

**Metadata xxx**

**Daniel Mocek**

**Nuruljannah Hamizan**

**x**

**School of Informatics and Cybersecurity, Faculty of Computing, Digital & Data,**

**Technological University Dublin**

**Submitted to Technological University Dublin in partial fulfilment of the requirements for the degree of**

**Bachelor of Science in Computing in Digital Forensics & Cyber Security**

**Supervisor:**

**Peter Alexander**

**August 2025**

# Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Degree of **Science in Computing in Digital Forensics & Cyber Security** in the Technological University Dublin, is entirely my own work except where otherwise stated, and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed: Nuruljannah Hamizan Dated: 18/7/2025

Signed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Dated: \_\_\_\_/\_\_\_\_\_/\_\_\_\_\_

# Abstract

Metadata refers to the information regarding another piece of information, data of another data to help us further understand such data. Due to the evolving technology, the creation and use of metadata has tremendously changed, and metadata can be easily created by simply taking a picture on a mobile device. This opens the door to many risks for both users and creators of digital content to understand the different types of metadata that can be obtained from a single piece of data. This project explores the security implications of metadata, particularly the exploitation and misuse by malicious individuals.

Lei Zing Marcia and Jian Qin believes there are three features of any information objects, whether physical or digital, which are content, context and structure. These are the principal set of information that are attached to any information objects that can be reflected through metadata, answering questions such as who, what, when, why and whereabouts of the object [11].

Where the three features of information objects are embedded in a data structure, this project addresses the issues faced with metadata in today’s time by the implementation of a web-based tool aimed at identifying and removing metadata from digital files, primarily images, MP3 and PDF files. Furthermore, this paper will highlight the notable cases of metadata exploitation and evaluate the limitations of metadata handling across different online platforms on the Internet.

This project addresses the issue faced with securing digital files by identifying and removing embedded metadata that could potentially expose users’ privacy or file authenticity. Metadata information is embedded within files such as Images, PDFs and Documents, Metadata can unintentionally reveal sensitive information including but not limited to, File creation time, GPS location, Device Information and Author Details. The aim of this project is to develop a website-based tool that enables users to easily upload files and selectively strip them of such information.

XXXX

The result is a simple, yet effective web application that allows users, especially those in sensitive fields such as activists, journalists or government officials, to clean their files of compromising information. The website is easy to use to those with no previous knowledge of what metadata is, it allows users to select specific actions, such as viewing or removing metadata.

Through this metadata focused application, users gain control over their digital footprint, and it allows them to perform privacy preserving file management in a simple and secure manner.

# Acknowledgments

We would like to express our sincere thanks to Peter Alexander. Peter assisted us thought the project with engaging conversations, helped us guide us in the right direction that eventually grew into the project being presented. As well as that he gave us the opportunity to improve our previous work.

# Contents

[Declaration 2](#_Toc203754357)

[Abstract 3](#_Toc203754358)

[Acknowledgments 5](#_Toc203754359)

[Contents 6](#_Toc203754360)

[1. Introduction 8](#_Toc203754361)

[1.1 Background 8](#_Toc203754362)

[1.2 Problem Definition 8](#_Toc203754363)

[1.3 Aims and Objectives 8](#_Toc203754364)

[1.4 Research Questions 8](#_Toc203754365)

[1.5 Scope and limitations 8](#_Toc203754366)

[2. Literature Review 10](#_Toc203754367)

[2.1. Introduction to Metadata 10](#_Toc203754368)

[2.2. Types of Metadata 10](#_Toc203754369)

[2.3. Risk associated with Metadata 11](#_Toc203754370)

[3. System Analysis and Design 12](#_Toc203754371)

[3.1. Requirements 12](#_Toc203754372)

[3.2. Expected Outcomes 12](#_Toc203754373)

[3.3. Potential Outcomes 12](#_Toc203754374)

[3.4. Use case diagram 12](#_Toc203754375)

[3.5. System architecture 12](#_Toc203754376)

[3.6. Technologies used 12](#_Toc203754377)

[3.7. User interface Design considerations 12](#_Toc203754378)

[4. Implementation 13](#_Toc203754379)

[4.1. Frontend Development 13](#_Toc203754380)

[4.2. Backend Framework 13](#_Toc203754381)

[4.3. File Upload Validation 13](#_Toc203754382)

[4.4. Metadata Extraction and removal 13](#_Toc203754383)

[4.5. Security and Privacy Consideration 13](#_Toc203754384)

[5. Technical Requirements 14](#_Toc203754385)

[6. Testing 15](#_Toc203754386)

[6.1. Testing Strategy 15](#_Toc203754387)

[6.2. Functional Testing of Metadata 15](#_Toc203754388)

[6.3. Comparison of Other Tools 15](#_Toc203754389)

[6.4. User Feedback 15](#_Toc203754390)

[6.5. Evaluation 15](#_Toc203754391)

[6. Results 16](#_Toc203754392)

[6.1. Summary of Outcomes 16](#_Toc203754393)

[6.2. Analysis of Metadata accuracy 16](#_Toc203754394)

[6.3. Impact on File Integrity 16](#_Toc203754395)

[7. Challenges 17](#_Toc203754396)

[8. Conclusion 18](#_Toc203754397)

[8.1. Summary of project 18](#_Toc203754398)

[8.2. Limitations of current application 18](#_Toc203754399)

[8.3. Future improvements 18](#_Toc203754400)

[8.4. Suggestions for Future Research 18](#_Toc203754401)

[9. References 19](#_Toc203754402)

[10. Appendix 20](#_Toc203754403)

# Introduction

## Background

In an age of digitalisation, data privacy has become increasingly important and ensuring that sensitive information is correctly handled and not susceptible to exploitation is becoming increasingly necessary. Many users are unaware that when a file is created, it often contains embedded metadata that may include sensitive personal information.

Metadata refers to such information embedded in digital files, including creation time, author’s identity, geographical location, as well as device information. While metadata can be useful for managing content by larger organisations, it poses significant privacy and security risks, especially when digital files are shared publicly. Professionals in sensitive roles including government officials, activists and journalists, as well as those with a social media background are especially vulnerable to metadata exposure which can lead to disclosure of their personal information.

To address these issues, this project aims to develop a web-based tool to identify and remove certain metadata found within an image file of the user’s choice as well as the ability to manipulate certain data that the user may choose to change to conceal certain sensitive data. The tool supports common image formats including JPEG, JPG, WEBP and PNG. Furthermore, this project aims to raise awareness concerning the exploitation and misuse of metadata embedded in certain image files as well as the importance of knowledge surrounding the risks of metadata privacy.

## Problem Definition

Many existing metadata removal tools often provide limited file types as well as lack usability and are often too complex that proves to be inaccessible for non-tech savvy users. This project addresses these challenges by proposing a web-based system designed to allow users to upload digital files of various file types and identify and selectively remove metadata. By prioritising on accessibility and efficiency, this application enables users to easily and efficiently enhance their digital privacy.

## Objectives

The aim of our web application is to enable users especially those in sensitive fields to analyse and secure their digital files through metadata handling and steganographic methods, as well as that to enhance the privacy by selectively stripping certain metadata, in a user friendly, accessible and user-friendly manner without the need for any prior technical knowledge.

## Research Questions

The following research study of metadata proposes for the following questions and will be answered throughout this paper:

1. Which metadata are considered the biggest risk?
2. What are the most effective ways to protect metadata without altering the file?
3. Can you retrieve removed metadata and what is the size limit of metadata?
4. How do different social media platforms handle metadata?

## Scope and limitations

* + 1. **Scope**

The scope of this project surrounds the study of the different metadata that can be obtained through a single image file, specifically in formats such as JPEG, JPG as well as PNG. The scope of this project revolves greatly on the study of the metadata found within an image file, including device information, timestamp, author, GPS location data as well as authenticity of the picture. Additionally, the project involves the development of a web-based tool that will allow the user to identify, remove and modify the metadata found within image file. The web-based tool is intended for general users, ensuring the ease of use by implementing checkboxes and guided information, as well as privacy enhancements by implementing a login option, allowing users to safely and securely use the tool.

A web-based application to allows users to upload digital files such as PDF or PNG.

The application will provide users with the ability to view their metadata and selectively remove it

The application will be a website based on Flask and SQL, it will use Python libraries such as Pillow and use tools like MAT2.

The feel of the website will be designed for users with very little technical knowledge, using checkboxes and guided information to simplify the process.

Testing will involve gathering a dataset of files and comparing each file before and after processing though the application.

* + 1. **Limitations**

The scope of this documentation is heavily limited due to the following considerations:

* The data submitted will have to adhere to the accepted file formats as listed: JPEG, PNG, WEBP and PDF.
* The application depends on already existing libraires and tools such as Pillow and MAT2 and thus may have certain constrains in terms of metadata handling or steganographic depth.
* Browser limitations may affect speed, performance and file quality depending on the user’s system and file size
* Metadata detection and removal accuracy may vary on different operating system and file formats as well as detect false positives, especially with non-standard files.

# Literature Review

## Introduction to Metadata

Metadata, a term referring to ‘data about data’ plays a foundational role in general developments where a library is utilised, whether digitally or physically. According to NISO Primer, metadata is something that we create to help us manage and interact with to further gain the knowledge that we need.

Metadata can be found everywhere and comes in many forms, the music app we listen to, the videos we watch, the photos we take and upload, metadata can be found on everything, regarding when the photo was taken, topic that the user likes, the type of music that is favoured and much more. Images hold EXIF metadata such as the date and place of capture, office files keep author details, creation/modification dates, and editing history, emails feature sender and recipient identifiers, subject lines, and timestamps, social media updates track posting times, user IDs, hashtags, and location data (Riley, 2017).

## Types of Metadata

In the digital world, metadata is the ‘key to the functionality of the systems’ to keep the favourability of the user and record significant information about the user. The favourability of the user is maintained by providing relevant content to the users, as well as disregarding the topics that the user has no interest in. In May of 2012, Google released the Knowledge Graph, a database created to allow users to easily obtain public information about any topic that is released on the Internet. Furthermore, the Knowledge Graph contains an additional functionality where users are able to see the connection that others made towards that topic to another to provide for an interconnected result. This is done through the record of metadata found in the patterns of which users connect multiple information at once. This section will focus on the primary types of metadata relevant to cybersecurity and digital forensics, including descriptive metadata, system metadata and embedded metadata.

(<https://www.happiestminds.com/insights/knowledge-graph/#:~:text=The%20Knowledge%20Graph%20is%20a,in%20response%20to%20the%20query>.)

2.2.1 Descriptive / Substantive Metadata

Substantive metadata revolves around the discovery and identification of a documentation, such as the author, type of document, the font, the size, the modifications made, comments, and many other. An example of a substantive metadata is within the function in Microsoft Word that allows users to track the font, modification history, timestamps of changes and the author of the file. Substantive metadata can also reveal information such as documentation of legal protocols as well as location information, this metadata is critical for forensic investigations for establishing the rightful ownership of the file, detecting tampering, misconstrued timeline as well as contents of the file.

(<https://research-ebsco-com.tudublin.idm.oclc.org/c/sr37fs/viewer/pdf/qszw7czazr?route=details>)

<https://www.getty.edu/publications/intrometadata/setting-the-stage/>)

2.2.3 Administrative Metadata

Administrative metadata is a type of metadata concerning the technical details surrounding the protection of a digital object. To ensure protection of data, there are four types of administrative metadata: preservation metadata, provenance metadata, technical metadata and intellectual property rights metadata. Preservation metadata concerns the digital information regarding the file for future preservation and to distinguish between its later versions. Provenance metadata is the information of the measures taken to preserve the file to observe the changes made while preserving to accurately compare from the original file. Furthermore, provenance metadata concerns the authority of the file, as well as the ownership of the file. Technical metadata refers to information regarding the file’s format, user characteristics of the file, as well as any decoding process that is required for the file. Lastly, the intellectual property rights metadata is information regarding the copyright and licensing surrounding the file. Though there are many similarities between administrative and descriptive metadata, the stark difference is that administrative metadata focuses on the technical information surrounding a file, whereas descriptive metadata surrounds information for discovery and identification.

<https://www.secoda.co/glossary/an-introduction-to-administrative-metadata>

<https://socialhistoryportal.org/bestpractices/administrativemetadata>

2.2.2 Structural Metadata

Structural metadata refers to the information surrounding the structure and relationships of data. Data is organised in a certain system unbeknownst to investigators, therefore the study of the structural metadata is used to provide insight of the data structure and reveal how the data elements are interlinked. Structural metadata is considered vital for understanding the organisation of data and enhances the understanding of each data piece. An example of structural metadata is the addition for a table of contents, where readers are easily able to determine the page numbers of each chapter, or the structure of XML documents where each data is organised in a certain rule.

<https://www.secoda.co/glossary/an-introduction-to-structural-metadata>

<https://www.connectedautomateddriving.eu/data-sharing/ccam-data-sharing-framework/dsf-4-data-and-metadata-description/4-3-2-structural-metadata/>

2.2.4 File System Metadata

This is known as the file system metadata as its created and stored in the file system of an electronic device, whether it’s a mobile device, a laptop or a computer. In a real-life crime investigation scenario, this type of metadata is the first to be examined to reveal the files of interest to establish an accurate timeline. The file system metadata can reveal many things regarding the access permission of a file as well as when the file was created and modified. The metadata that is stored in a device’s file system is proven to be more substantial than many other types of metadata as the rest lack “evidentiary value because it is not relevant”. Furthermore, whereas a substantive metadata can be tampered with, with the file system metadata, it is much harder to be tampered with due to the file system metadata not being stored in a digital file that can be edited but rather in the file system of the device.

File system metadata can branch off into a subtype known as MAC (Modified, Accessed, Created). This type of metadata is commonly referred to as the timestamps as it records the date and time of when a file is created or modified. The MAC metadata is often created in a format that is not readable for humans as it’s created by the computer system for computing purposes. An example of this is a timestamp for a MAC operating system, 3442503385 represents 2013/01/31 10:56:25 Thu PST. Computers use a system for tracking time known as the ‘epoch’ time, that measures the number of seconds that has elapsed since the beginning of the creation of the UNIX time on January 1st, 1970, at 00:00:00, hence the large value in the metadata structure representing time. On the Windows computer, there is an additional timestamp known as the NTFS timestamp that is stored in the Maste File Table (MFT) which records the MAC values for files within the device that is represented in a 100-nanosecond interval since January 1st, 1601. An example value of a timeline in the NTFS format is 132742726640000000, which translates to 2021/09/01 12:44:24.

## Risk associated with Metadata

While metadata can improve functionality, it also introduces significant security and privacy risks.

* + 1. **Information Leakage**

Metadata can reveal sensitive information such as usernames, GPS coordinates and software used (ThreatNG, 2021). For example, sharing a photo without removing the EXIF data can expose the user’s location, most social media sites do this for you now days to reduce file size, but it could still expose you to potential privacy risks.

* + 1. **Targeted Exploitation**

Attackers can use metadata to carry out reconnaissance, for example, knowing what software and version was used to create a file can help identify possible vulnerabilities. Document metadata has been used before for spear phishing attacks (Symmetry Systems, 2022).

* + 1. **Real Life Examples**

There have been real world consequences due to metadata leak. In 2012, anti-virus founder John McAfee location was exposed through EXIF data in photos published by journalists (Greenberg, 2012). In 2017 a NSA contractor was identified through printer tracking dots embedded in leaked documents (EFF, 2019). Individuals in sensitive fields, such as journalists, government officials and whistleblowers should take extra precautions to scrub metadata from files prior to publication (Security in a Box, 2021).

For many places of the world, the charm and beauty of the country is heavily dependent in their coastlines and beaches. Places such as Spain, Italy, Australia and Mexico not only take pride in their coastlines for its beauty and cultural identity, but also the large amount of tourism it attracts. Coastlines open the opportunity for spiritual identity, habitats for a variety of wildlife, as well as providing for different recreation activities such as sunbathing, kayaking, snorkelling and scuba diving. Coastal regions with warm climates highlight the importance of protecting and preserving its coastal beaches due to the high attraction rates of tourism each year. With Australia’s tourism gross domestic product (GDP) estimating around 75 billion in 2023/24 and enabling 1.6 million jobs, the call for research and study of how to improve such coastlines for coastal management surrounding tourism increases.( <https://www.statista.com/statistics/1193089/australia-tourism-gdp-time-series/#:~:text=Australia's%20direct%20tourism%20gross%20domestic,due%20to%20the%20coronavirus%20pandemic>. )( <https://wttc.org/news/australias-travel-tourism-sector-set-to-reach-record-315bn-in-2025#:~:text=In%202024%2C%20Australia's%20Travel%20%26%20Tourism,lion's%20share%20of%20the%20recovery>. )

This puts a great focus on understanding the target audience and their needs. Facilities such as hotels, saunas, parking, public transport and much more must be considered to aid in the improvement of tourism within a coastline. To obtain such data, however, requires a lot of resources, time and money, and the result may be limited and vulnerable to bias. Furthermore, data collection can be affected by many variables such as the time of day, weather and season of the year with people that are reluctant to answer survey questions during their leisure activities.

In response, researchers are utilising social media understanding tourism patterns. The vast amount of people using social media promoting the coastline, as well as giving feedback about the amenities can greatly increase the country’s popularity due to the growing amount of public data on the Internet. The geolocation provided often consists of geolocation, time, and device information allows researchers to analyse patterns of visitation, as well as the amenities utilised in different landscapes.

Social media platforms such as Flickr are notable in this regard due to their inclusive feature of users able to add their geolocation, providing researchers an abundant of data for research purposes to aid in tourism and recreational improvements. With over 10 billion images, Flickr is one of the only platforms where users are able to freely add their metadata onto a picture. By contrast, social media platforms such as Twitter and Instagram have implemented stricter geolocation policies due to privacy concerns.

In recent years, there has been numerous studies demonstrating the beneficial effects and applications of metadata analysis in areas such as suicide prevention. In Ireland alone, there has been a reported amount of 436 suicide cases in 2022, highlighting the importance of education within mental health, as well as intervention and prevention strategies. While there are many preventative measures being taken, the best way to prevent these cases are to look for warning signs within their behavioural patterns. These warning signs are hard to detect as people tend to not talk about their suicide plans to others, however, this is not the case for social media, where people feel safer to talk about their emotions and feelings behind a screen. While warning signs are difficult to notice upon verbally talking to them as many are reluctant to divulge into their suicidal thoughts in person, social media often provides a sense of safety, allowing people to openly express their emotions. This presents an opportunity for suicidal detection within social media by utilising metadata.

In 2016, Braitwaite, Giraud-Carrier, West, Barnes and Hanson created a machine learning model to detect suicidal messages within Twitter utilising the Linguistic Inquiry and Word Count (LIWC) tool, consisting of 90 psychological and grammatical variables where messages are divided into those variables, and a percentage is given based on the variables within the sentence. For example, if a sentence had 10 verbs, the verb category would have a score of 10. In 2017, Cheng et al. utilised the metadata from this technology to create a Suicide Probability Scale (SPS) to measure the ‘suicidality of respondents’ on the posts published on a social media platform known as Sina Weibo, a commonly used platform in China. This technology is further utilised in additional studies where the metadata is utilised for suicide prevention, highlighting the need for advanced methods to identify the early warning sighs of mental health crisis through online behavioural patterns.

<https://www-tandfonline-com.tudublin.idm.oclc.org/doi/epdf/10.1080/13811118.2021.1955783?needAccess=true>

While metadata embedded images are often used for benefitting purposes such as digital forensics, tourism research and content management, there are many ways in which metadata embedded images can compromise privacy and security. Information surrounding the geolocation, timestamps and device models are examples of EXIF data that are stored onto images automatically upon taking them that can be exploited when sharing images. A notable moment where the EXIF data of a picture was compromised occurred in 2012 where a businessman, John McAfee was a fugitive running away from the law following the death of his neighbour. Alongside him was a reporter, recording his every move while on the chase, published a picture of John on Vice magazine containing the metadata including his geolocation. After extracting the geolocation of the photo, John’s location was compromised and was swiftly arrested in Guatemala, highlighting the risks of unstripped metadata published on the Internet and the need for awareness and policies while sharing digital content.

(<https://www.theguardian.com/world/2012/dec/03/john-mcafee-location-revealed-vice>)

# System Analysis and Design

## Requirements

This section Outlines the key requirements for the proposed web application designed to create a simple to use, user friendly and guided tool to allow for the extraction, removal and viewing of metadata. This system aims to address the concerns users might have in relation to potential metadata exposure risks, particularly those in sensitive fields.

## Expected Outcomes

**Metadata Extraction:** The application must be able to scan, extract and remove metadata from various file types.

**Metadata Scrubbing:** The application must be able to selectively remove metadata after upload based on user’s needs,

**User Authentication:**

## Potential Outcomes

## Use case diagram

## System architecture

## Technologies used

## User interface Design considerations

# Implementation

## Frontend Development

## Backend Framework

## File Upload Validation

## Metadata Extraction and removal

## Security and Privacy Consideration

# Technical Requirements

## 5.1 Device Specs

* Laptop Windows 11
  + Processor :12th Gen Intel(R) Core(TM) i9-12900H 2.90 GHz
  + System type : 64-bit operating system, x64-based processor

## 5.2 Software / Library Requirements

* PIL – Acquire images
* Exiftool – Metadata extraction and Editing
* Python – Website functionality
* HTML – Website Frontend
* Flask – Lightweight web application framework
* CSS – Website Styling
* JSON – Obtaining Python variables

# Testing

## Testing Strategy

## Functional Testing of Metadata

## Comparison of Other Tools

Through our research in metadata extraction tools, we have concurred that there are many other metadata extraction tools out there. However, most are CLI based, where the output is displayed in the CLI and is not beginner friendly. Furthermore, it does not allow for those on different device such as a smartphone to access the tool.

Mat2 is a well-known tool for erasing metadata of different digital medias such as JPEG, PNG and GIF, as well as audio medias namely MP3, WAV and OGG and certain documents in PDF forms. Though the different formats in which the Mat2 accepts are starkly different to StegoSafe, we excel in other ways in which Mat2 does not. This includes the option to include and exclude the areas of metadata in which the user chooses to include in the photograph, as well as the ability to further manipulate the values within the metadata fields, whereas Mat2 does not have the option to select the values that gets deleted, but rather deletes every field of the metadata. Additionally, while Mat2 also has the functionality to view the metadata, it is not editable, whereas StegoSafe allows for all the fields to be editable with an ease of access as StegoSafe includes the different options to edit certain vital parts of an image including geolocation, phone model, date created, author and an added function to embed a comment into the image.

## User Feedback

## Evaluation

# Results

## Summary of Outcomes

## Analysis of Metadata accuracy

## Impact on File Integrity

# Challenges

# Conclusion

## Summary of project

## Limitations of current application

## Future improvements

## Suggestions for Future Research

# References

1. Duval, E., 2001. Metadata Standards: What, Who & Why. J. Univers. Comput. Sci., 7(7), pp.591-601.
2. Hart, T.R. and de Vries, D., 2017. Metadata provenance and vulnerability. Information Technology and Libraries, 36(4), pp.24-33.
3. Hart, T.R. and de Vries, D. (2017) Understanding Metadata in the Digital Age. Journal of Digital Curation, 12(2), pp. 11–21.
4. NIST (2015) NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements. National Institute of Standards and Technology. Available at: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-3r1.pdf> (Accessed: 18 July 2025).
5. Riley, J., 2017. Understanding metadata. Washington DC, United States: National Information Standards Organization <http://www.niso.org/publications/press/UnderstandingMetadata.pdf> (Accessed: 18 July 2025), 23(2017), pp.7-10.
6. Getty Research Institute (2018) Introduction to Metadata. 3rd edn. Los Angeles: Getty Publications.
7. IBM (2021) *Metadata Explained.*
8. ThreatNG (2021) *Metadata in Cybersecurity* (Accessed: 17 July 2025).
9. Symmetry Systems (2022) *The Metadata Problem: Why You Should Care*.
10. Security in a Box (2021) *Guide to Removing Metadata*. Tactical Tech.
11. “Introduction to Metadata.” *Google Books*, 2016, books.google.ie/books?hl=en&lr=&id=xgZVDQAAQBAJ&oi=fnd&pg=PP1&dq=metadata+&ots=bbRnjJ0jR8&sig=IpU-UnueX7TOi4r-yzJRmjTSL9w&redir\_esc=y#v=onepage&q=metadata&f=false.

# Appendix

# 