Google 検索演算子を用いた収集手法による 日本の国立大学で公開されている電子文書から漏洩しうる情報の調査・分析

岩本 実樹[1†] 戸田 季那[2] 渡辺 康介[1] 宮地 麟[1] 齊藤 泰一[2]

概要：IPA が公表する「情報セキュリティ 10 大脅威 2025［組織編］」において，「機密事項等を狙った標的型攻撃」は第 5 位にランクインされており，10 年連続で脅威として選出されている．このような標的型攻撃では，攻撃者は特定の組織を狙い，攻撃の足がかりとなる脆弱な情報を得るために，事前の情報収集を行う．攻撃者は組織内部に侵入することなく，インターネット上から多くの情報を取得可能である．その情報源の一つとして，組織が Web 上で公開する電子文書が挙げられる．PDF や Microsoft Office ファイルなどの電子文書には，メタデータと呼ばれる非表示情報が含まれており，氏名，メールアドレス，使用ソフトウェア名などが意図せず公開されるリスクがある．本稿では，日本の国立大学全 86 校を対象に，Google 検索演算子を用いて PDF ファイルおよび Microsoft Office ファイルを収集し，それらに含まれるメタデータの実態を調査した．その結果，Microsoft Office ファイルの方が，PDF ファイルよりもメタデータに文字列を含む割合が高かった．

キーワード：Google 検索演算子，電子文書，隠れ情報，メタデータ，標的型攻撃

Research and analysis of information that could be leaked from electronic documents published at Japanese national universities using collection techniques based on Google search operators.

Miki Iwamoto[1†] Tokina Toda[2] Kousuke Watanabe[1] Rin Miyachi[1]
Taiichi Saito[2]

Abstract: In the "10 Major Security Threats 2025 [Organizational Edition]" published by the IPA, "Targeted Attacks Aiming at Confidential Information" is ranked fifth and has been selected as a threat for ten consecutive years. In such targeted attacks, adversaries conduct prior information gatherings to obtain vulnerable information that can serve as a foothold for their attack on a specific organization. Attackers can acquire a significant amount of information from the internet without infiltrating the organization's internal network. One source of this information is electronic documents that organizations publish on the web. Electronic documents such as PDF and Microsoft Office files contain hidden information called metadata, which poses a risk of unintentionally disclosing names, email addresses, and software names. In this paper, we investigated the actual state of metadata by collecting PDF files and Microsoft Office files from all 86 national universities in Japan using Google search operators. The results showed that Microsoft Office files had a higher percentage of metadata containing character strings than PDF files.

Keywords: Google Search Operators, Electronic Documents, Hidden Information, Metadata, Targeted Attack

1. はじめに

近年，特定の組織や個人を狙って機密情報の窃取やシステム侵入を試みる標的型攻撃が深刻化している．IPA（独立行政法人情報処理推進機構）の「情報セキュリティ 10 大脅威 2025［組織］」[1]では，「機密情報等を狙った標的型攻撃」が第 5 位に位置付けられており，2016 年以降，毎年脅威として取り上げられている．PDF や Microsoft Office ファイルなどの電子文書には，ファイルを開いたときにユーザインタフェース上に表示されない情報が存在する．本稿では，このようなデータを「隠れ情報」と呼ぶ．隠れ情報は，特別なツールを使用しないと確認できないため，Web サイトに電子文書を公開する者は，意図せずそれを含んだ状態で公開している可能性がある．その中には，攻撃者が標的型メールを作成する際の手掛かりとなる情報が含まれる場合がある．

電子文書に含まれる隠れ情報については，国内外で調査が行われている．海外の組織を対象とした調査では，Google 検索演算子を収集手法とした提案がされている[2]．国内の組織を対象とした調査では，クローラーツールを用いた収集を行い，警察庁と日本の大学を対象とした調査が行われている[3][4][5]．しかし，日本の組織を対象とした調査では，Google 検索演算子を用いた調査は行われていない．また，日本の大学を対象とした調査では，大学の学生数を基準として調査対象を限定していた．本研究は，これらの課題を踏まえ，日本の全ての国立大学を対象に，Google 検索演算

1 東京電機大学大学院，〒120-8551 足立区千住旭町 5 番
Tokyo Denki University, 5 Senju Asahi-cho Adachi-Ku
2 東京電機大学，〒120-8551 足立区千住旭町 5 番
Tokyo Denki University, 5 Senju Asahi-cho Adachi-Ku
† 25kmc07@ms.dendai.ac.jp

子を用いて公開電子文書を収集し，隠れ情報の実態を調査する．

2. 関連研究

まず，本調査を行う上で先行研究にあたる 3 つの研究を以下に示す．

1 つ目は，Karl Mendelman らによるエストニア政府機関の Web サイトで公開されている電子文書(PDF, MS バイナリファイル，MS OOXML ファイル)の調査である[2]．この調査では，3 つの異なる検索エンジンとその検索機能を用いてファイルの収集が行われた．本稿における調査手法は，彼らの収集方法の 1 つである Google 検索と Google 検索演算子を用いた収集方法を基に行った．

2 つ目は，長谷川らによる調査である．長谷川らは，日本の大学が Web サイトで公開している電子文書を対象にメタデータを含む隠れ情報の調査を行っている[3]．本稿のメタデータの調査は，この研究を参考にして，メタデータの値が含まれているか確認した．一方，この研究では電子文書の収集にクローラーツールを用いていたが，我々の研究では，収集方法として Google 検索演算子を採用している点で異なる．

3 つ目は，長谷川らによる日本の警察が発行する PDF ファイルに関する調査である[4][5]．長谷川らは，PDF ファイルのメタデータを対象に調査を行った．その研究の中で，電子文書が含むソフトウェアに関する隠れ情報から，どのようなソフトウェアが使われているのか調査している．本研究ではこの成果を参考にし，大学の電子文書を対象としてソフトウェア情報の分析を拡張し，利用傾向をより詳細に検討する．

3. 調査手法

3.1 収集について

3.1.1 調査対象

本稿において調査対象とする大学は，2024 年 12 月時点において，文部科学省のホームページに記載のある日本の国立大学全 86 校とした[6]．

本稿で収集対象としたファイル形式は，PDF ファイルおよび Microsoft Office ファイルである．PDF ファイルは拡張子が“.pdf”であるものとし，Microsoft Office ファイルは，MS バイナリ形式では拡張子が“.doc”，MS OOXML 形式では拡張子が“.docx”，“.xlsx”，“.pptx”であるものとした．さらに，調査対象はメタデータ内に filename キーが存在するファイルに限定した．この filename キーにはファイル名に関する情報が格納される．

本稿では，調査対象の大学を 5 つのグループに分類してグループ毎に調査・分析を行った．分類の基準として，文部科学省の Web サイトにて用いられている分類方法である，大学の設置地区における分類を用いた．この分類方法では，以下の 5 つのグループに分類される．

- 北海道・東北地区
- 関東・甲信越地区
- 東海・北陸・近畿地区
- 中国・四国地区
- 九州・沖縄地区

3.1.2 収集ツール

調査に使うツールやコマンドは，ファイルの収集に使用するツールとファイルの分析に使うツール・コマンドがある．

先行研究では，クローラーと呼ばれる，URL を指定して Web サイト内を巡回し，ファイルをダウンロードできるツールを使用していた．しかし，本研究では Google 検索を行い，以下に示す二つの手法を組み合わせて収集を行った[7]．

1 つ目は検索演算子を使用した．"site: "と"filetype: "である．"site: "は，指定したドメインの Web サイトに限定して検索結果を表示することができる．"filetype: "は，特定のファイル形式に限定して検索結果を絞り込むことができる．"site: "と"filetype: "の検索演算子を組み合わせて用いることで，指定したドメインの Web サイトで公開されている，指定したファイル形式のみを検索結果に表示することができる．

本稿では，文部科学省のサイトに掲載されている各国立大学の Web サイトの URL より，ドメインを抽出して調査対象とした[6]．

```
site:dendai.ac.jp filetype:doc
```

図 1. 検索演算子の利用例

2 つ目はファイルの実際の収集のために，DownThemAll!を用いた[8]．DownThemAll!とは，ダウンロードマネージャーとして機能し，ダウンロードの一時停止，再開，キュー管理などを行うことができる．このツールはウェブページ上の多数のファイルや画像を一度にダウンロードするために使用される．今回は，電子文書をダウンロードするために用いた．

これらの二つの手法を用いてファイル収集を効率化した．例外として次の事例は手動で行った．

- 検索結果を表示させる
- 次のタブを開く
- エラーが発生した際の対応

3.2 隠れ情報の調査手法

本稿では，隠れ情報の一種であるメタデータの分析を行った．メタデータとは，ファイルに関する情報がまとまったデータのことである．ファイルのメタデータは，キーと値の形式からなり，作成者名や作成日時，使用言語など，

様々な情報を保持している．

メタデータの分析には exiftool（バージョン 12.97）を用いた．exiftool はファイルのメタデータを取得することができるツールである．exiftool を用いてメタデータを取得した際の例を以下に示す．

```
Page Count                      : 5
Language                        : ja
Tagged PDF                      : Yes
XMP Toolkit                     : 3.1-701
Producer                        : Microsoft® Word for Microsoft 365
Creator                         : IWAMOTO MIKI
Creator Tool                    : Microsoft® Word for Microsoft 365
Create Date                     : 2025:07:30 17:09:10+09:00
Modify Date                     : 2025:07:30 17:09:10+09:00
Document ID                     : uuid:8A018802-7D49-43A9-9B51-F87B7B78A286
Instance ID                     : uuid:8A018802-7D49-43A9-9B51-F87B7B78A286
Author                          : IWAMOTO MIKI
```

図 2. メタデータの例

本稿では，隠れ情報の調査を PDF ファイル，MS バイナリファイル，MS OOXML ファイルに分類して実施した．これは，ファイル形式によって含まれるメタデータの種類が異なるためである．各形式において，隠れ情報として格納されるメタデータのキーを抽出し，調査対象とした．抽出する情報および対応するメタデータキーについては，関連研究[3]を参照しつつ，本稿独自のキーも一部追加した．

なお，各項目において複数の該当キーに値が存在する場合でも重複してカウントせず，1 回として扱った．

3.2.1 PDF ファイルの調査手法

まず，PDF ファイルに付与されたメタデータから抽出可能な情報について調査を行った．抽出できる情報とそれに対応するメタデータのキーについては下記である．

- 作成者名(/Author，/Creator，/Tag Author Email Display Name)
- ソフトウェア名(/Producer，/Creator，/Creator Tool)
- OS 名（/Producer，/Creator，/Creator Tool）
- メールアドレス（/Author，/Creator*，/Creator Work Email*，/Current User Email，/Tag Author Email）

これらの項目を分析した．なお，「*」を付したキーは，関連研究[3]で示されたキーに，本研究で新たに追加したものである．

作成者名は，/Author，/Creator，/Tag Author Email Display Name のいずれかのキーに値が存在する場合にカウントした．ただし，OS 名，メールアドレス，ソフトウェア名が含まれている場合は，作成者名が存在するとは判定しない．

ソフトウェア名については，/Producer，/Creator，/Creator Tool のいずれかのキーの値にソフトウェアの名称を含んでいればカウントする．判定に用いた正規表現については，付録に示す．また，/Producer，/Creator，/Creator Tool のキーの値に Windows, Macintosh, Linux, Mac OS X，macOS の文字列が含まれている時がある．この場合に OS 名が存在するとカウントした．

メールアドレスは，/Author，/Creator，/Creator Work Email，/Current User Email，/Tag Author Email のいずれかにメールアドレスが含まれている場合にカウントした．メードアドレスの判定には図 3 の正規表現を用いた．

```
[A-Za-z0-9_.+-]+@[A-Za-z0-9-]+¥.[A-Za-z0-9-.]+
```

図 3. メールアドレスの正規表現

さらに，先行研究[4]を参考に，ソフトウェア名が検出されるごとに，/Author，/Creator，/Tag Author Email Display Name に文字列が含まれている割合を調査した．調査した結果を 4.1 の表 2 にまとめた．

3.2.2 MS バイナリファイルの調査手法

まず，MS バイナリファイルに付与されたメタデータから抽出可能な情報について調査を行った．抽出できる情報とそれに対応するメタデータのキーについては以下である．

- 作成者名（/Author，/Tag Author Email Display Name*）
- 最終編集者名（/Last Modified By，/Last Saved By*）
- ソフトウェア名（/Software，/Identification*）
- バージョン情報（/App Version）
- メールアドレス（/Tag Author Email，/Author，/Last Modified By，/Last Saved By*）

これらの項目を分析した．なお，「*」を付したキーは，関連研究[3]で示されたキーに，本研究で新たに追加したものである．

作成者名，ソフトウェア名，メールアドレスについては，PDF ファイルの調査と同様の基準に基づいて判定を行った．

最終編集者名は，/Last Modified By，/Last Saved By のキーに，文字列が含まれているか調査した．なお，メールアドレス及びソフトウェア名の値が含まれている場合は除外した．

バージョン情報については，/App Version のキーに値が存在する場合にカウントした．

さらに，ソフトウェア名が検出されるごとに，/Author，/Tag Author Email Display Name に文字列が含まれている割合を調査した．調査した結果を 4.2 の表 4 にまとめた．

3.2.3 MS OOXML ファイルの調査手法

まず，MS OOXML ファイルに付与されたメタデータから抽出可能な情報について調査を行った．抽出できる情報とそれに対応するメタデータのキーについては以下である．

- 作成者名（/Creator,/_AuthorEmailDisplayName*）
- 最終編集者名（/Last Modified By）
- ソフトウェア名（/Application, /Creator*）
- バージョン情報（/App Version）
- メールアドレス（/_AuthorEmail*, /Creator, /Last Modified By, /Tag Author Email）

これらの項目を分析した．なお，「*」を付したキーは，関連研究[3]で示されたキーに，本研究で新たに追加したものである．

MS OOXML ファイルのメタデータの調査においては，対応するキーは異なるが，それぞれの情報を含むかの判定基準は MS バイナリファイルと同様の基準とした．

さらに，ソフトウェア名が検出されるごとに，/Creator, /_AuthorEmailDisplayName に文字列が含まれている割合を調査した．調査した結果を 4.3 の表 6 にまとめた．

4. 調査結果

4.1 PDF ファイルの調査結果

PDF ファイルにおけるメタデータの分析結果について述べる．表 1 は，国立大学の PDF ファイルにおけるメタデータの調査結果である．調査した PDF ファイルは，24,496 件だった．

表 1. 国立大学の PDF ファイルにおけるメタデータの調査結果

	北海道・東北地区	関東・甲信越地区	東海・北陸・近畿地区	中国・四国地区	九州・沖縄地区	全体
作成者名	64.7%	48.6%	60.3%	52.6%	53.8%	55.7%
ソフトウェア名	94.4%	89.1%	93.2%	89.9%	94.4%	91.9%
os名	17.9%	17.1%	16.0%	13.7%	16.9%	16.5%
メールアドレス	0.17% (7)	0.35% (26)	0.66% (45)	0.25% (7)	0.66% (21)	0.43% (106)

作成者名に文字列が含まれているファイルの割合は，全体で 55.7%である．最も数値が高いグループでは北海道・東北地区の 64.7%であった．

ソフトウェア名を含むファイルの割合は，いずれのグループでも 9 割前後であり，他項目と比較しても，ソフトウェア名が含まれている割合が高いことがわかる．

OS 名を含むファイルの割合は，最も数値が低いグループは中国・四国地区の 13.7%であり，最も数値が高いグループは北海道・東北地区の 17.9%であった．また，いずれのグループでも 20%以下の割合であることがわかる．

メールアドレスを含むファイルの割合は，いずれのグループでも 1%未満である．1%未満と低い割合ではあるが，全体で 106 個の PDF ファイルにメールアドレスが含まれていた．

PDF ファイルのメタデータの調査結果から，メタデータが含まれる割合は地区間で大きな差がなかった．

先行研究[4]を参考に，ソフトウェア名が検出されるごとに，/Author，/Creator，/Tag Author Email Display Name に文字列が含まれている割合を調査した結果を表 2 にまとめた．

表 2．ソフトウェア名が検出される場合における作成者名が検出される割合（PDF ファイル）

ソフトウェア名	/Author	/Creator	/Tag Author Email Display Name
Adobe PDF Library(5293)	42.8%	42.7%	0.1%
Acrobat Distiller(4919)	67.0%	68.8%	0.1%
Word(3457)	75.7%	40.4%	0.0%
Word 用 Acrobat Pdfmaker(2275)	75.8%	75.6%	0.2%
Just PDF(1854)	67.2%	13.6%	0.0%
Powerpoint(891)	79.8%	44.7%	0.0%
Quartz Pdfcontext(731)	50.6%	19.8%	0.0%
Excel(659)	81.5%	40.5%	0.0%
Adobe Indesign(579)	9.7%	9.3%	0.0%
iText(531)	24.9%	14.5%	0.0%

PDF ファイルにおける作成者名の検出割合は，利用されたソフトウェアによって大きく異なることが分かる．Microsoft Office 製品に関連するソフトウェアでは，/Author に作成者名が記録される割合が高く，特に Excel（81.5%），Powerpoint（79.8%），Word 用 Acrobat Pdfmaker（75.8%）において高い割合が確認された．一方，Adobe Indesign では 9.7%，iText では 24.9%と，/Author に作成者名が含まれる割合は低く，ソフトウェアの種類によって情報の残りやすさに大きな差があることが分かる．/Creator に関しても同様に，Acrobat Distiller（68.8%），Word 用 Acrobat Pdfmaker（75.6%）では比較的高い割合で記録が確認されたが，Just PDF（13.6%），Indesign（9.3%），iText（14.5%）といったソフトウェアでは低い傾向にあった．さらに，/Tag Author Email Display Name が記録される割合はすべてのソフトウェアで 0.2%以下と低いことがわかる．

4.2 MS バイナリファイルの調査結果

MS バイナリファイルにおけるメタデータの分析結果に

ついて述べる．表 3 は，国立大学の MS バイナリファイルにおけるメタデータの調査結果である．調査した MS バイナリファイルは，13,987 件だった．

表 3. 国立大学の MS バイナリファイルにおけるメタデータの調査結果

	北海道・東北地区	関東・甲信越地区	東海・北陸・近畿地区	中国・四国地区	九州・沖縄地区	全体
作成者名	85.1%	82.7%	88.9%	86.6%	87.8%	86.3%
最終編集者名	87.2%	83.5%	89.3%	88.1%	88.8%	87.3%
ソフトウェア名	96.6%	97.4%	97.7%	96.4%	97.4%	97.2%
バージョン情報	97.0%	98.1%	97.9%	96.3%	97.7%	97.5%
メールアドレス	1.31% (25)	2.86% (98)	1.66% (67)	2.20% (54)	1.39% (30)	1.96% (274)

作成者名に文字列を含むファイルの割合は，いずれのグループでも 8 割以上である． 全体の割合に関して，PDF ファイルの結果と比較すると，MS バイナリファイルでは作成者名に文字列が含まれる割合が約 30%高いことがわかった．

最終編集者名に文字列を含むファイルの割合は，いずれのグループでも 8 割を超えており，最も数値が高いグループは東海・北陸・近畿地区の 89.3%である．

ソフトウェア名を含むファイルの割合は，いずれのグループでも 9 割を超えている．

バージョン情報を含むファイルの割合は，ソフトウェア名と同様に，いずれのグループでも 9 割を超えている．全体を見ると，バージョン情報が含まれている割合は 97.5%と，5 項目の中で最も高い割合である．

メールアドレスを含むファイルの割合は，いずれのグループでも 3%未満である．3%未満と低い割合ではあるが，全体で 274 個の MS バイナリファイルにメールアドレスが含まれていた．

MS バイナリファイルの調査結果でも，PDF ファイルの調査結果と同様に，メタデータが含まれる割合は地区間で大きな差はなかった．

表 4. ソフトウェア名が検出される場合における作成者名が検出される割合（MS バイナリファイル）

ソフトウェア名	/Author	/Tag Author Email Display Name
Word(13596)	86.6%	1.1%
Microsoft Office Word(11813)	86.4%	1.0%
Microsoft Word(1348)	90.7%	1.8%

ソフトウェア名が検出されるごとに，/Author， /Tag Author Email Display Name に文字列が含まれている割合の調査結果を表 4 にまとめた．

MS バイナリファイルにおいては，いずれのソフトウェア名が検出された場合でも，8 割以上のファイルで /Author に作成者名に文字列が含まれていることが分かる．特に Microsoft Word の場合，/Author に作成者名に文字列が含まれる割合が最も高く，90.7%に達していた．一方，/Tag Author Email Display Name では，すべてのソフトウェアにおいて 1%程度と割合は低く，最大でも 1.8%であった．

4.3 MS OOXML ファイルの調査結果

MS OOXML ファイルにおけるメタデータの分析結果について述べる．表 3 は，国立大学の MS OOXML ファイルにおけるメタデータの調査結果である．調査した MS OOXML ファイルは，26,681 ファイルである．

表 5 国立大学の MS OOXML ファイルにおけるメタデータの調査結果

	北海道・東北地区	関東・甲信越地区	東海・北陸・近畿地区	中国・四国地区	九州・沖縄地区	全体
作成者名	80.7%	77.7%	85.6%	85.6%	87.2%	83.3%
最終編集者名	80.6%	78.0%	84.8%	87.2%	87.5%	83.4%
ソフトウェア名	93.4%	92.0%	93.0%	95.8%	95.0%	93.6%
バージョン情報	99.7%	97.4%	99.6%	99.0%	99.8%	99.0%
メールアドレス	0.61% (18)	1.62% (111)	1.75% (145)	1.35% (58)	0.88% (38)	1.39% (370)

作成者名に文字列を含むファイルの割合は，いずれのグループでも 8 割前後である．

最終編集者名に文字列を含むファイルの割合は，いずれのグループでも 8 割前後であった．全体の割合に関して，MS バイナリファイルの調査結果と比較すると， 4%程度の差はあるものの，同程度の割合で最終編集者名が含まれていることがわかった．

ソフトウェア名を含むファイルの割合は，いずれのグループでも 9 割を超えている．

バージョン情報を含むファイルの割合は，ソフトウェア名と同様に，いずれのグループでも 9 割を超えている．最も高いグループの数値は，九州・沖縄地区の 99.8%であり，全体でも 99.0%と高い数値である．

メールアドレスを含むファイルの割合は，いずれのグループでも 2%未満と他の項目より低く，全体では 1.39%であった．

MS OOXML ファイルの調査結果でも， PDF ファイル，MS バイナリファイルの調査結果と同様に，メタデータが含まれる割合は地区間で大きな差はなかった．

ソフトウェア名が検出されるごとに，/Creator, /_AuthorEmailDisplayName に文字列が含まれている割合を調査した結果を表 6 にまとめた．

表 6. ソフトウェア名が検出される場合における作成者名が検出される割合（MS OOXML ファイル）

ソフトウェア名	/Creator	_AuthorEmailDisplayName
Microsoft Office Word(15116)	85.4%	0.3%
Microsoft Excel(7826)	83.3%	0.1%
Microsoft Office Powerpoint(1714)	91.8%	0.0%
Powerpoint(602)	92.4%	0.0%
Excel(320)	64.4%	0.0%
Word(262)	80.5%	0.4%
Adobe Indesign(36)	0.0%	0.0%
Adobe Illustrator Cs(16)	0.0%	0.0%
Adobe Indesign CC(15)	0.0%	0.0%
Microsoft Office Outlook(14)	85.7%	0.0%
Libreoffice(11)	45.5%	0.0%
Adobe Acrobat Pro(4)	0.0%	0.0%

Microsoft 製品が検出された場合には，/Creator に文字列が含まれる割合が高いことが分かった．中でも PowerPoint が最も高く，92.4%のファイルにおいて /Creator に文字列が含まれていた．一方で，/_AuthorEmailDisplayName に情報が含まれる割合は最大でも 0.4%であった．また，Adobe 製品についてもソフトウェア名として検出されたが，/Creator および /_AuthorEmailDisplayName のいずれにも文字列が含まれていることは確認されなかった．

また，Microsoft Office Word と Word など同一のソフトウェアであると考えられるが，別々のソフトウェア名として検出されるケースが確認された．

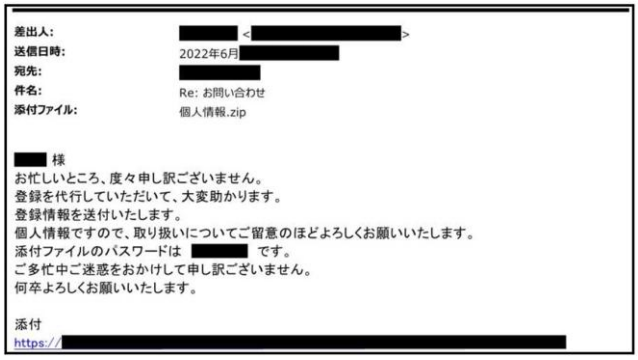
5. 考察

PDF ファイル, MS バイナリファイル, および MS OOXML ファイルの調査結果から, MS バイナリファイルおよび MS OOXML ファイルの方が PDF ファイルよりも多くの情報を含む割合が高いことが明らかとなった．特に，メールアドレスが含まれる割合は，いずれのファイルタイプにおいても他の項目に比べて低い傾向があり，3 つのファイル形式を比較した場合，PDF ファイルが最も割合が低く，件数も少ないことが分かった．

このように，PDF ファイルと MS Office ファイルに含まれる情報量に差が生じる背景には，ファイル構造の違いが関係していると考えられる．MS バイナリ形式および MS OOXML 形式は，文書編集に伴い詳細な属性情報を自動的に保存する機能を持つ．特に OOXML 形式は，ZIP 圧縮された複数の XML ファイル群で構成され，編集の過程で多様な要素が追加されるため，情報が残存しやすいと考えられる[9][10][11].

5.1 結果から考えられる攻撃シナリオ

調査結果から考えられる攻撃として，標的型メールを作成する足掛かりとなることが考えられる．


差出人: [redacted] <[redacted]>
送信日時: 2022年6月[redacted]
宛先: [redacted]
件名: Re: お問い合わせ
添付ファイル: 個人情報.zip

[redacted] 様
お忙しいところ、度々申し訳ございません。
登録を代行していただいて、大変助かります。
登録情報を送付いたします。
個人情報ですので、取り扱いについてご留意のほどよろしくお願いいたします。
添付ファイルのパスワードは [redacted] です。
ご多忙中ご迷惑をおかけして申し訳ございません。
何卒よろしくお願いいたします。

添付
https://[redacted]

図 4. シンクタンクに送信された標的型メールの文面[12]

警視庁は，シンクタンクを対象とした標的型攻撃メールの事例を公開している［12］．同事例では，不正プログラムを埋め込んだ添付ファイルを開封させるよう誘導するメールが送信されていた．氏名やメールアドレスは標的型メールの作成に利用され得る有効な情報であり，電子文書に作成者情報を残したまま外部に公開することはリスクを高める要因となり得る．本研究では，国立大学が公開している電子文書を対象に調査を行った結果，大学職員が作成したと推測されるファイルが多数確認された．特に，Web サイト上で非公開となっている大学職員の氏名が作成者情報に含まれていた場合，攻撃者は外部では入手困難な識別情報を取得できる．この情報を差出人情報に利用することで，受信者に高い信憑性を与えるメールを構築可能となり，結果として攻撃成功率の上昇につながると考えられる．さら

に，メタデータに記録されるソフトウェア名やバージョン情報，作成日時などから，対象組織がサポート期限を過ぎたソフトウェアを使用しているかを推定できる場合がある．この情報を悪用すれば，攻撃者は電子文書に含まれる情報を手掛かりとして脆弱性を特定し，当該脆弱性を利用可能な不正ファイルを送信する，あるいはリモートから脆弱性攻撃を実行するといったシナリオが成立し得る．

5.2 公開前データに対するサニタイズ処理の適用

電子文書を公開する前には，内容に不要な情報が含まれていないかを確認し，万が一含まれている場合にはサニタイズなどの適切な処理を行う必要がある．

英国データ保護機関（ICO）では，組織が文書を安全に開示できるよう，文書の一般公開に関するガイダンス[13]を公開している．ICO は，個人情報が文書に隠されている場合の偶発的な侵害を防ぐためのチェックリストを提供している[14]. 文書を公開する組織は，公開前に同チェックリストを参照し，意図しない情報が含まれていないか確認することが望ましい．

PDF ファイルにおけるサニタイズ方法として，Adobe Acrobat tool を使用する方法を挙げる[15]．Adobe Acrobat tool を使用すると，PDF ファイルに含まれる非表示情報を検索して削除することができる．

MS Office ファイルにおけるサニタイズ方法として，標準機能であるドキュメント検査を用いて行う方法を挙げる[16]．ドキュメント検査によりユーザはドキュメント内に個人情報などの情報がないか確認でき，不要な情報を削除できる．

6. おわりに

6.1 まとめ

本稿では，国立大学を文部科学省の分類に基づいて 5 個のグループに分け，PDF と Microsoft Office ファイルの調査・分析を行った．調査の結果，PDF，MS Office ファイル共に，電子文書内部に情報が含まれていた．地区ごとに調査・分析した結果，地区に関係なく，ほとんどのファイルにメタデータとして様々な情報が含まれていることが分かった．

外部から閲覧可能な状態となっている電子文書内にメタデータが含まれている場合，メタデータの情報が悪用される可能性がある．特に，メールアドレスは，直接的な攻撃を行うために悪用される恐れがあるため，意図的でない場合は，削除することが推奨される．

本調査により，電子文書には隠れた情報が含まれていることが確認できた. 本調査は Google 検索演算子を用いた収集を行っており，その調査方法で収集したファイルであってもメタデータを調査することが可能であるとも確認できた．これらのことから，電子文書をインターネット上に投稿する前に，適切な処理をし，意図しない情報を削除することが求められる．

6.2 今後の展望

本稿では，調査対象を日本の国立大学として調査を行ったが，各大学が保有する公式のドメインから得られるファイルのみを収集している．しかしながら，大学に関係する Web サイトは公式のドメインで運営されるもののみでなく，大学に所属する研究室が個別に作成している Web サイトや，大学同窓会に関連する Web サイトなどが存在することもある．そのため，これらの Web サイトで配布されるファイルは，本稿の調査では収集できなかったことが考えられる．そのため，それらの Web サイトも含めた収集を行うことを考えている．

本稿の調査は，電子文書に含まれる隠れた情報の一種であるメタデータを対象としたものである．しかし，収集した MS OOXML ファイルにおいてメタデータ以外の隠れ情報を参照したところ，内部情報を含むと考えられるファイルを複数の大学が公開している事例を確認した．これらについては，各大学の CSIRT 等に報告し，対応を依頼した．以上の結果から，電子文書にはメタデータ以外にも情報漏洩のリスクとなり得る隠れ情報が存在することが明らかとなった．

今後は，大学以外も調査対象に含め，メタデータ以外の隠れ情報についても体系的に調査を進めていく予定である．

謝辞　本研究において発見された事例への対応にご協力いただいた各大学関係者の皆様に感謝申し上げます．

参考文献

[1] “情報セキュリティ 10 大脅威 2025”, https://www.ipa.go.jp/security/10threats/10threats2025.html, （参照 2025-8-20）
[2] Karl Mendelman, “Fingerprinting a Organization Using Metadata of Public Documents”, https://core.ac.uk/download/pdf/237084149.pdf,
[3] 長谷川太一，高木泉希，齊藤泰一，佐々木良一，“大学の Web サイトで公開される電子文書から漏洩しうる情報の調査・分析”,2024 Symposium on Cryptography and Information Security
[4] Taichi Hasegawa, Taiichi Saito，Ryoichi Sasaki，“Analyzing Metadata in PDF Files Published by Police Agencies in Japan”, 2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)
[5] 長谷川太一，齊藤泰一，佐々木良一，“Web で公開される PDF ファイルの Hidden data の調査―日本の警察を対象として―”，マルチメディア，分散，協調調とモバイルシンポジウム 2022 論文集, 2022
[6] “国立大学”, https://www.mext.go.jp/b_menu/link/daigaku1.htm, （参照 2025-4-30）
[7] “Google 検索演算子の概要”, https://developers.google.com/search/docs/monitor-debug/search-operators, （参照 2025-8-20）
[8] “What is DownThemAll?”, https://www.downthemall.org/, (参照 2025-08-20)
[9] Hyunji Chung, Jungheum Park, Sangjin Lee, “Forensic Analysis

of Residual Information in Adobe PDF Files”, https://arxiv.org/pdf/2003.10546

[10] “2 Structures”, https://learn.microsoft.com/en-us/openspecs/windows_protocols/ms-cfb/28488197-8193-49d7-84d8-dfd692418ccd, （参照 2025-8-21）

[11] Espen Didriksen, “Forensic Analysis of OOXML Documents”, https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/198656/EDidriksen.pdf

[12] “令和４年上半期におけるサイバー空間をめぐる脅威の情勢等について”, https://www.npa.go.jp/publications/statistics/cybersecurity/data/R04_kami_cyber_jousei.pdf，(参照 2025-08-20)

~~[12]~~[13] “Disclosing documents to the public securely: hidden personal information and how to avoid an accidental breach”, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/security/disclosing-documents-to-the-public-securely/, (参照 2025-08-20)

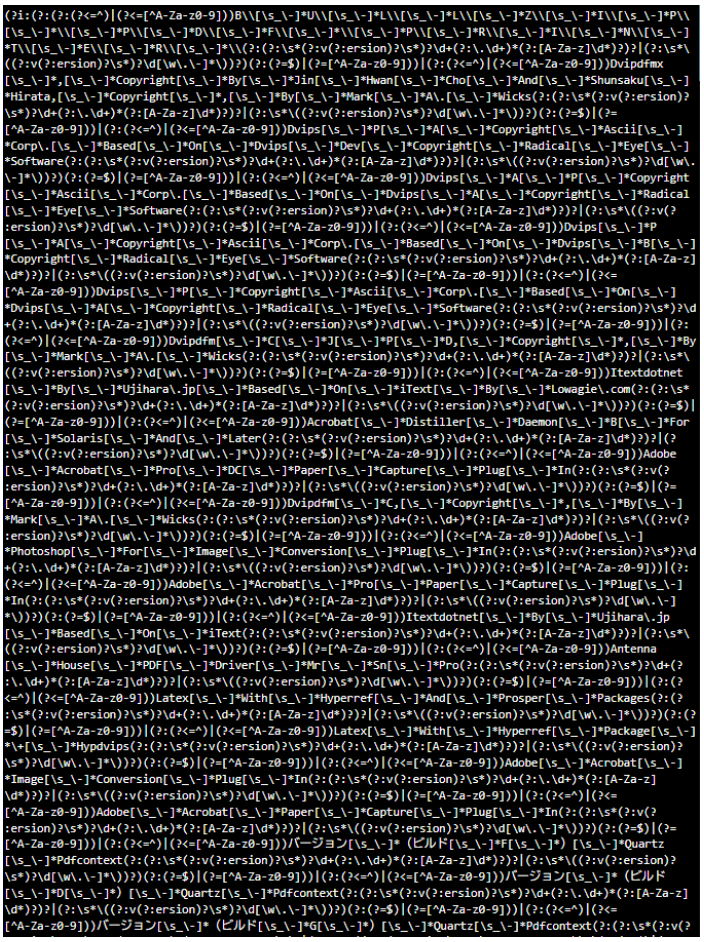
~~[13]~~[14] “Checklist: How to avoid an accidental breach when personal information is ‘hidden’ in documents”, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/security/disclosing-documents-to-the-public-securely/checklist-avoid-an-accidental-breach-when-personal-information-is-hidden-in-documents/, (参照 2025-08-20)

~~[14]~~[15] “墨消しと非表示情報”, https://helpx.adobe.com/jp/acrobat/using/securing-pdfs-passwords.html，(参照 2025-08-21)

~~[15]~~[16] “ドキュメント検査を使用する”, https://learn.microsoft.com/ja-jp/office/vba/library-reference/concepts/using-the-document-inspector, (参照 2025-08-20)

付録

本稿のソフトウェア名の判定に用いた正規表現の一部を以下に示す．ソフトウェア名が格納されるメタデータキーから抽出した値を集計し，作成したものである．



```
(?i:(?:(?:(?<=^)|(?<=[^A-Za-z0-9]))B\\[\s_\-]*U\\[\s_\-]*L\\[\s_\-]*L\\[\s_\-]*Z\\[\s_\-]*I\\[\s_\-]*P\\
[\s_\-]*\\[\s_\-]*P\\[\s_\-]*D\\[\s_\-]*F\\[\s_\-]*\\[\s_\-]*P\\[\s_\-]*R\\[\s_\-]*I\\[\s_\-]*N\\[\s_\-]
*T\\[\s_\-]*E\\[\s_\-]*R\\[\s_\-]*\\(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\
((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Dvipdfmx
[\s_\-]*,[\s_\-]*Copyright[\s_\-]*By[\s_\-]*Jin[\s_\-]*Hwan[\s_\-]*Cho[\s_\-]*And[\s_\-]*Shunsaku[\s_\-]
*Hirata,[\s_\-]*Copyright[\s_\-]*,[\s_\-]*By[\s_\-]*Mark[\s_\-]*A\.[\s_\-]*Wicks(?:(?:\s*(?:v(?:ersion)?
\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=
[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Dvips[\s_\-]*P[\s_\-]*A[\s_\-]*Copyright[\s_\-]*Ascii[\s_\-]
*Corp\.[\s_\-]*Based[\s_\-]*On[\s_\-]*Dvips[\s_\-]*Dev[\s_\-]*Copyright[\s_\-]*Radical[\s_\-]*Eye[\s_\-]
*Software(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.
\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Dvips[\s_\-]*A[\s_\-]*P[\s_\-]*Copyright
[\s_\-]*Ascii[\s_\-]*Corp\.[\s_\-]*Based[\s_\-]*On[\s_\-]*Dvips[\s_\-]*A[\s_\-]*Copyright[\s_\-]*Radical
[\s_\-]*Eye[\s_\-]*Software(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?
:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Dvips[\s_\-]*P
[\s_\-]*A[\s_\-]*Copyright[\s_\-]*Ascii[\s_\-]*Corp\.[\s_\-]*Based[\s_\-]*On[\s_\-]*Dvips[\s_\-]*B[\s_\-]
*Copyright[\s_\-]*Radical[\s_\-]*Eye[\s_\-]*Software(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]
\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=
[^A-Za-z0-9]))Dvips[\s_\-]*P[\s_\-]*Copyright[\s_\-]*Ascii[\s_\-]*Corp\.[\s_\-]*Based[\s_\-]*On[\s_\-]
*Dvips[\s_\-]*A[\s_\-]*Copyright[\s_\-]*Radical[\s_\-]*Eye[\s_\-]*Software(?:(?:\s*(?:v(?:ersion)?\s*)?\d
+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:
(?<=^)|(?<=[^A-Za-z0-9]))Dvipdfm[\s_\-]*C[\s_\-]*J[\s_\-]*P[\s_\-]*D,[\s_\-]*Copyright[\s_\-]*,[\s_\-]*By
[\s_\-]*Mark[\s_\-]*A\.[\s_\-]*Wicks(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\
((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Itextdotnet
[\s_\-]*By[\s_\-]*Ujihara\.jp[\s_\-]*Based[\s_\-]*On[\s_\-]*iText[\s_\-]*By[\s_\-]*Lowagie\.com(?:(?:\s*
(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|
(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Acrobat[\s_\-]*Distiller[\s_\-]*Daemon[\s_\-]*B[\s_\-]*For
[\s_\-]*Solaris[\s_\-]*And[\s_\-]*Later(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?
:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Adobe
[\s_\-]*Acrobat[\s_\-]*Pro[\s_\-]*DC[\s_\-]*Paper[\s_\-]*Capture[\s_\-]*Plug[\s_\-]*In(?:(?:\s*(?:v(?
:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=
[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Dvipdfm[\s_\-]*C,[\s_\-]*Copyright[\s_\-]*,[\s_\-]*By[\s_\-]
*Mark[\s_\-]*A\.[\s_\-]*Wicks(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?
:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Adobe[\s_\-]
*Photoshop[\s_\-]*For[\s_\-]*Image[\s_\-]*Conversion[\s_\-]*Plug[\s_\-]*In(?:(?:\s*(?:v(?:ersion)?\s*)?\d
+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:
(?<=^)|(?<=[^A-Za-z0-9]))Adobe[\s_\-]*Acrobat[\s_\-]*Pro[\s_\-]*Paper[\s_\-]*Capture[\s_\-]*Plug[\s_\-]
*In(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]
*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Itextdotnet[\s_\-]*By[\s_\-]*Ujihara\.jp
[\s_\-]*Based[\s_\-]*On[\s_\-]*iText(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\
((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Antenna
[\s_\-]*House[\s_\-]*PDF[\s_\-]*Driver[\s_\-]*Mr[\s_\-]*Sn[\s_\-]*Pro(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?
:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?
<=^)|(?<=[^A-Za-z0-9]))Latex[\s_\-]*With[\s_\-]*Hyperref[\s_\-]*And[\s_\-]*Prosper[\s_\-]*Packages(?:(?
:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?
=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Latex[\s_\-]*With[\s_\-]*Hyperref[\s_\-]*Package[\s_\-]
*\+[\s_\-]*Hypdvips(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?
\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))Adobe[\s_\-]*Acrobat[\s_\-]
*Image[\s_\-]*Conversion[\s_\-]*Plug[\s_\-]*In(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]
\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=
[^A-Za-z0-9]))Adobe[\s_\-]*Acrobat[\s_\-]*Paper[\s_\-]*Capture[\s_\-]*Plug[\s_\-]*In(?:(?:\s*(?:v(?
:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=
[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))バージョン[\s_\-]*（ビルド[\s_\-]*F[\s_\-]*）[\s_\-]*Quartz
[\s_\-]*Pdfcontext(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]\d*)?)?|(?:\s*\((?:v(?:ersion)?
\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=[^A-Za-z0-9]))バージョン[\s_\-]*（ビルド
[\s_\-]*D[\s_\-]*）[\s_\-]*Quartz[\s_\-]*Pdfcontext(?:(?:\s*(?:v(?:ersion)?\s*)?\d+(?:\.\d+)*(?:[A-Za-z]
\d*)?)?|(?:\s*\((?:v(?:ersion)?\s*)?\d[\w\.\-]*\))?)(?:(?=$)|(?=[^A-Za-z0-9]))|(?:(?<=^)|(?<=
[^A-Za-z0-9]))バージョン[\s_\-]*（ビルド[\s_\-]*G[\s_\-]*）[\s_\-]*Quartz[\s_\-]*Pdfcontext(?:(?:\s*(?:v(?
```

図 5. ソフトウェア名の判定に用いた正規表現