The ultimate guide to redaction: What works and what doesn't.

This industry guide takes a closer look at data leaks that have happened because of poor or improper redaction. It evaluates typical advice on methods for electronic redaction based on two key criteria: efficiency and success. It also asks whether a PDF redaction tool is a must-have for businesses.

Big data. Big problems.

We live in the age of big data. Whether it's the amount of our personal data being collected by businesses and advertisers; the government passing new data privacy laws; or, it's news about our personal data being stolen as a result of a hack or an inadvertent disclosure—there seems to be no escaping it.

The term big data is a relatively new one, but it refers to the amount of information businesses collect on their customers from a variety of sources; sales transactions, credit card information, social media activity, online behavior, or websites visited. The goal for businesses in collecting big data is so they can get to know their customers better and to sell them goods and services, wherever they are, and on whatever device they use.

While big data brings great insight, it also comes with a high level of risk. Governments around the world have responded to this unprecedented growth in data collection by passing laws aimed at protecting an individual's personal information and data.

In most cases, companies have to ensure that they have robust processes and systems in place to prevent data leaks in the first place and to inform individuals if a leak has occurred. Severe penalties can be imposed if either of these requirements are not met.

Despite new laws and regulations, we continue to hear and read about data breaches that result in millions of private records and sensitive information being compromised or stolen. Unintentional data leaks account for a significant amount of such breaches. We can break these leaks down into the following categories:

Missent emails – sending confidential emails to the wrong person.

Hidden metadata – failure to remove sensitive document or author properties from documents before emailing.

Improper redaction – sharing confidential documents that have been improperly redacted.

What does redaction have to do with big data?

A decade ago, the term 'redacted' would not have been widely used or understood unless you worked in the legal profession or had been involved in a court case. Searches on Google for 'redacted' and other similar terms have spiked since 2005, a trend that is likely to increase given a) the growing number of reported data breaches caused by improper or poor redaction technique and b) the growing body of government legislation aimed at protecting people's personal information.

The **Duhaime Legal dictionary** explains redaction as follows;

> *"Redaction is generally justified for reasons of privilege. Although relevant documents have to be disclosed between litigants, some documents, in whole or in part, may contain references, parts or elements which are not subject to disclosure. An example might be a long, relevant document which has a few paragraphs that contain a summary of legal advice protected by the client-attorney privilege, jeopardize state security or reveal the identity of a state informer. If practicable, the document should be disclosed but "redacted for privilege," with the confidential portion blacked-out or whited-out or otherwise removed. Generally, the Courts prefer a party to redact segments of a document for privilege, as opposed to the complete nondisclosure of a document, as it fosters full disclosure."*

The courts are also mandating the redaction of personal information in court filings. Attorneys now have an obligation to redact many sorts of confidential information before submitting documents to courts, including social security numbers, financial account numbers, names of minors, dates of birth, home addresses, and other sensitive information.

Recent redaction fiascos

Over the past 10 years, we have seen spectacular redaction fiascos exposing national security secrets, business deals, and everything in between. And no one appears to be immune. The following individuals, businesses, and government departments all accidentally disclosed information they thought had been safely redacted.

- Paul Manafort's lawyers,
- the British Ministry of Defence,
- Facebook,
- UK Department of Transport,
- US Transportation Security Administration,
- The New York Times, and
- Rob Blagojevich (former Illinois governor).

Even though redaction software is widely available, you have to ask, "why is it so many people continue to get redaction so wrong?"

Google, how do I redact a document?

To answer the question "*why do so many people continue to redact documents incorrectly,*" you need to look at the advice and guidance people are getting. There is a wealth of information freely available on the Internet that explains how you should redact information using a variety of techniques ranging from flattening and round tripping files to redacting content and copying information from one application to another.

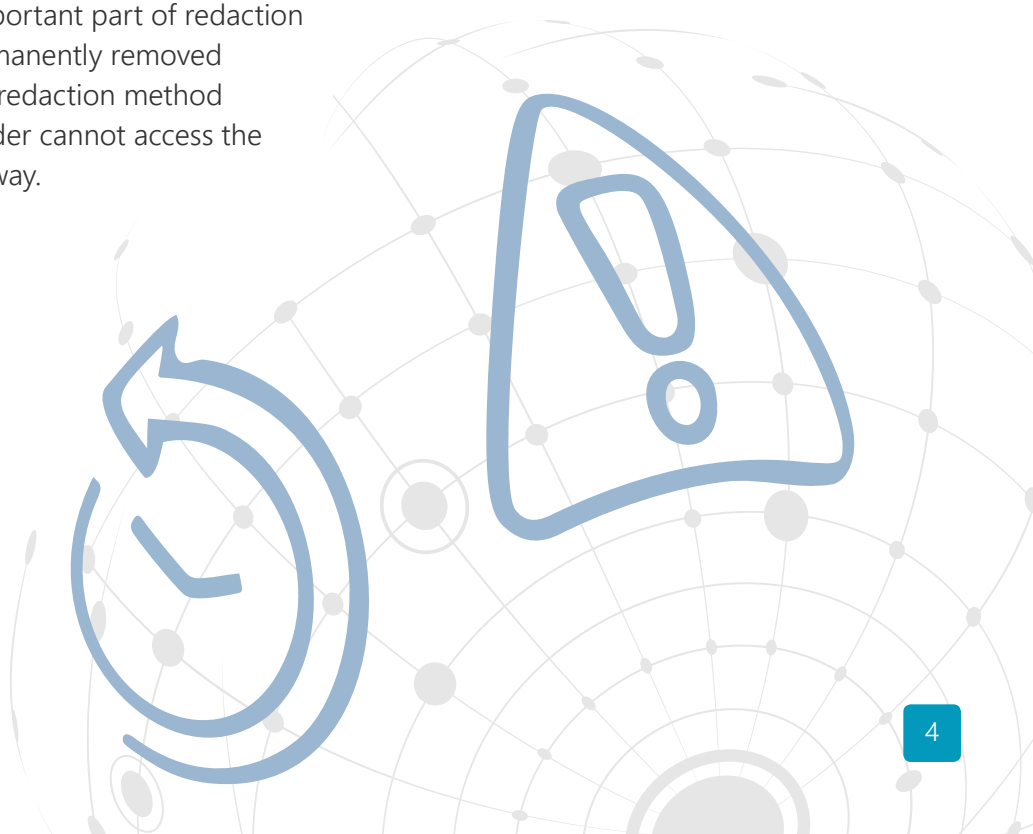
In the following section we will look at the most commonly recommended methods for electronic redaction found on the web. You can also consult these authoritative sites:

- The Case Management/Electronic Case Files (CM/ECF) in the United States
- The Office of the Victorian Information Commissioner (OVIC) in Australia
- The National Archives in the United Kingdom

This guide will evaluate advice on redaction based on two criteria: efficiency and risk mitigation.

Efficiency: the redaction process must be streamlined so that it can be undertaken with minimum disruption or effort. A convoluted manual process is not only time-consuming but tends to lead to errors.

Risk mitigation: the important part of redaction is that the content is permanently removed from the document. The redaction method must ensure that the reader cannot access the redacted content in any way.



Electronic redaction methods

MICROSOFT WORD – FIND AND REPLACE

You can redact information by deleting the text in the word processing version of the document. Replace text with "Text Redacted."

METHOD

Create a new copy of the document. Using Word's Find and Replace functionality, replace the text string with "Redacted Text." Save the new document as a copy.

EFFICIENCY

This is more efficient than hardcopy redaction and can usually be done quickly.

RISK

While the sensitive information has been removed from the visible layer of the document, it may not have been removed from the document metadata.

Unlike plain text files, Microsoft Word and other word processing software provide change histories, audit trails, and metadata, which can reveal deleted text.

ROUND TRIPPING MICROSOFT WORD – NOTEPAD – MICROSOFT WORD

You can redact information by deleting the text in the word processing version of the document. Replace text with "Text Redacted." Copy the text from Word to Notepad and back to Word.

 METHOD	Using Word's Find and Replace functionality, replace text or a string of text with "Redacted Text." Copy the contents of the document and paste it into Notepad or similar text editor. Open up a blank Word document and copy the content from Notepad into the new metadata-free Word document.
 EFFICIENCY	This is not an efficient process. You will have to reformat the document as styles and headings will have been lost.
 RISK	Minimal risk - the redacted content has been permanently removed from the document. There is no metadata or version information available for the document

WORD – CHANGE FONT COLOR – CONVERT TO PDF

Another method of redaction for Word documents is to match the font color of the redacted text to the background color to hide it.

METHOD

Create a copy of the document in Word. Find a text string for redaction and change the font color to white to match the background. Repeat for all other text to be redacted. Convert the document to PDF.

EFFICIENCY

More labor-intensive than the find and replace method.

RISK

Sensitive information has not been removed from the document. It is simply hidden. Select the text in the PDF and copy into a new Word document. Change the font color to expose the hidden text.

Also, sharing a redacted PDF using the above method will expose it to web search engines that may be able to read the text. This is an unnecessary risk.

PDF – MASK TEXT WITH A BLACK BOX

A commonly used method is to mask content in a PDF document using the annotation tools.

METHOD

Convert the document to PDF. Draw a black box over the text to cover up the sensitive information. Repeat for all other instances of redacted text. Save the PDF as a new document. Flatten the file as the final step in the process to merge the layers of the PDF, i.e., the text layer and the annotation layer.

EFFICIENCY

This is not efficient as you have to review the document for text that needs to be redacted. There are also too many steps.

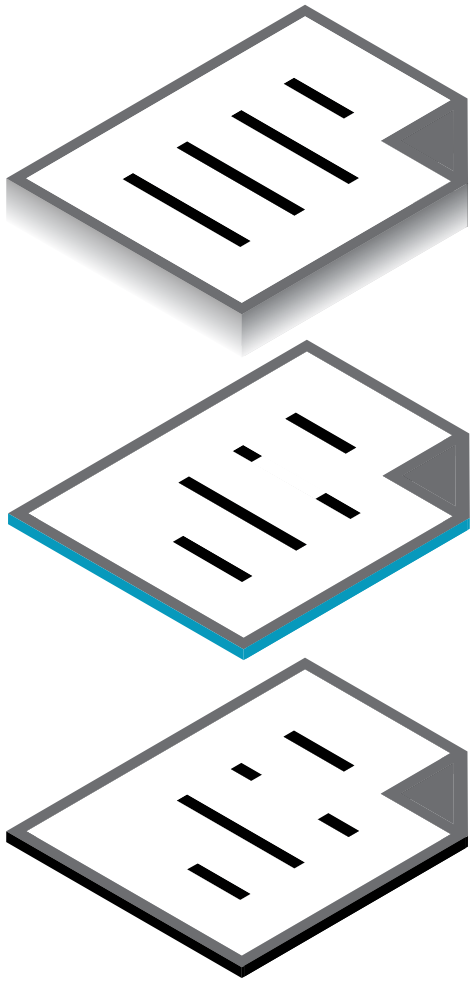
RISK

There is the risk that text will be missed.

The information has not been removed from the document, it is merely hidden. More importantly, PDFs are constructed in layers. One layer for text, another for graphics, and another for annotations. Copying the text from the PDF and pasting it into a new Word document will expose the hidden text.

Flattening a PDF is not the same as burning in a redaction. Flattening simply merges everything onto the text layer, which can be copied despite the black boxes.

Redaction should always be performed in PDF



PDF is the only format to meet the key requirements of a redaction tool: efficiency and risk mitigation.

1 **Remove content from the document** – a PDF redaction tool removes the information from the document once the redaction(s) has been applied. It cannot be undone or exposed later because it is no longer in the file.

2 **No metadata** – when converting documents to PDF from Word, the metadata is not carried over to the PDF. PDF creators or editors do not generate a lot of document metadata and certainly not any metadata on versions.

10 features your PDF redaction tool should have

1. **Remove content, don't mask it:** your redaction tool must burn out the text from the PDF to ensure it is completely removed from the document.

2. **Handle all PDF types:** When you examine a PDF file, there is no conclusive way to visually know which PDF format you're viewing (Text-PDF, Image-PDF, Image-OCR PDF). It is vital that you use an application with tools specifically designed for redaction. The tool must (a) understand which type of PDF file it is working with; (b) remove any text in a redaction zone if the text exists and (c) permanently affix any redaction marking.

3. **Redact text, graphics, images:** your redaction software should be able to permanently remove text, graphics, and images from the PDF document. So, if someone copies the content into another application, the content will not display.

4. **Search and redact:** this feature works much like the find and replace function in Word. Type in a word or string of words to locate them in the document. Then you can apply the redaction to one or all of them.

5. **Pattern search and redact:** this feature enables you to quickly find and redact information in patterns such as credit card numbers, dates, email addresses, and social security numbers.

6. **Exemption codes to explain redaction:** when you redact a word, image or an area of a document, the reader will want to know why this information is blacked out. There are some standard exemption codes that are required by certain courts and other regulatory authorities. However, your software should give you the ability to write on the redacted area explaining the redaction.

7. **Redaction review workflow:** it is not unusual for redactions to be first marked up by a paralegal or legal assistant and then sent to a lawyer for final approval. The lawyer can then apply the redactions as needed or required.

8. **Full-page and multi-page redaction:** should you want to redact an entire page; your redaction solution should provide you with the ability to redact a full page as well as a range of pages.

9. **Protect the original document:** when the redaction is applied or burned into the PDF document, it should not overwrite the original.

10. **OCR workflow for paper documents:** your PDF application should provide OCR workflows for capturing paper documents copied or scanned and converting them to text-searchable PDFs as part of an automated process.

In summary

One of the greatest risks a business can face is the inadvertent disclosure of privileged or confidential information. This can result in huge fines and loss of professional reputation. Permanently removing content from a document to ensure confidentiality is a common practice that is all too frequently attempted with tools intended for other types of annotation. There is no substitution for a true redaction solution; look for one that is simple to use, efficient, and eliminates risk.

Sponsored by:

DocsCorp

Work smart

DocsCorp designs easy-to-use software and services for document professionals who use enterprise content management systems. We provide solutions for metadata removal, document processing, PDF manipulation, and document comparison.

The DocsCorp product suite is built to drive business efficiency and increase the value of existing technology investment. We help transform any organization's slow and outdated processes using sophisticated solutions that integrate with most major document management systems.

DocsCorp is a global brand with customers located in the Americas, Europe, and Asia Pacific. More than 500,000 users in 67 countries rely on DocsCorp software every day.

SYDNEY
LONDON
PITTSBURGH

www.docscorp.com