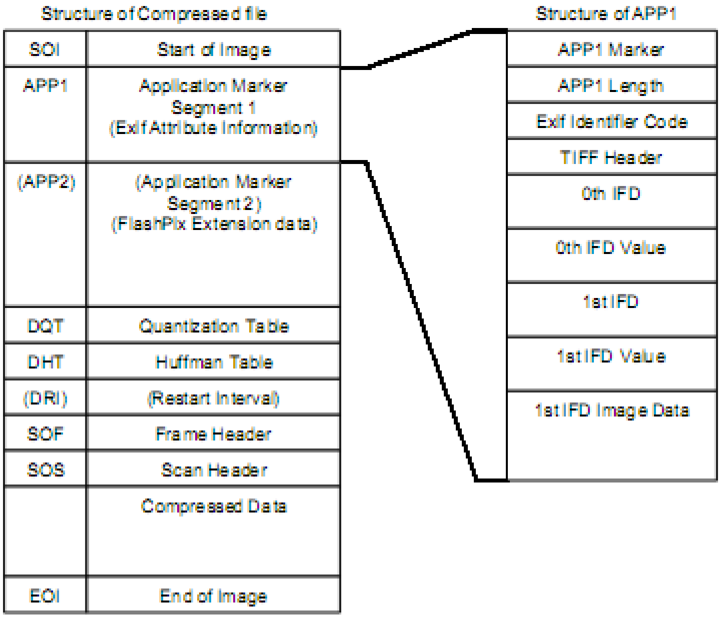


*Figure 13: Basic structure of compressed data files [39].*

The following example, found in Table 1, shows Exif data for a typical photo taken with a mobile device, and is meant to illustrate the type of information which can be derived from the metadata associated with the photo.

*Table 1: Example of Exif information contained in a photo.*

| Tag | Value |
|---|---|
| ImageDescription | SAMSUNG |
| Make | SAMSUNG |
| Model | GT-I9000 |
| Orientation | Horizontal (normal) |
| XResolution | 72 |
| YResolution | 72 |
| ResolutionUnit | inches |
| Software | fw 05.15 prm 07.53 |
| ModifyDate | 2010:07:17 19:02:33 |
| YCbCrPositioning | Centered |
| ExposureTime | 1/142 |
| FNumber | 02.jun |
| ExposureProgram | Program AE |
| ISO | 50 |
| ExifVersion | 220 |
| DateTimeOriginal | 2010:07:17 19:02:33 |
| CreateDate | 2010:07:17 19:02:33 |

| | |
|---|---|
| ComponentsConfiguration | Y, Cb, Cr, - |
| ShutterSpeedValue | 1/142 |
| ApertureValue | 02.jun |
| BrightnessValue | 06.jun |
| ExposureCompensation | 0 |
| MaxApertureValue | 02.jun |
| MeteringMode | Center-weighted average |
| LightSource | Unknown |
| Flash | No flash function |
| FocalLength | 3.8 mm |
| FlashpixVersion | 100 |
| ColorSpace | sRGB |
| ExifImageWidth | 2560 |
| ExifImageHeight | 1920 |
| InteropIndex | R98 - DCF basic file (sRGB) |
| InteropVersion | 100 |
| SensingMethod | One-chip color area |
| SceneType | Directly photographed |
| ExposureMode | Auto |
| WhiteBalance | Auto |
| DigitalZoomRatio | undef |
| FocalLengthIn35mmFormat | 0 mm |
| SceneCaptureType | Standard |
| Contrast | Normal |
| Saturation | Normal |
| Sharpness | Normal |
| GPSVersionID | 2.2.0.0 |
| GPSLatitudeRef | North |
| GPSLatitude | 0°0.0000 |
| GPSLongitudeRef | East |
| GPSLongitude | 0°0.0000 |
| GPSAltitudeRef | Above Sea Level |
| GPSAltitude | 0 m |
| Compression | JPEG (old-style) |
| ThumbnailOffset | 1272 |
| ThumbnailLength | 8986 |

## 2.5. The Cloud

Today's users demand access to their data from anywhere and at any time, resulting in a migration from local storage devices to cloud based storage services. The term cloud, or cloud computing, refers to the current trend of gathering more and more services on servers, typically hosted by third parties, on the internet ensuring accessibility and availability from any location and usually through a multitude of different technologies such as laptops, desktops and mobile devices.

### 2.5.1. The NIST definition of cloud computing

The National Institute of Standards and Technology (NIST) [14] defines cloud computing as:

> *a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models [40].*

***Essential Characteristics***

***On-demand self-service:*** *A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.*

***Broad network access:*** *Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).*

***Resource pooling:*** *The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacentre). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.*

***Rapid elasticity:*** *Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out, and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.*

***Measured Service:*** *Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.*

### *Service Models*

***Cloud Software as a Service (SaaS):*** *The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.*

***Cloud Platform as a Service (PaaS):*** *The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.*

***Cloud Infrastructure as a Service (IaaS):*** *The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).*

### *Deployment Models*

***Private cloud:*** *The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.*

***Community cloud:*** *The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.*

***Public cloud:*** *The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.*

***Hybrid cloud:*** *The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).*

### 2.5.2. Cloud services

There exists a vast amount of cloud services on the internet today, and more are made available every day. This thesis will look closer at three different cloud based services for storing and sharing images on the internet. These services are Windows Live SkyDrive [41], Windows Azure Platform – Storage service [42] and Flickr [27]. This section will give a brief introduction to these services.

### Windows Live SkyDrive

Windows Live SkyDrive [41] is offered through Microsoft's Windows Live concept, which is a collection of web based – or web integrated – services including Hotmail, Messenger, Essentials, Office and SkyDrive, and is a common portal for storing and sharing information through the internet. SkyDrive is a universal storage and synchronisation element of the Windows Live suite, and can be used to store photos, documents or any files a user wishes to make available across a multitude of platforms and locations. Through the Windows Live Photos service, which is part of the SkyDrive functionality, users can store and share images through a web based interface. The user gets a certain amount of storage space (25GB) in which to upload and share images with the world or a restricted audience. The service also lets users share the associated Exif metadata.

### Windows Azure Platform – Storage

The Windows Azure Platform [42] is a web based operating system environment which offers three main components: Compute, Storage and Fabric. The storage service offers three main types of storage methods: Binary Large Object (blob) storage, table storage and queue storage [43]. The blob storage is the preferred method of storing any binary files within the Windows Azure Platform, and it consists of a varying amount of containers, which in turn can contain a varying amount of blobs. The blobs are simple raw byte arrays. The containers can be chosen to be shared or private, and could therefore be used to share files over the internet. Figure 14 aims to illustrate these principles.

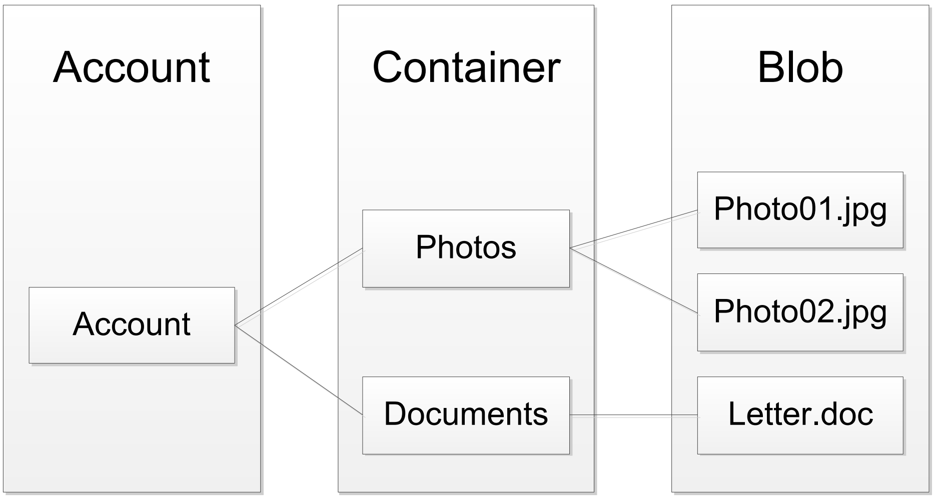


*Figure 14: Windows Azure Platform Storage service blob storage principles[44].*

### Flickr

Flickr [27] is an online service for image and video storage and sharing that encourages the user to share their photos and update the story behind the images – generate metadata. Flickr allows for easy integration with other web services, and is widely used to host images for other web services such as Facebook [30] or various blogs, in addition to other Flickr users. Through Flickr's web interface, the associated Exif metadata can be viewed, if the user has chosen to share this information along with the image itself.

# 3. Method approach and experimental design

## 3.1. Description of model/approach

Josh Brunty with Digital Forensic Investigator News has summarised the main considerations for validation of forensics tools and software in the article: "Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner" [8], partially reproduced below:

> *According to the National Institute of Standards and Technology (NIST), test results of the testing must be repeatable and reproducible to be considered admissible as electronic evidence. Digital forensics test results are repeatable when the same results are obtained using the same methods in the same testing environment. Digital forensics test results are reproducible when the same test results are obtained using the same method in a different testing environment (different mobile phone, hard drive, and so on). NIST specifically defines these terms as follows:*
>
> ***Repeatability*** *refers to obtaining the same results when using the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time.*
>
> ***Reproducibility*** *refers to obtaining the same results being obtained when using the same method on identical test items in different laboratories with different operators utilizing different equipment.*

In the Daubert ruling [5], described in chapter *1.1.1 Daubert criteria of admissibility*, The Court defined scientific methodology as *"the process of formulating hypotheses and then conducting experiments to prove or falsify the hypothesis."* The Scientific Method refers to a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge. To be termed scientific, the method must be based on gathering, observing, or investigating, and showing measurable and repeatable results. Most of the time, the scientific process starts with a simple question that leads to a hypothesis, which then leads to experimentation, and an ultimate conclusion.

With Daubert and NIST as premise providers for the experimental setup, a hypothesis test can be used in the experiment. A hypothesis test is a procedure that summarises data making it possible to detect differences among groups, and is used to make comparisons between two or more groups [45].

The starting point is the null hypothesis ($H_0$), and an alternate hypothesis ($H_a$). The null hypothesis is the least radical state, where there is no difference between groups, whereas the alternate hypothesis can be that the groups are different.

The approach suggested in this thesis uses quasi-experimental methodology as its scientific foundation. More precisely a modified version of a nonrandomised control group pretest-posttest design [46]. To fulfil the criteria of being qualified as "true experimental design", there should be random assignment to groups. However, in the case presented in this thesis, random assignment would be unpractical because mixing the real world images with the synthetically generated images would not provide any additional value. Instead the synthetically generated images uploaded serve as the control set to the real world images. The method itself should include a control set with no treatment (Tx); however in the example of uploading images to cloud services, it is not really practically useful to have a control set with no treatment (Tx), as it would yield no difference from simply observing the metadata before uploading to the cloud services.

Using a quasi-experimental design does require an added focus to other alternate explanations to the results, as this method cannot completely rule out other explanations. In the test setup described below, the images within the two different groups, $\text{Group}_{\text{Synthetically\_generated}}$ and $\text{Group}_{\text{Real\_World}}$, are first observed (Obs) before being uploaded to any of the cloud based storage services, where a treatment (Tx) is being introduced, before a new observation is made to determine the effect of the different cloud based storage services. This design makes alternate explanations than what is being introduced by the different cloud based storage services unlikely. The complete method can then be represented as described by Table 2.

*Table 2: Test methodology: a modified version of a nonrandomised control group pretest-posttest design.*

| Group | Time → | | |
|---|---|---|---|
| **$\text{Group}_{\text{Synthetically\_generated}}$** | Obs | $\text{Tx}_{\text{SkyDrive}}$ | Obs |
| **$\text{Group}_{\text{Synthetically\_generated}}$** | Obs | $\text{Tx}_{\text{Azure}}$ | Obs |
| **$\text{Group}_{\text{Synthetically\_generated}}$** | Obs | $\text{Tx}_{\text{Flickr}}$ | Obs |
| **$\text{Group}_{\text{Real\_World}}$** | Obs | $\text{Tx}_{\text{SkyDrive}}$ | Obs |
| **$\text{Group}_{\text{Real\_World}}$** | Obs | $\text{Tx}_{\text{Azure}}$ | Obs |
| **$\text{Group}_{\text{Real\_World}}$** | Obs | $\text{Tx}_{\text{Flickr}}$ | Obs |

The entire experiment can, for each image file and each cloud based storage service tested, be described by the steps given in Figure 15.

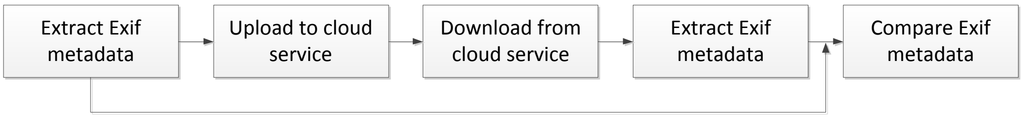


*Figure 15: Block diagram of method described to test metadata in the cloud.*

Dual tool verification will be utilised when extracting the Exif metadata, to eliminate the tool used as a potential error source.

Leedy et. al. [46] describes internal validity of a research study as:

> *The extent to which its design and the data it yields allow the researched to draw accurate conclusions about cause-and-effect and other relationships within the data.*

In order to ensure that the internal validity of the proposed method is as high as possible, the experiments are conducted as controlled laboratory studies such that only the measured variables – what happens when images are uploaded to the different cloud based storage services – are monitored, and have a chance to impact the results.

External validity is described by Leedy et. al. [46] as follows:

> *The external validity of a research study is the extent to which its results apply to situations beyond the study itself – in other words, the extent to which the conclusions drawn can be generalized to other contexts.*

To enhance the external validity of the research presented in this thesis, in addition to carrying out the experiments in a controlled lab environment using synthetically generated image files with controlled metadata attached, a real world sample will also be used. This will help simulate a real-life setting. This is further done by using a representative sample of JPEG image files with varying amounts of Exif metadata attached, found in “the wild” on the internet.

## 3.2. Experiment setup

For the first segment of the testing phase, it is desirable to identify what kind of information people share freely (or unwittingly) on the Internet, contained in Exif metadata. Specifically, analysing Exif metadata in raster images found on local computers and stored using cloud services, using a selection of different forensics tools. In the experiments the main focus will be on JPEG images, as this turns out to be the most widely used image format for storing photography images today, as discussed in chapter *2.3 The most commonly used image file formats today*.

To be able to answer this question, a selective sample of shared images has to be acquired. To get the sample size of images needed, a confidence level of 95% is chosen and taken into account with a confidence interval of 5%. Given that the possible population of image files shared on the internet is inconceivably large, but that the population really only is relevant for relatively small populations it really does not matter. As an example, a population size of 20,000, gives a sample size of 377 images. If the population is changed to 900,000,000, this gives a sample size of 384. With these numbers taken into account, a total sample size of 400 images of each tested file type was chosen, and will give a representative selection of the images shared on the internet [46].

To actually get images with a representative amount of metadata attached, different web based services were considered as sources. This was done because a wide variety of cameras and mobile devices were unavailable during the testing phase of this thesis. The main alternatives considered were Flickr [27], Picasa [28], PhotoBucket [29] and Facebook [30], as all the services are widely in use and contain images shared by users from all around the world. As described in chapter *2.3 The most commonly used image file formats today*; All four services allow uploading of images stored in JPEG, GIF, PNG or TIFF file formats, but not all of the services allow downloading the original unaltered uploaded image.

Out of the four web services considered, Flickr was chosen as the main source of image files. None of the services has a feature to easily download multiple images of a given quality or sorted by file type. The services all do some alterations of the metadata provided, as described in chapter *4.1.2 Behaviour after storage in cloud services*; however where the users have not explicably kept the original uploaded image private, the original metadata is retained. As with most of the services, the uploaded images are resized into different sizes and in varying degree has their original metadata stripped off, or even altered. This does not apply to the original uploaded image in Flickr, where the original metadata is retained. Flickr also has image files stored in JPEG, GIF and PNG formats available for download, and with the help of a third party bulk-downloader tool called "portable Flicka" [47], came to be the most time effective solution to get the required amount of image files for the experiments.

An observation made was that TIFF files were not as easily obtained as the other formats, due to limitations in searching options in all of the selected test sites. However, an image database from an office environment with TIFFs was made available for use in the research, and was used for general TIFF control tests. The images were provided under a non-disclosure agreement, and were thus not part of the testing through the cloud services, to avoid infringing this agreement.

For each of the acquired JPEG images, the associated metadata was extracted using the selected tools. Multiple tools were used to minimise possible errors with one piece of software, and is in accordance with the dual tool verification principle in digital forensics. The test is done to see what types of information is freely available, and to answer the first research question.

To verify that neither Flickr nor "portable Flicka" introduces any error, synthetically generated test images with varying amounts of metadata are uploaded to the service, and then re-downloaded using "portable Flicka" and compared to the original image to check for alterations. A control data set with information in all Exif fields will be used as a baseline test to verify that this is true for all Exif fields.

The most interesting information that will be searched for are potentially identifying information like GPS coordinates, model- or serial numbers for the camera used, the type and make of camera used, and so on.

The real world data set consisted of 400 JPEG images containing Exif information, downloaded from the internet. Analysis of the images revealed that the images varied both in physical/pixel size and with the amount of Exif information associated with each image, giving a varied selection.

The second experiment performed is to see what metadata information is altered or lost after sharing images through cloud based storage services. The main focus of these tests is to see if any Exif information is altered or lost. This test is further limited to JPEGs only, due to the earlier findings where JPEGs turned out to be the most widely used format for storing images on the internet today.

For this test, a selection of different cloud based storage services was considered. The cloud storage services provided by SkyDrive [41], the Windows Azure Platform [42], Flickr [27], Picasa [28], PhotoBucket [29], Amazon Cloud Drive [48], Google Docs [49] and Dropbox [50] were considered. For the experiment, the cloud storage service in Sky Drive, the Windows Azure Platform and Flickr was selected. This is mainly due to an observation made while testing the SkyDrive service and Flickr, and stumbling over the change in the image uploaded and the Exif information from before and after storage through the services. As a check to see if this was the case for all Microsoft cloud services, the Windows Azure Storage service was selected as a reference service.

The tests performed for this part of the experiment will be similar to the tests to make sure neither Flickr nor “portable Flicka” manipulated the images in any way; a control set of JPEG images with varying Exif information will be uploaded to the cloud services, and later re-downloaded and compared to the original photographs to see what sections were changed, if any.

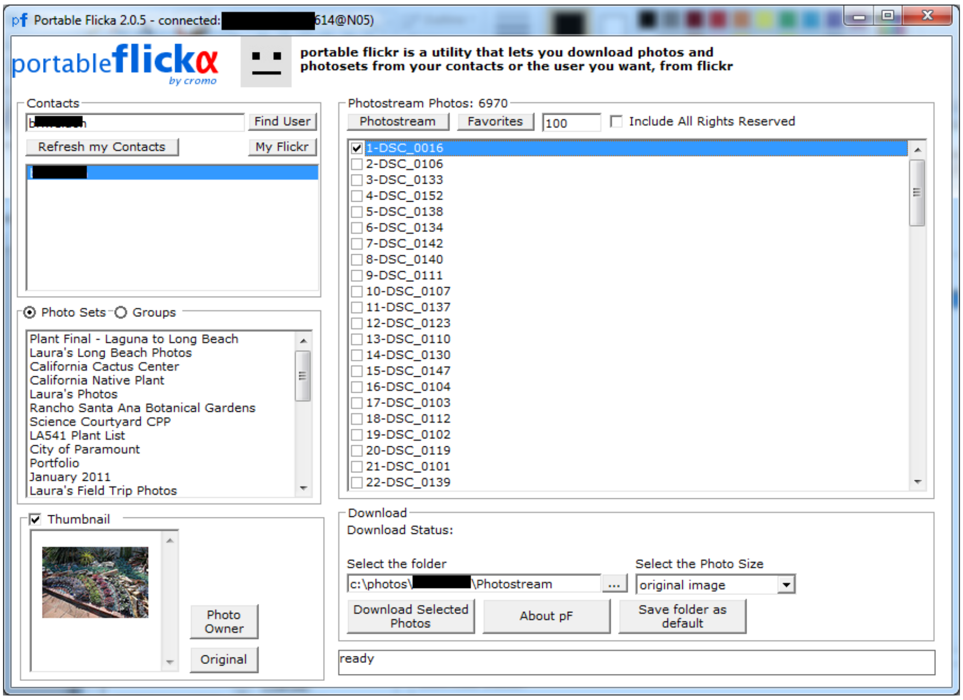
The tools selected to extract the metadata were ExifTool by Phil Harvey [51] and EnCase Forensic by Guidance Software [52]. These tools are introduced and explained in chapter *3.4 Tools used in experiments*, and were selected due to their high standing in the forensic community as tools for extracting Exif information.

## 3.3. Description of test environment

To keep possible interference from other sources to a minimum, the tests performed were carried out in a lab environment on a dedicated computer. For the experiments requiring processing through the cloud services, an internet connection was added.

The hardware used was an Acer Aspire Timeline X 5820TG, with a dual core Intel Core i5 processor 430M running at 2.26GHz [53], with 4GB RAM and 500GB hard drive. The computer had a clean install of 64bit Windows 7 Ultimate with Service pack 1 installed as its operating system.

To acquire the necessary image files, portable Flicka [47] version 2.0.5 was utilised.



Portable Flicka 2.0.5 - connected: [redacted] 614@N05)

portable flickr is a utility that lets you download photos and photosets from your contacts or the user you want, from flickr

Contacts: [redacted]

[redacted]

● Photo Sets ○ Groups

Plant Final - Laguna to Long Beach
Laura's Long Beach Photos
California Cactus Center
California Native Plant
Laura's Photos
Rancho Santa Ana Botanical Gardens
Science Courtyard CPP
LA541 Plant List
City of Paramount
Portfolio
January 2011
Laura's Field Trip Photos

☑ Thumbnail

Photostream Photos: 6970

100 ☐ Include All Rights Reserved

☑ 1-DSC_0016
☐ 2-DSC_0106
☐ 3-DSC_0133
☐ 4-DSC_0152
☐ 5-DSC_0138
☐ 6-DSC_0134
☐ 7-DSC_0142
☐ 8-DSC_0140
☐ 9-DSC_0111
☐ 10-DSC_0107
☐ 11-DSC_0137
☐ 12-DSC_0123
☐ 13-DSC_0110
☐ 14-DSC_0130
☐ 15-DSC_0147
☐ 16-DSC_0104
☐ 17-DSC_0103
☐ 18-DSC_0112
☐ 19-DSC_0102
☐ 20-DSC_0119
☐ 21-DSC_0101
☐ 22-DSC_0139

Download
Download Status:

Select the folder: c:\photos\[redacted]\Photostream

Select the Photo Size: original image

ready

*Figure 16: Screenshot of tool portable Flicka.*

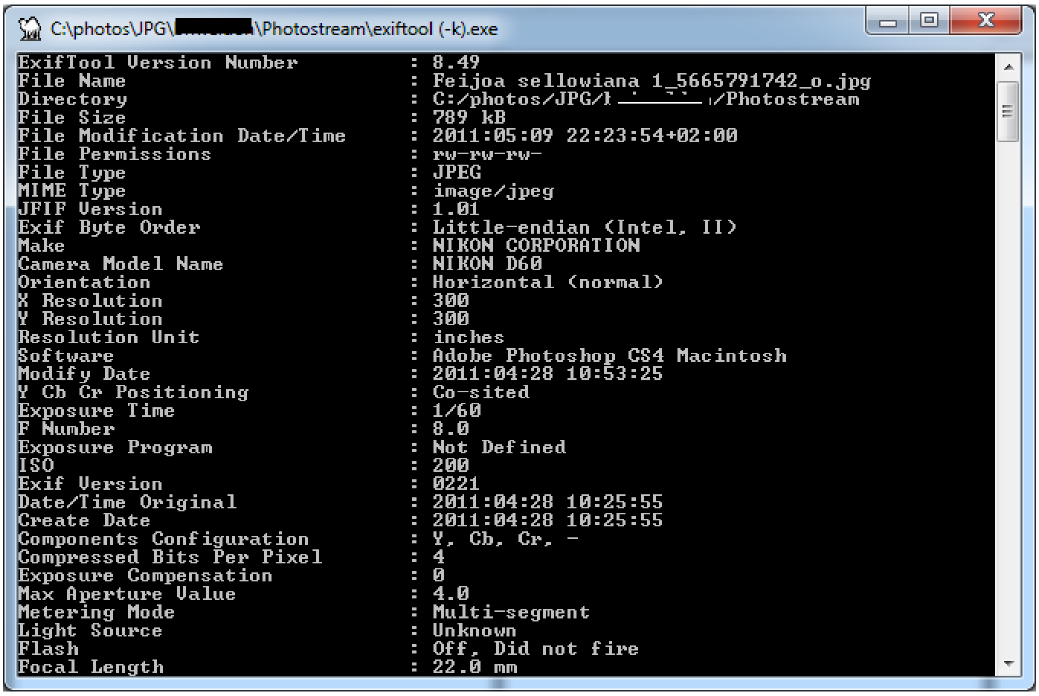
Portable Flicka uses the Flickr Application Programming Interface (API) to connect to a user's Flickr account, and enables bulk-downloading of images with a selected photo size. To be able to extract Exif information from the images, the original image was required, due to Flickr removing the metadata for all of its other image size alternatives, as described in *Flickr* chapter under *2.5.2 Cloud services*.

## 3.4. Tools used in experiments

This section will give a description of the forensics tools used in the experiments performed.

### 3.4.1. ExifTool

To extract Exif metadata, ExifTool by Phil Harvey [51] was used. The version utilised in the experiment was 8.49.



```
ExifTool Version Number         : 8.49
File Name                       : Feijoa sellowiana 1_5665791742_o.jpg
Directory                       : C:/photos/JPG/[redacted]/Photostream
File Size                       : 789 kB
File Modification Date/Time     : 2011:05:09 22:23:54+02:00
File Permissions                : rw-rw-rw-
File Type                       : JPEG
MIME Type                       : image/jpeg
JFIF Version                    : 1.01
Exif Byte Order                 : Little-endian (Intel, II)
Make                            : NIKON CORPORATION
Camera Model Name               : NIKON D60
Orientation                     : Horizontal (normal)
X Resolution                    : 300
Y Resolution                    : 300
Resolution Unit                 : inches
Software                        : Adobe Photoshop CS4 Macintosh
Modify Date                     : 2011:04:28 10:53:25
Y Cb Cr Positioning             : Co-sited
Exposure Time                   : 1/60
F Number                        : 8.0
Exposure Program                : Not Defined
ISO                             : 200
Exif Version                    : 0221
Date/Time Original              : 2011:04:28 10:25:55
Create Date                     : 2011:04:28 10:25:55
Components Configuration        : Y, Cb, Cr, -
Compressed Bits Per Pixel       : 4
Exposure Compensation           : 0
Max Aperture Value              : 4.0
Metering Mode                   : Multi-segment
Light Source                    : Unknown
Flash                           : Off, Did not fire
Focal Length                    : 22.0 mm
```

*Figure 17: Screenshot of tool ExifTool.*

*ExifTool is a platform-independent Perl library plus a command-line application for reading, writing and editing meta information in a wide variety of files. ExifTool supports many different metadata formats including EXIF, GPS, IPTC, XMP, JFIF, GeoTIFF, ICC Profile, Photoshop IRB, FlashPix, AFCP and ID3, as well as the maker notes of many digital cameras [51].* ExifTool is widely acknowledged in the digital forensic community as one of the leading tools when it comes to processing metadata information in different file types, and was a natural tool choice when handling Exif information.

In the experiment performed, the following commands were executed for the folder containing the image files:

*Table 3: exiftool command 1.*

| **Command** |
|---|
| *exiftool –H –w! txt –a –e –ee –u –U –r <folder containing image files>* |

| **Parameter** | | **Description** |
|---|---|---|
| -H | (-hex) | Show tag ID number in hexadecimal |
| -w[!] EXT | (-textOut) | Write output text files |
| -a | (-duplicates) | Allow duplicate tags to be extracted |
| -ee | (-extractEmbedded) | Extract information from embedded files |
| -u | (-unknown) | Extract unknown tags |
| -U | (-unknown2) | Extract unknown binary tags too |
| -r | (-recurse) | Recursively process subdirectories |

The command given in Table 3 will generate output files for all images contained within the folder, enabling parsing of the fields to see what information is present in the images. See the online manual-page for further details [54].

To get only the aggregated Exif information for all images contained in the subfolder where the JPEG images are stored, the command presented in Table 4 was issued:

*Table 4: exiftool command 2.*

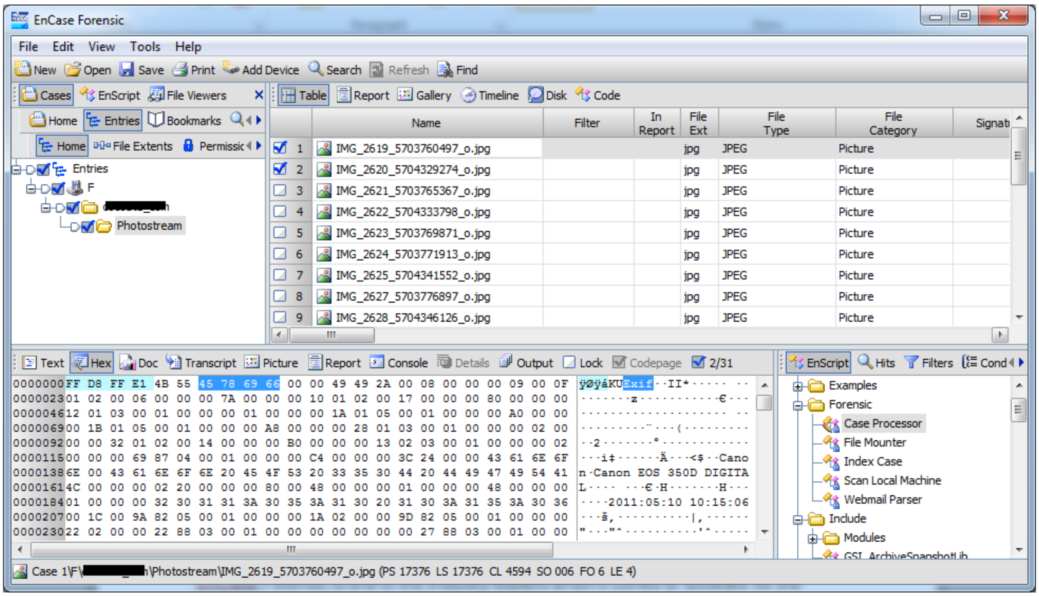
| **Command** |
|---|
| *exiftool -r -exif:all <folder containing image files> > <outputfile>* |

| **Parameter** | | **Description** |
|---|---|---|
| -r | (-recurse) | Recursively process subdirectories |
| -exif:all | | Extract all Exif information only |

This will extract only the Exif information contained in all images in the selected folder and any subfolders, and merge all information to one file, which can be further processed and easily searched.

### 3.4.2. EnCase Forensic

To validate that the extracted Exif metadata was complete, EnCase Forensic from Guidance Software [52] was used to verify the results gained by using ExifTool [51]. The version utilised in the experiment was 6.8.0.51.


EnCase Forensic
File Edit View Tools Help
New Open Save Print Add Device Search Refresh Find
Cases EnScript File Viewers
Home Entries Bookmarks
Home File Extents Permissic
Entries
F
[redacted]
Photostream
Table Report Gallery Timeline Disk Code
Text Hex Doc Transcript Picture Report Console Details Output Lock Codepage 2/31
EnScript Hits Filters Cond
Examples
Forensic
Case Processor
File Mounter
Index Case
Scan Local Machine
Webmail Parser
Include
Modules
GSI_ArchiveSnapshotLib

| | | Name | Filter | In Report | File Ext | File Type | File Category | Signat |
|---|---|---|---|---|---|---|---|---|
| ☑ | 1 | IMG_2619_5703760497_o.jpg | | | jpg | JPEG | Picture | |
| ☑ | 2 | IMG_2620_5704329274_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 3 | IMG_2621_5703765367_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 4 | IMG_2622_5704333798_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 5 | IMG_2623_5703769871_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 6 | IMG_2624_5703771913_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 7 | IMG_2625_5704341552_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 8 | IMG_2627_5703776897_o.jpg | | | jpg | JPEG | Picture | |
| ☐ | 9 | IMG_2628_5704346126_o.jpg | | | jpg | JPEG | Picture | |

```
0000000FF D8 FF E1 4B 55 45 78 69 66 00 00 49 49 2A 00 08 00 00 00 09 00 0F  ÿØÿáKUExif··II*······ ··
000002301 02 00 06 00 00 00 7A 00 00 00 10 01 02 00 17 00 00 00 80 00 00 00  ·······z···········€···
000004612 01 03 00 01 00 00 00 01 00 00 00 1A 01 05 00 01 00 00 00 A0 00 00  ······················ ··
000006900 1B 01 05 00 01 00 00 00 A8 00 00 00 28 01 03 00 01 00 00 00 02 00  ··········¨···(·········
000009200 00 32 01 02 00 14 00 00 00 B0 00 00 00 13 02 03 00 01 00 00 00 02  ··2·······°·············
000011500 00 00 69 87 04 00 01 00 00 00 C4 00 00 00 3C 24 00 00 43 61 6E 6F  ···i‡·····Ä···<$··Cano
00001386E 00 43 61 6E 6F 6E 20 45 4F 53 20 33 35 30 44 20 44 49 47 49 54 41  n·Canon EOS 350D DIGITA
00001614C 00 00 00 02 20 00 00 00 80 00 48 00 00 00 01 00 00 00 48 00 00 00  L···· ···€·H·······H···
000018401 00 00 00 32 30 31 31 3A 30 35 3A 31 30 20 31 30 3A 31 35 3A 30 36  ····2011:05:10 10:15:06
000020700 1C 00 9A 82 05 00 01 00 00 00 1A 02 00 00 9D 82 05 00 01 00 00 00  ···š,···········|,······
000023022 02 00 00 22 88 03 00 01 00 00 00 00 00 00 00 27 88 03 00 01 00 00  "···"ˆ···········'ˆ·····
```

Case 1\F\[redacted]\Photostream\IMG_2619_5703760497_o.jpg (PS 17376 LS 17376 CL 4594 SO 006 FO 6 LE 4)

*Figure 18: Screenshot of tool EnCase Forensic.*

EnCase Forensic is one of the industry leaders when it comes to software for the digital examiner. It is used in all stages of the evidence handling, from acquisition to analysis, and to reporting. Although containing a large amount of features, EnCase Forensics functionality can also be expanded by writing customised add-ons using the object oriented language EnScript. Different EnScripts can as an example be used to automate time consuming investigative tasks.

For JPEGs, EnCase Forensic has the built-in features needed, and the procedure for extracting the Exif metadata is as follows:

- Select “EXIF viewer” from “Case Processor”:
    - Located under the path “EnScript programs -> EnScript -> Forensic -> Case Processor”
    - Right click, and select “Run”
- In the case processor window, select a bookmark name, select the case number where the image files are stored and the export path where the result should be stored, and click “Next”.
- From the Modules-section, choose “EXIF Viewer”, and click “Finish”.
    - Located under the path “Modules -> Nodes -> File parsers”

This process will export all Exif information for all images found in data set to an excel workbook with separate sheets for each file.

# 4. Results, discussion and conclusion

## 4.1. Results

This section will provide the observations noted during the experiments performed.

### 4.1.1. General level of potentially sensitive information

When analysing the data gathered in the initial selection of JPEG images, the first step was to categorise the different Exif information contained in the population. The tool ExifTool was run to extract all Exif metadata contained in all images in the population. Based on the output, a mapping of what information was contained in the images was created. Table 6 shows the distribution of all identified Exif tags found within the population. Table 5 shows a sub-selection of Table 6 with only the most interesting Exif fields. The selected Exif fields shown in Table 5 are the fields which turned out to contain information which could be used to help identify the person taking the photo, the time and date of when the photo was taken, the make and model of the equipment used, location of where the photo was taken and comment fields possibly containing additional information regarding the image.

*Table 5: Distribution of interesting Exif fields found in population.*

| Exif field | Value contained in no. of images out of 400 | Value contained in % of population |
|---|---|---|
| Camera Model Name | 177 | 44,3 % |
| Date/Time Original | 177 | 44,3 % |
| Make | 166 | 41,5 % |
| Software | 148 | 37,0 % |
| Subject Distance Range | 63 | 15,8 % |
| Copyright | 29 | 7,3 % |
| Artist | 29 | 7,3 % |
| GPS Version ID | 28 | 7,0 % |
| Subject Distance | 23 | 5,8 % |
| Host Computer | 13 | 3,3 % |
| User Comment | 8 | 2,0 % |
| Image Description | 6 | 1,5 % |
| Lens Model | 2 | 0,5 % |

Remarkably enough, GPS latitude and longitude information, as contained in the example given in Table 1 in chapter *2.4.1 Exif metadata description*, was not discovered in any of the photographs in the initial selection. Neither was any serial number information of the equipment used.

What is seen, though, by these observations is that more than 40% of the images in the population contain information about the camera make and model used to take the pictures, and that the original time and date when the picture was taken is shared as frequently.

These observations are further discussed in chapter 4.2 Discussion and conclusion, specifically in chapter 4.2.1 General level of potentially sensitive information.

*Table 6: Distribution of all Exif fields found in population.*

| **Exif field** | **Value contained in no. of images out of 400** |
|---|---|
| Modify Date | 190 |
| Exif Version | 178 |
| F Number | 177 |
| Exposure Time | 177 |
| Create Date | 177 |
| Camera Model Name | 177 |
| Flash | 177 |
| Date/Time Original | 177 |
| Focal Length | 177 |
| Scene Capture Type | 176 |
| ISO | 176 |
| Exposure Mode | 176 |
| Custom Rendered | 175 |
| X Resolution | 171 |
| Resolution Unit | 171 |
| Y Resolution | 171 |
| Exposure Program | 167 |
| Make | 166 |
| Color Space | 159 |
| Exif Image Height | 159 |
| Exif Image Width | 159 |
| Orientation | 157 |
| Software | 148 |
| Thumbnail Length | 139 |
| Thumbnail Offset | 139 |
| Y Cb Cr Positioning | 136 |
| Flashpix Version | 135 |
| Components Configuration | 134 |
| Compression | 132 |
| Exposure Compensation | 128 |
| Metering Mode | 124 |
| Aperture Value | 116 |
| Shutter Speed Value | 116 |
| Max Aperture Value | 105 |
| File Source | 100 |
| Sensing Method | 92 |
| Focal Plane X Resolution | 90 |
| Focal Plane Resolution Unit | 90 |
| Focal Plane Y Resolution | 90 |
| Interoperability Version | 87 |
| Scene Type | 85 |
| Interoperability Index | 84 |
| Focal Length In 35mm Format | 78 |
| Sharpness | 66 |
| Sub Sec Time Digitized | 65 |
| Sub Sec Time Original | 65 |
| Digital Zoom Ratio | 65 |
| Light Source | 65 |
| Compressed Bits Per Pixel | 64 |
| Subject Distance Range | 63 |
| Saturation | 61 |
| Contrast | 60 |
| Sub Sec Time | 56 |
| Gain Control | 56 |
| Sharpness | 66 |
| Sub Sec Time Digitized | 65 |
| Sub Sec Time Original | 65 |
| Digital Zoom Ratio | 65 |
| Light Source | 65 |
| Compressed Bits Per Pixel | 64 |
| Subject Distance Range | 63 |
| Saturation | 61 |
| Contrast | 60 |
| Sub Sec Time | 56 |
| Gain Control | 56 |
| White Balance | 53 |
| Copyright | 29 |
| Artist | 29 |
| CFA Pattern | 28 |
| GPS Version ID | 28 |
| Padding | 25 |
| Offset Schema | 25 |
| Photometric Interpretation | 24 |
| Samples Per Pixel | 24 |
| Subject Distance | 23 |
| Host Computer | 13 |
| Y Cb Cr Coefficients | 12 |
| Primary Chromaticities | 12 |
| White Point | 12 |
| Gamma | 11 |
| Related Image Height | 10 |
| Related Image Width | 10 |
| User Comment | 8 |
| Reference Black White | 6 |
| Image Description | 6 |
| Brightness Value | 5 |
| Sensitivity Type | 2 |
| Lens Model | 2 |
| XP Keywords | 1 |

### 4.1.2. Behaviour after storage in cloud services

#### Windows Azure Platform

Uploading images to the storage service provided through the Windows Azure Platform, using the recommended blob storage method has proven to provide an exact copy of the file uploaded, as expected. Exif metadata is intact, and running a binary diffing tool reveals all files to be identical after being re-downloaded after being stored in the cloud.

#### Windows Live SkyDrive

When uploading images to an account at SkyDrive through the web interface, there are three options for storing images. The options are to upload the original image, a "Large" image resized to 1600 pixels, or a "Medium" image resized to 600 pixels. The default setting is for the images to be resized to 1600 pixels.

When images are uploaded using the default setting, they are resized to 1600 pixels, but all Exif metadata is retained, even though the image size and shape is altered. The image size is only altered if the image uploaded is larger than 1600 pixels. Running the images through a binary diffing tool reveals that only the image files larger than 1600 pixels are altered, while images with a smaller size than 1600 pixels are identical.

The same observation is true for uploading using the "Medium" setting. All Exif metadata is retained, while the image size and shape is altered. The image size is only altered if the image uploaded is larger than 600 pixels. Running the images through a binary diffing tool reveals that only the image files larger than 600 pixels are altered, while images with a smaller size than 600 pixels are identical.

While uploading using the "Original" setting, the image file itself is unchanged no matter how large the image file is. Running both the extracted Exif information and the image file itself through a binary diffing program, reveals all files to be identical.

Figure 19 gives a graphical overview of how the image files are altered using the different upload sizes in Windows Live SkyDrive. White portions are identical parts of the file; yellow portions are alterations, while grey portions are sections not contained in the other file.

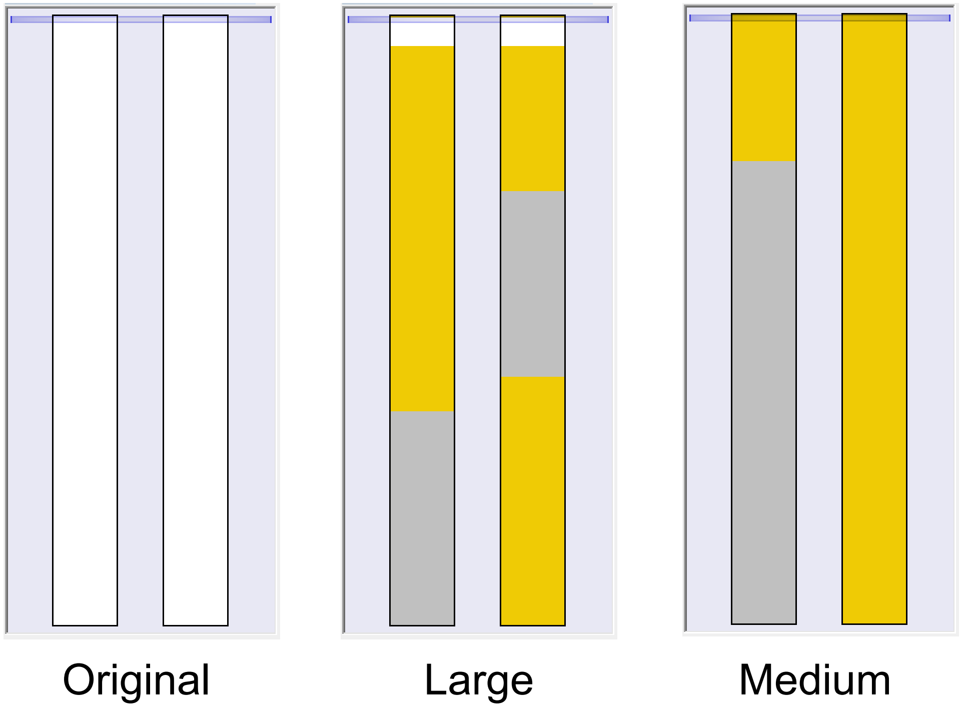


*Figure 19: Graphical representation from binary difffing tool showing what parts of an image file has been altered after storage in SkyDrive.*

### Flickr

When uploading the images to Flickr, Flickr automatically creates different sizes of the original image as well. For a regular user account the original image is not published. Only the resized images are available for download. This is true even for the owner of the account. The different resized images made available through Flickr are "Square" (75x75), "Thumbnail" (67x100), "Small" (161x240), "Medium 500" (336x500), "Medium 640" (430x640) and "Large" (688x1024), depending on the size of the original uploaded image. If the image size is between "Small" and "Medium", only sizes from "Square" up to "Small" will be available. The option to download the file in its original uploaded size is only provided if the user has a "Pro" account at Flickr, and explicitly has allowed sharing of the original file through the privacy settings. The Exif information is available through Flickers own "View Exif Info" function, but is stripped from all images available for download, except the original file size.

Although the Exif information is stored separately, the Exif information is not shared with the rest of the world without the user explicitly allowing the sharing of Exif information through the privacy settings. Even after this is done, GPS information is not mapped unless a second permission is given.

## 4.2. Discussion and conclusion

In this section, the results and procedures regarding the performed tests will be discussed, and a conclusion will be given.

### 4.2.1. General level of potentially sensitive information

As described in the results noted in chapter *4.1.1 General level of potentially sensitive information*, there were certain fields which were expected to turn up in the population that did not. This was mainly GPS latitude and longitude information and serial numbers of camera devices.

This could be the result of the randomness in the selection of the initial population, or that the images from the population were taken using equipment not capable of storing GPS coordinates and serial number information. A deeper look at the equipment used, shows that all images with this information available were taken with equipment not capable of storing GPS information without attaching other equipment to the camera or manually adding the location when editing the image.

Figure 20 shows the distribution of camera models used in the population, where such information was available. This information was available in 177 of the 400 images.

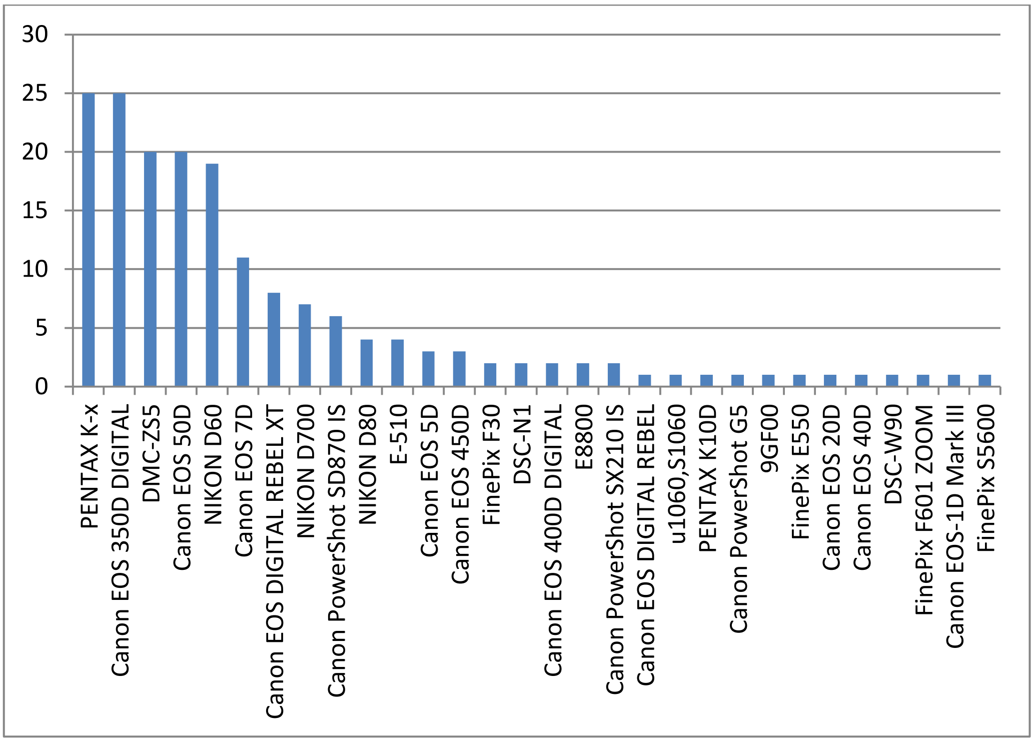


*Figure 20: Images taken with different camera types.*

Based on these observations, a better approach to selecting the initial population might be to look at a selection of published images taken by mobile devices which for the most part contain GPS functionality as well. While randomly selecting images, as in the experiment carried out in this thesis, gives an overall feeling for the general state of the sensitivity level of information contained in the images published. Another observation could be that mobile devices, or other devices, containing a GPS receiver are not as frequently used to share images on the internet as camera devices without this functionality built-in today.

Another option for generating the population from which to test would be to get a much larger control set of physical digital camera devices than was made available during the writing of this thesis, to get a feel for what information the different devices actually store. This would help in knowing what to expect in regards to the contained Exif information from the different camera models.

Another important aspect to be aware of is the introduction of different bias in regards to the experiments performed. It is not possible to say with complete certainty that the experiments performed contains absolutely no trace of any bias what so ever. In the experiments performed through this thesis, there might quite possibly exist some sampling bias in the way the images for the real world samples were chosen. The fact that images were chosen only from one source, although through many different users, a complete picture might not have been gathered. A perhaps better approach could have been to select images from different websites as well as from different users to get a wider selection of images containing Exif metadata.

### 4.2.2. Behaviour after storage in cloud services

While making it convenient for the user who publish their images in the cloud, altering the images themselves could mean trouble for a forensic investigator trying to find out about the origin of an image. As discovered through the work of this thesis, several of the cloud services tested performs some alterations to the images uploaded by the user. The cloud services tested do, however, retain the original image uploaded so that they can be provided to law enforcement agencies upon request, whilst some of the services tested hide them from the users. The amount of openly available information varies greatly from service to service, and depending on the users' privacy settings. As a forensic examiner this is well worth being aware of in the process of an investigation, where the examiner may not be able to contact the cloud service provider to have the original uploaded image released.

Although some cloud storage providers shields the common user by restrictive default settings regarding sharing of Exif metadata, other services retain the metadata in different sections, but strip them away from the images themselves, so that the information is not retained simply by downloading the images from the service. Other services simply keep all original uploaded information intact, making it easier for the forensic investigator, but might expose potentially sensitive information contained in the image metadata to the world, for the unwitting user.

The results presented in this thesis could be practically useful for forensic practitioners as a means to demonstrate the need to carry out experiments before using information gathered through cloud services, as a business case in the scoping of a forensics assignment. Further, the results can be used by cloud service providers to help develop their cloud service to be able to aid forensics practitioners in tracing the origin of an image.

### 4.2.3. Presented model/approach

The approach presented in this thesis fits into the general approach provided by NIST [15], as described in chapter *2.1.1 NIST Computer Forensics Tool Testing Project*, in point *6: develop testing procedures and method*, and uses cloud based storage services as the services to be tested.

By using the approach suggested in this thesis, the forensic examiner will be able to fulfil the criteria set by Daubert, due to the fact that the approach is based on scientific methodology, and through discussions following this publication will be subject to peer review, and follows the NIST guidelines for forensics tool testing.

The approach presented in this thesis harmonises well with the recommendations for Validation Testing provided by SWGDE [13], and most closely fits in under point *1c: Methodology* and can further be argued to support the SWGDE recommendations for general validation testing by providing a reproducible and repeatable method for testing metadata handling in cloud based storage services.

The approach could also be put in the context of Batten and Pan's framework for Testing digital forensic software tools used in expert testimony [19], in point *1: Selecting a testing design*. This makes the suggested approach very versatile in its area of use and application.

The validity of the method has been addressed by conducting the method in a controlled laboratory environment, to help increase the internal validity of the method by providing a means to more strictly control that only the variable of the cloud based storage service is measured. To increase the external validity of the method, a real world sample is introduced to help simulate a real-life setting. The method described could also be used to test how metadata is handled in cloud based storage services for different document types and other media files as well.

A weakness in the design is that it does not take into account and calculate the error rates involved with the cloud services, and it is thus difficult to say with certainty to what extent there might exist other factors rather than the cloud services themselves manipulating the image files uploaded.

Although the method presented will give an overview of what metadata is altered using various cloud based storage services, it is important to note that Exif information itself should not be considered reliable information, as it is very easily editable. There are no guarantees that the information contained in the image metadata is in fact unaltered original information. The values are just as easily overwritten and altered as they are extracted, and can thus not be used on its own to draw definitive conclusions about the origin of the image. It can, however, be used to further support other evidence, or give a starting point to narrow down a list of suspects for further investigation.

# 5. Summary and further work

## 5.1. Summary

Through this thesis, a general overview of what information which could potentially reveal sensitive information through image metadata has been gathered. An overview of what information is publically shared has been presented, by looking at a selective section of images published on the internet through image sharing services in the cloud.

It was found that although available, far from all images published contain information within the Exif metadata which could potentially contain personally identifiable information. However, more than 40% of the images in the population contain information about the camera make and model used to take the pictures. This is information that could be used by a forensic examiner to help identify the person who took the photograph, or narrow down the list of suspects.

Information revealing GPS coordinates and serial numbers of the equipment used to take the photographs were observed in a much lesser extent than in the initial hypothesis for this thesis. It is, however, an existing possibility for devices to attach such information to an image file, and the potential for revealing information which could be used to pinpoint the exact location where an image was taken, and the exact equipment being used, exists.

Further, this thesis has focused on what information is lost or altered when uploading images to cloud based storage services. A selection of cloud based services were used in the experiments, and it was found that there was a varying practice amongst the services tested by what information was altered or removed, both in regards to the physical image file itself and the Exif metadata contained within these images. The findings presented confirms the need to verify what information each cloud based storage service actually retain, and what is altered, before relying on information found attached to the images.

The approach in the experiments was based on a quasi-experimental methodology, more precisely a modified version of a nonrandomised control group pretest-posttest design. The model being used has been specified to be both repeatable and reproducible through following the steps presented in the thesis, and should thus fulfil the Daubert criteria of admissibility as well as adhere to the NIST guidelines for forensics tool testing.

The main contributions made through this thesis has been to provide an overview of what information regular users give away while sharing images through sharing services on the internet, either willingly or even unwittingly, as well as provide an overview of how cloud based services handle Exif metadata today, and an approach to test cloud based storage services has been presented.

Limitations in the work carried out through the practical experiments performed through this thesis were mainly limited to looking at Exif metadata contained in JPEG image files. Only a selection of cloud based storage services were tested.

It is, however, worth noting that although Exif information can be used in an investigation to help identify the source of the image, that Exif information in itself is not a reliable source of evidence gathering, as the Exif information could easily be edited. As such, this information should be used with care.

## 5.2. Further work

Through the writing of this thesis a selection of cloud based services were used in the experiments; however, there is an ever increasing pool of cloud based services to choose from. It would be useful to conduct similar experiments using a multitude of these services, to generate a reference list for the forensic practitioner while navigating through the yet largely uncharted terrain of cloud forensics.

While this thesis has focused mainly on Exif metadata contained in JPEG image files, metadata is also contained within a variety of other file formats widely shared on the internet through various cloud based sharing services. It would be interesting to perform similar testing on metadata contained in various document formats such as PDFs, Word, Excel, PowerPoint and others.

As the internet is filled with more and more user content, videos are published in an astounding and ever increasing capacity. Similar experiments as carried out in this thesis could be performed on video formats as well, to get an overview of what information is contained in these formats.

As Exif information is easily editable, work should be done to incorporate a feature to positively identify if the image or metadata has been altered in any way. This could perhaps be done using some form of digital watermarking and checksums as starting points. If incorporated in a default way by cameras, could prove to be invaluable to forensic examiners.

# Bibliography

[1] Franke K. Computational Forensics: Trends and Challenges in Applying Artificial Intelligence Methodologies to Digital Forensics.

[2] Saks M, Koehler J. The coming paradigm shift in forensics identification science. Science 3092005. p. 892-5.

[3] United States vs. Starzecpyzel, 880 F.Supp. 1027 (S.D.N.Y.) (1995).

[4] Frye v. United States, 293 F. 1013 (D.C. Cir. 1923), (1923).

[5] Daubert et ux., individually and as guardians ad litem for Daubert, et al. V. Merrell dow pharmaceuticals, inc, 509 u.s. 579 (1992).

[6] Frye. Frye Standard. Available from: http://en.wikipedia.org/wiki/Frye_standard.

[7] L. Dixon, Gill B. Changes in the Standards for Admitting Expert Evidence in Federal Civil Cases Since the Daubert Decision. RAND Institute for Civil Justice; 2001.

[8] Brunty J. Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner. 2011 [01.05.2011]; Available from: http://www.dfinews.com/article/validation-forensic-tools-and-software-quick-guide-digital-forensic-examiner.

[9] Dejnoñka J. Daubert vs. Frye: Logical Empiricism and Reliable Science. 1996; Available from: http://www.members.tripod.com/~Jan_Dejnozka/daubertvsfrye.pdf.

[10] Huber PW. Junk Science in the Courtroom. Valparaiso University Law Review. 1991;26(723).

[11] NIST. National Institute of Standards and Technology, Computer Forensics Tool Testing Project [01.03.2011]; Available from: http://www.cftt.nist.gov/.

[12] ISO/IEC. International Organization for Standardization and the International Electrotechnical Commission, General Requirements for the Competence of Testing and Calibration Laboratories, ISO/IEC 17025:1999. Geneva Switzerland1999. p. 26.

[13] SWGDE. Scientific Working Group on Digital Evidence, SWGDE Recommendations for Validation Testing Version v1.1. [01.03.2011]; Available from: http://www.swgde.org/documents/current-documents/2009-01-15%20SWGDE%20Recommendations%20for%20Validation%20Testing%20Version%20v1.1.pdf.

[14] NIST. National Institute of Standards and Technology. [01.03.2011]; Available from: http://www.nist.gov/.

[15] NIST. National Institute of Standards and Technology, General Test Methodology for Computer Forensic Tools. 2001.

[16] NIST. National Institute of Standards and Technology, Verification of Digital Forensic Tools. In: Lyle J, editor. Montana Supreme Court Spring Training Conference; 14 May 20102010.

[17] SWGDE. Scientific Working Group on Digital Evidence. [01.03.2011]; Available from: http://www.swgde.org/.

[18] Li C-T. Handbook of research on computational forensics, digital crime and investigation: methods and solutions. Hershey, PA: Information Science Reference; 2010.

[19] Lynn M. Batten, Pan L. Testing Digital Forensic Software Tools Used in Expert Testimony: Deakin University, Australia.

[20] Jain R. The art of computer systems performance analysis: Techniques for experimental design measurement, simulation and modelling. Hoboken, NJ: Wiley; 1991.

[21] Taguchi G. Introduction to quality engineering: Designing quality into produces and processes. White Plains, NY: Quality Resources; 1986.

[22] Taguchi G, Chowdhury, S, & Wu, Y. Taguchi's quality engineering handbook. Hoboken, NJ: Wiley; 2004.

[23] Mario Hildebrandt, Stefan Kiltz, Dittmann J. A common Scheme for Evaluation of Forensic Software. Proceedings of the 6th International Conference on IT Security Incident Management and IT Forensics (IMF2011); Stuttgart: Universitaet Magdeburg, Germany; 2011. p. 92-106.

[24] Murray JD, vanRyper W. Encyclopedia of graphics file formats. Bonn: O'Reilly; 1996.

[25] Riumplus. Rgb-raster-image.png. Wikimedia Commons; 2005 [01.03.2011]; Available from: http://en.wikipedia.org/wiki/File:Rgb-raster-image.png.

[26] Starbro D. VectorBitmapExample.svg. Wikimedia Commons; 2010 [01.03.2011]; Available from: http://en.wikipedia.org/wiki/File:VectorBitmapExample.svg.

[27] Yahoo! Flickr. Available from: http://www.flickr.com/.

[28] Google. Picasa. Available from: http://picasa.google.com/.

[29] Corporation P. Photobucket. Available from: http://photobucket.com/.

[30] Facebook. Facebook. Available from: http://www.facebook.com/.

[31] William B. Pennebaker, Mitchell JL. JPEG still image data compression standard. 3rd Ed. ed: Springer; 1993. p. 291.

[32] CCITT. The International Telegraph and Telephone Consultative Committee, Terminal Equipment and Protocols for Telematic Services, Information Technology - Digital Compression and Coding of Continuous-tone Still Images - Requirements and Guidelines, Reccommendation T.81. Available from: http://www.w3.org/Graphics/JPEG/itu-t81.pdf.

[33] Welch TA. A Technique for High-Performance Data Compression. 6 ed. Computer1984. p. 8-19.

[34] Deutsch P. DEFLATE Compressed Data Format Specification version 1.3. RFC 1951: Aladdin Enterprises; 1996.

[35] NISO. National Information Standards Organization. [01.03.2011]; Available from: http://www.niso.org/.

[36] NISO. National Information Standards Organization, Understanding Metadata. NISO Press; 2004 [01.03.2011]; Available from: http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

[37] Chaney M. Understanding Embedded Image Info. [01.03.2011]; Available from: http://www.steves-digicams.com/knowledge-center/understanding-embedded-image-info.html.

[38] JEIDA. Japan Electronic Industries Development Association. Available from: http://www.jeita.or.jp/english/.

[39] JEITA. Japan Electronics and Information Technology Industries, CIPA DC-008-Translation-2010 Exchangeable image file format for digital still cameras: Exif Version 2.3. 2010.

[40] NIST. National Institute of Standards and Technology, The NIST Definition of Cloud Computing (Draft). NIST Special Publication; 2011.

[41] Microsoft. Windows Live SkyDrive. Available from: http://explore.live.com/windows-live-skydrive-photos-videos.

[42] Microsoft. Windows Azure Platform. Available from: http://www.microsoft.com/windowsazure/.

[43] Microsoft. Windows Azure Platform - Storage service. Available from: http://www.microsoft.com/windowsazure/storage/.

[44] Nakashima J. Windows Azure Walkthrough: Blob Storage Sample. 2008 [01.03.2011]; Available from: http://blogs.msdn.com/b/jnak/archive/2008/10/29/walkthrough-simple-blob-storage-sample.aspx.

[45] George ML, Rowlands D, Price M, Maxey J. The lean Six Sigma pocket toolbook: a quick reference guide to nearly 100 tools for improvingprocess quality, speed, and complexity. New York: McGraw-Hill; 2005.

[46] Leedy PD, Ormrod JE. Practical research: planning and design. Boston: Pearson Educational International; 2010.

[47] cromo. portable Flicka. Available from: http://www.softpedia.com/get/PORTABLE-SOFTWARE/Multimedia/Graphics/Windows-Portable-Applications-Portable-Flicka.shtml.

[48] Amazon. Amazon Cloud Drive. Available from: https://www.amazon.com/clouddrive/learnmore.

[49] Google. Google Docs. Available from: http://docs.google.com/.

[50] Dropbox. Dropbox. Available from: https://www.dropbox.com/.

[51] Harvey P. ExifTool. [01.03.2011]; Available from: http://www.sno.phy.queensu.ca/~phil/exiftool/.

[52] GuidanceSoftware. EnCase Forensic. [01.03.2011]; Available from: http://www.guidancesoftware.com/forensic.htm.

[53] Intel. Intel® Core™ i5-430M Processor (3M Cache, 2.26 GHz) Specifications. [01.03.2011]; Available from: http://ark.intel.com/Product.aspx?id=43537.

[54] Harvey P. ExifTool Application Documentation. [01.03.2011]; Available from: http://www.sno.phy.queensu.ca/~phil/exiftool/exiftool_pod.html.