Workshare

Document Metadata:

What You Can’t See Can Actually Hurt You

Published: 2015

Contents

1.0 Introduction

Most employees go about their working day unaware that the business documents they are sharing outside the organization contain hidden, sensitive data. Referred to by security experts as *document metadata*, and not visible to the naked eye, it isn’t something we normally think about.

However hidden metadata left within documents can potentially disclose confidential information, such as document properties, redacted or deleted text, or notes made during review, and expose employees and the company to serious risk. As a minimum, document metadata leaks can lead to loss of confidence, clients, or disciplinary action. At worst, serious reputational damage, and lawsuits, especially if social security numbers or personally identifiable information is involved.

On the face of it, many may think that this is an issue for their company’s security experts – and to an extent it is. But as we’re increasingly using our personal devices and applications to share business documents, employees must take responsibility for protecting sensitive company information and intellectual property from accidental leaks.

So what exactly is metadata? And what are the best ways to ensure that documents released outside of an organization don’t carry sensitive or confidential hidden information? This paper aims to help safeguard professionals who work with sensitive information in order to prevent them from inadvertently sharing it, by explaining what document metadata is, what forms it can take, and more importantly, how to deal with it.

2.0 Hidden data

The volume of documents being shared electronically between business professionals and with their clients/customers is increasing. This rise has been driven by a need for faster review cycles, distribution of information among dispersed teams, and greater collaboration among these teams as groups of workers contribute to a document.

With a seismic shift in the last few years toward working outside of office hours and outside the office, this changing culture of work means there is now a higher propensity for risky sharing behavior, without people even knowing they are doing it.

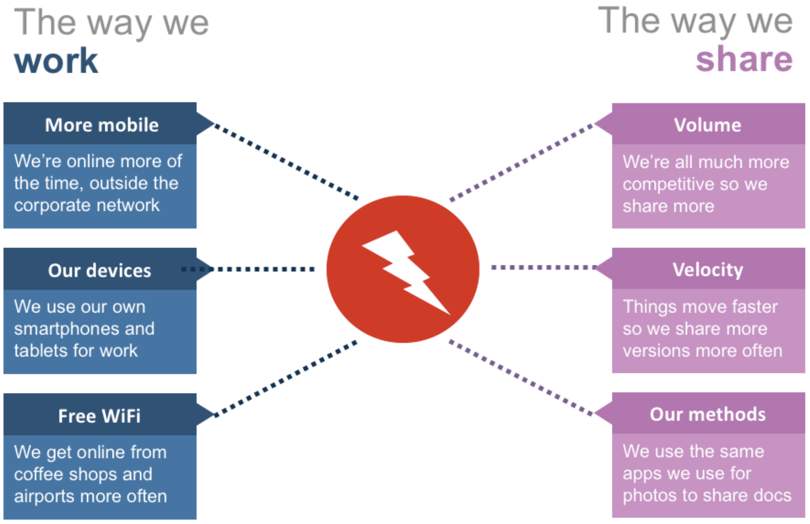
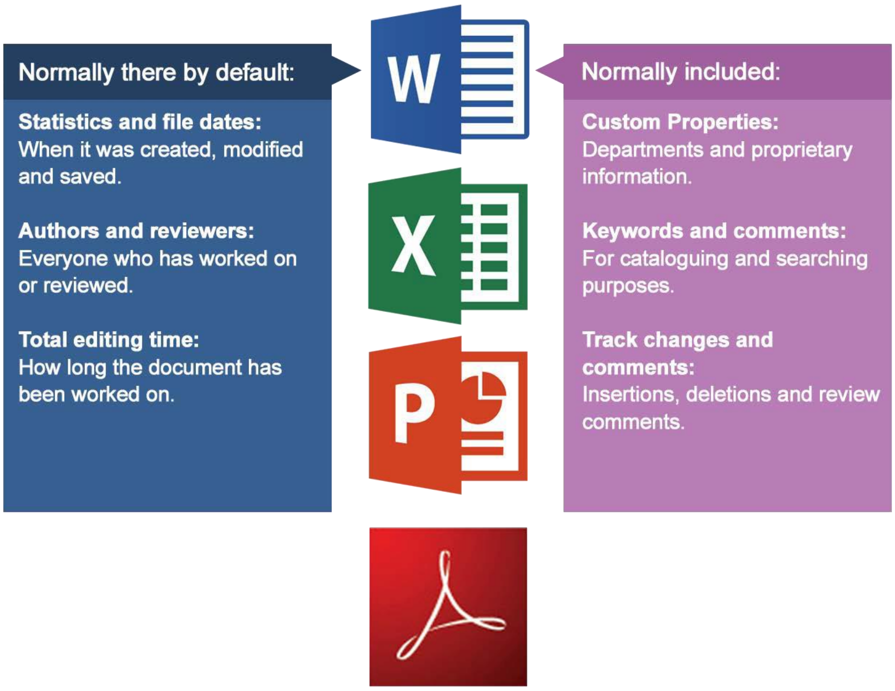

The way we work
More mobile
We're online more of the time, outside the corporate network
Our devices
We use our own smartphones and tablets for work
Free WiFi
We get online from coffee shops and airports more often
The way we share
Volume
We're all much more competitive so we share more
Velocity
Things move faster so we share more versions more often
Our methods
We use the same apps we use for photos to share docs

Figure 1: The perfect storm

As can be seen in figure 1, a perfect storm, made up of an increased volume and velocity of document sharing, coupled with the cultural change of a more mobile workforce, has led to an increase in risk when sharing or storing a document online. This, along with recent media coverage about surveillance, snooping, and information leaks, is bringing the issue of hidden metadata to the fore and highlighting the need for awareness around the inherent security risks.

2.1 What is metadata?

A document’s metadata is information about the document, including changes made during the development of that document, and is defined as the “data providing information about one or more aspects” of the document. Whenever we create, edit, or save a document, a rich set of metadata is automatically added to it – behind the scenes. This can include information about how long the document has been worked on, how the file was created, time and date, who the original author was, when the document was last saved, tools used, and a short summary of the document. As the document evolves, undo and track changes are normally added and included automatically.



Normally there by default:

Statistics and file dates: When it was created, modified and saved.

Authors and reviewers: Everyone who has worked on or reviewed.

Total editing time: How long the document has been worked on.

Normally included:

Custom Properties: Departments and proprietary information.

Keywords and comments: For cataloguing and searching purposes.

Track changes and comments: Insertions, deletions and review comments.

Figure 2: Metadata mechanics

Originally conceived to make it easier to track and find document data, no one argues that metadata is useful when used properly and with an appropriate level of awareness. However, such rich information “hidden” as a document is stored, modified, and shared can reveal more information than we would ideally want to share beyond the original team that created the document.

For example, edit and review features in Microsoft Word, such as reviewer comments, track changes, and undo typing result in a significant amount of dangerous metadata being included.

In practice, not every document starts as new. Workshare’s experience is that up to 70% of documents are recycled, i.e., starting life as a copy of another document. People tend to base a new document on a similar one that already exists and then make edits. But imagine that we are billing a customer for a custom report or proposal. We would not want the document’s origins to be revealed, especially if the first version was made by a junior member of the team, clearly labelled as “author” in the track changes.

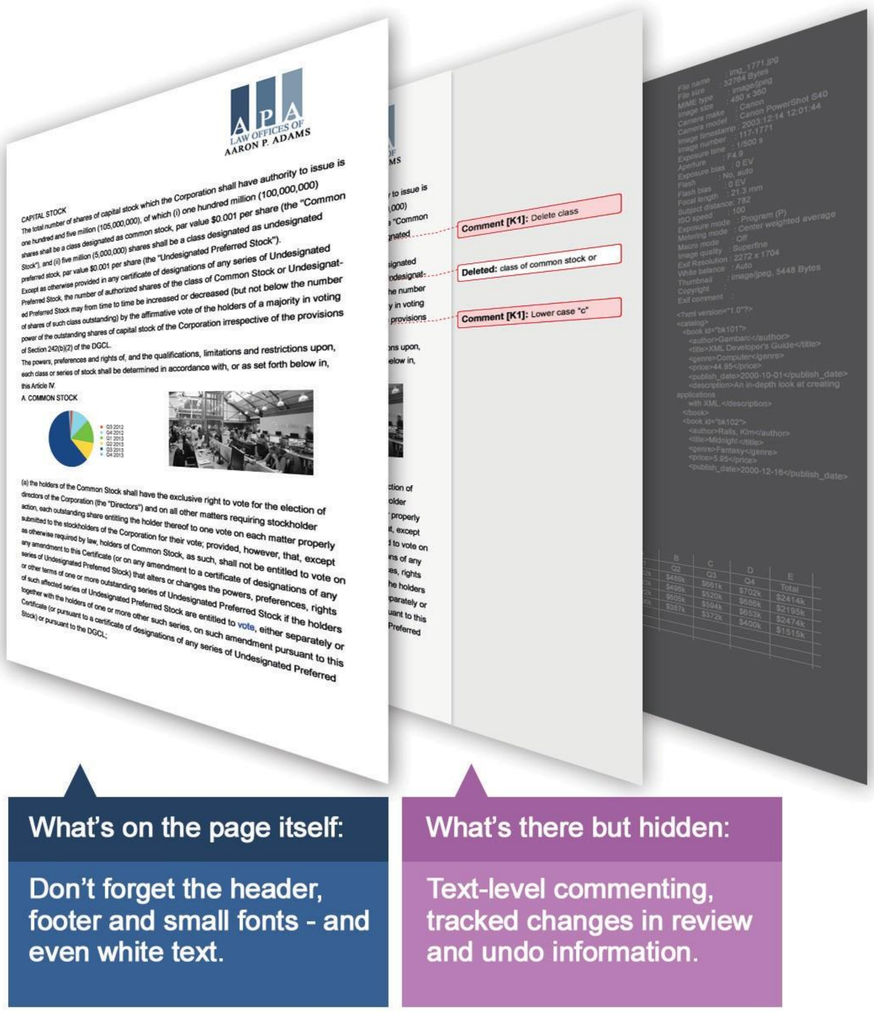
Forgetting or ignoring the fact that metadata exists when documents are shared means people can access privileged and confidential information. There have been many high-profile cases involving leaked document metadata, and the effect it has had on businesses has been catastrophic in some cases.

Document Metadata:

What You Can’t See Can Actually Hurt You

Here’s a very common scenario. During a contract negotiation, a team will normally start off with a standard document, possibly cloned from a previous contract project to save time. The team will replace “Company A” with “Company B”, have an internal review, and make changes with track changes enabled to help co-workers track the progression of the document. During the internal review, a clause that the team doesn’t think is necessary is removed. They may even remove some statements they had put in to the dealings with Company A that they don’t want to offer straight away to Company B. Later, they add a clause to address some particular internal concern about dealing with that company, and some notes are made as comments explaining why. How damaging would it be if all that was inadvertently shared as metadata with Company B?

By imagining “layers” of text, document information, and hidden text, this graphic shows just how much document metadata is potentially risky in a business context.



APA
LAW OFFICES OF AARON P. ADAMS
Comment [K1]: Delete class
Deleted: class of common stock or
Comment [K1]: Lower case "c"
What’s on the page itself:
Don’t forget the header, footer and small fonts - and even white text.
What’s there but hidden:
Text-level commenting, tracked changes in review and undo information.

Figure 3: Layers of a document – from visible to hidden

2.2 Eliminating metadata risk

The key to the problem is not that metadata is added to a document but that it is difficult to fully identify and remove. For example, in Microsoft Word, adding comments and tracking changes are very helpful to people working on a document together. However, when a change is not accepted, it remains within the document even though it appears to be invisible. These changes can easily be displayed by turning on the “Show Markup” view, which can result in damaging situations where external parties see information that was not intended for their eyes.

Accidental data leak helps wipe $22bn off stock value

Join thousands of others, and sign up for Naked Security's newsletter

you@example.com

I.S.S. Investigates Claim of Data Leaks

BY MICHAEL J. DE LA MERCED

Institutional Shareholder Services, the biggest player in the shareholder advisory world, disclosed on Monday that it was conducting an investigation into a claim that an employee improperly leaked confidential investor voting data in exchange for cash and gifts.

MSCI, the parent of I.S.S., disclosed the internal inquiry in a regulatory filing, adding that it had placed an unidentified employee on admin

investigation.

The filing came aft
Post on Sunday re
had filed a complai
Exchange Commis
by a midlevel exec

Proxy advisory fir
reports that urge

(TS//SI//NF) PRISM Collection Details

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!

What Will You Receive i
(Surveillance and Store
It varies by provider.

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target acti
- Online Social Networking
- Special Requests

ete list and details on PRISM web page:
Go PRISMFAA

Figure 4: Sensitive data leaks make the headlines

There have been many high-profile cases involving the exposure of metadata in documents:

- A UK Bank: An Excel spreadsheet containing 179 contracts within hidden columns was converted to PDF, revealing the hidden data, and was accidentally submitted in a buyout offer of Lehman Brothers assets. The bank ended up paying $1.35bn for these assets as well as taking on responsibility for more of Lehman’s trading positions than it intended because of the formatting error.
- Telecoms firms: Two firms branded journalists hackers after they found the personal identifiable data of over 170,000 customers stored on an unsecure and publicly accessible server. It was kept on the company servers and posted to an open file sharing area, and the journalists could access it via an online search engine.
- Global bank: 150,000 customers who went into bankruptcy between 2007 and 2011 had their personal information exposed after the bank failed to properly redact court records before they were put on the Public Access to Court Electronic Records system. The bank were ordered to notify all the affected parties, offer a year of free credit monitoring, and redact the sensitive information at its’ own expense.
- London Borough Council: The Council published hidden sensitive metadata related to 2,375 residents online in an Excel document. This metadata leak cost them £70,000 in fines.
- Broadcaster, Australia: The salaries of broadcasters were inadvertently leaked in response to a freedom of information request by a South Australian MP. This led to outrage due to the amount some were being paid and discrepancies in pay between colleagues, with a number of employees demanding pay rises.

2.3 Document metadata elements and their risk

The following lists examples of areas in Microsoft Office documents where metadata problems may arise:

Review Details and Change History applies to: Microsoft Word, Microsoft Excel, and Microsoft PowerPoint documents.

Track changes, comments, and document revisions, including the last "undo", help the author understand what additions and deletions have been made by others working on the document. These are tagged with initials of the co-authors. Comments are included to help reviewers make suggestions to the person collating all the review information. Previous versions and fast saves can also fall into this category.

- Risks: Identifiable comments can be devastating. For example, a senior member of the team recommending that something be removed or changed before it reaches the intended recipient of the document can survive as metadata in the document that is sent. Not only will the recipient see the item that was intended to be removed but also see that the reviewer wanted it to be removed. Even if comments don't accompany deletions or insertions, the text or figures that were intended to be deleted or added will be seen.

Document properties apply to: Microsoft Word, Microsoft Excel, and Microsoft PowerPoint documents.

Document properties are details about a file that help identify it, including a descriptive title, subject, author, manager, company, category, keywords, comments, hyperlink base, server, network names, and anything that reveals a blueprint for a hacker. Document properties display information about a file to help us organize them in order to find them easily.

- Risks: The names of authors and the name of the organization can display sensitive information. If a document has been sent outside your own organization, the author name and company name contained in the built-in properties could be a name other than your own. Also, if documents are repurposed or used as a template for a new document, information specific to a previous client (for example, pricing, terms or client's name) can be stored as hidden information within the new document, which can then be revealed to a client.

Document statistics & file dates apply to: Microsoft Word documents only.

Document statistics include information on when the document was created, modified, accessed, and printed. In addition, document statistics display the name of the person it was last saved by, the revision number, and the total editing time. Other statistics include number of pages, paragraphs, lines, words, and characters.

- Risks: Document statistics can create embarrassing situations. For example, the "last saved by" metadata shows the last person who edited the document and can create discrepancies over who worked on a document.

Document reviewers apply to: Microsoft Word documents only.

Document reviewers consist of a list of users that have added or accepted document changes.

- Risks: Document reviewers' metadata exposes who has suggested what changes. Removing the names of reviewers can be as important as removing the changes they have suggested.

Hidden text or macros apply to: Microsoft Word, Microsoft Excel, and Microsoft PowerPoint documents.

Especially in Microsoft Word, metadata issues arise in hidden text, footnotes, white text, and small text remaining in documents.

- Risks: Either added to documents wilfully or unintentionally, these items are not visible to the naked eye, but can be revealed easily if they remain in the document.

Custom properties apply to: Microsoft Word, Microsoft Excel, and Microsoft PowerPoint documents.

Custom properties include any property fields added to a document manually, or by various programs, to help manage and track files.

- Risks: Custom properties are normally specific to an organization. Common types of custom properties are document ID, department, and status. Custom properties can reveal proprietary information or competitive business practices.

What about PDF?

PDF documents also present similar metadata challenges. For example, a document converted to a PDF file may include metadata in the form of document properties that allow the document to be indexed more fully by archiving systems, as well as searched with more granularity. Some tools, such as Adobe Acrobat and the Mac OS X Preview function, allow annotations and comments to be stored in PDF documents.

While inadvertently including metadata in a PDF file is not as likely as with other types of files, it can happen as we have seen with the “UK Bank” example. People often convert Word documents to PDF to eliminate comments and tracked changes. However, if these changes are displayed in the Word document when the PDF is created, the changes will also appear in the resulting PDF file. Similarly, if the ‘Print Hidden Text’ option is selected in Word, hidden text will appear when the PDF file is created.

3.0 Three simple sharing considerations

In light of these risks, Workshare recommends taking these three key considerations into account before sharing documents.

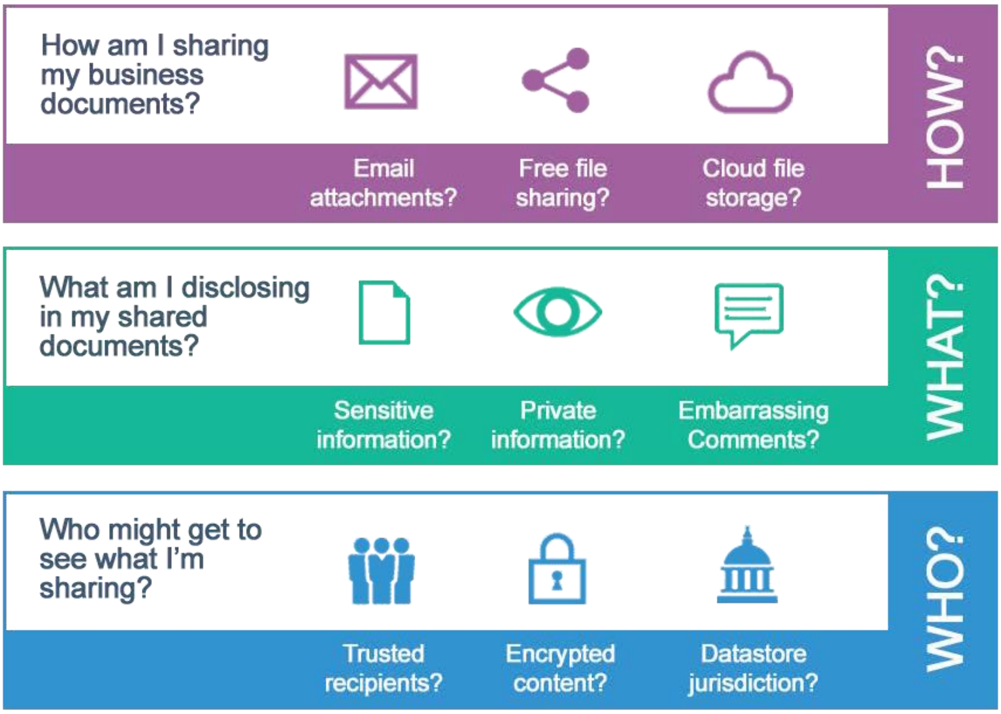


Figure 5: Levels of sharing consideration

How: Consider the methods you use to share information

The traditional method for sharing company documents has been as email attachments. But as there have been massive changes in how people work, especially with the surge of the Bring Your Own Device trend, more and more work on documents is being carried out on mobile devices, much of it outside of the traditional confines of the office. So while IT departments may have policy or technology in place to protect email attachments once they are sent, we’re now sharing in so many different ways, including via consumer-grade file sharing services or personal email accounts which are not being protected. As such, there’s a higher chance for information to fall into the wrong hands or be exposed.

Therefore, we must become more conscious of the risks inherent with sharing documents in this way, and further still, understand how to mitigate those risks. IT sanctioned tools and processes must be deployed company-wide, across devices. The only way to ensure that employees are using these applications is to deploy solutions that give them the ability to share as they are used to working. This includes protecting email and providing a secure alternative to the consumer file sharing apps they are using, with the same ease of use, but fully sanctioned by IT.

What: Consider what is being shared when you share a document

The documents that professionals share will contain high-value intellectual property, confidential information, and even highly sensitive or personally identifiable information. But inside those documents, behind the text or a clickable embedded table, there’s hidden metadata – track changes in a Word document, notes in a PowerPoint presentation, or confidential financial information in an Excel table. Think about that. Do we really intend someone to be able to click through to the complex Excel data and formulas behind what is meant to be a simple graph in PowerPoint? By being aware of all the data we may be sharing, we can ensure we have systems in place to selectively remove the metadata we don’t want to share.

Who: Consider snooping and surveillance

Recently, the term “metadata” has entered mainstream vocabulary. With Wikileaks and scandals around national security and intelligence agencies making headlines, we are all more aware of the need to protect our digital footprint, be that business or personal. But it’s confusing too. Our phone provider has knowledge and patterns about our phone usage, and they store that information as metadata. While our calling behavior creates that metadata, we don’t have direct control over how it’s protected. But with document metadata hidden inside reports, spreadsheets, or presentations that we share and store online, we do have control. As the controversy around privacy and snooping continues, protecting our own documents and the metadata inside them is our own responsibility.

4.0 Conclusion

Document metadata can serve useful purposes for identifying, indexing, and managing documents. It is critical for us all to understand how metadata is created, where it is stored in a document, and how it changes, especially when collaborating. All types of document metadata can reveal confidential information that may result in discrediting incidents, competitive disadvantage, or outright legal action against your organization.

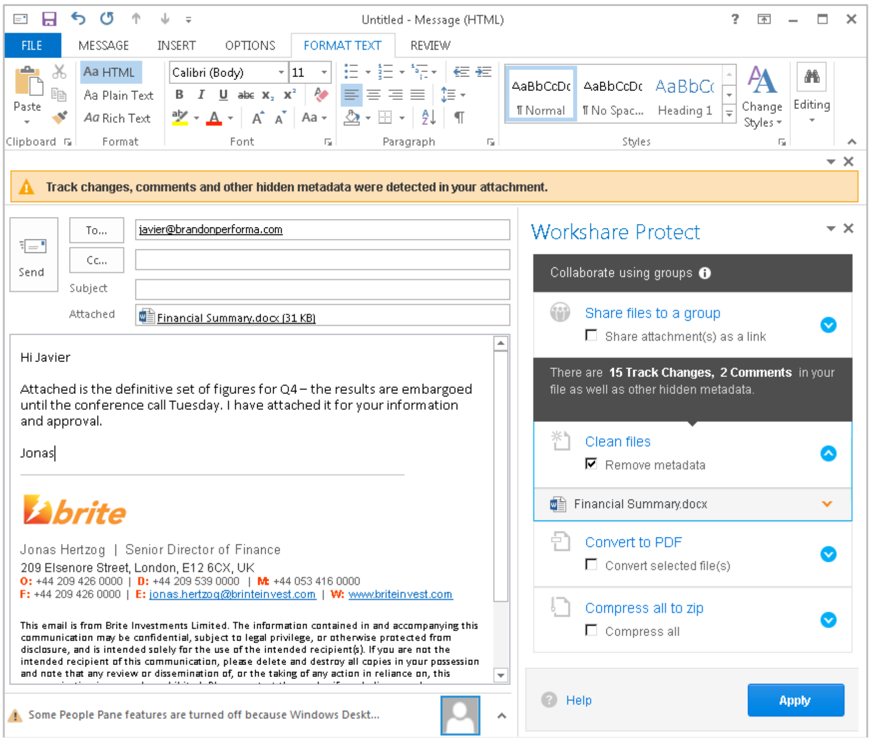
Be aware of document metadata – understand what it is and how to selectively remove it from documents that are shared. The methods professionals’ use, especially when using mobile devices and networks outside of the office, must be sanctioned by IT and fall under company compliance policy. Equally as importantly, they must provide tools for metadata cleaning.

Take control and mitigate the risks around metadata that could be leaked or exposed before it’s too late. The first step is to become more conscious of the risks and how to manage them. Metadata can and generally should be considered for removal before distributing a document outside the organization, so check in with your team to see if this best practice is being implemented and, if not, take steps to ensure your organization and everyone within it is protected.

5.0 Workshare’s heritage

Workshare has been providing secure document sharing methods and the means to identify and strip metadata from files for over a decade. In fact 62% of the Fortune 1000 Legal Services corporate counsels use Workshare today, as do 98% of the legal firms in the US.

Workshare document collaboration and protection software/services analyze documents as soon as they are attached to an email, giving a detailed list of the hidden data (metadata) associated with that document, and presenting a clear and obvious way to deal with those potential issues. For example, in a Word document, cleaning the track changes history or removing notes and comments that have been made in the document.



⚠ Track changes, comments and other hidden metadata were detected in your attachment.

To... javier@brandonperforma.com

Cc...

Subject

Attached Financial Summary.docx (31 KB)

Hi Javier

Attached is the definitive set of figures for Q4 – the results are embargoed until the conference call Tuesday. I have attached it for your information and approval.

Jonas

brite

Jonas Hertzog | Senior Director of Finance
209 Elsenore Street, London, E12 6CX, UK
O: +44 209 426 0000 | D: +44 209 539 0000 | M: +44 053 416 0000
F: +44 209 426 0000 | E: jonas.hertzog@brinteinvest.com | W: www.briteinvest.com

This email is from Brite Investments Limited. The information contained in and accompanying this communication may be confidential, subject to legal privilege, or otherwise protected from disclosure, and is intended solely for the use of the intended recipient(s). If you are not the intended recipient of this communication, please delete and destroy all copies in your possession and note that any review or dissemination of, or the taking of any action in reliance on, this

⚠ Some People Pane features are turned off because Windows Deskt...

Workshare Protect

Collaborate using groups

Share files to a group
□ Share attachment(s) as a link

There are **15 Track Changes, 2 Comments** in your file as well as other hidden metadata.

Clean files
☑ Remove metadata

Financial Summary.docx

Convert to PDF
□ Convert selected file(s)

Compress all to zip
□ Compress all

Figure 5: Workshare’s Interactive Protect removing metadata in Outlook

For more information on metadata removal, see **http://www.workshare.com/products/protect**.

Or contact Workshare at **sales@workshare.com**, or via **www.workshare.com**.

6.0 APPENDIX A: Recommendations

As well as being aware of the metadata types and potential implications of document metadata, it’s important for us to actually deal with it. Workshare recommends that technology is consistently used to identify, manage, and clean metadata, making it an essential, habitual part of our daily sharing routine.

This table shows recommendations on whether to leave, remove, or perform another option on document metadata before sharing a document. In the 'Comments' column, we spell out the actions you should take with Workshare’s metadata management technology – either selected in the Interactive Protect panel, or automated by a centrally enforced policy.

Metadata Type	Recommendations (Remove/Leave/Options)	Comments
Built-In Document Properties	REMOVE	Built-in document properties include summary properties (author, category, comments, company, keywords, manager, title, subject, etc.) Some are automatically set by Word when saving a document. If “Remove” is selected, Workshare removes all summary properties author, category, comments, company, keywords, manager, title, subject and custom properties (text, date and number).
Document Statistics	REMOVE	Document statistics keep track of date/time information of creation, edit, and print as well as total editing time. If “Remove” is selected, Workshare resets all of the document statistics including total edit time, revision number, last authors, and file dates.
Custom Document Properties	OPTION	Custom properties are added in through many 3rd party applications such as the DMS. This should be reviewed within your organization to setup exclusions to maintain the integrity of the documents. You can prevent certain custom properties from being cleaned, for example, DMS Doc ID Property, by excluding them in the Protect module. If this is selected, Workshare removes any custom properties that have been added to the document (and this is turned ‘on’ by default).
Document Variables	REMOVE	Document variables are values stored in Microsoft Word documents that are used by either field codes or macros. These variables may contain confidential information like company names, or file locations. Even if field codes and macros are removed, the variables used may remain in the document. Variables can be viewed in Microsoft Word in the Visual Basic editor. You can prevent certain document variables from being cleaned by excluding them in the Protect module. If “Remove” is selected, all document variables are removed.

Metadata Type	Recommendations (Remove/Leave/Options)	Comments
Track Changes	OPTION	Track Changes contains not only the changes made in the document but also the editor’s name. Depending on how the view options are setup within Track Changes, the user will see differing levels of data. Many times a user will not even know that track changes are in the document. There are many times where Track Changes are needed for the review cycle with a client or a colleague, the user is alerted to the existence of Track Changes metadata. If “Remove” is selected, all revisions made to the document are accepted. The revisions are therefore no longer displayed as revisions but rather as text in the document. Track Changes is also turned off so that further revisions are not tracked. This option is set by default.
Hidden Text	REMOVE	Hidden text is often used for hiding data from being printed or being viewed by the recipient of the document. This data can be easily viewed with the Show/Hide button in Word potentially allowing access to confidential information. If “Remove” is selected, Workshare removes all text that has been formatted as hidden.
Comments	OPTION	Comments are added by reviewers of documents and are part of the collaboration process. Comments should be assessed before sending. If this option is selected, Workshare removes any comments embedded in the document. It is on by default.
Embedded Objects	OPTION	Word can embed objects like Excel, PowerPoint, PDF and equations into a document. These objects contain their own metadata and are therefore should be converted or removed to protect the underlying data. Workshare sets this to off as default.
Small Font	OPTION	Small font is often used as a formatting tool for applying spacing and is used in many templates. Therefore, templates should be reviewed for the use of small fonts to avoid loss of formatting. If this option is selected, Workshare removes all text that has been formatted with a font size smaller than 5pt. It is on by default.
White Text/Font Color Matching Background	OPTION	White text on matching background is typically used for hiding text, however, the text can be revealed when pasted into a new document. This is set to be removed by default.
Content Controls	OPTION	If customers are using Content Controls in their templates then the ability to remove them is an advisable option to protect the integrity of the document. Although Workshare sets this on by default, you may want to pay attention to this option.

Metadata Type	Recommendations (Remove/Leave/Options)	Comments
Bookmarks	REMOVE	Bookmarks should be removed.
Field Codes	OPTION	Field codes have many different applications in a Word document, including hyperlinks, table of contents and index. Converting field codes helps to prevent any field code updates to the document after you have distributed it. It also prevents errors for fields that reference built-in or custom properties that have been removed. You can prevent certain field codes from being cleaned, for example TOC or page numbers by excluding them in the Protect module. If this option is selected, Workshare converts any field codes that exist in a Microsoft Word document to text, for example hyperlinks, table of contents, index. This conversion is on by default.
Headers/ Footers	OPTION	Headers and footers can be used to identify a document but in some circumstances might give away more than the author intended. If they contain sensitive information then the default should be on.
Footnotes/Endnotes	OPTION	Footnotes and Endnotes contain data that can be part of the document, however, a choice is given to remove this data from the document. If this option is selected, Workshare removes any footnotes or endnotes included in the document.
Smart Art	LEAVE	Smart art is built into Word and allows for illustrations in a document, including flow charts. We do not remove SmartArt from documents.
Hyperlinks	OPTION	Hyperlinks fall under the Field Codes section. You can prevent hyperlinks from being cleaned, by explicitly excluding them from removal in the Field Codes options. If “Remove” is selected, Workshare converts hyperlinks that exist in a Microsoft Word document to text.
Custom XML	OPTION	“Remove” is set to on by default.
Hidden Objects (selection and visibility pane)	OPTION	“Remove” is set to off by default.
Attached Template Name	REMOVE	A template name other than “Normal” can reveal information about your internal IT structure and at times cause Word to hang when it is looking for a template that does not exist. If “Remove” is selected, Workshare converts the attached template to normal .dot. Automatic style updating is disabled before the template is removed. Therefore the formatting and styles in your document will not be affected by removing the attached template.

Metadata Type	Recommendations (Remove/Leave/Options)	Comments
Legacy Information(e.g. Routing Slip, Fast Saves, and Personal Information related to .doc fileformat)	REMOVE	Routing slips can hold email addresses of colleagues/ clients and can unknowingly be distributed. If "Remove" is selected, Workshare removes all entries from a routing slip, as well as the message subject and text. This can prevent email addresses of colleagues from being unknowingly distributed. This also deletes any envelope information, such as recipients, subject, and introduction, which are used when sending to a mail recipient. Fast Saves carry data of every version of a document which can be easily reviewed by a recipient. If "Remove" is selected, any previous versions of the document that may have been saved, but often contain confidential information that should not be shared, are removed.
Track Changes/ Comments/Name of User Making Changes	OPTION	As opposed to a unilateral removal, cleaning reviewers but not Track Changes may be useful if you are collaborating on a document with an external party who uses Track Changes. So with Workshare, you can retain the actual track changes made in the document, but remove confidential information about the author who made the change. If "Remove" is selected, information about all document reviewers who have made changes in the document are removed. Track changes are not removed but information about the user who made the change is.
Track Changes/ Comments – Date and Time of Changes	OPTION	Set to be on by default, Workshare alerts you to the existence of this type of metadata.
Password Protected Documents	OPTION	Password protected documents are cleaned like any other document and fall under the same standards of metadata removal. As the document is being processed, the user receives a prompt from Workshare to enter the password so that cleaning can take place.
Embedded/ Inserted Objects' Metadata	OPTION	With this option, metadata in embedded or inserted objects can be cleaned like any other metadata contained in the document.