Review -3

For

Data Cleansing Tool

(Randomization, Anonymization & Suppression)

Prepared by

Nikhil Pinnamaneni	**18BCI0053**	**pinnamaneni.nikhil2018@vitstudent.ac.in**
Sumanth Dodda	**18BCI0067**	**dodda.sumanth2018@vitstudent.ac.in**
Sai Charan Muvva	**18BCI0073**	**muvvasai.charan2018@vitstudent.ac.in**
Manohar Gogireddy	**18BCI0077**	**gogireddymanohar.2018@vitstudent.ac.in**

Faculty : Vijaya Kumar K

Abstract:

Metadata consist of information that characterizes data. Metadata are actually used to provide the documentation for the data products. In the essence the metadata can answer: who, what, when, where, why, and also how about every facet of data that are being documented in a machine.

Metadata within a file type(can be anything) can actually tell a lot about you. Cameras can record the data about when a picture was taken and also what camera was used at the time. Office documents which are: PDF or Office automatically adds the author and the company info to the documents and also the spreadsheets. Maybe you don't want to disclose those information. That is when this data cleansing tool comes into picture . you can have all the metadata of the documents you share either randomized, anonymized or even both. This tool essentially focuses on the location, source of the data anonymization.

There are many notable drawbacks to Exif metadata. The most challenging drawback is ease in which data can be manipulated and lack of audit methodology for tracking the changes to metadata. Much of media source material found on Internet will contain a very little metadata. Practically all of social networking sites now routinely strip the metadata from images that are uploaded to their websites. original reason for removing the metadata is that the Internet trolls and the stalkers were harvesting the GPS data from the images of the celebrities and the officials to determine their travel habits and then also intercepting them at the expected locations.

Introduction:

After many years of the massive digitization activities, the important libraries hold now the large collections of the digitized books and also the journals. Some among these collections are also available in the Internet and are accessible for absolutely free download. These days, most documents which are being distributed by the publishers are digital born and the need for the retro conversion of the document contents is also reduced. However, to perform the automatic information extraction from the PDF documents(or any type of data) the files must be thoroughly processed so as to identify relevant metadata. We will also do the automatic extraction of the administrative metadata from the PDF documents(or any other file type) that are now standard de-facto for the documents in the digital libraries. Metadata extraction is actually very useful for retrospective annotation of the digital born works which are produced by the publishers.

Exchangeable Image File Format: Exif metadata which is one type of embedded technical metadata, which includes the rights, management and the provenance fields, follows object as it also travels through the web. This research project tried determine, how or if the embedded metadata followed digital object as it is being shared on the social media platforms by using the EXIFTool, a variety of the social media platforms and also user profiles, embedded metadata

extracted from the selected New York Public Library NYPL and also the Europeana images, PDFs from the open access, science journals, and the captured mobile phone images. goal of project was to clarify which embedded fields of metadata, if there is any metadata, which is migrated with object as it was being shared across the social media.

We can cleanse the following data types:

- Audio Video Interleave (.avi)
- Electronic Publication (.epub)
- Free Lossless Audio Codec (.flac)
- Graphics Interchange Format (.gif)
- Hypertext Markup Language (.html)

- Portable Network Graphics (PNG)
- JPEG (.jpeg, .jpg, ...)
- MPEG Audio (.mp3, .mp2, .mp1, .mpa)
- MPEG-4 (.mp4)
- Office Openxml (.docx, .pptx, .xlsx, ...)
- Ogg Vorbis (.ogg)
- Open Document (.odt, .odx, .ods, ...)
- Portable Document Fileformat (.pdf)
- Tape ARchive (.tar, .tar.bz2, .tar.gz)
- Torrent (.torrent)
- Windows Media Video (.wmv)
- ZIP (.zip)

Required Modules:

- python3-mutagen for audio support
- python3-gi-cairo and gir1.2-poppler-0.18 for PDF support
- gir1.2-gdkpixbuf-2.0 for images support
- gir1.2-rsvg-2.0 for svg support
- FFmpeg, optionally, for video support
- libimage-exiftool-perl for everything else
- bubblewrap, optionally, for sandboxing

Literature review:

Microsoft offers a free Document Inspector for removing the personal or sensitive information before we share Office file. Microsoft Support site provides the specific information on deleting the metadata from Microsoft Word 2013 and Microsoft Word 2010. Likewise even the Adobe's Help site explains how we can remove metadata from our PDFs in the Acrobat Pro, the Acrobat X Standard. The thread on the forum provides step-by-step instructions for the excising metadata, or hidden content from the PDFs using Examine Document tool in the Acrobat 9. Windows Explorer

lets you to view and also delete the metadata from file via Properties dialog box. The most quickest way is to click the Properties, then Details , Remove Properties and Personal Information : Create a copy with all the possible properties removed. Another metadata removal tool for the Windows is image files is free JPEG & PNG Stripper from the Steel Bytes. For the Mac OS X users can delete the EXIF data from the image files by using free Image Optimize program. **[9]**

Facebook, Twitter and other social sites and image-sharing services will automatically add location information and other metadata to the images you upload. However, you are still sharing data with the service. Maybe you're just fine with it and may not know the services when you take photos.
To disable location services for camera on the iPhone or iPad (basically IOS), select Settings> Privacy> Location Services and turn off the camera. **[4]**

The simplest way to remove the GPS data from the images on Android phones is to use free image privacy app which adds the strip metadata option to Android Share menu. After you select the option, the menu will reappear after a short delay. Now when you select a service (Twitter, Facebook, Dropbox, Flickr, etc.), the image will be removed and uploaded with its location data.Another option is the free VerExif.com service to remove location information from images. (The site opens by default in Spanish, but can also be viewed in English using google translate.) Select the file option, select the image, navigate

and click Open. You can delete or view EXIF data, The delete option also appears on view screen.
VerExif.com only works on one image file at a time, unlike programs mentioned above, you can delete location information from multiple image files at once.The widespread growth in the creation, storage and exchange of digital information affects all aspects Of human activity. The exponential growth of digital Information means that digital evidence is now prevalent in both criminal and also the justice systems. **[5]**

The legal sector is largely dependent on digital systems. The huge increase in digital evidence presents many challenges and opportunities for legal practitioners. Digital evidence often contains a hidden piece of metadata form that legal practitioners should be aware of. Each digital file contains a variety of metadata. Every communication sent through digital communication, such as email and social media messages, contains metadata. Participants in online conversations can retrieve metadata such as social media from their smartphones and computers. The growing importance of digital evidence and metadata requires legal practitioners to gain knowledge about this evidence. Digital evidence, also known as electronic storage information (ESI), was added to the Federal Rules of Civil Procedure in 2006 to make it clear that ESI is searchable. Metadata is often an important form of ESI. **[11]**

Information on message metadata such as routing and addressing should ensure that the message is transmitted across the Internet. Email servers generate metadata about each message, sending messages to their intended destinations and storing email messages for each email account. Some of these routing metadata will appear, such as sender and recipient email addresses. Most email metadata will be hidden until certain steps are taken to expose it. Message header is an important

type of hidden metadata. Email titles contain information about the email server that handles a specific message when it is transmitted from the sender to the receiver, with the message ID (the specific number that identifies the specific email message) being the first email to handle the message. Internet Protocol (IP) addresses are used to guide messages and time stamps folded by the server. Each email message ID is unique. The IP address is an important part of the TCP (Transmission Control Protocol) for solving and routing Internet communications. Therefore, IP addresses are an important part of email addresses, although email titles and their contents may vary depending on the email provider. For example, outgoing messages in Gmail do not contain the sender's IP address. The log file of the email server that manages the message also contains various metadata records. Metadata is the underlying infrastructure of Internet communication and the record of the travel of an email message from the sender to the receiver.

At the very least, the Defense Council should insist that the metadata be generated intact, especially e-mail messages in the original format with the message header. Metadata can sometimes be used to identify Internet Protocol (IP) addresses and therefore the location where a specific e-mail message was sent. Additionally, message titles are comprehensive in verifying the authenticity of an e-mail message. Message routing information and cryptographic signatures can often be used in the authentication process embedded in message titles. Additionally, webmail providers such as Google, Yahoo and Microsoft maintain login records that reveal the IP address the user logs into their e-mail. In some cases, metadata may indicate whether an e-mail account has been accessed from a mobile device or other computer device. Also related to the metadata authentication, which provides an important context for the e-mail messages. For example, metadata is when an e-mail message is opened, answered or forwarded, whether the message is flagged for further action, the names of the recipients of the identity and blind copies, the presence of attachments, and the name of the folder where the e-mail appeared. The metadata associated with the e-mail attachment includes additional data about the attachment creator, the computer on which it was created, the last written and last accessed dates, and the history of changes to the file.

Getting knowledge about metadata is important because metadata is a fundamental part of digital evidence. Computers, phones, and other digital devices automatically generate metadata as part of the process of generating, storing, and communicating digital information. Metadata is not only a by-product of creating and storing digital files; This is actually a possible form of evidence: if the evidence proves or disproves any claim, it is clearly metadata evidence. Metadata sheds light on the source, context, authenticity, reliability and distribution of electronic evidence, as well as provides evidence for human behavior. It is an electronic equivalent of DNA, ballistic and fingerprint proofs

comparable in strength and detail.

A computer file contains information (or data) for creating a file, as well as data related to the metadata or properties of the file. For example, a file usually contains the name of the document creator, the date the file was last saved, and the metadata on which the file was last printed. Additionally, some files, such as the Microsoft Word document, may contain hidden information that is not clearly visible to the user. For example, a Word document may contain hidden

information in the form of tracked changes and comments that may not appear when the document is opened in certain document views. When documents are shared between individuals, information hidden in metadata and / or documents can be problematic. For example, many important documents can be communicated between individuals by attaching documents to electronic mail. However, when a document is shared in its electronic form, the document contains embarrassing or problematic metadata and hidden information. For example, the consultant may retrieve a document previously created for the first client and edit parts of the document with information about a project for the second client. The advisor may not notice that the "Track changes" option is turned on before editing, especially if the advisor is set to hide tracked changes in the Document View screen.

When the consultant forwards the edited document to another electronically the client will only see the document that can change the document to see the information about the first client that the consultant deleted during the editing of the second customer document. Before accidentally disclosing confidential information of the client to another customer. Tools are available to assist the user in removing hidden data from metadata and files. However, such tools are usually manual in nature and file. The user needs to remember the tool for deleting metadata and hidden information from the file before sharing it with others. , Such tools may not work effectively in a wide variety of collaborative environments.

The increased use of electronic communication has increased awareness in disclosing unwanted information to third parties. A type of disclosure caused by the electronic communication through the exposure of metadata: Metadata, or data about data, is information about an electronic file in a form that is automatically created and stored and travels with the electronic file, either automatically or by manual means.

Some users are reluctant to disclose metadata that occurs with artifacts. In view of this need, methods and programs for removing metadata from artifacts have been developed. One way to remove metadata from artifacts is by locally based metadata deletion applications. The solution is stand-alone metadata removal software, which is installed on the computing device and removes the metadata from the artifacts. There are disadvantages to the application of locally based metadata deletion. Some disadvantages of a locally based metadata deletion application The metadata deletion application requires user intervention for the setup and use of the local computing device. Additionally, the metadata deletion application occupies a portion of the computing device's storage capacity. Natively based metadata deletion applications are also slow and can affect other applications or programs running on the local computing device. Still, deleting metadata is limited to artifacts on a local computing device.

Another method of removing metadata from an artwork is the remote metadata removal system. The Remote Metadata Removal System allows you to delete metadata from artifacts associated with email. Emails are first created on local personal computers, an internal email service, or a mobile communication device. The email and attached artifact will be sent from the sources of the communication server, which will describe the email and prepare the email and attached artwork to be delivered to the recipient. Prior to delivery, email and artwork will be sent to the Metadata

Removal Server. Metadata Removal Server removes metadata from artifacts based on previously provided instructions or a purge procedure (which is often performed). The cleaned artwork and email will then be sent to the intended recipient. There are also disadvantages to remote metadata removal systems. The disadvantage of the remote metadata removal system is that a centralized cleaning method is used to categorize each artwork and remove the metadata according to the general instructions provided to the metadata removal server. Therefore, there is a need for a method and mechanism intended to cross the boundaries of pre-art solutions.

Software for embedded systems written in C / C ++ is overloaded with memory vulnerabilities, which, despite

significant academic and industrial efforts, pose a security threat to computer systems. The consistent approach to detecting vulnerability is very slow and often loses a lot of vulnerabilities in the product software. Therefore, several runtime protections have been proposed to protect against a variety of attacks during runtime. In these rescuers, memory protection can be considered as a strong defense that can detect memory errors as soon as it is triggered. However, it can provoke significant runtime / memory overhead and often suffer from compatibility issues with legacy software. Other methods, such as Control Flow Integrity (CFI) and ASLR, target specific types of attacks to reduce the workload, but their limited defense weakens programs against advanced attacks or new types of attacks. The Multi-Variant Execution Environment (MVEE) has been proposed as a robust defense that can add a distinction between security and performance. It uses the practice of multi-core environments in modern embedded computing systems to execute multiple variations of a program and monitor their deviations as indicators of security breaches. MVEEs monitor system call limits for deviation signals based on the fact that attackers must eventually use certain system calls to use their malicious intentions on the system. The data layout in the variants is random so that the variations run the same under normal conditions and show different behavior when attacked.

Imagine these daily activities of a the media specialists taking the digital photographs, the videos, and also recordings. While taking this possibly hundreds of photos during a particular session, photographer would have to stop at some point and also record the technical elements of image being recorded. Metadata such as the shutter speed, date or time, the location, the focal length, etc. would definitely have to be manually transcribed for each set of the images, taking the time which could be better if applied to the photographing a far better image. The automatic recording of these available metadata has been dramatically improved the very professional capabilities of those who would actually use these digital tools for the recording.

Now practically all of the digital recording devices also track the metadata about recording which includes where it was actually made, what exact time it was made, device used, and other valuable information about recording. This collection of the data saves the time and effort as well as improving the accuracy in preserving the data about recording. exchange image file format (Exif) standard defines specification for how to store the metadata for the image, video, and the audio files. Typically this is output format generated by the digital cameras (including the smart phones) and the scanners. These are the metadata about image itself. There are a great deal of the metadata

which is potentially available about any image or video, or the audio file. These data are exploitable by the researchers in attempting to provide attribution for web based sources of the information that we are able to uncover. The Each element within metadata is identified by a TAG. **[12]**

There are numerous of metadata tags embedded in each of the audio, video, or image file, but actually not all available tags are included in with each file. It actually depends upon whether tag is actually set by device creating the file. Typically these metadata tags cover a broad spectrum of data including the date and time information; the camera settings (such as GPS location of where image was actually taken, the shutter speed, the focal length, the image orientation, if a flash was used, the camera make and model, etc) the thumbnail preview; the descriptions; and also the copyright information. Originally, the metadata tags and the content rules were developed by manufacturers who created and distributed media equipment, which remains prime case for most of the metadata. However, Dublin Core Metadata Initiative is an open effort that provides the metadata design and the best practices. This is an international and also neutral technology organization established to manage the long term curation and the development of the metadata standards that also include digital media accessible through ExifTool . Members and also volunteers have manage an ongoing discussion of the metadata themes and also setup international and regional events to promote an open discussion for establishment and the curation of these standards.

Recent years have seen an increase in attention paid to the embedded metadata by information profession. Foundational research has explored advantages of embedding the metadata into the digital images and objects. embedded metadata can also include the technical, descriptive, and also administrative elements. They wrote: the

properly applied, the embedded descriptive metadata can be easily understood and also used as the technical metadata.

Knowing who created object shown in digital image can be easy knowing when that image file was actually created. while the technical metadata is automatically recorded by capture device, the descriptive and the administrative metadata can be manually added and manipulated using the software designed for this purpose. the Embedded metadata also comes with the limitations, including:

(a) it’s actually not always persistent
(b) it can be very easily removed “during the actions of uploading and downloading the digital files into and out of the social media platforms”

(c) the embedded descriptive metadata... can also be incorrect, incomplete, or even missing entirely” **[3]**

The Previous groups have completed studies on the embedded metadata. Some are mainly focused on developing the standards for capturing and populating the embedded metadata elements. A team at Smithsonian Institution identified the core minimal the embedded metadata fields for their digital

image production studio They wrote that using the existing standards for the embedded metadata, whether in form of the descriptive, the technical, the structural or even administrative can aid in the searchability, provenance, rights management, interoperability, and also data repurposing".

the Another project, funded by Library of Congress National Digital Information Infrastructure and Preservation Program which is led by American Society of Media Photographers (ASMP), designed and published the "guidelines for the refined production workflows, the archiving methods, and the best practices for the digital photography based on a variety of capture methods and the intended image use". The guidelines contained the recommendations and the commentary on the embedded metadata, including the IPTC, Exif, XMP, and the Global Positioning System (GPS). They are closely linked to author's very own research project, IPTC Photo Metadata Working Group study have investigated how the embedded metadata is being shared across the social media. reveal the image metadata is inconsistently supported across the social media sites and that two most popular sites for sharing the digital images, Flickr and the Facebook, remove the embedded metadata from image file header during the procedures for uploading a digital image to social media platform and downloading the digital image onto desktop from a social media platform**[15]**

In the current time when the amount of data that is being stored is huge, the problem of keeping it secure arises. Sensitive information must be protected from being leaked to unauthorized entities. To protect data some techniques that are being used are dissociation and encryption.

Dissociation techniques include anonymization that can be achieved by depersonalization. This method systematically removes a persons' identifiable information from a document such that the data can't be traced back to the person. K-anonymity deals with the quasi-identifiers present in the data by using generalization and suppression techniques. Some other forms of k-anonymity deal with the same data in a different manner by giving each quasi-identifier of different persons different values to further reduce the risk of identifying a person just by identifying another person with same quasi identifier values. Anonymization is non-reversible but pseudo anonymization is reversible because the data that has been stripped from the given document or a data-set has been stored somewhere and can be leaked in some circumstances leading to a failure of the whole mechanism.

While encryption is a good approach to avoid data leakage it is not possible to use that encrypted data to produce a meaningful result without decrypting it which defeats the whole purpose of collecting that data. To combat this the data can be sent over to a secure machine where it is decrypted and processed as usual but this method is not very efficient. An efficient method would be usage of a special encryption method or some sort of precalculated indexing. The simplest way to do this would be a hash-based or encryption-based indexing over individual values.
Both methods discussed above can be circumvented by bypassing the architecture. One person like a database admin

will have access to all the information and they can't be trusted.

To overcome these individual weaknesses of the above described methods a new method has been proposed called Perimeter which refers to Pseudo-anonymization and personal metadata Encryption. In this method the data is categorized into non-critical and critical data. The non-critical data is left unmodified. Pseudonymize the fragments and assign access identifiers and access authorization tickets to each fragment of a number of pseudonyms. Save the original document structure and query able document descriptions and arbitrary keywords in the form of an XML. Encrypt just the sensitive content of this metadata so that only the document owner or a person with key can create an extended table of this data. Split the data before doing this so that only the people with key can identify a link between the segments and fragments of the data sent. Use the information from XML to create an efficient and content related query process.

According to the author of the paper "The Devil is in the Metadata – New Privacy Challenges in Decentralized Online Social Networks". To solve the privacy issues in centralized social networking services Decentralized networks are used. In completely decentralized networks the users are connected to each other via a peer-to-peer network. There are wide range of designs spanning from centralized ones to completely de-centralized ones. Most of the data shared online other than text has metadata in some form or the other. Metadata leakage can lead to discovering some personally identifiable information like the location where you took a photo and at what time the photo was taken. Using these values alone we can find out the personnel details like address of the person that took that picture. Metadata also has a lot of sensitive data and quasi identifiers. Metadata can be used to find out the access control on that particular data. In an example given in the paper "The Devil is in the Metadata – New Privacy Challenges in Decentralized Online Social Networks" the user could find out the total number of pictures taken a person on a holiday from the limited access to some certain pictures of that person on holiday. Such information was not meant to be shared in the first-place but gets shared without the users consent or knowledge. **[13]**

In the paper "Metadata Protection Scheme for JPEG Privacy & Security using Hierarchical and Group-based Models" the authors have proposed 2 different models for uploading images to social networking sites. Two models have been proposed because each service has different needs and they can't be generalized under a single model to fit the requirements of all services. So, two models will be proposed for and any service can choose a model that they can choose from. One of the models is called hierarchical model it is based on lattice-based access control. This model has hierarchical access control that dictates what amount of access a person at a certain hierarchical level has. The lower they are in the hierarchical tree the lower access they have. The other model proposed in this is called group-based access control. In this model all the people with access are divided into groups and the amount of access to this information is defined by the group that they are in. This model has role based predefined contextual groups. Privacy settings for each group will be different and access to the service metadata depends on the group settings. **[8]**

The hierarchical model does not allow an individual to set a higher level of security even if they want to. Even if the user has a concept of security they don't know which metadata is being shared at what level of the hierarchy. To mitigate this issue an extension was developed which allows users to make changes to the hierarchy manually. Social networking sites like Facebook have users with different sets of contacts to manage the access to their data, but the main goal of

social networking sites is to deliver the content rather than just protecting it. In this use case the hierarchical model might not be the correct one to choose. So Facebook can operate well with group model because users can finely control the access to the whole group without having to assign a hierarchy to a user or adjust each users settings. The group based model falls behind in terms of providing transparency to the user regarding which data is shared in which group. And personal participation is available only if the user opts in to define their groups and the related metadata permissions themselves.

In the paper “DSPM: A Platform for Personal Data Share and Privacy Protect Based on Metadata” author proposes a prototype that improves the probability of data discovery and the ability of the individual to control the data to balance the sharing and privacy. The data owner sends data without metadata to a centralized server where everyone can

access and if anyone requests the metadata their request is looked into by the data owner and he/she can decide what metadata they want to send to the requester. This model only works if the data owner has some idea of security. **[6]**

Summary :

Avatars of current innovation are identifying systems and methods and removing metadata and hidden information in files. One aspect of innovation relates to computing systems, including the multiplicity of files. Computing systems may have a review module configured to interpret the extensible markup language component of the file to identify metadata and hidden information, and a scrubber module to remove metadata and hidden information from the file.
Another aspect of the invention relates to the method of removing metadata and hidden information from a file linked to the electronic mail, including: Check the file type of the file to see if the file type is detected; If the file type is detected, search for metadata and hidden information by interpreting the expandable markup language portion of the file; Warn the user if metadata and hidden information are found; And scrubbing the file to remove metadata and hidden information. Another aspect of innovation relates to the method of automatically deleting metadata and hidden information from a file uploaded to the server, including: uploading the file to the server; After uploading, check the file type to see if the file type is detected; If the file type is detected, search for the metadata and hidden information in the file; Deleting metadata and hidden information; And making the file available on the server.

Metadata is also gaining rapid attention in data visualization and irrational / interactive data mining research Industry. However, despite a lot of work in the field of advanced data mining techniques and visualizations, integrated Methods that solemnly combine interactions with advanced visualization and / or data mining techniques are rare. A framework based on forced randomization, which allows users to identify high-dimensional data through information-based information. Two dimensional data visualization: Users are presented with 'interesting' expectations, allowing users to express their comments using visual interactions that update the background pattern indicating the user's trust status. From this background pattern, The data is guided by a projection-search

algorithm that uses randomization to calculate the new-interesting 'projection'. By providing this, we increase the likelihood that users will hit new ideas with information that contradicts the background pattern. **[1]**

Technique:

The strategy used by this tool is to process all the metadata that can be removed: any piece of the file that is not a data, and can be removed, is considered as a threat and so is deleted. The tool’s output is short on purpose, since it is intended for non-technical people. The goal of the "metadata listing" functionality is to give a global view of present compromising metadata.

Metadata fields are suppressed when possible. Otherwise, numerical data are set to 0, dates to Epoch, and strings to an empty string. Filling fields with random values or real-looking ones may seem to make sense. **[2]**

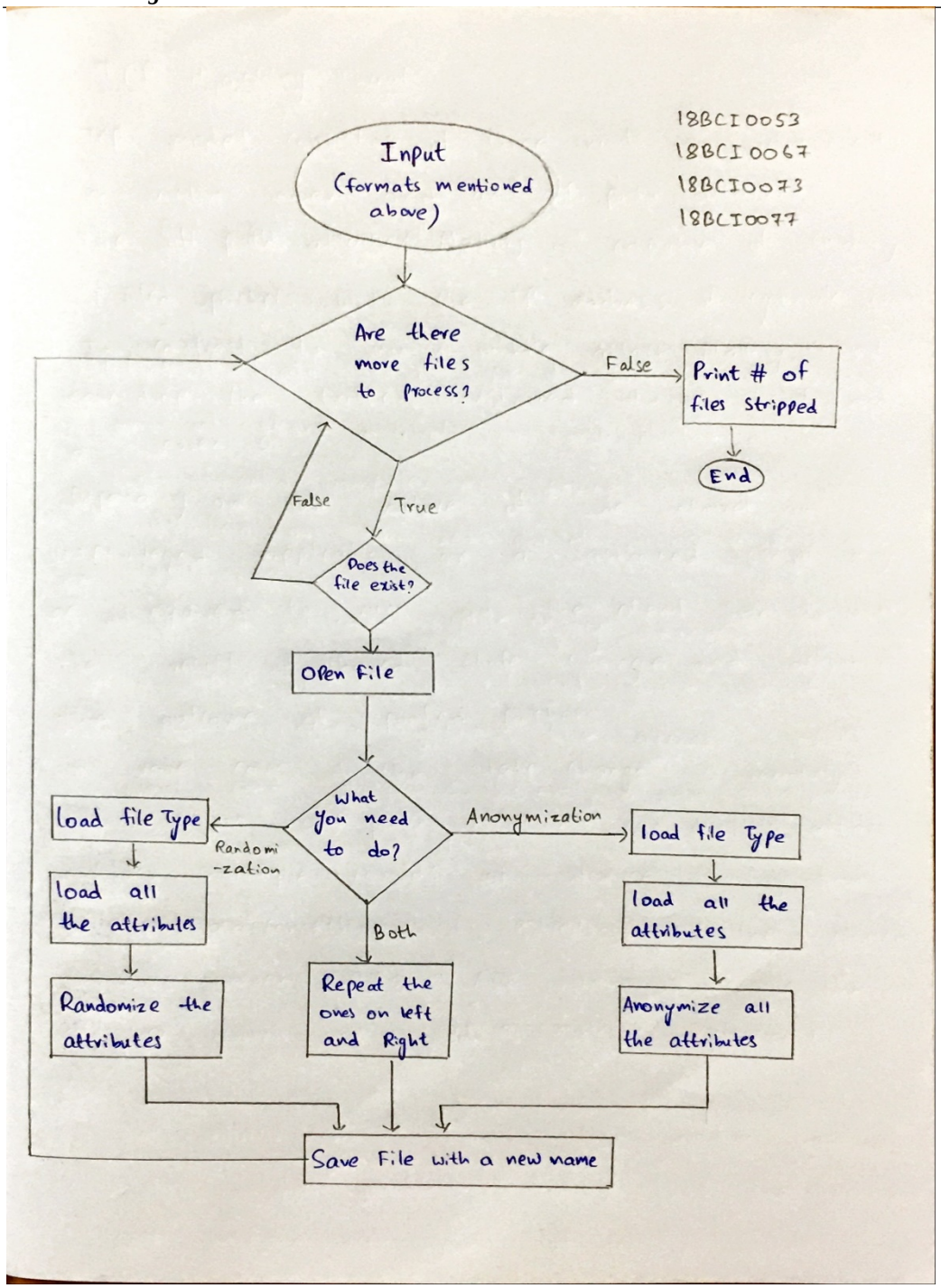

Fig 1: the flow chart for the proposed technique

References:

[1] B. Kang, K. Puolamaki, J. Lijffijt, and T. De Bie, “A Constrained Randomization Approach to Interactive Visual Data Exploration with Subjective Feedback,” *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1–1, 2019, doi: 10.1109/tkde.2019.2907082.

[2] D. Hwang, J. Shin, J. Kim, and Y. Paek, “Data Randomization for Multi-Variant Execution Environment,” *Proc. - 2019 Int. SoC Des. Conf. ISOCC 2019*, pp. 291–292, 2019, doi: 10.1109/ISOCC47750.2019.9027747.

[3] S. Thompson and M. Reilly, “Embedded Metadata Patterns Across Web Sharing Environments,” *Int. J. Digit. Curation*, vol. 13, no. 1, pp. 223–234, 2018, doi: 10.2218/ijdc.v13i1.607.

[4] S. Tayeb *et al.*, “Toward metadata removal to preserve privacy of social media users,” *2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018*, vol. 2018-January, pp. 287–293, 2018, doi: 10.1109/CCWC.2018.8301741.

[5] B. M. J. Hannon, “Metadata The Importance of Metadata in Digital Evidence for Legal Practitioners Metadata,” vol. 34, no. 10, 2017.

[6] X. Dong, B. Guo, X. Duan, Y. Shen, H. Zhang, and Y. Shen, “DSPM: A platform for personal data share and privacy protect based on metadata,” *Proc. - 2016 13th Int. Conf. Embed. Softw. Syst. ICESS 2016*, pp. 182–185, 2017, doi: 10.1109/ICESS.2016.10.

[7] W. Dai, L. Chen, M. Qiu, A. Wu, and B. Chen, “A Privacy-Protection Data Separation Approach for Fine-Grained Data Access Management,” *Proc. - 2nd IEEE Int. Conf. Smart Cloud, SmartCloud 2017*, pp. 84–89, 2017, doi: 10.1109/SmartCloud.2017.20.

[8] J. Lepsoy, S. Kim, D. Atnafu, and H. J. Kim, “Metadata protection scheme for JPEG privacy & security using hierarchical and group-based models,” *2015 5th Int. Conf. Inf. Commun. Technol. Access. ICTA 2015*, 2016, doi: 10.1109/ICTA.2015.7426905.

[9] B. Toevs, “Processing of metadata on multimedia using exiftool: a programming approach in python,” *Proc. - 2015 Annu. Glob. Online Conf. Inf. Comput. Technol. GOCICT 2015*, pp. 26–30, 2016, doi: 10.1109/GOCICT.2015.14.

[10] R. Parekh, R. Armañanzas, and G. A. Ascoli, “The importance of metadata to assess information content in digital reconstructions of neuronal morphology,” *Cell Tissue Res.*, vol. 360, no. 1, pp. 121–127, 2015, doi: 10.1007/s00441-014-2103-6.

[11] R. Forest and D. Grove, “(19) United States (12),” vol. 1, no. 19, 2013.

[12] J. Heurix, M. Karlinger, and T. Neubauer, “Pseudonymization with metadata encryption for privacy-preserving searchable documents,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 3011–3020, 2012, doi: 10.1109/HICSS.2012.491.

[13] B. Greschbach, G. Kreitz, and S. Buchegger, “The devil is in the metadata - New privacy challenges in decentralised online social networks,” *2012 IEEE Int. Conf. Pervasive Comput. Commun. Work. PERCOM Work. 2012*, no. March, pp. 333–339, 2012, doi: 10.1109/PerComW.2012.6197506.

[14] P. Examiner and T. Dinh, “(12) United States Patent,” vol. 2, no. 12, 2009.

[15] S. Marinai, “Metadata extraction from PDF papers for digital library ingest,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 251–255, 2009, doi: 10.1109/ICDAR.2009.232.

[16] T. Aura, T. A. Kuhn, and M. Roe, “Scanning electronic documents for personally identifiable information,” *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 41–49, 2006, doi: 10.1145/1179601.1179608.

[17] D. Filtration, “Digital Filtration,” pp. 285–287, 2006.

[18] I. Lazarovich and Y. Nikolaychuk, “Method of randomization and its application for adaptive data compression,” *Proc. 2nd IEEE Int. Work. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS 2003*, pp. 362–364, 2003, doi: 10.1109/IDAACS.2003.1249587.

[19] V. Raman and J. M. Hellerstein, “Potter’s wheel: An interactive data cleaning system,” *VLDB 2001 - Proc. 27th Int. Conf. Very Large Data Bases*, pp. 381–390, 2001.