IMPORTANCE OF FILE DOCUMENT METADATA IN COMPUTER FORENSICS

Aashish Kumar Purohit[1], Naveen Hemrajani[2], Ruchi Dave[3]

[1]SGVU, Jaipur, India

[2,3]SGVU,Jaipur, India
Email: [1]purohit2aashish@gmail.com

ABSTRACT

The metadata has wide range of applications in real world. But the importance of metadata in computer forensic is quite large. In this paper we will discuss what kind of information exists on current common document file types and how it can be useful for a computer forensic expert to reach at the guilty person. Also we will list the famous cases in which computer related crime is solved using document metadata information. A brief introduction of current status of cyber law of India is listed in this paper.

Key words: metadata, computer forensics, cyber law, office document, pdf.

I. INTRODUCTION

Metadata is structured data which describes the characteristics of a resource. The "meta" term originates from the Greek word indicating a character of a higher order or more elementary kind. A metadata record consists of numerous embedded elements representing precise attributes of a resource, and each element can have one or more values. It is used for two essentially different concepts or types. Although the expression "data about data" is frequently used, it does not apply to both in the same way. Structural metadata, the design and specification of data structures, cannot be about data, because at design time the application does not contain any data. In this case the correct description would be "data about the containers of data". Descriptive metadata, on the other hand, is about individual examples of application data, the data content. In this case, a useful explanation would be "data about data contents" or "content about content" thus metacontent[1]. Metadata may appear safe to store within the document; however, it can be a convicting piece of proof against the document's publisher, owner, reviewer, author, and may offer information concerning the network storage place as well as the unique identifier of the computer on which the document was created.

At a fundamental level, computer forensics is the investigation of information contained within and created with computer systems and computing devices, normally in the interest of finding out what happened, when it happened, how it happened, and who was involved in it. Metadata can disclose sensitive information not supposed for public use, which can be destructive to a company or individual, but at the same time offer a lot of useful information for a forensic investigator. For example, forensic researcher investigating a case of unsuitable material may be able to track down the original owner of the file by simply examining its metadata.

Investigators can use the metadata stored with documents to help in the overall forensic analysis process. This paper will address one small portion of the document forensics process and focus on the detection, analysis, and utility of metadata saved with common document formats used these days. Also it will look at the current cyber law of India.

II. FUNCTIONAL METADATA

A. Risks (User) / Advantages (Investigator)

Metadata stored in documents can reveal important information about the document relating to the investigator. Though, many features provided by Microsoft Office, OpenOffice.org and PDF documents can create problems for a user but assist a forensic researcher in the performance of his or her duties. The Track Changes feature of Microsoft Office and OpenOffice.org will be discussed in this section as well as the commenting feature, which is common to Microsoft Office, OpenOffice.org, and PDF documents. Two extra features of Microsoft Office which provides useful information are the metadata stored with Macros and Fast-Saves. In addition, Older versions of Microsoft Office used to store computer specific information, is also included in discussion, since it revealed much

more user-specific information. PDF documents also contain metadata which is considered as incriminating information.

Track Changes and Commenting

Microsoft Office and OpenOffice.org both provide the Track Changes and commenting features that are significant features when used correctly and cautiously. Neither feature is enabled by default in both the products; though, use of these features can give valuable information for a forensic specialist. The Track Changes feature enables us to view the history of all the changes made to a file. If this feature is left enabled by the user, all viewers of the document will have the potential to view all changes made since the last round of changes were accepted by the document creator [3]. Both products (Microsoft Office and OpenOffice.org) commenting system similarly can disclose the same type of information. Through this mechanism users can make comments to the documents without affecting the data contents. However, if those comments are left in, each subsequent viewer can review all the comments made by document reviewers. Excel and PowerPoint do not warn a user of the comments that are embedded in a document and it is a little bit of worry [3]. The name of the reviewer is also stored with comments or changes, if any, made to the document while using the track changes and commenting feature.

It can be highly disgusting to publicly release this information, as a company or a firm may not want others to know every person who has reviewed the document for release. As stated previously, the document properties can disclose a great amount of information. The name of the document creator is included in both Microsoft Office and OpenOffice.org documents. If this document is used again as a template for another document, the creator's name may remain the same; however, the author may not. In this case, a published document may come from one organization, but in truth started off from another totally separate entity.

Macros and Fast-Saves

Macros which are used in office items can be very useful and also can raise productivity in many situations. However, Microsoft Office also provisions an extra bit of information to any macro used inside of the document, spreadsheet, or presentation – the name of the macro author. Fast-saves feature is also given in Microsoft Office, which is a suitable way for the user to make sure that if the computer crashes, a recent backup is just a click away. This can give significant information to a forensic investigator. Similar to other metadata, changes saved during a fast save can disclose sensitive information to an investigator when viewed using a text or hex-editor. In the electronic files text can remain after the deletion. According to the Gartner Group's Research Note on Metadata in Office, "users do not remember that metadata exists when they send the document to any other person. Some metadata is never noticeable, such as portions deleted by users but not really deleted by Microsoft Office when operating with fast save turned on" [3].

GUID

A unique identifier was involved in Microsoft Office 97, that could identify the system and installation on which the document was generated. Microsoft's dispute for involving this feature in the file was to provide the cooperation with third-party programs [4]. Microsoft further confirmed that the Globally Unique Identifier (GUID) number, which has been defined in RFC 4122 [5], could not be used to "discover the author of a document without thorough knowledge of the personal computer on which the document was originally created." Though, the method used to create the GUID proved significant during the hunt for the author of the Melissa virus.

Ethernet Media Access control (MAC) address is used in the generation of GUID number. Since MAC addresses are unique and using the GUID number an researcher can produce evidence directly related to the computer with which the original document was created. But this GUID number also contains a weakness, that is it is only stored once when a document is created by the user. Therefore it will always contain the GUID of the original document creator whether it is modified by the other user or not.

PDF documents

Documents created in PDF format are usually used as a process of distributing documents in a universal format readable cross-platform. Features given by the PDF documents also present a challenge to the forensic researcher and a certain amount of danger to the user. Plug-in involved in Adobe can work flawlessly with OpenOffice.org and Microsoft office products. This flawless feature is of a huge benefit to the user, but can also produce a challenge regarding

stored metadata. When a document is transformed to a PDF document, all the metadata that was stored with the original document such as creator name, version, as well as the change tracking information is also stored in the PDF document. Also the other risks discussed above regarding the commenting and track changes features remains in the PDF documents. Commenting feature is also associated with Adobe that can add huge amount of wealth of metadata stored in the documents.

The benefits achieved even if the user want not to copy over to source document's metadata when performing the file conversion would be canceled if the user did not understand the capabilities and risks related with storing the source document with the PDF document.

B. METADATA AND REAL LIFE COMPUTER FORENSIC CASES

The BTK Killer

Dennis Rader, also well-known as the BTK killer, which stands for bind, torture, and kill, started his killing chain in 1974 in Wichita, Kansas. He continued to live a deceiving life of a married father of two children and president of church until his arrest in 2005. He was responsible for 10 murders. There were DNA evidence and also a witness left at the crime scenes but the police and law enforcement were not able to capture him until 2005.

After 30 years of deceiving life his murder chain ended in 2005 due to his casualness and lack of knowledge of computer forensic. He sent an email to a Wichita TV station about his crime. After inspecting the file, police and forensic experts were able to identify the author, Dennis and the organization, Christ Lutheran Church, from the metadata involved in file. Upon more investigation more details about the church, forensic experts were able to discover Dennis Rader, who was the president of the church [6].

Merck Report

Merck is a pharmaceutical company which has been involved in various charges related to its arthritis drug medication. In 2005, important information regarding the medication's danger of causing heart attacks was deleted from a document sent by Merck was discovered by the New England Journal of Medicine. The track changes feature was used by the Merck while preparing the informational paper to be sent to the journal. The authors at Merck did not remember to accept the changes made to the original document and therefore released the document with all the metadata intact. Upon detail inspection, the Journal discovered the deleted text and after that released the information regarding the Merck's blunder [7].

DNC

A document originated from the Democratic National Committee (DNC) in October 2005 with a few not so enjoyable things to say about Supreme Court candidate Judge Samuel Alito. After detail inspection of the document's metadata, a few of the author's names were exposed as well as the creation date of the document. In particular, two user-ids were discovered and associated with two members of the DNC and the document was created right after Justice O'Connor resigned, meaning that the document was created before Judge Alito was nominated [2]

Weapons of Mass Destruction Report

One of the most famous blunders that have taken place in modern years was brought to light in the year 2003. The United Kingdom courts summoned documents from its government on the topic of Iraq's weapons of mass destruction program. Among the documents submitted to the courts was one that emerged to be genuine and made a case for Hussein's ownership of these weapons. However, upon further examination, it was discovered a large part of the document was actually taken from a twelve year old PhD thesis [8].

The Infamous Melissa Virus

Melissa was originated in the early 1999. Melissa was a micro virus that mass-mailed itself to the first 50 addresses found in the Microsoft Outlook Address Book. It was initially circulated on a newsgroup (alt.sex), this macro virus quickly spread in March 1999. The Melissa virus was spread in a Microsoft word file called list.doc, which consisted of passwords for X-rated websites. If any user download that file and open it in Microsoft Word then a macro inside the document will execute and e-mail the list.doc file to other users. Because of its widespread impact the federal government looked to investigate and prosecute the creator of the virus. It was very difficult task but it was supported by two independent experts who discovered some destructive information [9].

These experts looked at the metadata of the document containing the virus and exposed the GUID. The experts carried out more research and at last located a website belonging to a malicious hacker, on which other documents with the same GUID were exposed. In the end, the discovered information was passed to ZDNet, who published several articles and the federal authorities. This valuable information helped in efforts to find and convict David Smith of Aberdeen, for creating the virus.

III. METADATA IN MICROSOFT OFFICE

The most widely used office product today is Microsoft Office. Because of this fact, the metadata that is used by its ingredient products, Word, Excel, and PowerPoint, has been more and more inspected in previous years and expansively addressed by Microsoft. According to Microsoft's Knowledge Base [10], the following data can be saved as hidden information inside of Microsoft Office documents, spreadsheets, and presentations

- name
- initials
- company or organization name
- The name of computer
- The name of the network server or hard disk
- Other file properties and summary information
- Non-visible portions of embedded OLE objects
- The names of previous document authors
- Document revisions
- Document versions
- Template information
- Hidden text or cells
- Personalized views
- Comments

A designation was made to distinguish the information that may be exclusive to each Microsoft Office program. Where the metadata type applies to all programs, the designation reflects that fact. Information such as this can prove very useful to a forensic expert as well as any other person involved in examination activities. In both the cases, a person could possibly find out the author of the document, as well as his or her user-id, and the network path location of the original document. A small part of network's layout can be mapped using this information. Through the use of track changes and comments features in Microsoft Word, Excel, PowerPoint cooperation among document authors is also facilitated. If a document author publishes a document without remembering to delete all comment and changes from the document, then he or she could expose all the contributors to the documents as well as its original data; all of which could lead to the disclosure of data that was not meant to be seen by anyone, much less the forensic expert.

These types of services can be very useful to the document authors. It can facilitate teamwork among several authors as well as it allows the reader to take notice of the creator. All information was stored in proprietary binary format in the versions prior to Microsoft Office 2003. Therefore the user had no means to know what metadata was stored in the document. Microsoft Office 2003 and its successor's offers the capability to save the metadata in more widely known and accepted XML format, but the binary format is still the default format and the users which are non- technical may not be aware of the total capabilities of the newer format.

IV. METADATA IN OPEN OFFICE

OpenOffice.org's office product is gaining status day by day and it is a major challenger of Microsoft Office product. OpenOffice.org has taken a different concept in its file format. OpenOffice.org stores metadata and as well as document data in a chain of XML files that are associated with the name of the document given by the user. OpenOffice.org outlines in its XML file format arrangement [11] the capacity to store a lot of metadata inside of its documents, spreadsheet, and presentations to include:

- Generator of the document
- Title of the document
- Description
- Subject
- Keywords
- Initial Creator
- Creator
- Printed By
- Creation Date and Time
- Modification Date and Time

- Print Date and Time
- Document Template
- Language
- Editing Duration
- User-defined Metadata
- Document Statistics

According to the specification document [11], all of the above modules are stored, or have the potential to be stored, with each document, spreadsheet, and presentation created in OpenOffice.org. This information is quite similar to that stored in Microsoft Office; however, OpenOffice.org does not make guesses about what personal information the user would like related with the file. For example, if a user does not enter a name in the application's user information widget, the program will not attempt and get the registration information from the operating system to facilitate the name or initials of the user. OpenOffice.org also has the capability to store information facilitating the use of the Track Changes service and commenting function. OpenOffice.org offers easier access to a files metadata, which provides user and examiner discovery, given he or she knows what they are looking for. OpenOffice.org stores the metadata associated with the overall document, spreadsheet, or presentation in a separate plain text.

V. PDF METADATA

PDF document stores metadata also that, as stated above, when a document is transformed to a PDF document, all the metadata that was stored with the original document such as creator name, version, as well as the change tracking information is also stored in the PDF document. The types of data and format are outlined in the PDF Reference Manual [12] version 1.6 specify that the information is stored in a document information library. Furthermore, the information is to be found inside of an optional Info entry in the trailer of the PDF file. PDF document has the capability to store much of the same information as the products described above, which includes:

- Title
- Author
- Subject
- Keywords
- Creator
- Producer
- CreationDate
- ModDate

Adobe Systems has also joined in the regulation of the metadata stored in documents. This feature will facilitate greater relieve in the searching and storage of the information in databases as well as web-based environments. This attempt is named as Extensible Metadata Platform (XMP) [13] and one definite function is to allow various programs that process Adobe files to add their own categories of metadata. The nature of information stored as metadata will not change, but only the format in which it is stored will change. It will be provide ease of access by having XML type format.

VI. INDIAN CYBER LAW-LACKING BEHIND

A considerable development regarding the computerization of traditional courts and their procedures has been realized in India. Now many important features regarding Indian litigation like case status, case list, case status, certified copies, etc are available online. This has also significantly reduced the accumulation of cases in India.

On the other hand, in spite of this growth India has not succeeded on almost all other faces which include E-courts (Electronic courts), online dispute resolution (ODR) functionality, good cyber law, good laws concerning cyber forensics, etc. For instance, till July 2011 we are still waiting for the establishment of first Electronic Court in India. We also do not have any Online Dispute Resolution (ODR) mechanism in India; we also have a criminal friendly and outdated cyber law. Also the Information technology act, 2000 requires urgent revision; we have no laws regarding cyber forensics in India. In short, legal enablement of ICT (Information and Communication Technology) systems in India has failed so far.

A latest development about legal and judicial reforms pertains to National Litigation Policy of India (NLPI). Law Minister of India has initiated this policy and very soon the same may be completed. However, NLPI also not succeeded to consider legal enablement of ICT systems in India accurately. It failed to address the requirements of Electronic courts and ODR mechanisms [14].

The Bar Council of India (BCI) is the major authority that is accountable for improving standards of legal professionals in India. Till now, BCI has been failed to materialize its purpose of providing scientific and technical professional training to lawyers in India. The problem seems to be short of management and skill in this regard in India. Even Law Ministry of India has failed to contribute considerably in this direction. There is also lack of proper cyber law and cyber forensics training for law enforcement agencies of India.

On the other hand, other countries are working really hard in this regard. Take the example of U.S. Attorney's. The U.S. Attorney's Office has funded its first "Cyber Crime, Electronic Evidence, and Cyber Security Training" for local and federal law enforcement officials. The aim of the training is to provide law enforcement agencies of US to deal with cyber crimes [15].

VII. CONCLUSION

Metadata is data that describes other data. As discussed in this paper, metadata can lead to the discovery of an overabundance of information. Assessment of document metadata can guide us to the sighting of information such as: document author names; names of contributors as well as their recommended changes and comments; network storage path locations; user-ids of the document author; as well as computer specific information such as the GUID.

A variety of different real-world conditions were presented demonstrating how metadata was exploited to disclose hidden information. In some cases, it was admitted as evidence to refute or support claims. In others, it resulted in an embarrassing situation for companies and governments alike. This paper discussed the types of metadata information stored in documents, spreadsheets, and presentations created in Microsoft Office and OpenOffice.org applications, PDF document. As shown in this paper, a move towards XML file formats has provided forensic investigators easier access to this type of information.

The cyber law of India is very weak and requires urgent revisions and the establishment of E-courts and ODR mechanism is yet to be done. Also there is no provision of scientific and technical professional training to lawyers in India.

ACKNOWLEDGEMENT

We express our sincere and deep gratitude to all the faculty members of our college and also our parents for their valuable help and guidance, which has enabled us to complete this paper.

REFERENCES

[1] Metadata – Wikipedia, http://en.wikipedia.org/wiki/Metadata

[2] Byers, S. "Information Leakage Caused by Hidden Data in Published Documents," IEEE Security and Privacy Magazine. March/April 2004.

[3] "Dangers of Document Metadata", http://www.metadatarisk.org/ document_security/dangers_of_docmetadata_overview.htm.

[4] Microsoft Knowledgebase Article 222180, "How and why unique identifiers are created in Office documents," Revision: 2.2 January 24, 2007, http://support.microsoft.com/kb/222180/

[5] Leach, P., Mealling, M. and Salz, R. Request for Comments (RFC) 4122: A Universally Unique Identifier (UUID) URN Namespace, July 2005, http://www.ietf.org/ rfc/rfc4122.txt.

[6] Use of computer forensics technology crime investigation by Hyechin Blakeslee, http://acsupport.europe.umuc.edu/~sdean/ProfPaps/Bowie/S09/Blakeslee.pdf

[7] Ewalt, D."When Words Come Back from the Dead," http://www.forbes.com/2005/12/13/microsoft-word-merck_cx_de_1214word.html?partner=yahootix.

[8] Matt Loney, "'Dodgy-dossier syndrome' rife in the workplace" ZDNet UK. http://www.zdnet.co.uk/news/systems-management/2003/11/14/dodgy-dossier-syndrome-rife-in-the-workplace-39117905/

[9] "Melissa Creator Sentenced" About.com http://antivirus.about.com/library/weekly/aa050102a.htm

[10] Microsoft Knowledgebase Article ID 223396, "How to minimize metadata in Office documents," Revision3.4,http://support.microsoft.com/kb/223396

[11] Sun Microsystems, OpenOffice.org XML File Format 1.0 Technical Reference Manual, Version 2, http://xml.openoffice.org/xml_specification.pdf.

[12] Adobe Systems Incorporated, PDF Reference- Adobe Portable Document Format, 5th ed., Version 1.6,http://partners.adobe.com/public/developer/en/pdf/PDFReference16.pdf.

[13] Adobe Systems Incorporated, "Extensible Metadata Platform (XMP) homepage", http://www.adobe.com/products/xmp/

[14] Legal Framework for Information Society in India,http://barexamsindia.noads.biz/blog/?cat=40

[15] Legal and Judicial Fraternity of India Needs Scientific Knowledge, http://barexamsindia.noads.biz/blog/?cat=40

Aashish Kumar Purohit is a M.Tech. scholar at Suresh Gyan Vihar university . His area of interest includes metadata, computer forensic, data mining, and software engineering.