Digital Crime Investigation: Unleashing the Secrets of Data by Attributing the Origination and Authorship

Eswara Sai Prasad Chunduru
Assistant Director, Digital Forensic Division,
Central Forensic Science Laboratory,
Hyderabad, India

Nagendar Rao Koppolu
Inspector of Police (In-charge State Cyber Vertical),
Telangana Police Department,
Hyderabad, India

Abstract — **The data about data commonly referred to as, metadata, plays a vital role in the forensic analysis of content that is electronically stored in a digital storage medium. Metadata, along with MAC (Modified, Accessed, and Created date and time) details of files and folders, provides vital supporting information as to the nature, characteristics, genesis, handling and legitimacy of the data. As Internet usage is widespread, the forensic analysis of metadata is becoming vital as it can be used to link a digital artifact with their original source, which can be a web download, or an email attachment. Metadata can be considered as the electronic equivalent of DNA. The collection and analysis of such metadata enhances the quality of forensic reports, which are essential in our Criminal Justice System.**

I. INTRODUCTION

Metadata is information stored within a digital artifact that automatically adds identifying characteristics to it. It is a type of electronic fingerprint. Metadata of a file consists of a collection of fields. There will be specific metadata fields associated with every digital artifact. Each field stores a particular type of data/information. A field will have a label (name) and associated value. The term metadata is used to represent a collection of fields, with each field having a name and a corresponding value. Metadata associated with a digital artifact depends on the type of a file in which the artifact present. Most of the File Types contain additional metadata fields compared to simple file formats such as text files. These additional fields are created and embedded in the files by the software applications that have generated these files. Hence metadata may be different between different file formats and depend on the software that is used to create or modify the file.

II. TYPES OF METADATA

The different types of metadata [3] [5] [10] and their functioning is as follows:

A. Descriptive Metadata:

These types of metadata are used to provide detailed information about any file or source. These files include keywords, titles, and descriptions. These are mainly used on websites. When a user downloads a video file, the runtime of the film would be descriptive metadata. It provides basic information such as titles, author name, date, etc. The descriptive metadata becomes more advanced or complex when it is used to identify unique elements, such as code-driven projects and websites.

B. Structural Metadata:

This type of metadata clears up the details of how files or objects are managed. This metadata example is the table of contents on the book. The first page, Index Page of a book defines the details about content, chapter pages. Structural metadata records information about how a particular object or resource might be sorted.

C. Administrative Metadata:

Administrative Metadata is the information about the types of resources, permission for other users, etc. These metadata also help users to identify how the data was created. The two sub-classes of Administrative Metadata are -

a. Rights management metadata, which deals with intellectual property rights.
b. Preservation metadata, which contains information needed to archive and preserve a resource.

The relationship between the Administrative, Descriptive and Structural metadata can be seen in the below fig. 01 (Source: https://research.csc.fi/metadata-and-documentation).

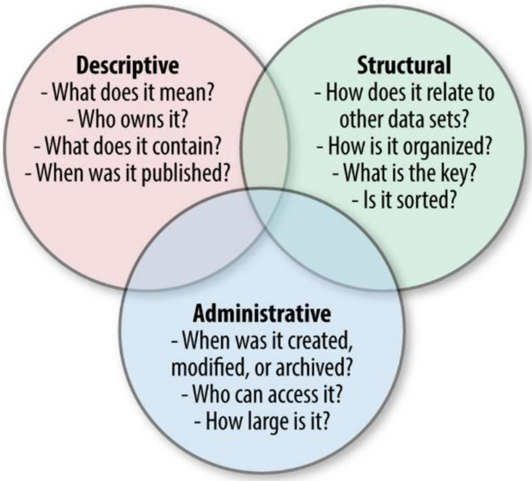

Fig. 1: Relationship between the Administrative, Descriptive and Structural Metadata

D. Statistical Metadata:

These types of metadata are used to collect process and give final output for any information. For example, we enter data on the computer, then the computer process, and finally, it gives us results. This metadata is generally used in data analytics, data clustering and bigdata for providing a bird's eye view as well as in-depth statistical view of the data/information as well as in Artificial Intelligence for development of decision-making algorithms. This type of Metadata is widely adopted in the online e-Commerce Platforms for ascertaining the likes and dislikes of the customers and recommending the products basing on their previous purchase or interested history.

- **Another classification of metadata is as follows:**

1) *Application Metadata:*

Also termed as OLE metadata/Substantive metadata, is automatically generated by application software used to create that content and is embedded in every file created or edited using that application software. This data is associated with the file as the file transcends from one location to another.

2) *System Metadata:*

It is generated by the Operating system/s that control individual digital system. This includes file allocation table fields (file name, location of the file, date and time of creation, length, and the date the content was modified) to all files stored on the digital system so that the operating system can identify and locate that file for future use. System metadata resides in the System Registry, $MFT, $USNJRNL etc., of the digital system and is used in the file management.

3) *Embedded Metadata:*

It is the text, numbers, content, data, or other information directly or indirectly contributed to a Native File by a user and is not typically visible to the user viewing the output display of the Native File on screen or as a printout. Examples for this type of metadata include formulas used in a spreadsheet, hidden columns rows, hyperlinked files, references and fields, and certain database information.

III. STANDARTS AND ELEMENTS OF METADATA

Large number of metadata models are available. They can be classified as either general models or content-specific models, as depicted below Fig.2 (Source: https://livebook.manning.com/book/tika-in-action/chapter-6/)

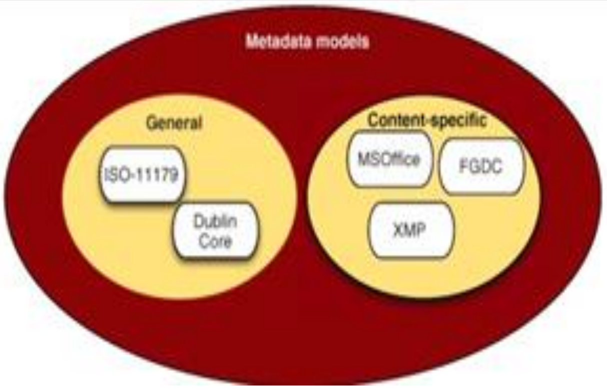

Fig. 2: Metadata Models

The International Standards Organization (ISO) has published a reference standard for the description of metadata elements as part of metadata models. The standard, ISO-11179, defines a mechanism for generating metadata models. Apart from this there exists fifteen elements of Metadata endorsed in the following standards:

1. ISO Standard 15836:2009 of February 2009 [ISO15836]
2. ANSI/NISO Standard Z39.85-2012 of February 2013 [NISOZ3985]
3. IETF RFC 5013 of August 2007 [RFC5013]

The document "DCMI Metadata Terms" [DCTERMS] provides an abbreviated reference version of the above fifteen elements as follows, popularly known as DC Element Set (Dublin Core Metadata Element Set) [10]:

1. Content**:** Coverage, Description, Type, Relation, Source, Subject, Title, etc.
2. Intellectual Property**:** Contributor, Creator, Publisher, Rights, etc.
3. Instantiation**:** Date, Format, Identifier, Language, etc.

A. *Image File Standards*:

Digital images are stored in various standard file formats such as PNG, JPEG, TIFF, and PSD as well as proprietary formats such as RAW. In addition, there are three metadata formats widely used in the industry:

- Exif

"Exchangeable Image File Format" – a standard for image file metadata, is jointly managed by Japan Electronics and Information Technology Industries Association (JEITA) and Camera and Imaging Products Association (CIPA)

- IPTC-IIM

"International Press Telecommunications Council" is the creator and maintainer of metadata standard "Information Interchange Model".

- XMP

"Extensible Metadata Platform" is a multimedia metadata standard introduced by Adobe.

Each metadata container format has unique rules regarding how metadata properties must be stored, ordered and encoded within the container. The inference each metadata container provides and the relation among them is depicted in the below Fig.03 (Source: https://exiftool.org/forum/index.php?PHPSESSID=29c1e3139bcc824cb87b186623dcddb5&topic=3065.msg13734#msg13734):

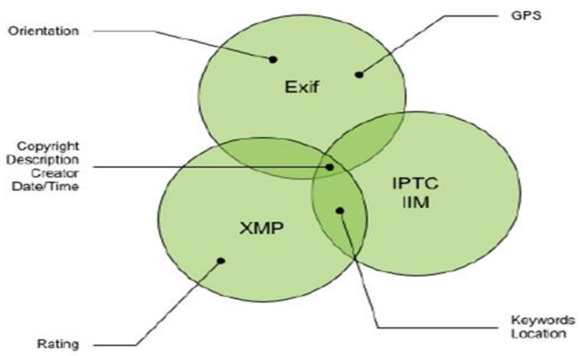

Fig. 3: Relationship among the various containers of image Metadata Properties

IV. Characteristics of Metadata

What so may be the type of metadata, it is portable across systems and devices, which helps in various ways of analysis methodologies. Metadata is used by newer file systems to store information about the data that is saved on the disk. For instance, the Linux ext3 file system logs all file system data and metadata changes in a journal. Microsoft file systems, such as NTFS, also store this type of information. In addition, older file systems, such as Linux ext2 and Windows FAT16, store metadata of files such as timestamps, size, and data location. Ext2 also has the capability to store information regarding permissions of files for different users and groups.

Investigators should be able to differentiate between system, application and embedded metadata elements so that they can track relevant metadata fields that are beneficial for an investigation. System metadata, such as file MAC data times, usernames attributed to the files and original sources, can help an investigator trace the origins of a file and other relevant information such as the dates when the file was created. All these are useful in an investigation involving cybercrime. Application metadata can help an investigator obtain additional details such as the authorship of a file and other proprietary information stored by the applications as metadata within the file. Embedded metadata can be extremely valuable to understand any suggested changes to a document or the underlying formulae in a spreadsheet or hyperlinks to external sources and so on. These types of metadata may even help in understanding the thought processes or intentions of document authors.

V. FUNCTIONS OF METADATA

The various functions of metadata [9][10] that can be helpful in the process of Unleashing the Secrets of Data and Attributing the Origination and Authorship are as follows:

A. *Resource discovery*

1) Allowing resources to be found by relevant criteria
2) Identifying resources
3) Bringing similar resources together
4) Distinguishing dissimilar resources
5) Giving location information

B. *Organizing e-resources*

1) Organizing links to resources based on audience or topic
2) Building these pages dynamically from metadata stored in databases

C. *Facilitating interoperability*

Using defined metadata schemes, shared transfer protocols and crosswalks between schemes, resources can be searched more seamlessly across the network.

1) Cross-system search, e.g., using Z39.50 protocol
2) Metadata harvesting, e.g., OAI protocol

D. *Digital identification*

1) Elements for standard numbers, e.g., ISBN
2) The location of a digital object may also be given using:
   a) a file name
   b) a URL
   c) some persistent identifiers, e.g., PURL (Persistent URL); DOI (Digital Object Identifier)
3) Combined metadata acts as a set of identifying data, differentiating one object from another for validation purposes.

E. *Archiving and preservation*

1) Challenges:
   a) Digital information is fragile and can be corrupted or altered.
   b) It may become unusable as storage technologies change.
2) Metadata is key to ensuring that resources will survive and continue to be accessible into the future. Archiving and preservation require special elements:
   a) To track the lineage of a digital object.
   b) To detail its physical characteristics.
   c) To document its behavior to emulate it in future technologies.

VI. IMPORTANCE OF METADATA IN A FORENSIC INVESTIGATION

The main area of digital forensics is the retrieval of data from a digital storage system. The process of digital forensics contains broadly two phases: forensic imaging and analysis. The second phase of digital forensics involves retrieval/recovery of data and scrutiny of the retrieved/recovered data for its authenticity, authorship and attributes.

For example, during a digital investigation that involves the analysis of collection(s) of digital images, many forensic questions can be raised, some of which are listed below [5]:

- What is the camera used for picturizing the image?

- What is the actual date and time when the image was taken?
- Any Geo-Location related information is tagged to the image so that ascertaining the same becomes easy.
- How many sources can be identified from the file metadata?
- How many files belong to each of these identified sources?
- How many files show evidence of being doctored?
- How many Internet downloaded files show evidence of doctoring?
- What was editing software used?
- Are there other “similar” files where source metadata is incomplete?
- How many other metadata matches for such files?
- Which of the files were downloaded from the Internet?
- If so, can the source of these files be identified?

It is a general practice to report the results of analysis, the attributes of the file/s containing details such as when the file was created, last accessed, last modified, its logical size, physical size on the disk/media, file description, etc., are added to the report.

The metadata of a file provides valuable evidence to the investigator on its sources, authorship and timestamps. While the contents of a file form the primary evidence, metadata provides the supportive information which helps in strengthening the case.

The list of the Metadata fields is not limited but vast and dependent on various factors like Application, System and User Customization. Some of the Metadata fields that can be utilized in a forensic investigation are given as below:

CREATION DATE →→ FILE SIZE →→ FILE NAME →→ FILE TYPE & EXTENSION →→ IMAGE WIDTH & HEIGHT→→ RAW HEADER →→ COLOR COMPONENTS →→ AUTHOR →→ EXIF DATA →→ FILETIME →→ SYSTEM TIME →→ CREATION DATE.

The following table shows [5] [10] some of the metadata parameters that the analysis applications use during Forensic analysis:

TABLE I: METADATA PARAMETERS USED IN FORENSIC ANALYSIS [5] [10]

Source of Input	S/No	Input Parameter	Data Type
Log File Entries (Security Logs)	1	Event ID	Integer
	2	User Name	String
	3	Date Generated	Date
	4	Time Generated	Time
	5	Machine (Computer Name)	String
File System Metadata Structures (NTFS)	6	Modification Time	Date & Time
	7	Access Time	Date & Time
	8	Creation Time	Date & Time
	9	File Size	Integer
	10	Directory flag (File/Folder)	Boolean
	11	Filename	String
	12	File Type	String
	13	Path	String
	14	File Status (Active, Hidden etc.)	Enumeration
	15	File Links	Integer
Registry Information	16	Key Name	String
	17	Key Path	String
	18	Key Type	String
	19	Key Data / Value	String or integer
Application Logs	20	Name	string
	21	Version no.	Integer
	22	Timestamp	Date & Time
Network packet	23	Packet length	Integer
	24	Class	String
	25	Priority	String
	26	Source IP	Integer
	27	Destination IP	Integer

VII. METADATA OF WIDELY USED APPLICATION FORMATS

Most applications do not create application-specific or embedded metadata. For files created by such applications, which do not create application-specific or embedded metadata, system metadata is the source of metadata. Following are some important applications/file formats that contain metadata created by applications within the files. These applications are widely used, and cybercrime investigators must understand the metadata generated by these applications.

A. *Office Documents*

The metadata stored in the documents, spreadsheets, and presentations can reveal information about the document beneficial for the investigation [5][10]. As per Microsoft, there are four types of properties associated with any document file. They are:

1) Standard properties (author, title, and subject): These properties can be used to identify the person responsible for creating the content.

2) Automatically updated properties include file system properties, like file size or the dates when a file was created or last changed and statistics like the number of words or characters in a document. One cannot specify or change the automatically updated properties. These properties can be used to search the files based on criteria like date of creation, modification, etc.

3) Custom properties – One can define additional custom properties for Office documents. One can assign a text, time, or numeric value to custom properties and also assign them the values yes or no.

4) Document library properties are associated with documents in a document library on a website or in a public folder. When you create a new document library, you can define one or more document library properties and set rules on their values. When you add documents to the document library, you are prompted to include the values for any required properties or update any incorrect properties. For example, a document library that collects product ideas can prompt the person for properties such as Submitted By, Date, Category, and Description. When you open a document from a document library in Word, Excel, or PowerPoint, you can edit and update these document library properties by clicking File > Info. All required properties from the document library are outlined with red borders on the Info tab in Word, Excel, and PowerPoint,

The following metadata can be saved within Microsoft Office/Open Office documents, spreadsheets, and presentations:

- Description: Title, Subject, Tags, Categories, Comments.
- Origin: Authors, Last saved by, Revision number, Version number, Program name, Company, Manager.
- Content: Content created, Date last saved, Last printed, Total editing time, Content status, Content type, Pages, Word count, Character count, Line count, Paragraph count, Template, Scale, Links, Language.
- File: Size, Date created, Date modified, Date accessed, Availability, Offline status, Shared with, Computer.

The fields under the section "Description" are user-editable. Apart from the above, the features and capabilities in MS Word such as Track Changes, Commenting, Macros and Fast-Saves of Microsoft Office and OpenOffice documents can provide useful information for an investigation.

Track Changes and Commenting:

The Track Changes and Commenting features of Microsoft Office and OpenOffice provide the investigator/forensic analyst valuable details [1][10] regarding the history of the changes, creator's details and subsequent details of the users who reviewed the file. Neither of the features is enabled by default in either of the products; however, its use by document creators and reviewers can reveal additional information such as suggested modifications and so on. The Track Changes feature provides a facility to view the history of all the changes made to a document that is yet to be "accepted" by the reviewer. If this feature is left enabled by the user, all the viewers of the document will have the capability to view the document creator accepted the changes made since the last round of changes. It allows users to make comments to the document without changing any of the underlying data content. However, if those comments are left as is, each subsequent viewer can review all the comments made by document reviewers. The drawback with this feature is that Excel and PowerPoint do not warn a user of the embedded comments in a document. When using the Track Changes and commenting features, the reviewer's name is also stored with any comment or change made to the document.

Macros and Fast-Saves:

The Macros feature used within office products can be very useful to increase productivity in many situations. Macros are scripts, which are used to automate operations within the office products.

Microsoft Office adds the name of the macro's author to any macro used inside a document, spreadsheet, or presentation. This feature of the office is a goldmine for a forensic examiner investigating macro viruses. A well-known example of a macro virus is the Melissa Virus. Anyone who opens a document with the virus in Microsoft Office can potentially become a victim of the virus. The virus would then send itself by e-mail to the first 50 contacts in the person's e-mail address book. This enabled the virus to replicate at a fast rate.

B. *PDF Documents*

PDF documents are widely used as a method of distributing documents in a common format readable across platforms. As such, the capabilities and features of this document format also present a forensic challenge to the examiner and risk to the user. Adobe, the creators of PDF format and its viewers, such as Adobe Acrobat Reader, has a plugin that is designed to work seamlessly with Microsoft Office and OpenOffice products. This seamless functionality is of great benefit to the user and poses a challenge regarding stored metadata. When a document is converted to PDF format, by default, it is enabled to store all of the metadata stored with the original document, such as version information, the document's creator and so on. All of the risks mentioned above regarding the Track Changes and Commenting features are also an inherent risk to PDF documents. Adobe also has a commenting feature that can also add to the amount of wealth of metadata stored with the documents. One additional feature, which is not enabled by default, is the ability of PDF documents to store the source document inside the PDF file.

PDF files can contain two types of metadata [1]. The first is the Document Information Dictionary, a set of key/value fields such as author, title, subject, creation and update dates. This is stored in the optional Info trailer of the file. Using the Extensible Metadata Platform (XMP), metadata, as used in XML-based file formats such as information about the embedded fields, can be added to the Metadata of PDF files. It gives a good insight into the various sources from which the content of the embedded multimedia files originates, thereby giving a broad spectrum of scope to the forensic examiner in the identifications of the origin of the embedded fields.

C. *Image Files*

The basic queries regarding the images a forensic examiner comes across are:

- Who is involved with this image? (who took it, who owns it and, who/what is in it?)
- Where is this image from?
- When was this image created or modified?

Forensics of metadata from the images addresses queries related to the location and MAC times of the image [4] [5] [10]. That said, the location information can only be available for supported image formats. PNG and JPEG are the most popular formats that support storing of location information within image metadata. The device used to capture the image should also have the capabilities to store location information in an image. The Exif section of the metadata store's location and other details, such as the make and model of the camera and settings within the camera.

The queries of an investigator related to the content of an image (objects and/or persons within the image) and the image's original creator are not available in the metadata. Hence, the investigator should use alternative methods to obtain these (viewing images in a gallery mode that can provide an efficient means of navigating a huge number of images or using AI-based methods to extract information within images and so on). IPTC Photo Metadata Standard provides the details of various metadata fields that can be relied on for traceability, authenticity and authorship.

D. Audio files

The metadata of an audio file can reveal information such as: Title, Subtitle, Rating, Comments, Duration, Bit Depth, Sample rate, Number of channels, Audio data format type, Audio data orientation, Sound channel map, Sound field, Sound channel assignment, Channel number, Sound map location, Bit Rate Reduction, Codec Name, Codec Name Version Codec Creator Application, Codec Creator Application Version, Codec Quality, Data Rate, Data Rate Mode, Offset to first relevant stamp, block size, First valid byte in block, Last valid byte in block, Word size, Time Stamp Start, Time Stamp End, Software used [1][6].

E. Video files

Two types of video metadata exist:

- Automatically generated Metadata based on file properties and other details such as the equipment used, location information and so on.
- User-generated Metadata, which can be any additional metadata fields added to the content to enable better visibility on the Internet and so on.

F. E-mails

Following are some of the important metadata fields that can be extracted from e-mails:

- Address fields – To, Reply to, From, CC, BCC
- Date fields – Sent date and time
- Data fields – Accent of the language, metadata from the attachments
- E-mail headers –IP addresses of servers that handled the e-mail

Using this metadata, investigators should reconstruct a graph of e-mail and contact history to demonstrate how certain emails were sent across different contacts.

VII. CONCLUSION

Metadata is data that describes other data. As discussed in this paper, if the metadata that is embedded in various file formats is retrieved and used in the process of forensic analysis of electronic data, it can open new avenues of investigating angles in precisely identifying the authorship, sources and time periods.

Metadata is created both by the operating system and applications. The operating system creates standard file system level metadata for file management purposes. The same metadata can help investigators in understanding the modified, accessed and created date time stamps of files and also the original source locations and authorship. With much of the content being either downloaded from the Internet (from websites or social media) or shared via email, identifying the download location will be of great help to investigators in certain cases. Many widely used applications, such as MS Office, Adobe Reader and so on, add additional application and file format specific metadata. These metadata may help in identifying the author of the document and many related details. Images, audio and video files contain additional metadata that may even help in identifying the location where these files were created and also the specific devices and models that were used in creating these files. Certain applications, such as MS Word and OpenOffice may also include embedded metadata such as comments and suggested changes. Such metadata may help identify additional sources of information, based on the participants in a conversation via comments to a document. The three types of metadata – system generated, application generated and embedded, may together help an investigator identify important people and resources from where additional evidence can be collected.

Metadata can also be used to filter files based on date time stamps and file format types. Digital forensics products provide such filtering mechanisms. This process of filtering metadata can help improve efficiency in an investigation. By filtering the metadata and searching within the content of the resulting files can help an investigator link file attributes (date and time stamp, authorship, file location and so on) with the content. This approach will help in substantiating the evidence (identified based within the content) with specific persons and the date and time of the crime being investigated. Therefore locating, extracting and mapping the metadata with content in files is a critical part of a cybercrime investigation.

REFERENCES

[1] Document Metadata and Computer Forensics, Spring 2006 Term Paper for CS633 (Computer Forensics) by Jeffrey R. Jones, INFOSEC Master's Program, James Madison University, jonesjr@jmu.edu

[2] Mohammed, Hussam; Clarke, Nathan; and Li, Fudong (2016) "An Automated Approach for Digital Forensic Analysis of Heterogeneous Big Data," Journal of Digital Forensics, Security and Law: Vol. 11 : No. 2 , Article 9.

[3] W. Lawrence Wescott II, The Increasing Importance of Metadata in Electronic Discovery, 14 RICH. J.L. & TECH. 10, http://law.richmond.edu/jolt/v14i3/article10.pdf.

[4] Raghavan S, Raghavan S V (2013c). Determining the origin of downloaded files using metadata associations, J Commun, ISSN: 1796-2021, 8(12):902–910

[5] Raghavan, S., Raghavan, S.V. Eliciting file relationships using metadata-based associations for digital forensics. CSIT 2, 49–64 (2014). https://doi.org/10.1007/s40012-014-0046-4

[6] AES31-3-1999: AES standard for network and file transfer of audio – Audio-file transfer and exchange – Part 3: Simple project interchange (PDF, 228KB) <http://aessec.aessc.aes.org/pub/aes31-3-1999.pdf>

[7] B. Carrier, File System Forensic Analysis, Addison-Wesley, Boston, Mass, 2005.

[8] Forensic Implications of Metadata in Electronic Files By John Ruhnka and John W. Bagby

[9] Role of metadata in cyber forensic and status of Indian cyber law by Aashish Kumar Purohit, Naveen Hemrajani, Ruchi Dave M. Tech Scholar, SGVU, Jaipur, Dept. of CSE, SGVU, Jaipur

[10] Web Page: https://www.lter.uaf.edu/metadata_files/ UnderstandingMetadata.pdf