UpGuard

February 2021

How public documents can be used for attack reconnaissance

UpGuard research team analyses risk profiles of Fortune 500 companies, revealing how the metadata in public PDFs and Office documents can be used for attack reconnaissance

UpGuard

Table of Contents

UpGuard

PDF metadata report

Introduction

As a business's digital operations expand, so does the size of a business's attack surface. New technologies for businesses emerge, and once deployed, they become part of the potential attack surface. The number of components comprising that attack surface increases over time: from employee workstations to the portfolio of customer-facing applications a business produces, each with its own particular qualities and potential to contribute to a data breach.

Documents published on companies' websites, such as PDFs and Word documents, are another such component of the overall attack surface, a collection of data that has, however, been overlooked in the studies of cyber risk.

At the center of the tug-of-war between the data mining industry and privacy regulators in the 21st century are users' metadata – data about data that provides useful insights into individuals' personalities and behavior without necessarily having the content a user is actively (and voluntarily) posting online. Businesses make a conscious decision to upload and publish documents and the content within them as part of the ordinary course of their business, but attackers could leverage the lesser-known metadata in these documents.

For this report, we investigated specific metadata that virtually all businesses produce but often leave unmonitored: the metadata in publicly available PDFs and Microsoft Office documents. Like other digital artifacts, office documents (such as PDF and Word documents) can contain metadata that is not obvious to the user, and metadata can unintentionally leak valuable information about that person, their organization, and the company's structure, its software, and its vendors.

UpGuard conducted a study of Fortune 500 companies and reviewed the metadata associated with public documents on their websites. These files included PDF, Word, Excel, and PowerPoint documents. This report will investigate how much reconnaissance information an attacker can obtain from downloadable website documents based on the study.

Cyber attackers can use metadata for reconnaissance

The MITRE ATT&CK[1] framework gives us a standard for the tactics used by cyber attackers during each phase of a breach. The framework begins with the “Reconnaissance” stage, during which adversaries use any source to “gather information they can use to plan future operations,” specifically including information about the target’s IT systems, their personnel, and their organization. That information is abundantly available in the metadata of PDF and Office documents available on companies’ websites (the focus of this report).

This research report and study measures how much of the reconnaissance information (as described by Mitre ATT&CK) an attacker can gain from public documents. We looked at documents publicly hosted on the websites of Fortune 500 companies, and we looked at the document contents and metadata to discover information described as the objective of the ATT&CK Reconnaissance phase.

1 MITRE ATT&CK® is a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations.

Research findings

How big is the public document attack surface area?

Reconnaissance campaigns begin with an evaluation of a victim's attack surface area. One way to measure the document surface area for any website is to use a search engine to locate files of a certain type (such as PDF or Word documents) associated with the primary domain.

An analysis of our collected data revealed two interesting findings.

Finding 1

Digital document attack surface area is large

The total digital document surface area was large, and has the potential to be enormous. Across the public websites of the Fortune 500 companies included in this study, the average number of PDFs per company was over 9,700.

9,700+

Average number of PDFs found on Fortune 500 companies' websites.

This extensive volume contains more than enough data for uncovering key trends and anomalies for reconnaissance campaigns.

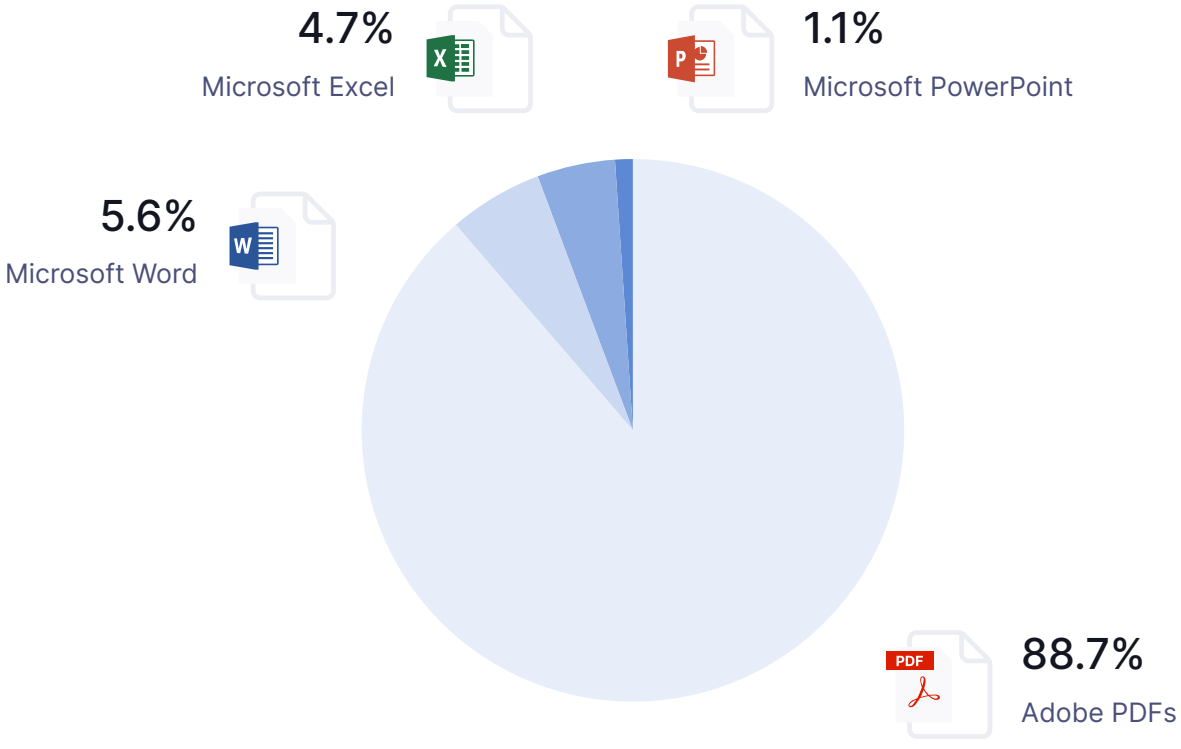
Finding 2

Public documents are predominantly PDF files

Microsoft Office documents - such as Word, Excel, PowerPoint documents - were surfaced but in far smaller numbers.

This distribution makes sense since downloadable website documents are intended to be read by website visitors and not edited. Since PDF documents are read-only, they are the predominantly offered document on websites. As a result of this finding, this study primarily relates to PDF metadata.

However, this does not discount other document types from the attack surface area. On the contrary, as discussed further in this study, Office documents contain deeper data leaks making them highly coveted reconnaissance discoveries.



Key takeaways

- Cyber professionals and attackers know that they can collect a large number of documents for analysis when planning their cyber attack (mostly PDF documents).
- Even though there are fewer public Office documents, it is as important to potential attackers. In fact, a previous data exposure that UpGuard reported began with an Office document that had been used for internal documentation, and was hosted on a site that had an exposed dashboard of employee information.

Research findings

Victim identity and organization data in the metadata

PDFs and Office documents have metadata fields for information such as the author of the document and the software used to produce it. Such information may be used by attackers for reconnaissance into specific areas of responsibility of a target organization's employees[2].

51%

Public PDF documents are not following this best practice and had a value set for the optional author field

Finding 1

51% of public PDF documents are not following this best practice and had a value set for the optional author field

PDFs and Office documents have metadata properties describing the author of the document and the software used to produce it. The best practice industry recommendation is to omit metadata revealing authorship and associated production software where possible.

However, we discovered that **51% of publicly available PDF documents are not following this best practice and had a value set for the optional author field populated**, potentially with sensitive information such as employees' names.

2 Section T1589 of ATT&CK describes identity information that attackers may collect during reconnaissance, including names and email addresses, and T1591 of ATT&CK describes victim organization information

It should be noted that documents with exposed author values do not necessarily give away useful information. If you are running an up to date web server, revealing that server version is not particularly damaging. Similarly, in reviewing the document data, a common practice is to set the author to be the name of the company, which provide minimal information about what documents inside the company look like, but nothing novel to a potential attacker.

However, in analyzing the documents we collected, the information in the author fields were far from uniform, and contained more than just the company name or employees names. Often those fields would also contain identifying information about the individual's place in the organization, like the abbreviation used for their business unit, job titles, and the name of their office's city. Some even included numbers that appear to be employee IDs.

Key takeaway

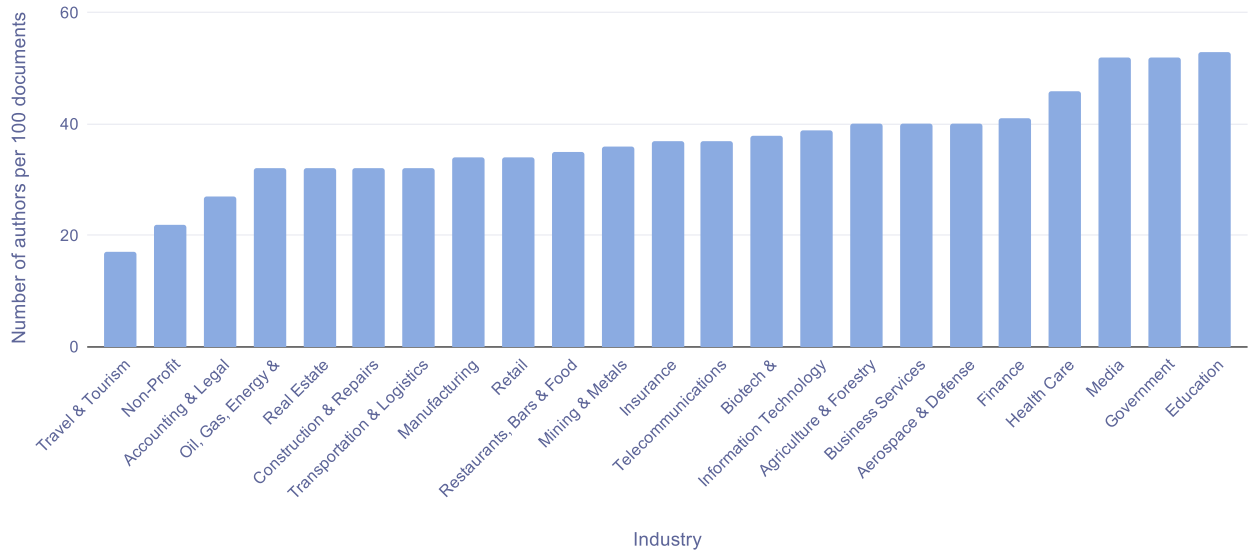
Limit the information included in metadata of public documents. Don't give anything away that you don't have to — omit the author information where possible.

Finding 2

There are differences in populated author fields across different industries

We looked into the control that any given company had over its document metadata by measuring how uniform author values were across documents. In the figure below, we show the number of unique author values we would have found if we sampled 100 documents, classified by industry. Higher numbers indicate more inconsistency (less uniformity), while lower numbers indicate more consistency (more uniformity) — something that may indicate rigorous procedures are used more often in these industries with lower numbers. From our sample data, industries that have the most revealing documents (in terms of populated author fields with varying information being leaked, such as different unique authors) are media, government, and education.

Number of authors (per 100 documents), by industry



The higher the number of authors in the figure above, the more inconsistent author fields were across a company's documents (i.e., the more unique values in author fields were found): potentially more metadata leaked in public documents for these industries.

The higher the number of authors per 100 documents analyzed for those industries, the more inconsistent author fields were across a company's documents (i.e., the more unique values in author fields were found): potentially more metadata leaked in public documents for these industries. This confirms to us that while there is a general lack of process controls, we see a difference of these process controls across industries, and certain industries are faring worse than others.

Key takeaways

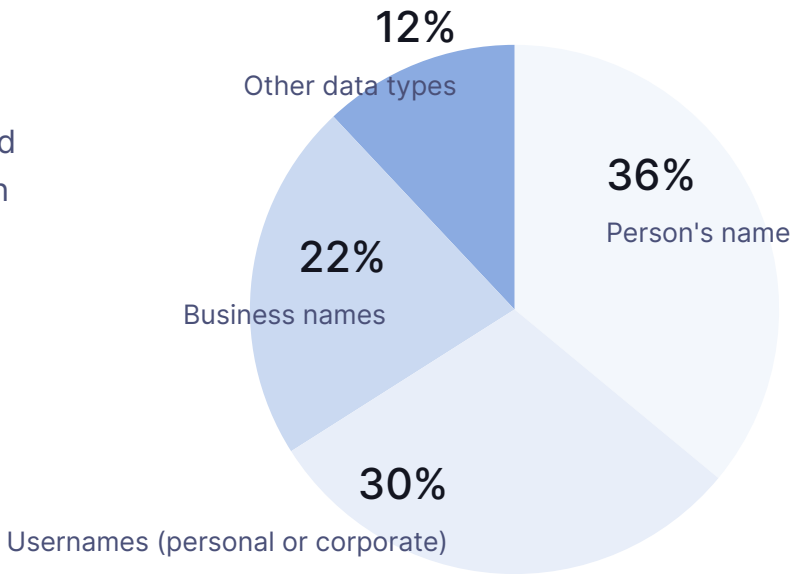
- While there is a general lack of process controls, security and IT professionals should review their processes when users generate new documents and the metadata that is auto-populated with it.
- For InfoSec professionals in industries with more revealing documents in the figure above, it's worth another look at your processes and how much data you are leaking. Consolidating processes for PDF generation across the company would go some way to address the problem for these industries.

Finding 3

Author fields in PDFs reveal personal names, usernames, and more

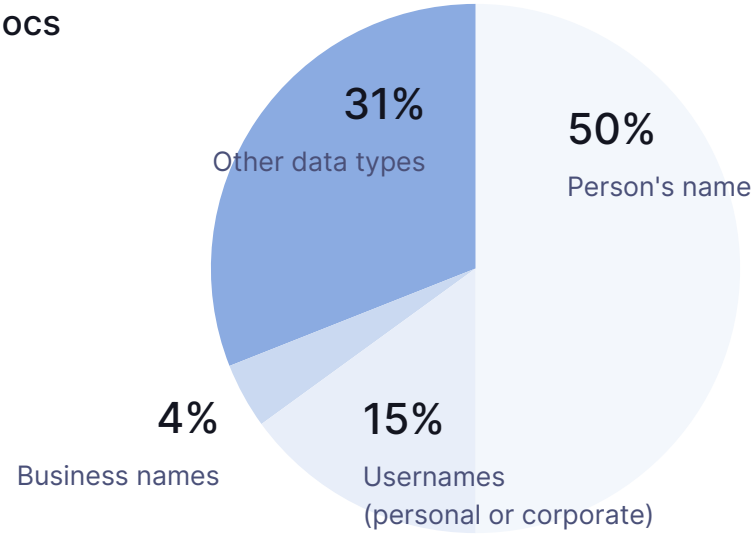
Found in author fields in PDFs

Of the PDFs sampled with some value in the author field, we found the following types of information populated.



Found in author fields in Office docs

For Office documents, the overall trends are similar to the above.

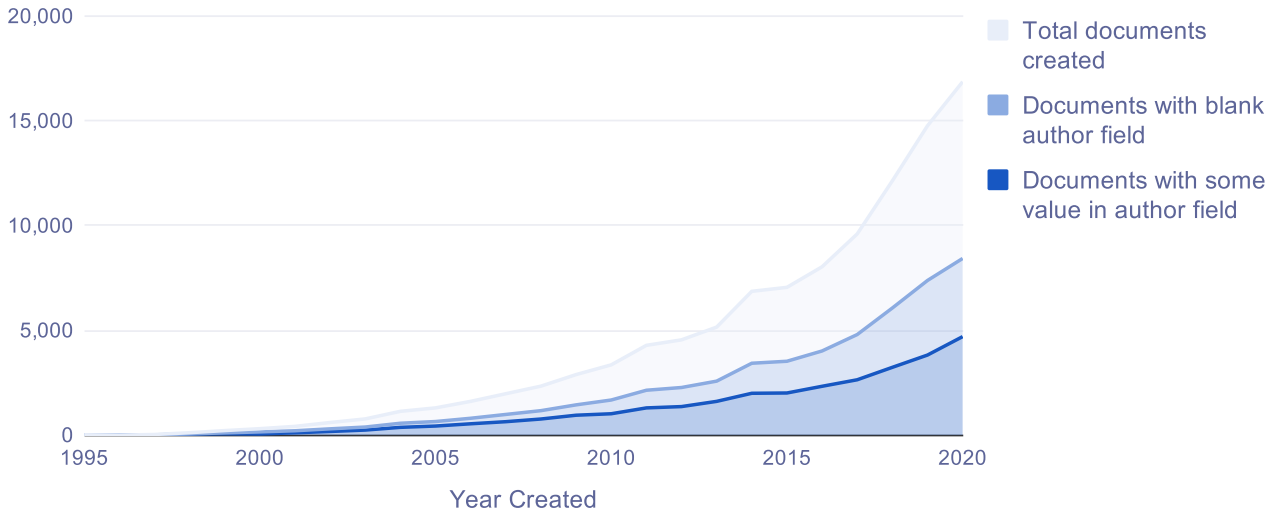


Finding 4

Author sanitization is getting worse over time

Based on our research, the leaking of personal or company identifiers through document metadata is getting worse over time. As with digital information in general, the number of documents added to the internet each year continues to increase. Despite all being well known document types, the solutions for sanitizing metadata are not gaining ground. Every year, more documents with unsanitized author metadata are being added to the internet.

Documents on Fortune 500 websites with blank author field vs populated author field, by document create date



While the proportion of documents with blank author fields to non-blank author fields are roughly consistent year-on-year, since we are generating more total documents each year, the raw number of documents with author fields being set is increasing.

Key takeaway

- Security and IT professionals should audit the public PDF and Office documents on their websites for any potential leak in metadata to review their attack surface.
- Review how your author fields are being populated at your organization. One way to measure the control that a company has over its document metadata is to look at how consistent those author fields are when populated. So if your process is to set the author as your company name, we would expect to see more consistency across your published documents.

Research findings

Host data

Now an attacker knows who works on which documents, their work username, email address, business unit, what the standard procedures are (if any) for naming documents. Their next step is to attempt an attack to understand the hardware and software used by these employees.

For this, there are multiple pieces of information that can be harvested from document metadata.

Finding 1

Software information and online converters found in author field

The author field (discussed earlier in the report) is not intended to hold software version information, but because any value can be inserted into each of the metadata properties, it is up to the user (and software used) to conform to convention.

Among the files we sampled, only 23 files listed identifiable software as the author, such as Adobe Acrobat. More concerning, however, were the over 100 documents with non-identifiable software such as an online software provider as the author (typically PDF converters). There are many such sites, some of which convert the document as a one-time transaction and some which continue to host it publicly.

23

Files listed an identifiable software as the author

100+

Documents with non-identifiable software such as an online software provider as the author.

This increasing prevalence of free online PDF converters is a considerable (and potentially overlooked) concern, and this is something that security and IT professionals should look into how their employees are using PDF converters. If frequently used, security and IT professionals should consider restricting access to such tools.

It is a possibility that free online PDF converters are being used not only due to their convenience but because certain users at these Fortune 500 companies do not have access to Adobe or Microsoft PDF creators for cost or security reasons. Often these online tools are certain users' only resort to do their job.

How an attacker can use this

Identifying that employees of a target use such a site is useful for either tricking them into using a malicious document converter or targeting the converter site itself, known as a watering hole attack[3].

3 In this study, we only observed documents that were more or less intentionally public; there are likely far more private documents that pass through such online converters.

Finding 2

Most documents reveal what software was used to create them

To learn about which software is used at a target company, we can look at the fields designed to hold such information.

About 75% of PDFs sampled had a value in the “creator” field. For Word, Excel, and PowerPoint documents, between 98-100% of those sampled had information in the comparable App or AppVersion fields.

As one might expect, most of the programs used are the native producers for PDFs and Office documents: Adobe and Microsoft products, respectively. Within those product lines, however, are many versions, and those many variations were represented in these documents. Out of 58,000 PDFs with a software version, there were 2,795 unique values, suggesting thousands of software types and versions.

How an attacker can use this

Given the metadata could reveal the software as well as version, attackers could find employees with vulnerable software and versions with unpatched bugs / weaknesses.

75%+

PDFs revealed the software used to create the document.

98%+

Office documents revealed the software used to create the document.

Finding 3

End user operating systems

The software used can also reveal the OS version of the user's computer. The documents sampled contained 857 unique values matching operating system names, indicating the abundance and specificity of operating system data. 371 files also included versions matching hardware products, which appear to be printers used to scan documents.

How an attacker can use this

Coupled with finding employees with vulnerable software versions, attackers could locate employees with any vulnerable operating systems.

Key takeaway

- Security and IT professionals should look into the processes and how to scrub all the information of the metadata leaking hardware and software information.
- Make sure your employees operating systems and application software are patched to the latest version to ensure minimal vulnerabilities.
- Security and IT professionals should consider restricting access to free online PDF converters.

Research findings

Conclusions

The information found in document metadata is likely to be identity data that is useful as part of a social engineering attack on businesses' employees. Although public facing documents do not provide a direct interface to business systems and networks, the fact that documents cannot be directly exploited in that way puts public documents out of scope of penetration tests, potentially creating an unobserved pool of resources for social engineering. Since the level of threat does not in itself rise to the level of a bug bounty, risk from document metadata can accumulate overtime, confirming that document metadata is an overlooked threat for businesses of all sizes.

While document metadata is by no means a new problem – some of the documents we analyzed claimed to have been produced in the 1990s – it is a problem that is getting worse. By being aware of the common data leaks in PDF metadata, organizations can avoid the malpractices caused by leaked metadata and strengthen their public document security.

What can you do now

Action items for security and IT professionals

1. Organizations should be aware of all of the metadata present in public documents and secure these widely overlooked attack surfaces and to further mitigate the potential of data breaches.
2. Moreover, employees may not have the proprietary software needed to edit documents in their native formats, leading to the use of shadow IT with more potential for unmanaged metadata.
3. Train the employees on how to minimize the risk of leaked metadata by creating processes on how to scrub information from PDF and Office documents before uploading documents online.
4. Since this sits outside the scope of pentests, conduct periodic audits of public documents on your company websites to understand your attack surface.
5. Remove old public documents that are no longer used or relevant.

Appendix

A. Methodology

To find public documents, we used the same techniques anyone would use to find any public information: on a search engine. Google search syntax supports limiting a search to a particular domain and filetype, allowing us to measure the approximate number of such documents publicly available and to find the documents themselves. We searched for PDFs, Word, Excel, and PowerPoint documents on the primary domain of each of the Fortune 500 companies.

Due to the massive numbers of such public documents, we did not analyse all of them, but took between 100 and 400 documents per company in order to reach a 90-95% confidence for each vendor. For vendors with fewer than 100 documents for a given type, we sampled enough to reach 95-99% confidence. In some cases, Google reported having documents that exist only in cache (and thus not in the PDF or Office format) or which do not exist at all, making it impossible to collect the number reported in the search results. In any case, the number collected provides a high degree of statistical certainty, and the absolute numbers of significant data points leaked are sufficient to demonstrate the importance of this vector for data leakage.

B. What does the UpGuard research team do?

The cyber research team's work involves discovering new vectors for data leaks and measuring the prevalence of sensitive data exposed.

UpGuard

Questions? We have answers

We're here to help, shoot us an email at sales@upguard.com

Know your vendors. Secure yourself.

Looking for a better, smarter way to protect your data and prevent breaches?

UpGuard offers a full suite of products for security, risk and vendor management teams.

Contact sales | Free demo →

Trusted by hundreds of companies worldwide

www.upguard.com

+1 888-882-3223

650 Castro Street, Suite 120-387, Mountain View CA 94041 United States