Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.2, Apr. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.2.193

손상된 OOXML 파일에서의 데이터 추출 고도화 방안 연구*

김 지 윤,[1†] 김 민 수,[1] 박 우 빈,[2] 정 두 원[3‡]
[1,3]성균관대학교 (대학원생, 교수), [2]동국대학교 (대학원생)

Research on Advanced Methods for Data Extraction from Corrupted OOXML Files*

Jiyun Kim,[1†] Minsoo Kim,[1] Woobeen Park,[2] Doowon Jeong[3‡]
[1,3]Sungkyunkwan University (Graduate student, Professor)
[2]Dongguk University (Graduate student)

요 약

디지털 시대의 발전과 더불어, 디지털 자료의 중요성과 그로 인한 디지털 포렌식 수사의 필요성이 증가하고 있다. 그러나 디지털 증거의 수집 및 분석 과정에서 정보저장매체의 손상이나 안티포렌식 등으로 발생하는 주요 문제점인 손상된 파일의 데이터 식별 불가능으로 증거 수집이 원활하게 이뤄지지 않고 있다. 또한, 손상된 파일 복구를 위한 기존 도구의 기술적 한계로 인해 복구에 어려움이 존재한다. 따라서 본 논문은 디지털 자료 생성에 활용되는 손상된 MS Office 파일의 복구 방안을 제시하고자 한다. 복구 방안 제시를 위하여 MS Office 파일 구조인 OOXML 포맷을 분석하고, 기존 복구 도구의 한계를 극복하기 위한 새로운 접근 방식을 제시한다. 이를 통해 손상된 데이터를 보다 효율적으로 복구하고 식별할 수 있는 방안을 마련함으로써 디지털 포렌식 분야에서 증거 수집의 질을 향상하는 데 기여하고자 한다.

ABSTRACT

In tandem with the advancements in the digital era, the significance of digital data has escalated, necessitating an increased focus on digital forensics investigations. However, the process of collecting and analyzing digital evidence faces significant challenges, such as the unidentifiability of damaged files due to issues like media corruption and anti-forensic techniques. Moreover, the technological limitations of existing tools hinder the recovery of damaged files, posing difficulties in the evidence collection process. This paper aims to propose solutions for the recovery of corrupted MS Office files commonly used in digital data creation. To achieve this, we analyze the structure of MS Office files in the OOXML format and present a novel approach to overcome the limitations of current recovery tools. Through these efforts, we aim to contribute to enhancing the quality of evidence collection in the field of digital forensics by efficiently recovering and identifying damaged data.

Keywords: Digital Forensics, OOXML, PK ZIP, Data Extraction

Received(12. 28. 2023), Modified(02. 13. 2024), Accepted(02. 14. 2024)
* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임 (No.RS-2024-00398745, 디지털 환경에서의 증거인멸행위 증명 및 대응 기술 개발).
† 주저자, jiyoon4023@g.skku.edu
‡ 교신저자, doowon@g.skku.edu(Corresponding author)

I. 서 론

디지털 기기의 보급률 및 사용자 활용 수준이 높아짐에 따라 디지털 형태로 생성되는 자료의 규모가 꾸준히 증가하고 있다. 이러한 사회적 변화에 따라 범죄 발생 시, 정보저장매체에 저장된 대량의 전자정보를 분석하는 디지털 포렌식 수사의 중요성 또한 높아지고 있다.

특히 우리나라의 경우, 정보저장매체 등에서의 디지털 증거 확보를 위한 압수수색은 「형사소송법」 제106조에 의거하여 선별압수 원칙에 따라 사건과 유관한 정보만 한정하는 방식으로 이뤄지고 있다[1]. 디지털 포렌식 도구를 활용해 분석 대상에 저장 또는 삭제된 파일을 확인하고, 사건과의 관련성을 파악하여 유관하다고 판단되는 파일만을 선별적으로 압수하여 수사가 진행된다.

하지만 디지털 자료는 정보저장매체의 손상이나 멀웨어에 의한 삭제, 파일 시스템의 오류, 증거 인멸을 위한 안티포렌식 행위 등 여러가지 원인으로 인해 식별할 수 없어지기도 하는데, 이는 증거 수집의 어려움으로 이어진다. 이러한 경우에는 별도의 도구를 통한 복구 과정이 요구된다. 디지털 포렌식 조사관들은 파일 카빙 도구나 문서 복구 도구 등을 활용하여 디지털 기기에서 숨겨지거나 삭제된 파일 등을 수집하여 복구를 진행, 분석하고 있다.

그러나 기존의 카빙 도구 및 문서 복구 도구는 파일 시그니처 및 구조를 기반으로 동작한다는 특성으로 인해 유사한 구조를 지닌 파일과의 분류가 적절히 이뤄지지 않거나 복구된 대량의 파일을 직접 열람하여 확인해야 한다는 번거로움이 존재한다. 그로 인해 증거로서 가치가 있을 수 있는 데이터 식별 시 상당한 시간이 소요될 뿐만 아니라 증거 확보 자체에도 어려움이 존재한다.

이에 본 연구에서는 기존 도구의 한계를 보완하고자 문서 생성 시 사용되는 MS Office 파일 중 버전 2007 이후의 word, excel 파일 구조를 분석 및 해당 파일을 대상으로 카빙된 데이터를 분석하여 손상된 파일에서 남아있는 데이터를 추출할 수 있는 방안에 대해 연구하고자 하였다.

본 논문의 개요는 다음과 같다. 2장에서는 MS Office 파일이 갖추는 형식인 PK ZIP 아카이브와 OOXML(Office Open XML) 구조를 분석한 내용 및 관련 선행연구를 다루고자 한다. 그리고 2장에서 분석한 파일 구조 및 카빙된 데이터를 바탕으로 정의 내린 MS Office 파일의 손상 유형과 유형별 손상 파일을 활용하여 기존 복구 도구의 한계를 3장에 기술한다. 이를 기반으로 4장에서는 손상된 문서 데이터를 추출하는 방안을 제시하고, 5장에서 연구의 의의 및 한계를 다룬다.

II. Background

2.1 MS Office 파일 구조 분석

MS Office 파일은 2007 이후 버전부터 ISO/IEC 29500 표준에 정의된 OOXML 형식을 사용한다. 여기서 OOXML 형식은 최종적으로 표준화 패키징 기술인 OPC(Open Packaging Conventions)를 기반으로 하나의 PK ZIP 아카이브 형식의 패키지(package)로 구현된다[2]. 예를 들어, 문서에 기록된 텍스트 및 첨부된 이미지 파일, 스타일 정의를 비롯해 문서 구성과 관련된 여러 요소는 OOXML 형식을 사용하여 표현된다. 해당 요소들은 OPC 표준을 따라 패키징되어 하나의 ZIP 아카이브 구조를 띄는 방식이다.

즉, MS Office 파일은 사실상 ZIP 아카이브 내에 하나의 문서를 구성하는 파일 및 디렉터리가 OOXML 형식으로 작성 및 저장된다. 따라서 본 절에서는 MS Office 파일이 따르는 포맷인 PK ZIP 및 OOXML 형식에 관해 설명하고자 한다.

2.1.1 PK ZIP

PK ZIP 형식 파일이란, 파일 및 디렉터리 압축을 통해 하나의 파일로 패키징 및 구성하는 방식이다. 확장자 .zip, .jar, .war 등처럼 익히 알려진 압축 형식 파일뿐만 아니라 .docx, .xlsx, .pptx 등과 같은 MS Office 2007 버전 이후의 파일 구성에도 사용된다. 데이터 압축에는 여러 알고리즘이 활용되는데, 기본적으로는 deflate 압축 알고리즘이 사용된다. 그리고 악의적인 위변조로부터 파일을 보호하기 위해 데이터 암호화, CRC32를 통한 무결성 검증 등의 기술을 갖추고 있다. 해당 형식은 여러 파일을 하나로 압축하는 경우뿐만 아니라 응용프로그램이나 단일 파일 등 다양한 유형의 파일에서 컨테이너로 사용되기도 한다. 만약 응용프로그램에 PK ZIP 형식을 적용하는 경우, 파일 관련 정보가 기술된 Manifest 파일이 함께 저장되며, MS Office 파일

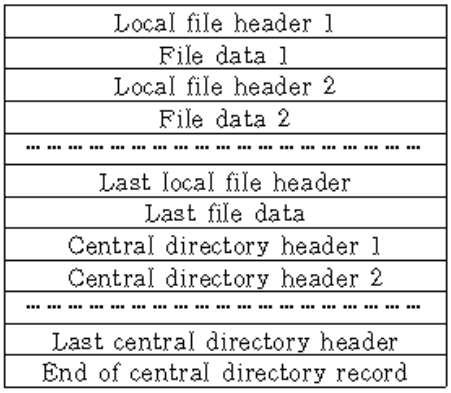


Fig. 1. The structure of PK ZIP

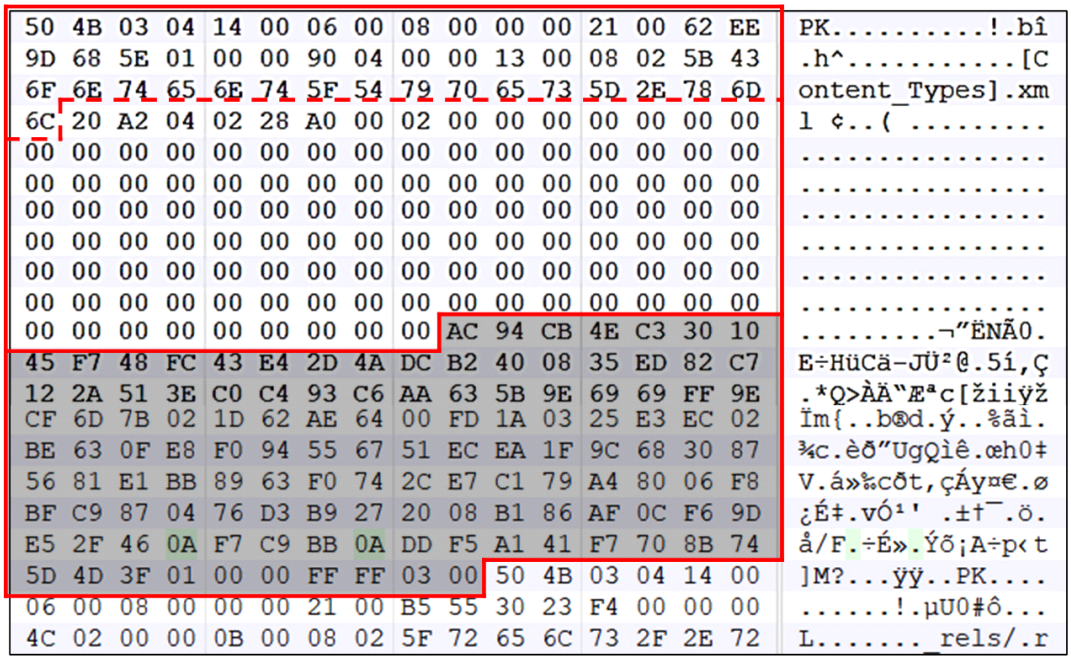
은 문서 내용 및 구조 정보 등이 저장된 하나의 파일이 PK ZIP 형식으로 패키징된다[3].

PK ZIP 형식은 일반적으로 문자 'PK(0x4B 50)'로 시작하여 end of central directory record로 끝맺는 Fig. 1.과 같은 구조를 갖추고 있으며, 레코드 file data를 제외한 모든 레코드의 시작에는 고유한 시그니처가 존재하며, 구체적인 값은 Table 1.과 같다.

Table 1. The signature of records

The name of record	Value
Local file header	0x 50 4B 03 04
Central directory header	0x 50 4B 01 02
Last central directory header	0x 50 4B 06 07
End ofcentral directory	0x 50 4B 05 06

Table 1.에서 local file이란, 하나의 PK ZIP 파일 구성을 위해 파일 내부에 위치하는 개별 파일을 의미한다. local file의 첫 부분에는 각 local file에 대한 메타데이터를 저장하는 local file header가 위치하며, local file header 뒤에는 실제 파일의 데이터인 file data가 존재한다. 이때 file data는



	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	OF
0x00	The signature of a local file (0x50 4B 03 04)				The version of decompressed		General bit flag		The method of compressed		The last modify time		The last modify date		Checksum-(CRC32)	
0x10	-Checksum (CRC32)		Compression size				Decompression size				Local file name length		Extra field length		Local file name	
0x20	Local file name (Variable)				Extra filed(Variable)											
0x30	Compressed file data(Variable)															

Fig. 2. An example of a local file structure

local file별로 압축 여부가 달리 적용되어 저장된다. 모든 local file이 나열된 이후, 파일 하단에는 local file header와 마찬가지로 파일에 대한 메타데이터가 기록된 central directory가 오게 된다. local file의 구조 및 각 필드에 대한 개괄적인 설명은 Fig. 2.와 같다.

PK ZIP 구조를 분석한 결과, 해당 형식은 크게 문서 내용과 관련된 데이터가 저장되는 개별적인 local file 영역과 레코드 central directory처럼 문서의 내용 식별에 필수적이지 않은 부분 두 영역으로 나눌 수 있다. 그리고 모든 영역은 고유한 시그니처 값으로 시작하고 있음을 확인했다.

즉, 해당 형식으로 패키징되는 파일이 손상으로 인해 일반적인 열람 방식으로는 내용 식별이 불가능하더라도, 문서 내용과 직접적인 연관이 있는 local file이 존재한다면 해당 file data에 접근하여 이를 추출하는 방식으로 내용 식별이 가능하다는 것을 파악하였다.

따라서 PK ZIP 형식으로 패키징된 MS Office 파일에서 문서 내용 관련 데이터 추출 방안 고도화를 위해서는 해당 파일의 본문 내용 확인에 필수적으로 존재해야 하는 local file을 파악하는 추가적인 분석이 필요하다.

2.1.2 OOXML

OOXML이란 MS Office 2007 버전 이후의 문서에서 사용되는 파일 구조로, 주로 XML 기반의 마크업 언어로 문서의 콘텐츠 및 구조가 작성되는 방식이다. 작성된 파일들은 표준 패키징 방식인 OPC 방식에 따라 ZIP 아카이브 형식을 갖추며 하나의 파일 구성에 필요한 local file은 2.1.1항에 기술된 방식을 따라 구성된다. Fig. 3.에서 문서 유형별로 가지고 있는 local file을 확인할 수 있으며, 각 local file별 내용은 Table 2.에서 확인 가능하다.

2.1.1항에서 분석한 결과에 따르면, 파일의 손상

Table 2. Descriptions of some Parts in OOXML File

File	The name of Local file	Description
Excel, Word	app.xml	A part where metadata information, such as details about the application that created the document and related settings is stored
	core.xml	A part containing document metadata, such as the document author, creation date, modification time, and revision count
	theme1.xml	A part containing information related to the document's theme (design, formatting, etc)
	fontTable.xml	A part storing information about fonts used in a document per package, limited to a maximum of two Font Table parts
	settings..xml	A section containing where general document settings information related to specific parts such as document templates, mail merge data sources, and XSL transformations
	webSettings.xml	A part storing content related to specific web settings, including information about the HTML 'frameset' element
	styles.xml	A part where style-related information, such as numeric and text formatting, alignment, font, color, and borders specified in the document
	[Content_Types].xml	A part defining the content types of each part within a package
Excel	sheet(n).xml	A part storing all data, formulas, and characteristics associated with a specific worksheet
	sharedStrings.xml	A indexed list of string values shared across the entire workbook, allowing implementations to store values only once
Word	document.xml	A part storing the document's body along with basic structures and formatting information such as paragraphs, fonts, styles, etc.

으로 인하여 문서 내용 확인을 위한 열람이 불가능하더라도, 해당 파일의 본문 내용을 담고 있는 local file이 존재한다면 해당 파일 내용 확인 또는 추측이 가능해진다. 따라서 잔여 데이터를 추출하기 위해서는 작성된 문서 내용이 저장되는 local file이 무엇인지 파악하고 식별하는 과정이 필요하다. excel 파일의 경우에는 Fig. 4.처럼 문서에 작성된 모든 데이터 및 기본 구조 등이 저장된 'xl/worksheets/sheet(n).xml'과 workbook 전체에서 공유되는 문자열 값이 저장되는 'xl/sharedStrings.xml'가 해당된다. 그리고 word 파일에서는 단락, 글꼴, 스타일 등의 기본 구조와 함께 문서 본문이 'word/document.xml'에 저장됨을 Fig. 5.에서 확인할 수 있다.

만약, 문서 본문 내용이 담겨 있는 local file의 손상 정도가 심하여 추출할 데이터가 존재하지 않더라도, Fig. 6.에서 확인할 수 있듯이 OOXML 형식의 파일에는 문서의 작성자, 작성일, 수정일 등의 메타데이터를 local file 'docProps/core.xml'에

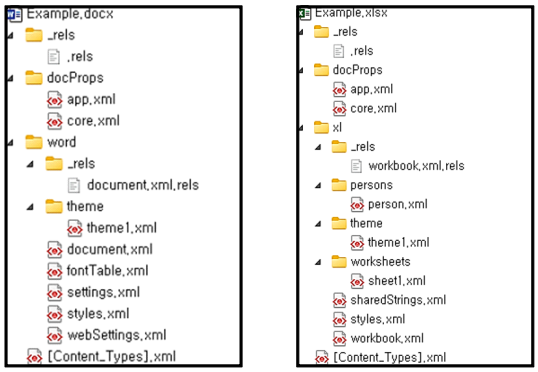


Fig. 3. The Structure of An Word File(.docx) and An Excel File(.xlsx)

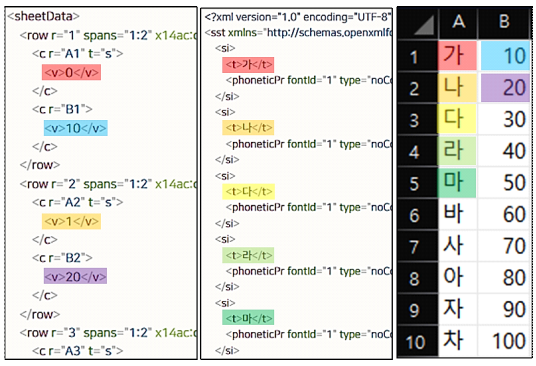


Fig. 4. Local files in Excel file

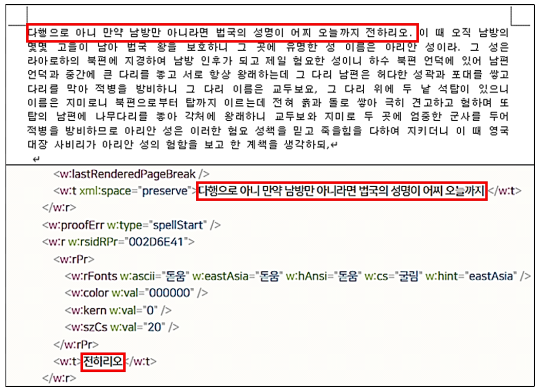


Fig. 5. Local file 'document.xml' in Word file

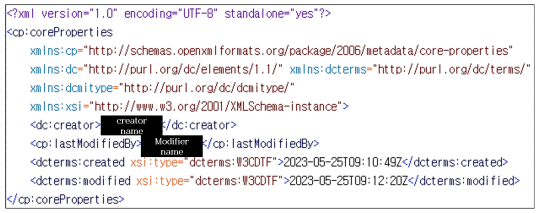


Fig. 6. Document metadata stored in core.xml

저장하기에 해당 문서에 대한 정보를 획득할 수 있게 된다.

2.2 선행연구

본 절에서는 MS Office 파일 구조 관련 연구 및 손상 파일 복구 방안에 관한 선행연구를 살피고자 한다.

2.2.1 손상된 ZIP 파일 복구 기법

정병준 외 2인[4]은 다양한 파일에서 활용되는 압축파일 형식인 PK ZIP 형식을 따르는 ZIP 파일이 손상되어 존재하는 경우, 해당 파일의 복구 기법을 제시하였다. 해당 연구에서는 Karl Wust[5]가 제시한 파일 손상에 대한 정의를 차용하여 손상된 ZIP 파일의 유형을 4가지로 분류, 유형별 복구 기법 연구를 진행하였다. 그리고 손상 파일 유형 분류는 ZIP 파일 구조에 기반, 필드별 손상을 기준으로 '파일 시작부터 local file 영역의 일부까지만 남은 Case 1', '파일 시작부터 central directory 영역의 일부가 남은 Case 2', '파일 시작 부분이 손상되고 local file 영역부터 시작하는 Case 3', 그리고

‘파일 시작 부분이 손상되고 central directory 영역 일부가 남은 Case 4’로 총 네 가지로 정의되었다. 모든 local file 영역이 온전히 존재하는 Case 2를 제외한 나머지 세 가지 Case는 손상되기 이전의 파일과 동일한 상태로의 복구는 불가능하며, 특히 central directory 영역의 내용만 존재하는 Case 4는 파일 관련 메타데이터를 추출하는 방식으로 연구를 진행하였다.

2.2.2 OOXML 형식 문서에서의 데이터 은닉 연구

해당 선행연구는 디지털 문서 작성 시 세계적으로 널리 활용되는 MS Office 파일의 구조인 OOXML 형식의 특성을 활용하여 안티포렌식 행위인 데이터 은닉이 가능한 사례를 제시 및 이를 탐지하는 알고리즘을 제시하였다[6].

해당 연구에서는 기존의 탐지 알고리즘의 한계를 발생 가능한 시나리오를 서술하여 제시하였는데, 먼저 은닉하고자 하는 기밀문서의 확장자를 MS Office 파일의 파트와 동일한 확장자인 ‘.xml’로 변경 및 파트별 관계를 정의하는 local file에 이를 추가한다면 이를 탐지하지 못 한다는 점을 한계로 지적하였다. 또한, local file header에서 문서 작성 및 수정 시간이 남는 레코드 정보를 활용하여 은닉된 데이터를 탐지하는 기존 탐지 알고리즘의 한계 또한 언급하며 이를 보완한 새로운 기법을 제시하였다.

먼저, 문서의 바이너리 값을 분석하여 각 파트에 저장된 시간 정보가 초기 값(0x00 00 21 00)과 일치하는지 그 여부를 파악, 일치하지 않는다면 은닉된 데이터가 존재한다고 간주한다. 다음으로는 문서를 압축 해제하여 어떠한 데이터를 은닉하였는지 이를 구분하고, 구분된 은닉 데이터 유형별로 이를 탐지하는 기술을 제시하였다.

2.2.3 그 외

2.2.1항과 2.2.2항에 기술한 선행연구를 비롯하여 손상된 파일과 관련된 연구는 다양하게 진행되어 왔다. 먼저, 원본 파일로의 완전한 복구의 어려움을 지적 및 예측할 수 없이 다양한 방식으로 발생하는 파일의 ‘손상’ 자체에 대한 정의를 제시하며, 파일 구조가 아닌 파일 뷰어의 실행을 변경하는 등의 방식을 통해 손상된 파일을 복구하는 연구가 진행된 바 있다[5]. 해당 연구에서는 원본 파일과 동일한 상태로의 복구가 어렵다는 점에 착안하여 파일 복구에 있어 완전한 정확성이 아닌, ‘원본과의 충분한 유사성(sufficiently similar)’을 목표로 지정하였다. 또한, 제시하는 복구 방안을 크게 세 가지의 지표를 기준으로 평가하는데, 그 기준은 (1) ‘복구한 파일이 프로그램을 사용하는 동안 오류를 발생시키지 않는다.’, (2) ‘복구된 파일에는 원본 파일에 존재하는 대부분의 정보가 유지되어 있어야 한다.’, (3) ‘복구된 파일에는 새롭거나 변형된 정보는 최소한으로 포함되어야 한다.’라는 점이다.

이뿐만 아니라 데이터 deflate 알고리즘을 통해 압축된 경우의 복구 가능성을 제안한 연구[7], 디지털 장치에서 증거물을 수집하는 과정에서 비지도 학습을 활용하여 주기억장치 RAM(Random Access Memory)에서 OOXML 형식 문서를 탐지 및 이를 추출할 수 있는 방안에 대한 연구[8] 등과 더불어 MS Office 파일뿐만 아니라 손상된 파일 복구를 목적으로 하는 연구 또한 진행되어 왔다.

그러나 기존 연구의 경우 주로 손상된 파일의 복구가 성공적으로 이루어졌는지를 확인하기 위해 손상 전 파일과 손상 후 이를 복구한 파일을 비교하는 방식으로 검증을 진행하였다. 하지만 이러한 방식의 경우 실험을 위해 제작된 데이터에서만 활용할 수 있는 방안이다. 실제 수사 과정에서 마주하게 되는 데이터는 대조 가능한 손상 전 파일이 존재하지 않는 경우가 다수라는 점에서 실질적으로 이를 활용하기 힘들다는 한계점을 지닌다. 따라서 본 연구에서는 기존 연구 및 도구에서 참조하는 문서의 메타데이터 활용을 최대한 배제 및 실제로 카빙된 데이터 또한 분석과 유형을 구분함으로써 고도화된 데이터 식별 및 추출 방안을 제시하고자 한다.

III. MS Office 파일 손상 유형 정의

본 장에서는 Karl Wust[5]가 제시한 ‘파일 손상이라는 정의 자체의 광범위함으로 인한 모호성’ 및 ‘원본으로의 완전 복구의 어려움’에 기반하여 데이터 추출을 위한 손상 파일의 유형을 정의하고자 한다. 이를 위하여 고의적으로 파일 손상을 진행하는 실험뿐만 아니라, 카빙 기술을 지원하는 도구의 출력 파일들의 유형 또한 함께 살피어 보다 구체적으로 유형 구분을 진행하였다. 이뿐만 아니라, 유형화한 손상 MS Office 파일을 활용하여 현재 상용화된 복구 도구의 한계 또한 확인하여 데이터 추출 고도화 방안

마련을 위한 고려 사항을 파악하였다.

3.1 파일 구조 기반 손상 유형 정의

먼저 실험을 통해 손상 유형을 정의하고자 MS Office 파일을 작성할 수 있는 여러 가지 응용프로그램으로 버전 2007 이후의 word, excel 문서를 생성하였다. 생성된 문서에서 local file의 file data와 메타데이터를 일정한 크기만큼 완전 삭제하거나 0x00 00 00으로 덮어쓰기하는 방식으로 손상을 가하여 손상 파일 데이터세트를 제작하여 local file별 손상이 문서 열람에 미치는 영향 또한 함께 살피었다.

구체적인 손상 방식은 세 가지로, (1) 'local file 일부를 완전히 삭제'하는 방식과 (2) '방식 (1)과 더불어 local file header에 기록된 메타데이터 덮어쓰기', (3) '방식 (1)과 더불어 central directory를 삭제'하는 방식이다. 손상을 가한 파일은 MS Office 공식 응용프로그램인 'Microsoft Word' 및 'Microsoft Excel'에서 해당 파일의 열람 가능 여부 및 내부 데이터 파악 가능 정도를 기준으로 손상 유형을 구분했다.

실험을 통해 구분된 손상 유형은 크게 두 가지로, Fig. 7.의 좌측에 해당하는 (1) '파일 열람 자체가 되지 않는 경우'와 우측의 (2) '파일 열람은 가능하나 저장된 내용이 모두 보이지 않는 경우'이다. 특히 손상 유형 (2)는 excel 문서 파일에 손상을 가한 경우 확인 가능한 유형으로, 이는 Fig. 4.에서 확인한 바와 같이 해당 파일의 내부 구조상 numeric data와 string data를 상이한 local file에 저장하여 그 값을 참조하는 방식을 사용한다는 특징에서 비롯된 유형이다.

local file별로 file data가 손상된 데이터세트를 활용하여 실험을 진행한 결과, 문서 내용을 포함하고 있는 local file이 데이터를 충분히 가지고 있음에도 불구하고 여타 local file의 손상으로 인한 파일 열람 및 내용 식별이 불가능할 수 있음을 확인하였다. 따라서 file data 및 메타데이터 등의 손상으로 인해 파일 열람 및 내용 식별이 되지 않을 시, local file이 파일 내부에 존재하는지를 탐지하고, 파일 내용이 저장되는 파트 또한 존재한다면 이를 압축 해제하는 방식으로 데이터를 추출이 가능함을 확인하였다.

3.2 카빙 데이터 분석 기반 손상 유형 정의

실험뿐만 아니라 수사 과정에서 활용되는 기존 포렌식 도구 중 카빙 기능을 지원하는 도구(Axiom, EnCase, Data Rescue PC3, Scalpel, Recover My Files)의 카빙 결과에 대한 분석을 진행하여 MS Office 파일의 손상 유형을 구체화하였다. 이는 PK ZIP 구조로 패키징된다는 MS Office 파일의 OOXML 형식적 특성으로 인해 동일한 구조를 사용하는 다른 확장자 파일이 카빙된 데이터에 혼용되어 존재하여 증거 확보에 어려움이 존재한다는 이유에서 착안한 방식이다. 해당 도구들의 카빙 데이터를 열람 및 분석한 결과, 크게 세 가지로 손상 유형이 구분

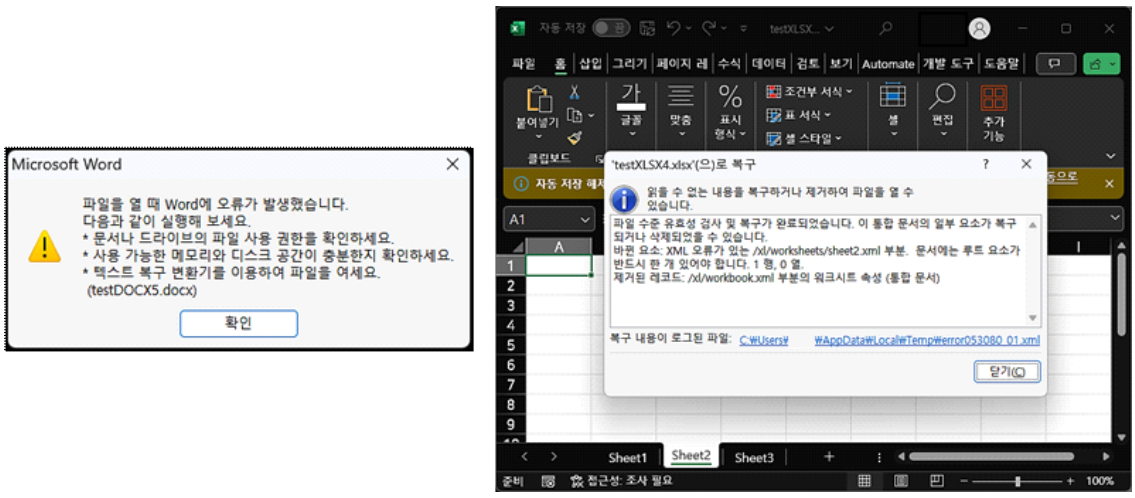


Fig. 7. Distinct Types of Damage Identified Through Experiments

가능했다.

먼저, word 및 excel 파일만을 대상으로 카빙을 진행하더라도 PK ZIP 방식으로 패키징되는 파일의 형식적 특성으로 인하여 해당 형식이 활용된 다른 확장자 파일과 구별이 어려워 여타 확장자 파일이 다수 혼동되어 추출되는 것을 확인하였다. 그리고 MS Office 파일임과 동시에 문서 내용을 추출할 수 있는 local file이 존재하지만, 문서 구성을 위한 관계값이 저장된 구조가 손상되거나 일부 local file의 유실 등으로 인하여 열람이 불가능한 경우, 저장된 내용이 존재하지 않는 파일(0x00 00 00…) 등을 발견할 수 있었다.

앞서 진행한 실험 결과와 기존 도구의 카빙 결과를 종합적으로 고려한 손상 유형 정의는 Table 3.과 같다.

Table 3.과 같이 손상 유형을 네 가지로 정의한다면 유형 ①과 ②는 복구를 통하여 내용 확인이 가능하지만, 유형 ③, ④는 원천적으로 복구 및 내용 파악이 불가능하다. 따라서 정의한 유형에 따른 복구 방안 마련 시, MS Office 파일이 아닌 유형 ③, ④에 해당하는 파일은 초기에 필터링하는 과정이 필요하게 된다. 그리고 데이터 추출이 가능한 유형 ①과 ②의 경우, 파일 내용이 저장되는 local file의 존재 여부를 파악 및 데이터를 압축 해제함으로써 내용을 식별하는 방식으로 접근할 수 있게 된다.

Table 3. Type of Corruption to MS Office Files

No.	Type of Corruption
①	An MS Office file but cannot be opened due to corruption
②	An MS Office file but the content is entirely unreadable due to corruption
③	The file is not an MS Office file
④	There is no content saved

3.3 기존 복구 도구 현황

평가를 진행할 도구 선정에는 PK ZIP 구조로 패키징되는 OOXML 형식을 고려하여, 도구 범주를 (1) ‘excel 또는 word 복구 도구’와 (2) ‘ZIP 복구 도구’ 두 가지로 지정하였다. 도구 범주가 다른바, 성

Table 4. Tool (1) Result - Partially Deleted Local File Data of Excel and Word Files

Corrupted Local file	DEFA		Stellar data Recovery		Cimaware [excel|word] fix		Kernelfor [excel|word]		Datanumen [excel|word] repair	
	excel	word	excel	word	excel	word	excel	word	excel	word
[Content_Types]	X	X	X	X	○	X	X	X	△	X
app	X	X	X	X	○	○	X	X	△	X
core	X	X	X	X	○	○	X	X	△	X
rels	X	X	X	X	X	X	X	X	△	X
styles	X	X	X	○	X	○	X	X	X	X
theme1	X	X	X	○	X	○	X	X	X	X
sharedStrings	X	-	X	-	X	-	X	-	X	-
sheet(n)	X	-	X	-	X	-	X	-	X	-
workbook	X	-	X	-	X	-	X	-	X	-
document	-	X	-	X	-	X	-	X	-	X
fontTable	-	X	-	○	-	○	-	X	-	X
document.xml.rels	-	X	-	○	-	○	-	X	-	X
settings	-	X	-	○	-	○	-	X	-	X
webSettings	-	X	-	○	-	○	-	X	-	X

○: Entire document content is viewable
X: Unable to view document content
△: Partial document content is viewable
-: Not applicable for local file

Table 5. Tool (2) Result - Overwriting Some Local File Data in Excel and Word Files

Corrupted Local file	ALZip		winrar		Zip2Fix		7-Zip		DiskInternals Zip repair	
	excel	word	excel	word	excel	word	excel	word	excel	word
[Content_Types]	X	X	X	X	○	X	X	X	△	X
app	X	X	X	X	○	X	X	X	△	X
core	X	X	X	X	○	X	X	X	△	X
rels	X	X	X	X	X	X	X	X	△	X
styles	○*	X	○	△	○	X	○	△	○	△
theme1	○*	X	○	△	○	X	○	△	○	△
sharedStrings	○*	-	○	-	○	-	○	-	○	-
sheet(n)	○*	-	○	-	○	-	○	-	○	-
workbook	○*	-	○	-	○	-	○	-	○	-
document	-	X	-	△	-	X	-	△	-	△
fontTable	-	X	-	△	-	X	-	△	-	△
document.xml.rels	-	X	-	X	-	X	-	X	-	X
settings	-	X	-	△	-	X	-	△	-	△
webSettings	-	X	-	X	-	X	-	X	-	X

○: Residual content of damaged local file is viewable
△: Recovery may be impossible depending on the extent of damage
X: Residual content of damaged local file is not viewable
○*: Unable to confirm residual file data in case of Central Directory loss

능 평가를 위한 기준 또한 상이하게 설정하였다. (1)의 경우, input data를 하나의 문서 파일 자체로 인식하기에 'MS 응용프로그램으로 열람 및 내용 확인이 가능한가?'를 기준으로 설정하였으며, 평가를 진행한 도구는 'DEFA, Stellar data Recovery, cimaware[word|excel] fix' 등이다. (2)에 해당하는 도구의 평가 기준은 압축된 local file 레벨에서 데이터를 처리하므로 '손상을 가한 local file의 잔여 file data 확인이 가능한가?'를 기준으로 하여 'ALZip, winrar, Zip2Fix' 등을 도구로 선정했다.

실험을 진행하여 분석한 결과는 다음과 같다. 도구 (1)의 경우, local file header 및 central directory에 기록된 메타데이터에 대한 높은 의존성으로 인하여 file data의 일부를 완전히 삭제하는 방식의 경우, Table 4.와 같이 손상된 파일의 복구가 제대로 진행되지 않고 있었다. 특히, 메타데이터가 덮여 쓰인 방식으로 손상된 파일은 복구가 완전히 불가능함을 확인할 수 있었다.

도구 (2)는 file data의 일부가 덮여 쓰인 파일에 대해서는 Table 5.의 결과처럼 일부 local file data를 확인할 수 있었으나, 해당 도구 또한 메타데이터가 손상된 경우에는 잔여 file data를 확인할 수 없었다. 특히, excel 파일의 local file data의 일부가 완전히 삭제되는 방식으로 손상된다면 사용한 상용 도구 모두에서 잔여 file data 확인이 불가능하였다.

기존 복구 도구 현황을 파악한 결과, 잔여 데이터가 존재하는 파일이더라도 내용 식별을 위한 복구가 제대로 이뤄지지 않고 있음을 확인하였다. 따라서 본 연구는 기존 도구의 한계를 극복하여 파일 내용 관련 데이터를 추출할 수 있는 방법을 모색하는 것에 초점을 맞추어 연구를 진행하였다.

IV. 데이터 추출 방법론

실험 및 기존 도구 결과 분석을 진행한 내용을 기반으로 본 연구에서는 다음과 같이 손상된 MS Office 파일에서의 데이터 추출 방안을 제시하고자 한다. 데이터 추출 방안에 대한 전체적인 흐름은 Fig. 8.과 같다.

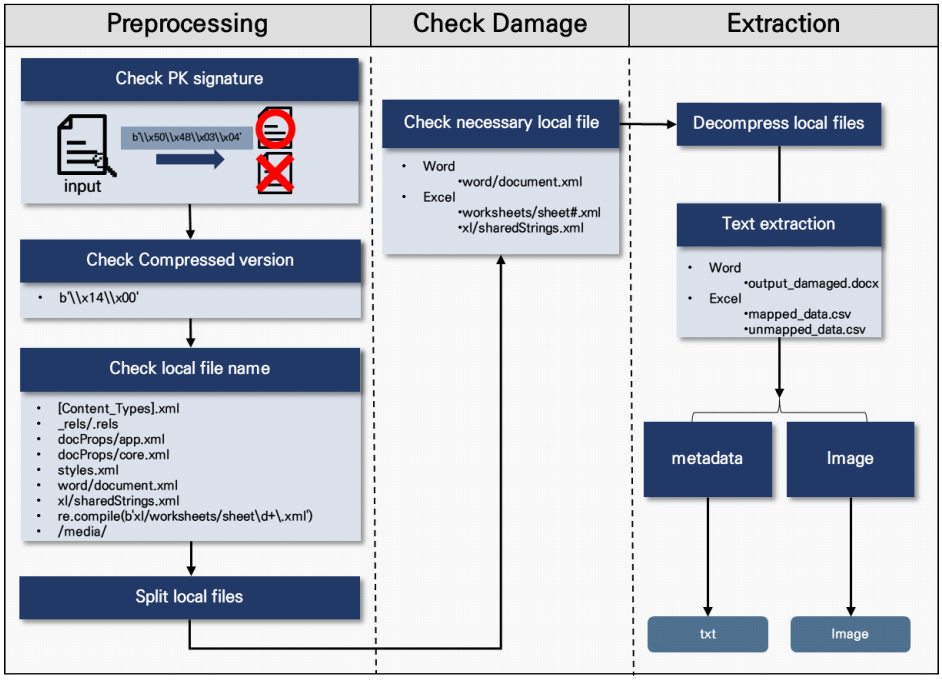


Fig. 8. Approaches for extracting data from corrupted OOXML files

4.1 Preprocessing

'Preprocessing' 단계는 실질적인 데이터 추출을 시도하기 전, 해당 파일이 실제 OOXML 형식을 따르는 MS Office 파일인지 검증하는 단계이다. 카빙 데이터 분석을 통해 손상 유형을 정의한 결과, 문서 내용을 담고 있는 local file이 존재함에도 불구, 구조적 유사성으로 인하여 MS Office 파일 자체가 아닌 확장자가 혼재되고 있음을 파악하였다.

따라서 해당 단계에서는 input data 필터링 과정을 거침으로써 실질적인 OOXML 형식의 MS Office 파일만을 대상으로 잔여하는 데이터를 추출할 수 있도록 설계하였다. 그와 동시에 문서 본문 내용과 관련된 데이터가 존재하는지도 함께 파악한다. 이를 통해 잔여 데이터가 없는 경우, 데이터 추출 과정으로 넘어가지 않고 도구를 종료하게 된다.

input 파일이 OOXML 형식의 MS Office word, excel 파일인지 확인하기 위하여 크게 세 가지 요소를 파악 및 필터링하게 된다. 먼저, input 파일이 local file header의 처음에 오는 PK 시그니처로 시작하는지 그 여부를 확인한다. 이는 OOXML 형식의 MS Office 파일이 PK ZIP 형식으로 패키징된다는 구조적 특성에서 기인한 것으로, 시그니처를 지니지 않은 파일은 구축하고자 하는 도구의 input data로 적절하지 않다.

그다음, 해당 파일 압축 버전이 MS Office 파일의 local file 압축 버전인지를 확인한다. MS Office 파일의 경우, file data의 압축을 위해 deflate 알고리즘이 적용되므로 해당 파일의 압축 버전이 기록되는 offset에 기록된 값이 deflate 알

Table 6. Local File Name for Input File Filtering

Selection Criteria	Corresponding Local File Name
Is it a local file that significantly impacts file viewing when damaged?	- [Content_types].xml - _rels/.rels - styles.xml
Is it a local file containing document content?	- word/document.xml - xl/sharedStrings.xml - xl/worksheets/sheet (n).xml
Does it have minimal impact on viewing but can assist in understanding the document?	- docProps/app.xml - docProps/core.xml - xl/media/ - word/media/

고리즘을 나타내는 '0x14 00'인지 확인하는 작업을 통해 input 파일의 적합성 여부를 판단한다. 마지막으로, 기본적인 MS Office 파일에서 확인할 수 있는 local file name의 존재 여부를 탐지함으로써 input 파일 검증 단계를 마무리한다. 이때 존재 파악을 위한 local file name은 앞서 진행한 파일 구조 분석 및 기존 복구 도구 현황 파악을 위해 진행한 실험의 결과로 도출한 내용을 토대로 선정되었으며, 구체적인 기준 및 각 기준에 해당하는 local file name은 Table 6.과 같다.

4.2 Check Damage

'Check Damage' 과정은 작성된 문서 내용 복구를 위해 필수적으로 존재해야 하는 데이터가 저장된 local file의 존재 여부를 확인하는 과정이다. 파일 구조 분석을 통해 word의 경우 'word/document.xml'에, excel은 'worksheets/sheet(n).xml' 및 'xl/sharedStrings.xml'에 문서 내용이 저장된다는 것을 확인하였다. 따라서 잔여하는 데이터를 추출하고자 손상 확인 기준을 문서의 내용이 저장되는 local file의 존재 여부로 설정하여 해당 local file이 존재하는 경우에만 Extraction 과정이 진행된다.

4.3 Extraction

'Extraction' 단계는 실질적인 데이터 추출 작업이 이뤄지는 단계이다. deflate 알고리즘으로 압축되는 local file의 file data의 내용 확인을 위해 압축 해제하는 작업이 필요하다. deflate 알고리즘 특성상 압축된 데이터의 앞부분이 해석 가능한 정상적인 상태로 온전하게 존재한다면 이를 해제하여 내용을 확인할 수 있게 된다. 따라서 압축 해제를 시도하려는 데이터가 손상된 채로 저장돼 있더라도, 손상 이전 포인트까지 위치하는 데이터의 추출이 가능해진다.

여기서 추출되는 데이터 형식은 크게 숫자를 포함하는 문자 형식 데이터와 이미지 데이터로 나뉜다. 이미지 데이터는 deflate 알고리즘으로 압축되지 않고 file data 위치에 데이터가 그대로 저장된다. 따라서 식별된 local file name이 이미지 파일에 해당한다면, 압축 알고리즘을 적용하지 않고 추출해야 한다. 이뿐만 아니라, 문서와 관련된 메타데이터를 저장하는 local file이 존재한다면 해당 데이터도 압축 해제함으로써 손상된 파일 관련 데이터를 최대한 추출 및 파일 내용 식별에 기여할 수 있게 된다.

4.4 제안하는 기법 검증

본 연구에서 제안하는 방법의 실효성을 확인하고자 MS Office 파일의 손상 유형을 정의 및 기존 도구의 한계 파악을 위해 제작하였던 손상 파일 데이터 세트 중 '기존 복구 도구가 복구하지 못한 파일' 131개 및 '카빙된 데이터' 445개를 3장에서 정의한 손상 유형별로 구분 및 이를 대상으로 검증을 진행하였다. 기존 복구 도구를 통해 복구되지 않는 파일의 경우, 파일 내용 파악을 위한 local file이 존재한다면 해당 local file의 file data 추출을 통해 내용 식별이 가능한지를 기준으로 개발한 도구의 실효성을 파악하였다. 그리고 카빙된 데이터를 활용하여 MS Office 파일이 아닌 확장자 파일이 input 파일로 지정될 시, 해당 파일은 데이터 추출 과정을 거치지 않고 올바르게 종료되는지를 확인하는 방식으로 검증을 진행하였으며, 개괄적인 검증 결과는 Table 7.과 같다.

먼저, Fig. 9.처럼 OOXML 형식의 문서임과 동시에 문서 내용을 확인할 수 있는 필수 local file가 존재하는 손상 유형 ① 또는 ②의 경우, 해당 local file의 데이터를 확인할 수 있도록 추출해 주었으며, 만약 존재하지 않을 때는 Extraction을 진행하지 않고 정상적으로 종료되었다.

이뿐만 아니라, 메타데이터의 손상으로 기존 도구에서 복구 및 데이터를 추출하지 못한 손상 유형 또한 Fig. 10.처럼 파일 내부에 존재하는 내용을 확인할 수 있도록 작동되었다.

그리고 Fig. 11.과 같이 MS Office 파일이 아님에도 해당 확장자로 카빙된 데이터인 손상 유형 ③의 경우, Extraction 과정까지 진행되지 않고 올바르게 종료된 것을 확인하였다.

실험을 통해 제안하는 기법이 기존 복구 도구에서

Table 7. Tool Results by Corrupted Type

Type of Corruption	Total Count of Files	Number of Recovered or Filtered Files
Type ①	116	108
Type ②	15	11
Type ③	385	301
Type ④	60	60

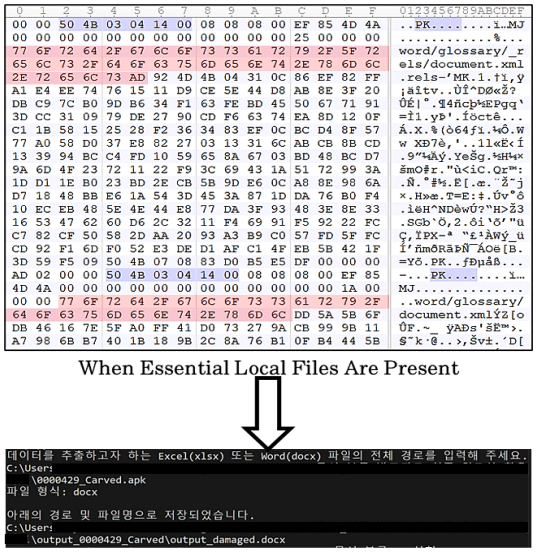


Fig. 9. When Essential Local Files Are Present for Document Content Identification

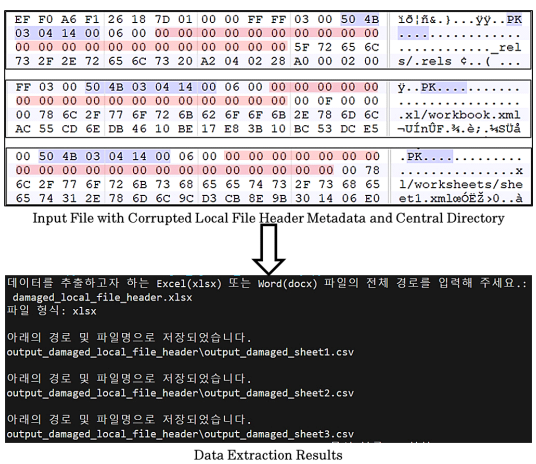


Fig. 10. Examples of Extraction Approaches Minimizing Metadata Dependency

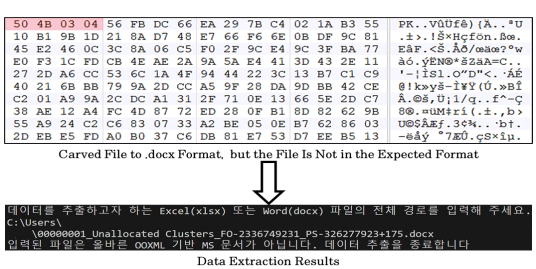


Fig. 11. Example of Filtering for Non-MS Office Files

데이터 추출이 불가하던 대다수의 손상 파일들에 대해서도 90% 이상의 성능으로 데이터 추출이 가능함을 확인하였으나 일부 파일에서는 복구 또는 필터링이 올바르게 되지 않는 것으로 나타났다. 예로, 데이터 추출이 가능하다고 판단되는 손상 유형①과 ②에 해당하는 excel 파일 중에서 sheet별 데이터 및 sheet의 수가 일정 수준을 초과하여 local file별 참조 값의 복잡성이 증가하거나 작성된 내용에 다수의 서식이 복합적으로 지정된 파일 등의 경우에는 올바르게 데이터 추출이 진행되지 않았다. 또한, PK ZIP 형식으로 패키징된 응용프로그램 등 필터링이 필요한 손상 유형③ 파일 중, 해당 파일 내부에 우연히 MS Office 파일의 local file name을 지닌 local file이 존재하는 경우에는 preprocessing 단계에서 필터링되지 않고 다음 과정으로 넘어가는 문제가 발생하기도 하였다.

V. 결 론

본 논문에서는 파일의 손상 및 기존 카빙 도구의 한계로 내용 확인이 불가능한 버전 2007 이후의 MS Office 파일을 대상으로 잔여하는 데이터 추출 방안 고도화 방안을 연구하였다. 이를 위해 MS Office 파일 구조를 분석 및 손상 유형을 정의하였고, 열람이 불가능하더라도 문서 내용 추측을 위한 내부 데이터가 존재한다면 이를 추출하는 알고리즘을 개발하였다. 제안하는 기법은 수사 과정에서 활용되는 기존의 포렌식 도구의 카빙 기술의 기술적 한계를 보완할 수 있는 기술로 활용될 수 있다.

이를 실무에 적용 및 법적 타당성을 갖추어 활용성을 높이기 위해서는 복구 결과물의 출처를 명시하는 방안에 관한 연구가 필요하다. 또한, 신뢰성 평가를 수행하여 추출한 정보를 디지털 증거로의 활용을 위한 추가적인 연구도 필요하다. 제안하는 기법으로도 복구하지 못한 파일 유형에 대하여 추가적인 분석을 진행 후, 도구에서 발견된 구현상의 문제를 개선, 도구의 성숙도를 높여 github 등에 오픈소스로 공개하여 해당 분야의 발전에 기여하고자 한다.

References

[1] Korean Law Information Center, "Criminal Procedure Law", https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%E

D%98%95%EC%82%AC%EC%86%8C%EC%86%A1%EB%B2%95, June. 2023.
〔2〕 ISO, "ISO/IEC 29500-1:2016", https://www.iso.org/standard/71691.html, May. 2023.
〔3〕 PKWARE Inc, "ZIP File Format", https://pkware.cachefly.net/webdocs/casestudies/APPNOTE.TXT, May. 2023.
〔4〕 Byunjoon Jung, Jaehyeok Han and Sangjin Lee, "A Method of Recovery for Damaged ZIP Files," Journal of The Korea Institute of information Security & Cryptology, 27(5), pp. 1107-1115, Oct. 2017.
〔5〕 Karl Wust, "Force Open: Lightweight black box file repair," vol. 20, pp. S75-S82, Jan. 2017.
〔6〕 Kiwon Hong, Jaehyung Cho, Soram Kim and Jongsung Kim, "Improved Data Concealing and Detecting Methods for OOXML Document," Journal of the Korea Institute of Information Security & Cryptology, 27(3), pp. 489-499, Jun. 2017.
〔7〕 Ralf D. Brown, "Improved recovery and reconstruction of DEFLATEd files", Digital Investigation, vol. 10, pp. S21-S29, Aug. 2013.
〔8〕 Noor Ul Ain Ali, Waseem Iqbal and Hammad Afzal, "Carving of the OOXML document from volatile memory using unsupervised learning techniques," Journal of Information Security and Applications, vol. 65, article. 103096, Mar. 2022.

〈저자소개〉

김 지 윤 (Jiyun Kim) 학생회원
2023년 2월: 동국대학교 경찰행정학부 학사
2023년 3월~2024년 2월: 동국대학교 일반대학원 경찰행정학과 석사과정
2024년 3월~현재: 성균관대학교 일반대학원 과학수사학과 석사과정
〈관심분야〉 정보보호, 디지털포렌식, 보안성 평가/인증 등

김 민 수 (Minsoo Kim) 학생회원
2024년 2월: 동국대학교 경찰행정학부 학사
2024년 3월~현재: 성균관대학교 일반대학원 과학수사학과 석사과정
〈관심분야〉 디지털 포렌식, 정보보호, 인공지능 등

박 우 빈 (Woobeen Park) 학생회원
2023년 2월: 동국대학교 경찰행정학부 졸업
2023년 3월~현재: 동국대학교 일반대학원 경찰행정학과 석사과정
〈관심분야〉 정보보호, 인공지능, 침해사고, 디지털포렌식 등

정 두 원 (Doowon Jeong) 정회원
2019년 2월: 고려대학교 정보보호대학원 공학박사
2020년 9월~2024년 2월: 동국대학교 조교수
2024년 3월~현재: 성균관대학교 과학수사학과 조교수
〈관심분야〉 디지털 포렌식, 정보보호, AI 등