Forensic authenticity examination of PDF documents

Jinhua Zeng[a], Xiulian Qiu[b*]
[a] Academy of Forensic Science, Shanghai 200063, China; [b] Forensic Science Center, East China University of Political Science and Law, Shanghai 200042, China

ABSTRACT

With the rapid development and widespread popularization of information technology, Portable Document Format (PDF) documents have gradually become a type of digital files that are easily accessible and closely related to the public. In digital forensics, digital files in the form of PDF documents are often encountered, and their authenticity needs to be determined. Therefore, the research on its key forensic techniques has important theoretical research significance and practical application value. However, through searching the literature, we can find that there is still a lack of systematic research on forensic authentication of PDF documents. Based on the above situation, this paper first studies the file structure and digital composition of PDF documents; The digital data characteristics and the contents for examination of PDF files produced by the mainstream scenarios are studied, including PDF generated by scanners, converting directly from images, and transforming from DOCX documents, etc. Finally, a case study is carried out to explain the key technologies and examination contents for authenticity examinations of PDF documents in detail. Our study will provide theoretical basis and practical guidance for forensic investigations of authenticity examinations of PDF documents.

Keywords: PDF, forensic authenticity examination, digital data, metadata

1. INTRODUCTION

With the rapid development and widespread popularization of information technology, office electronic documents have gradually become digital files that are easily accessible and closely related to the public. Among them, PDF document is a type of digital file that is independent of operating system platform and is a formal international standard, all of these advantages make it become an ideal carrier of electronic document distribution and digital information dissemination. PDF is first proposed by Adobe, in which the multimedia information, such as text, sound and image, and can cover hypertext links, typeset style and others, can be effectively integrated. Due to the extensive application in daily life, we often have to determine the authenticity of the PDF files which are submitted in courts to serve as digital evidence.

Many works[1-7] have focused on the forensics of the PDF documents. For example, Adhatarao and Lauradoux[8] used the coding styles of the PDF documents to identify the PDF producing tools. Through literature search, it is found that more attention is paid to the data recovery of PDF files[9, 10], there is still a lack of systematic study on the forensic authenticity examination of PDF documents.

Based on the above situation, firstly we discuss the PDF document format in the paper, and then the problems of the forensic authenticity examination of PDF documents generated by the common ways are further studied, including the PDF files generated by scanner devices, PDF files transformed from images directly, and the ones derived from the word processing softwares, such as WPS Office and Microsoft Office. Finally, a case study is carried out to demonstrate the key methods and contents for effective forensic authenticity examination of PDF documents.

2. PDF FILE FORMAT

2.1 PDF file structure

A PDF file transformed from a ".docx" file is used as an example to study the file structure of PDF documents, which is created and then converted into PDF documents by using WPS Office (version 11.1.0.11294).

Our results find that the PDF file mainly consists of four parts, that are file head, file body, cross-reference table and file tail. Its file header is the characters "%PDF-1.x", in which the last digit indicates the version number of the PDF file. The

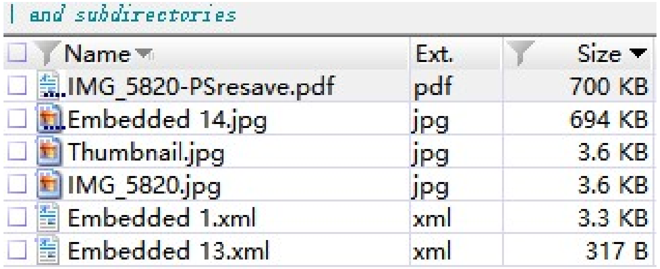
[*] qiuxiulian@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125064Y
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662214

highest version of the PDF updated by Adobe is PDF 1.7 and the subsequent versions are maintained and released by the International Standards Organization (ISO). The PDF file body consists of several “obj” objects, in which the first number “7” is used to uniquely identify the object number, and the second number “0” is used to indicate the number of changes that the object has undergone after being created, which is called the object generation number. The object generation number of a newly created PDF file is 0, indicating that the object has not been modified. The information of each “obj” object is contained between the charactericters “<<”and “>>”, and ends with the keyword “endobj”. The cross-reference table contains the location index information of each “obj” object, starting with the keyword “xref”. The number “0 15” in the second line demonstrates that the object number starts from 0 and there are 15 objects in total. Generally, the third row of the cross-index table is fixed characters “0000000000 65535 F”, in which the first field “0000000000” indicates the starting position of the object. The second field “65535” shows the maximum possible object generation number. The third field “f” demonstrateds that the object is a free object, and the character “n” indicates the object is in use and can be modified. In the file tail, the starting keyword is “trailer”. The “%%EOF" is used to indicate the end of the file, in which “/Size 15” indicates the total number of objects in the file, “/Root 1 0 R” indicates that the object number of the root object starts at 1, and the subsequent “startxref 22705” indicates that the offset address of the cross-reference table is “22705”. The access of each object is realized by combining the starting position of the object in the cross-reference table.

2.2 The digital composition of PDF files

We use a PDF file transformed directly from an image by using Adobe Acrobat X (version 10.1.16) as an example. The file size of the image is 707,352 bytes and the file name is “img_5820-psresave.jpg”. The size of the generated PDF file is 716,945 bytes. The X-Ways Forensics (version 20.0 SR-5 X64) is used to analyze its digital components. The file signature is used to analyze the different types of data embedded in the PDF file. The digital components of the above PDF file are shown in Figure 1. As we can see in Figure 1, the objects of the PDF file mainly consist of XML data and JPG images. In addition, a thumbnail image is embeded in the JPG image.



Name	Ext.	Size
IMG_5820-PSresave.pdf	pdf	700 KB
Embedded 14.jpg	jpg	694 KB
Thumbnail.jpg	jpg	3.6 KB
IMG_5820.jpg	jpg	3.6 KB
Embedded 1.xml	xml	3.3 KB
Embedded 13.xml	xml	317 B

Figure 1. The digital composition of the PDF file in the sample.

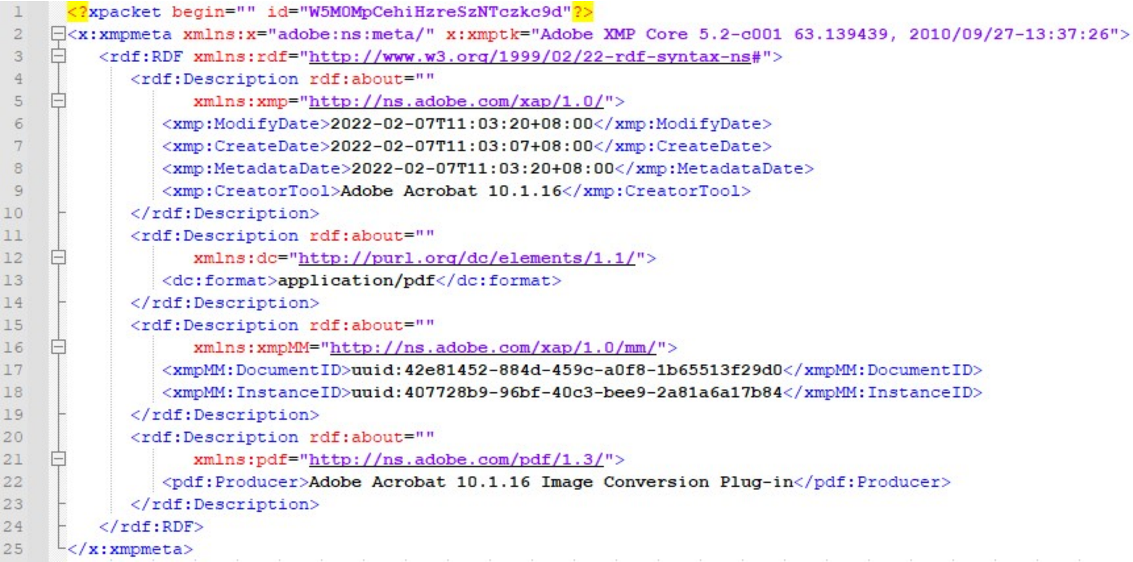
The XML data information viewed in Notepad++ (version 8.1.4) is shown in Figure 2. It can be found that the XML data mainly contains metadata information of PDF files, such as the PDF making program, creation time and modification time, among which the modification time is the saving time of the PDF file.

Comparing the digital data of the original image and the JPG image parsed from the PDF file with Beyond Compare 4 (version 4.1.2) software. we find that the metadata information in the header of the original image is basically preserved, but the digital data of the main body differ greatly, which may have been re-encoded.

3. FORENSIC AUTHENTICATION OF PDF FILES

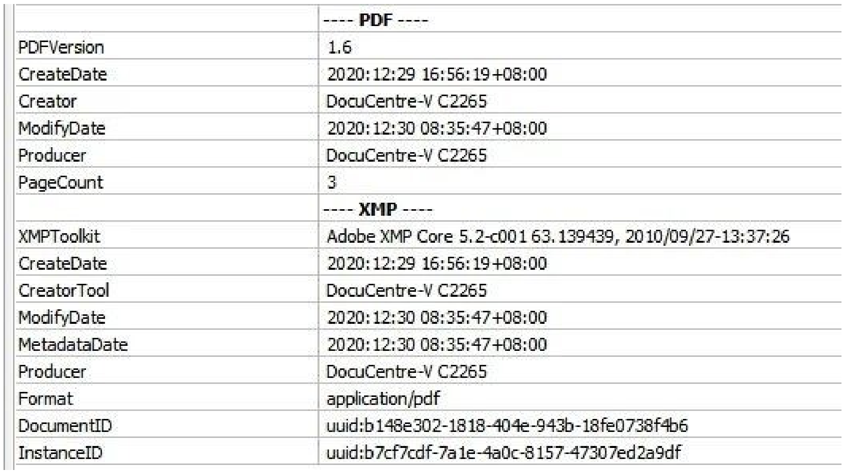
3.1 Forensic authenticity examination of PDF files generated by scanners

Due to the feature of optical transfer printing, the original digital data of images cannot be retained in the PDF files generated by scanners. Therefore, the effective contents for forensic authenticity examination of PDF files produced by scanners are mainly focused on the metadata information of the PDF files. In this paper, the EPSON Chops V330 Photo scanner and Fuji Xerox Docucente-V C2265 MFP are taken as examples to study the relevant points for examination. The results show that most metadata information of PDF files generated by EPSON Chops V330 Photo scanner was blank, and only PDF production program and PDF version is saved. The metadata information of PDF files produced by Fuji Xerox DocuCentre V C2265 all-in-one printer is relatively rich, as shown in Figure 3. In addition to basic information about the PDF maker, it also contains information about XMP-related fields, such as creation time and modification time.



```
<?xpacket begin="" id="W5M0MpCehiHzreSzNTczkc9d"?>
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="Adobe XMP Core 5.2-c001 63.139439, 2010/09/27-13:37:26">
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
      <rdf:Description rdf:about=""
            xmlns:xmp="http://ns.adobe.com/xap/1.0/">
         <xmp:ModifyDate>2022-02-07T11:03:20+08:00</xmp:ModifyDate>
         <xmp:CreateDate>2022-02-07T11:03:07+08:00</xmp:CreateDate>
         <xmp:MetadataDate>2022-02-07T11:03:20+08:00</xmp:MetadataDate>
         <xmp:CreatorTool>Adobe Acrobat 10.1.16</xmp:CreatorTool>
      </rdf:Description>
      <rdf:Description rdf:about=""
            xmlns:dc="http://purl.org/dc/elements/1.1/">
         <dc:format>application/pdf</dc:format>
      </rdf:Description>
      <rdf:Description rdf:about=""
            xmlns:xmpMM="http://ns.adobe.com/xap/1.0/mm/">
         <xmpMM:DocumentID>uuid:42e81452-884d-459c-a0f8-1b65513f29d0</xmpMM:DocumentID>
         <xmpMM:InstanceID>uuid:407728b9-96bf-40c3-bee9-2a81a6a17b84</xmpMM:InstanceID>
      </rdf:Description>
      <rdf:Description rdf:about=""
            xmlns:pdf="http://ns.adobe.com/pdf/1.3/">
         <pdf:Producer>Adobe Acrobat 10.1.16 Image Conversion Plug-in</pdf:Producer>
      </rdf:Description>
   </rdf:RDF>
</x:xmpmeta>
```

Figure 2. The content of XML data parsed in the PDF file.



	---- **PDF** ----
PDFVersion	1.6
CreateDate	2020:12:29 16:56:19+08:00
Creator	DocuCentre-V C2265
ModifyDate	2020:12:30 08:35:47+08:00
Producer	DocuCentre-V C2265
PageCount	3
	---- **XMP** ----
XMPToolkit	Adobe XMP Core 5.2-c001 63.139439, 2010/09/27-13:37:26
CreateDate	2020:12:29 16:56:19+08:00
CreatorTool	DocuCentre-V C2265
ModifyDate	2020:12:30 08:35:47+08:00
MetadataDate	2020:12:30 08:35:47+08:00
Producer	DocuCentre-V C2265
Format	application/pdf
DocumentID	uuid:b148e302-1818-404e-943b-18fe0738f4b6
InstanceID	uuid:b7cf7cdf-7a1e-4a0c-8157-47307ed2a9df

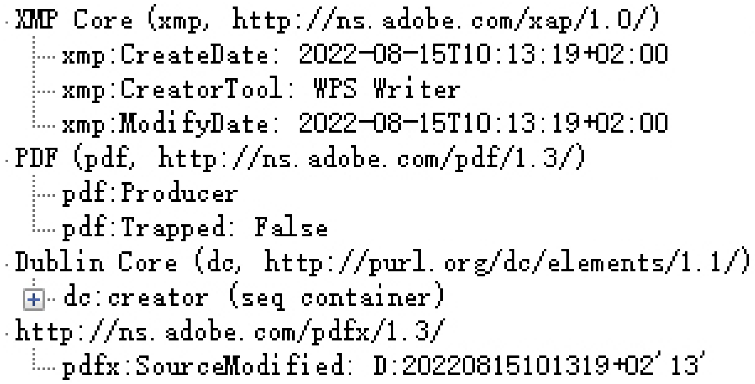
Figure 3. Metadata information of PDF files produced by Fuji Xerox DocuCentre-V C2265 MFP.

3.2 Forensic authenticity examination of PDF files converted from images

As mentioned in section 2.2, most of the file header data of the original image is retained in the process of the convertion from images into PDF documents by using Adobe Acrobat software. Therefore, in addition to the examination of the metadata information of the PDF file, the metadata information of digital images can be effectively examined according to the technical specification for forensic digital image metadata examination[11], and so on. Moreover, image data information in PDF files can be extracted and resaved as independent image files, and then the contents of imaging features, processing traces can be used for forensic authenticity examination according to related technical specification for forensic image authenticity examination[12, 13].

3.3 Forensic authenticity examination of PDF files transformed from DOCX documents

In the test, DOCX documents created by WPS Office (version 11.1.0.11294) and Microsoft Office 2019 are both studied. One image is embedded in the DOCX document, which is then converted into PDF documents by the built-in PDF converter in WPS Office and Microsoft Office. The results show that when the image is embeded in the DOCX documents, it will be re-compressed and re-encoded. The original data in the file header and the end of the file are basically erased. Therefore, forensic authenticity examination of PDF files in this case is mainly focused on the examination of the metadata information of PDF files. The metadata information of PDF files produced by WPS Office (version 11.1.0.11294) is shown in Figure 4, which consists of the information, such as PDF generation tool, creation time, modification time, author information and others.



```
XMP Core (xmp, http://ns.adobe.com/xap/1.0/)
  xmp:CreateDate: 2022-08-15T10:13:19+02:00
  xmp:CreatorTool: WPS Writer
  xmp:ModifyDate: 2022-08-15T10:13:19+02:00
PDF (pdf, http://ns.adobe.com/pdf/1.3/)
  pdf:Producer
  pdf:Trapped: False
Dublin Core (dc, http://purl.org/dc/elements/1.1/)
  ⊞ dc:creator (seq container)
http://ns.adobe.com/pdfx/1.3/
  pdfx:SourceModified: D:20220815101319+02'13'
```

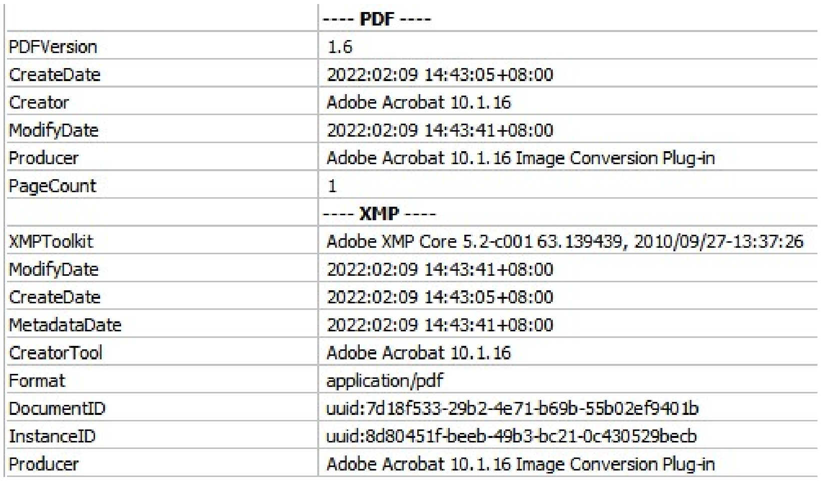
Figure 4. Metadata information of a typical PDF file produced by WPS Office (version 11.1.0.11294).

4. CASE STUDY

In a contract dispute case, it is necessary to verify the authenticity of the contract content which is in the form of PDF document. One party claims that the PDF document is created by scanning the original paper contract, while the other party claims that the original paper contract does not have its own signature, believing that the signature in the PDF document content is synthesized.

In the examination, “Specification for forensic authentication of digital documents (standart No. SF/Z JD0402004-2018)” and “Specification for forensic examination of digital image metadata (No. SF/T 0078-2020)” are followed. We use the tools including X-ways Forensics (version 20.0 SR-5 X64), Adobe Acrobat (version 10.1.16), ExifTool (version 8.02), etc.

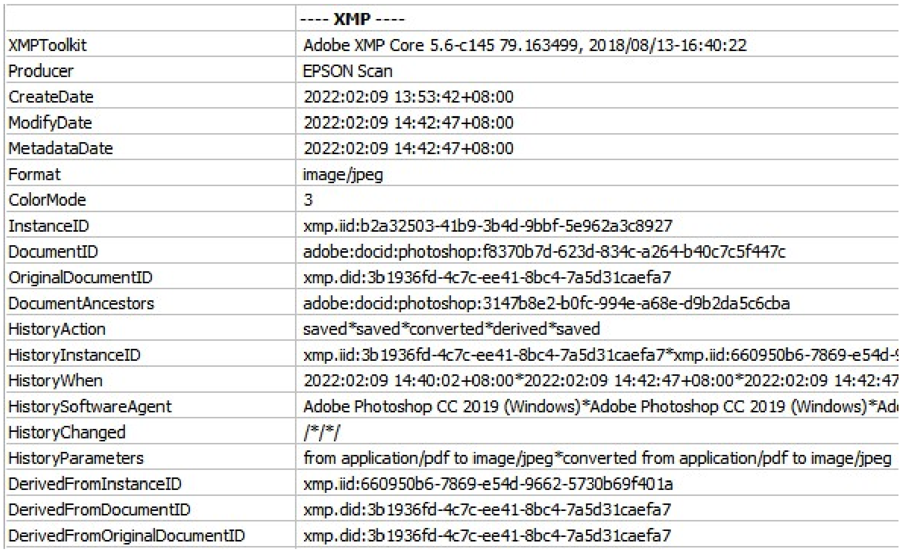
Firstly, the metadata of the PDF file is examined, as shown in Figure 5, which shows that the PDF file is created by using “Adobe Acrobat 10.1.16 Image Conversion Plug-in”. Its creation time is “2020:02:09 14:43:05”, and the modification time is “2020:02:09 14:43:41”.



	---- **PDF** ----
PDFVersion	1.6
CreateDate	2022:02:09 14:43:05+08:00
Creator	Adobe Acrobat 10.1.16
ModifyDate	2022:02:09 14:43:41+08:00
Producer	Adobe Acrobat 10.1.16 Image Conversion Plug-in
PageCount	1
	---- **XMP** ----
XMPToolkit	Adobe XMP Core 5.2-c001 63.139439, 2010/09/27-13:37:26
ModifyDate	2022:02:09 14:43:41+08:00
CreateDate	2022:02:09 14:43:05+08:00
MetadataDate	2022:02:09 14:43:41+08:00
CreatorTool	Adobe Acrobat 10.1.16
Format	application/pdf
DocumentID	uuid:7d18f533-29b2-4e71-b69b-55b02ef9401b
InstanceID	uuid:8d80451f-beeb-49b3-bc21-0c430529becb
Producer	Adobe Acrobat 10.1.16 Image Conversion Plug-in

Figure 5. The metadata of the PDF file.

After further digital analysis of the PDF file, we can find that the PDF file contains an image, which is then extracted and saved as a ".JPG" file. The metadata of the image file is shown in Figure 6, which shows that its production software is "Adobe Photoshop CC 2019 (Windows)", and the modification time is "2020:02:09 14:42:47". In addition, there is a "Photoshop" field information embedded in the metadata.



	---- **XMP** ----
XMPToolkit	Adobe XMP Core 5.6-c145 79.163499, 2018/08/13-16:40:22
Producer	EPSON Scan
CreateDate	2022:02:09 13:53:42+08:00
ModifyDate	2022:02:09 14:42:47+08:00
MetadataDate	2022:02:09 14:42:47+08:00
Format	image/jpeg
ColorMode	3
InstanceID	xmp.iid:b2a32503-41b9-3b4d-9bbf-5e962a3c8927
DocumentID	adobe:docid:photoshop:f8370b7d-623d-834c-a264-b40c7c5f447c
OriginalDocumentID	xmp.did:3b1936fd-4c7c-ee41-8bc4-7a5d31caefa7
DocumentAncestors	adobe:docid:photoshop:3147b8e2-b0fc-994e-a68e-d9b2da5c6cba
HistoryAction	saved*saved*converted*derived*saved
HistoryInstanceID	xmp.iid:3b1936fd-4c7c-ee41-8bc4-7a5d31caefa7*xmp.iid:660950b6-7869-e54d-
HistoryWhen	2022:02:09 14:40:02+08:00*2022:02:09 14:42:47+08:00*2022:02:09 14:42:47
HistorySoftwareAgent	Adobe Photoshop CC 2019 (Windows)*Adobe Photoshop CC 2019 (Windows)*Ad
HistoryChanged	/*/*/
HistoryParameters	from application/pdf to image/jpeg*converted from application/pdf to image/jpeg
DerivedFromInstanceID	xmp.iid:660950b6-7869-e54d-9662-5730b69f401a
DerivedFromDocumentID	xmp.did:3b1936fd-4c7c-ee41-8bc4-7a5d31caefa7
DerivedFromOriginalDocumentID	xmp.did:3b1936fd-4c7c-ee41-8bc4-7a5d31caefa7

Figure 6. The Metadata of the parsed image file in the PDF.

Based on the above results, it can be inferred that the PDF file is first produced by "EPSON Scan", then resaved as an image by using "Adobe Photoshop CC 2019 (Windows)". After that, the "Adobe Acrobat 10.1.16" is used to convert the image file into the final PDF document. Therefore, the statement that the PDF document is directly created by the scanner can not be hold up.

5. CONCLUSION

In this paper, we focus on the topic of forensic authenticity examination of PDF documents. Their file structure and digital composition are studied. The key points for forensic authenticity examination of PDF documents created by the forms of scanning by scanners, conversion from images, and transforming from DOCX docements are discussed. Then, a case study is carried out to describe the the above problem in detail. As we mention above, the wide usage of PDF documents makes PDF ducuments be more and more used as digital evidence in courts, and then we have to determine the authenticity of the submitted PDF documents. Our results can provide some insightful ideas for forensic authenticity examination of PDF docements and can be used as a practical guidance for practical case examination.

ACKNOWLEDGMENTS

This work was supported by Shanghai Science and Technology Commission Project (grant numbers 21DZ2200100), Shanghai Forensic Service Platform (16DZ2290900), and Ministry of Finance, PR China (grant numbers GY2021G-3 and GY2020G-8).

REFERENCES

[1] Castiglione, A., De Santis, A. and Soriente, C., "Security and privacy issues in the portable document format," Journal of Systems and Software, 83(10), 1813-1822(2010).

[2] Aminnezhad, A., Dehghantanha, A. and Abdullah, M. T., "A survey on privacy issues in digital forensics," International Journal of Cyber-Security and Digital Forensics, 1(4), 311-324(2012).

[3] Khitan, S. J., Hadi, A. and Atoum, J., “PDF forensic analysis system using YARA,” International Journal of Computer Science and Network Security, 17(5), 77-85(2017).

[4] Chung, H., Park, J. and Lee, S., “Forensic analysis of residual information in adobe PDF files,” Future Information Technology, Springer, Berlin, Heidelberg, 100-109(2011).

[5] Alanazi, F. and Jones, A., “The value of metadata in digital forensics,” Proc. IEEE 2015 European Intelligence and Security Informatics Conference, 182-182(2015).

[6] Maiorca, D. and Biggio, B., “Digital investigation of PDF files: Unveiling traces of embedded malware,” IEEE Security & Privacy, 17(1), 63-71(2019).

[7] Stevens, D., “Malicious PDF documents explained,” IEEE Security & Privacy, 9(1), 80-82(2011).

[8] Adhatarao, S. and Lauradoux, C., “Robust PDF files forensics using coding style,” IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, Cham, 179-195(2022).

[9] Povar, D. and Bhadran, V. K., “Forensic data carving,” Int. Conf. on Digital Forensics and Cyber Crime, Springer, Berlin, Heidelberg, 137-148(2010).

[10] Sitompul, O. S., Handoko, A. and Rahmat, R. F., “File reconstruction in digital forensic,” Telkomnika, 16(2), 776-794(2018).

[11] The Information Center of Ministry of Justice PRC., “Technical specification for metadata examination of digital images,” (2020).

[12] Photographic Inspection Sub-Technical Committee of National Technical Committee on Criminal Technology of Standardization Administration, “Technical specification of digital image authenticity identification-Image authenticity judge,” (2010).

[13] Bureau of Forensic Expertise, Ministry of Justice PRC., “Technical specification for forensic authentication of images,” (2015).