Web で公開される PDF ファイルの Hidden data の現状の調査 —日本の警察を対象として—

長谷川太一[1] 齊藤泰一[†1] 佐々木良一[†2]

概要：日本では 2012 年に電子行政オープンデータ戦略が決定し，オープンデータが増加している．オープンデータの公開方法は Web ページ，PDF ファイル，Excel ファイルなどがある．また広報など刊行物は PDF ファイルで公開されている．PDF ファイル，Excel ファイルなどでは作成者の意図しない情報（Hidden data）が含まれている可能性があり，Hidden data は標的型攻撃などに悪用される恐れがある．そのため Hidden data の確認，消去が必要である．Hidden data を消去することをサニタイズと呼ぶ．本研究では，日本の警察が Web サイトで公開している PDF ファイルの Hidden data の調査を行い，海外の先行研究との比較を行う．

Survey of Hidden data in PDF Files Published on the Web -Targeting Police Agencies in Japan-

TAICHI HASEGAWA[1] TAIICHI SAITO[†1] RYOICHI SASAKI[†2]

1. はじめに

日本では 2012 年に電子行政オープンデータ戦略が決定し，オープンデータが増加している[1]．オープンデータの公開方法は Web ページ，PDF ファイル，Excel ファイルなどがある．また広報など刊行物は PDF ファイルで公開されている．

PDF ファイル，Word ファイルなどでは作成者の意図しない情報（Hidden data）を含んでいる可能性があり，Hidden data は標的型攻撃などに悪用される恐れがある．そのため Hidden data の確認が必要であり，センシティブな情報であれば消去すべきである．Hidden data を消去することをサニタイズと呼ぶ．

本研究では，日本の警察が Web サイトで公開している PDF ファイルの Hidden data の調査を行い，海外の先行研究[2]との比較を行う．日本の警察を対象とした理由は，刊行物を定期的に発行していることやクローラーの使用を禁止していない場合が多かったことである．

2. PDF における Hidden data

PDF における Hidden data を説明するために，まずは PDF の構造について説明する．図 1 は PDF の構造を示しており，次の 4 つに分けられる．

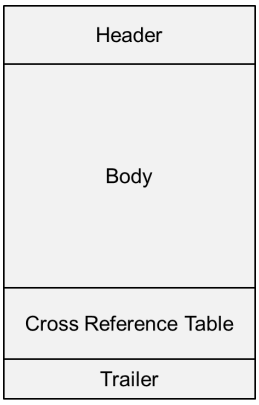

図 1 PDF の構造

Header

PDF のバージョンが記載されている．

Body

文書の内容を表し，複数のオブジェクトで構成される．

Hidden data は，オブジェクト内に存在する．

Cross Reference Table

ファイル内の各オブジェクトの位置を示したテーブルである．

各オブジェクトの表示，非表示も示している．

Trailer

クロスリファレンステーブルの位置やメタデータの位置を保持する．

1 東京電機大学 工学研究科 情報通信工学専攻，〒120-8551 東京都足立区千住旭町 5
†1 東京電機大学 工学部 情報通信工学科，〒120-8551 東京都足立区千住旭町 5
†2 東京電機大学サイバーセキュリティ研究所 客員教授，〒120-8551 東京都足立区千住旭町 5

図 2 は複数のオブジェクトで構成される Body の構造を示している．オブジェクトには作成者名，使用したソフトウェア名，作成日時など作成者の意図しない情報が含んでいるオブジェクトが存在する．

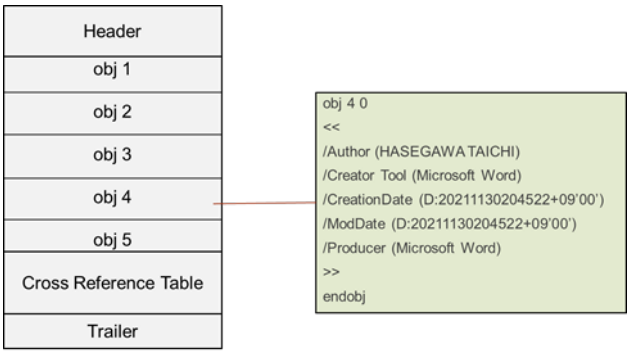

図 2　複数のオブジェクトで構成される Body

図 3 はクロスリファレンステーブルを示しており，クロスリファレンステーブルの各行は各オブジェクトと対応しており，1 つ目がオフセット位置，2 つ目が世代番号，最後が使用/未使用を表す．n が使用，f が未使用であり，未使用オブジェクトは表示されないがデータは残っている．

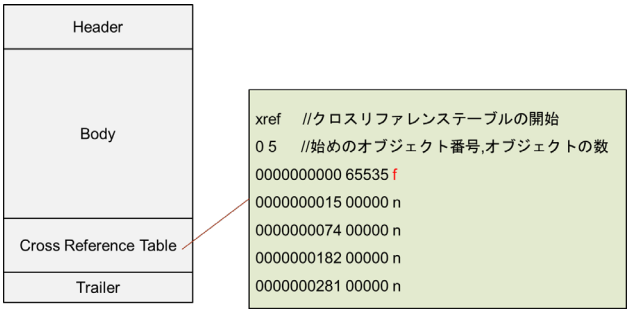

図 3　クロスリファレンステーブル

また PDF は変更・更新を行った際に Trailer の後ろに追記を行うため，サニタイズを行わないと変更前の情報が残ってしまう．図 4 は変更・更新後の PDF の構造を表している．

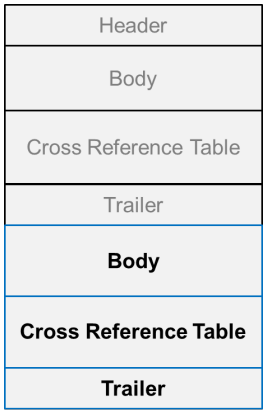

図 4　変更・更新後の PDF の構造

図 2 の右のオブジェクトはメタデータと呼ばれるデータであり，メタデータは Hidden data の一例である．図 5 はメタデータを示している．PDF のオブジェクトには辞書型と呼ばれる型があり，メタデータは辞書型で記述されている．辞書型はキーとバリューで構成され，キーは/Author，/Creator Tool などがあり，バリューに具体的な作成者名やソフトウェア名が記述される．メタデータは exiftool といったツールで見ることができる．図 6 は exiftool の出力例である．

```
obj 4 0
<<
/Author (HASEGAWA TAICHI)
/Creator Tool (Microsoft Word)
/CreationDate (D:20211130204522+09'00')
/ModDate (D:20211130204522+09'00')
/Producer (Microsoft Word)
>>
endobj
```

図 5　メタデータ

```
└─$ exiftool test1_word.pdf
ExifTool Version Number         : 12.16
File Name                       : test1_word.pdf
Directory                       : .
File Size                       : 49 KiB
File Modification Date/Time     : 2021:11:30 20:45:22+09:00
File Access Date/Time           : 2021:12:24 00:13:38+09:00
File Inode Change Date/Time     : 2021:11:30 20:50:11+09:00
File Permissions                : rwxrwxrwx
File Type                       : PDF
File Type Extension             : pdf
MIME Type                       : application/pdf
PDF Version                     : 1.7
Linearized                      : No
Page Count                      : 3
Language                        : ja-JP
Tagged PDF                      : Yes
XMP Toolkit                     : 3.1-701
Producer                        : Microsoft® Word for Microsoft 365
Creator                         : HASEGAWA TAICHI
Creator Tool                    : Microsoft® Word for Microsoft 365
Create Date                     : 2021:11:30 20:45:22+09:00
Modify Date                     : 2021:11:30 20:45:22+09:00
Document ID                     : uuid:B8AEE002-B475-4A38-AE7C-923986AFDF9D
Instance ID                     : uuid:B8AEE002-B475-4A38-AE7C-923986AFDF9D
Author                          : HASEGAWA TAICHI
```

図 6　exiftool によるメタデータの出力

3.　先行研究

Adhatarao，Lauradoux ら[2]は，47 か国 75 のセキュリティ組織が発行した PDF ファイルに対して正しく Hidden data がサニタイズされているかどうかの調査を行っている．日本の組織としては外務省（mofa.go.jp）が調査対象であった．彼らの調査では同じ人物によって作成された PDF を数年に渡って集めることで，作成者や組織の更新方針を観察している．彼らは以下のツールを用いている：wget，exiftool，pdfxplr[3]，strings，grep

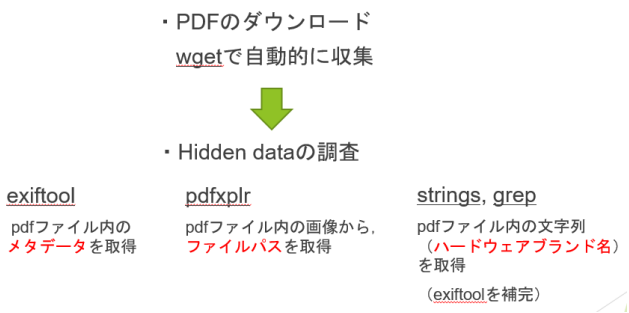

図 7　調査の概要とツールの用途

図 7 は調査の概要とツールの用途について表している．wget は Web サイトを巡回し，PDF をダウンロードするコマンドである．exiftool は PDF ファイル内のメタデータを取得することができる．pdfxplr は PDF ファイル内に存在する画像のファイルパスの抽出を行っている．strings と grep では簡単な文字検索を行うことができ，ハードウェアブランド名の取得を行っている．

彼らが収集した PDF 数は 39664 である．データセットの 13166 ファイル(33%)が，メタデータの/Author, /Creator, /Tag Author Email Display Name に作成者に関する情報を含んでいることが明らかになっている．データセットの 30155 ファイル(76%)は，メタデータの/Producer, /Creator, /Creator tool に，PDF ファイルを作成する際に使用したソフトウェア名を含んでいることが明らかになっている．ソフトウェア名は OS 名を含むことがあり，データセットの 16805 ファイル（42%）が OS 名を含んでいることが明らかになっている．データセットの 581 ファイルがハードウェアブランド名，52 ファイルがメールアドレス，1814 ファイルがファイルパスを含んでいることが明らかになっている．

4. 調査概要

日本における PDF ファイルの Hidden data の調査が必要であると考えられるため，本研究では日本の警察が公開する PDF ファイルを対象とした．

調査をするにあたり，先行研究と同様に以下のツールを用いた．

・wget
・exiftool
・pdfxplr[3]
・strings
・grep

まず PDF ファイルを収集するのには wget を用いた．クローラーの使用を禁止している場合は手動で収集した．そして Hidden data の調査には exiftool，pdfxplr，strings，grep を用いた．

5. 調査結果

5.1 調査数

先行研究の PDF 数は 39664，組織数は 75 である．それに対して我々が調査した PDF 数は 110989，組織数は警察庁+全都道府県警の 48 である．

5.2 Hidden data の調査結果

exiftool では PDF ファイルのメタデータを取得できる．メタデータは PDF の作成者名，作成日時，作成に使用したソフトウェア名などを含んでいる．PDF の作成者に関する情報はメタデータの/Author，/Creator，/Tag Author Email Display Name に含まれる．使用したソフトウェア名はメタデータの/Producer，/Creator，/Creator Tool に含まれる．

メタデータに Hidden data を含む PDF 数を表 1 に示す．作成者名は 62612 ファイル（56.4%）存在した．値は個人名，組織名，ユーザ名，組織内の番号などがあった．またソフトウェア名は 106690 ファイル（96.1%）存在した．これらは先行研究と比べると高い値となっている．OS 名，ハードウェアブランド名，メールアドレスはそれぞれ 17245 ファイル（15.5%），262 ファイル（0.2%），8 ファイル（0.01%）と先行研究より低い値であった．

人気のある PDF 作成ソフトウェア名を表 2 に示す．最も使用されたソフトウェアは JUST PDF で全体の 24%であった．順に Microsoft Office Excel, Acrobat Distiller, Microsoft Office Word，一太郎と使用率が高かった．使用されたソフトウェアは先行研究と比較すると日本向けのソフトウェア名が目立つ．

OS 名の具体的な割合を表 3 に示す．最も使用されている OS は Windows で 13.9%，続いて Mac OS で 1.6%，そして Linux は 0%であった．OS 名は以下のようにソフトウェア名に含まれていることが確認できた．

Acrobat Distiller 15.0（Windows）
Acrobat Distiller 5.0.5 for Macintosh
Mac OS X 10.4.5 Quartz PDFContext
Acrobat Distiller 20.0（macOS）

先行研究に比べて OS 名が低くなっている理由としては，上記のように OS 名はソフトウェア名に含まれているため，JUST PDF，一太郎など OS 名の記載がないソフトウェア名の使用率が高いことが影響していると考えられる．

表 1　メタデータに Hidden data がある PDF ファイル数

	PDF ファイル数（本研究）	PDF ファイル数（先行研究）
作成者名（/Author，/Creator，/Tag Author Email Display Name）	56.4%（62612）	33%（13166）
ソフトウェア名（/Producer，/Creator，/Creator Tool）	96.1%（106690）	76%（30155）
OS 名（/Producer，/Creator，/Creator Tool）	15.5%（17245）	42%（16805）
メールアドレス（/Tag Author Email，/Author，/Current User Email）	0.01%（8）	0.13%（52）

表 2　人気のある PDF 作成ソフトウェア名の比較

	PDF ファイル数（本研究）	PDF ファイル数（先行研究）
JUST PDF	24%（26834）	-
Microsoft Office Excel	13%（14743）	-
Acrobat Distiller	11%（12393）	23%（9054）
Microsoft Office Word	11%（12208）	12%（4850）
一太郎	9%（10195）	-
Microsoft Office PowerPoint	4%（4084）	-
Adobe PDF Library	3%（3789）	16%（6171）
CubePDF	3%（3166）	-
SkyPDF	3%（2953）	-
Antenna House	3%（2851）	-
その他	16%（17773）	49%（19664）

表 3　具体的な OS 名の割合

	PDF ファイル数（本研究）	PDF ファイル数（先行研究）
Microsoft Windows	13.9%（15447）	28%（11174）
Mac OS	1.6%（1798）	8%（3444）
Linux	0%（0）	6%（2187）

pdfxplr は PDF ファイル内の画像のファイルパスを取得できる．ファイルパスは日本語の文字化けの関係上，ファイルパスを含む PDF ファイルの判別ができておらず，ファイルパスの調査はできていない．

strings と grep は簡単な文字検索を行うことができる．本研究ではハードウェアブランド名は”Toshiba”，”Dell”，”Hewlett-Packard”，”Lenovo”の 4 単語で検索した．簡単な文字検索による Hidden data の取得数は表 4 に示す．

表 4　ハードウェアブランド名を含む PDF ファイル数

	PDF ファイル数（本研究）	PDF ファイル数（先行研究）
ハードウェアブランド名	0.2%（262）	1.5%（581）

5.3 組織で漏れている情報

組織別のソフトウェア名の使用率では，48 組織中 19 組織で JUST PDF が最も使用されている．表 5 は 5 つの組織による PDF 作成ソフトウェア名の違いを示している．組織 A，B，E は JUST PDF の使用率が最も高い．また組織 D のように特定のソフトウェアの使用率が高い組織が存在した．

表 5　組織による PDF 作成ソフトウェア名の違い

	A	B	C	D	E
Acrobat Distiller	0	3156	1143	49	533
Adobe PDF Library	46	824	1489	12	76
Antenna House	26	10	2	2087	0
CubePDF	1	9	3	0	0
JUST PDF	1344	5282	27	0	1597
Microsoft Office Excel	118	61	2470	19	107
Microsoft Office PowerPoint	17	170	344	7	52
Microsoft Office Word	33	406	492	206	77
SkyPDF	1	2613	0	0	2
一太郎	290	68	30	0	87
その他	338	1573	834	18	219

5.4 サニタイズに関する考察

サニタイズの方法は Adobe Acrobat tool を使用することが推奨される．Adobe Acrobat tool は先行研究でも推奨され

ている．

サニタイズについては，先行研究ではレベル 0 からレベル 3 までの 4 段階のサニタイズレベルを決めており，レベル 3 のサニタイズレベルが推奨されている．それぞれのレベルについては次のように述べられている・

レベル 0 は PDF ファイルが完璧なメタデータを持っており，サニタイズが行われてない．

レベル 1 は PDF ファイルが部分的なメタデータを持っており，いくつかのメタデータの項目は削除されている．

レベル 2 はメタデータが含まれてない PDF ファイルである．それらは exiftool でサニタイズを行っているかメタデータのない PDF ファイルを直接作成したものである．

レベル 3 は情報漏えいがなく適切にサニタイズされた PDF ファイルである．PDF ファイル内のセンシティブな情報を持つ全てのオブジェクトが削除されている．そしてレベル 3 は Adobe Acrobat tool を使用することで満たされる．

先行研究では Adobe Acrobat tool について，NSA のガイドライン[5,6,7]でしばしば言及される信頼性の高いサニタイズツールであり，彼らが使用したツールの中で最も完全なサニタイズツールであると述べられている．

そしてレベル 2 が不十分である理由も述べられており，彼らの調査ではレベル 2 の PDF ファイルの中には grep で復元を行えたものがあり，情報が洩れる可能性があるためレベル 2 では不十分であると述べられている．

6. おわりに

6.1 研究結果のまとめ

本研究では日本の警察が Web で公開している PDF ファイルの Hidden data の調査を行うとともに，海外の公的機関との比較を行った．

その結果，日本の警察においては作成者名が 56.4%，ソフトウェア名は 96.1%，メールアドレス，ハードウェアブランド名は 1%未満，OS 名は 15.5%という結果になった．ソフトウェア名が 96.1%であることから，レベル 2 または 3 の PDF ファイルは多くても 4%程である．

先行研究では作成者名，ソフトウェア名はそれぞれ 33%，76%であり，今回の調査では先行研究に比べて値が約 20%高いことが分かった．メタデータの/Tag Author Email，/Author, /Current User Email に含まれるメールアドレスやハードウェアブランド名は先行研究がそれぞれ 0.13%，1.5%であり，先行研究より値が低いことが分かった．そして OS 名は先行研究が 42%であり，先行研究に比べて値が 20%以上低かったが，これは OS 名がソフトウェア名に含まれるため，ソフトウェア名に OS 名の記載がないソフトウェア名の使用率が高いことが影響していることが考えられる．

メールによる標的型攻撃では氏名，メールアドレスが必要であり，作成者名，メールアドレスが残っていると標的型攻撃メールの足がかりになってしまう．そして使用したソフトウェア名がわかるソフトウェア名が残っていると脆弱性を狙われる可能性がある．このように Hidden data は悪用される恐れがあるため，警察だけでなく全ての組織で，5.4 で述べたような方法でサニタイズを行うことが望ましい．特にメールアドレスについては除去することが推奨される．

6.2 今後の研究計画

今回の調査では pdfxplr は日本語が文字化けしている．また strings では PDF 内の日本語のコメントなどを検索できないため，日本語に対応した調査の必要がある．

今回の調査はメタデータが対象であったが NSA のガイドラインでは Hidden data は 11 種類ある[5]とされており，メタデータ以外の Hidden data についても調査を行うことを考えている．

そして PDF ファイルだけでなく Word ファイルの Hidden data の研究[8]も行われており，その調査も行いたと考えている．

参考文献

[1] 政府 CIO ポータル：オープンデータ，入手先〈https://cio.go.jp/sites/default/files/uploads/documents/digital/opendata_lg_rate_20220112.pptx〉（参照 2022-05-02）

[2] S. Adhatara, C. Lauradoux, Exploitation and Sanitization of Hidden Data in PDF Files: Do Security Agencies Sanitize Their PDF Files?, *Proc. IH&MMSec '21 ACM Workshop on Information Hiding and Multimedia Security*, pp35-44, (2021).

[3] sowdust: Extract hidden data from pdf files, GitHub, 入手先〈https://github.com/sowdust/pdfxplr〉 (アクセス 2022-05-02)

[4] John Whitington（著），村上雅章（訳）: PDF 構文解析，pp.35-68，株式会社オライリー・ジャパン（2020）.

[5] NSA: Hidden Data and Metadata in Adobe PDF Files: Publication Risks and Countermeasures. Technical Report, National Security Agency.

[6] NSA: Redaction of PDF Files Using Adobe Acrobat Professional X, Technical Report I73-025R-2011, National Security Agency.

[7] Systems and Network Attack Center: Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word to PDF. Technical Report, National Security Agency.

[8] Catiglione, A., Santis, De.A. and Soriente, C. Talking advantage of a disadvantage: Digital forensics and steganography using document metadata, *J Syst Softw*, Vol.80, No.5, pp.750-764 (2007).

付録　Hidden data のランク付け

Hidden data のランク付けを標的型攻撃に使用される点から行った. そのため, 始めに標的型攻撃について説明する.

標的型攻撃は次のような流れである.

1. 標的型攻撃メールに必要な情報の収集
2. なりすましメールの作成，悪意のあるファイルの添付
3. 添付ファイルを開かせる
4. 悪意のあるプログラムが動作

標的型攻撃メールを行う上で，氏名とメールアドレスは必須であり，その手立てとなる Hidden data の作成者名，メールアドレスの重要度を考える.

作成者名は個人名，組織名やユーザ名や組織内番号などが存在した. 個人名はメールを送る際に悪用可能なため「非常に困る」とするが，組織名などは悪用されづらいと考えられるため，「あまり困らない」とする.

メールアドレスは個人のメールアドレス，非公開の組織のメールアドレス，公開しているメールアドレスの 3 つが考えられる. 個人のメールアドレスはビジネス用のメールアドレスではないことも考えられるため「特に困る」とする. 非公開の組織のメールアドレスは作成者名と合わせることで個人にメールを送ることができるため「困る」とする. 公開しているメールアドレスは漏れても「あまり困らない」とする.

表 6 に Hidden data のランク付けをしたものを示す. また今回調査した作成者名，メールアドレス以外の Hidden data も長期的に PDF の収集を行ったときの重要度も示す.

PDF を長期的に収集することで，組織全体が使用しているソフトウェア名や特定の作成者が古いバージョンなどを使用しているかなどがわかる. そのためソフトウェア名，バージョン記載がある OS 名は「非常に困る」とする. ファイルパスは氏名を含むことがあり，また画像やファイルの保存場所がわかるため「困る」とする. バージョンの記載がない OS 名，ハードウェアブランド名は攻撃者が得る情報は少ないと考えられるため「あまり困らない」とする.

表 6　Hidden data のランク

	Hidden data	Hidden data（長期的）
特に困る	**個人のメールアドレス**	—
非常に困る	**個人名の作成者名**	**ソフトウェア名，バージョンの記載がある OS 名**
困る	**非公開の組織のメールアドレス**	**ファイルパス**
あまり困らない	**組織名の作成者名，公開しているメールアドレス**	**バージョンの記載がない OS 名，ハードウェアブランド名**