Date 3/18/2008

Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word 2007 to PDF

Architectures and Applications Division
of the
Systems and Network Attack Center (SNAC)
Information Assurance Directorate

National Security Agency
9800 Savage Rd. STE 6704
Ft. Meade, MD 20755-6704
(410) 854-6191 commercial
(410) 854-6510 facsimile

Abstract

Prior to the distribution of Microsoft Word documents, users may want to remove content and assure that no data remnants persist in the file from the removed portion(s). Office 2007 offers a new interface and new features that change how users interact with Word. This paper describes the issue and provides a step-by-step procedure that provides reasonable assurance that information will not be unintentionally released following redaction of a Word 2007 document.

What This Document Addresses

This guidance intends to present Microsoft Word 2007 and Adobe Acrobat 8.1 (or higher) users with a procedure that can be followed in order to reduce the threat of hidden data in a document following redaction.

What This Document Does Not Address

This guidance does not intend to examine the effectiveness of the Microsoft Office 'Inspect Document' feature, the Adobe Acrobat Professional 8.1 'Examine Document' feature, or any possible vulnerabilities associated with Microsoft Office or Adobe Acrobat. This guidance enables users to reduce their level of risk in a uniform manner, although users should recognize that it does not intend to mitigate all possible hidden data risks.

Contents

1 Introduction 4
2 Background 4
3 Hidden Data and Metadata Removal Tools 5
4 Process Overview 6
4.1 Word 2007 Process 6
4.2 Acrobat 8.1 Process 6
5 Detailed Process Description 7
6 Conclusion 14
References 15

Figures

Figure 1 Correct: Tracking Area with Track Changes Off 7
Figure 2 Incorrect: Track Changes Turned On! 7
Figure 3 Accept All Changes 7
Figure 4 Review and Remove Comments 7
Figure 5 Text View Prior to Deletion 8
Figure 6 Text View Following Deletion 8
Figure 7 Redacted Text Where Formatting and Placement have Been Preserved 9
Figure 8 Redacted Image or Object Where Formatting and Placement have Been Preserved 9
Figure 9 AutoShapes 10
Figure 10 Document Views 10
Figure 11 'Browse By' Interface 10
Figure 12 Select 'Inspect Document' 11
Figure 13 Check all options and Inspect 11
Figure 14 Word Menu Ribbon for Adobe Acrobat Features 12
Figure 15 Basic Conversion Settings 13
Figure 16 Word-Specific Conversion Settings 13
Figure 17 Document Menu 13
Figure 18 'Examine Document' Interface 14

1 Introduction

Microsoft Word is used throughout the Department of Defense (DoD) and Intelligence Community (IC) for preparing documents, reports, notes, and other formal and informal materials. Commonly used versions of MS Word include Word 2000, Word XP, and Word 2003. The release of Office 2007 has added Word 2007 to this list.

The Adobe Portable Document Format (PDF) is used extensively by all parts of the U.S. Government and military services for disseminating and distributing documents. PDF provides excellent fidelity and portability, and allows for easy distribution of documents over computer networks and the Internet. PDF files are usually produced using commercial conversion software programs (known as "distillers") that accept source formats such as Postscript or MS Word and output to PDF. PDF is often used as the format for downgraded or sanitized documents.

As past evidence of inadvertent disclosure reveals, simply converting a Word document to PDF does not remove all metadata automatically. In addition, Adobe Distiller and the PDFMaker Add-in to Word (one of the most common ways to convert) translate much of the layering complexity from one format to the next. For example, images placed on top of text in Word will continue to hide (but not prevent access to) the text within the resulting PDF file.

2 Background

Inadvertent disclosure of sensitive information has become an increasingly common problem because complex file formats offer substantial avenues for hidden data. Once a user enters data into the document, effectively removing it can be difficult. In response to the growing problem, the National Security Agency released *Redacting with Confidence: How to Safely Publish Sanitized Reports Converted from Word to PDF* in 2006. This guidance applied to versions up to and including Microsoft Office 2003 and Adobe Acrobat 7. For those applications, the previous guidance still applies.

Microsoft Office 2007 and Adobe Acrobat 8 offer new features, capabilities, and user interfaces. In addition, the new default OpenXML format used by Word 2007 changes how data is stored and enhances the ease with which users can manually verify that metadata has been removed. For these reasons, this updated version addressing the most recent release of Office was produced.

The three most common types of insufficient or incorrect redaction relate to:

1. **Text and Diagrams**- Covering text, charts, tables, or diagrams with black rectangles or highlighting the text in black is a common and effective means of redaction for hardcopy printed materials. It is generally not effective, however, for documents distributed in softcopy format. The most common mistake is simply covering the text with an image or object that does not

effectively remove the underlying content.

2. **Images**- Covering up parts of an image with separate graphics such as black rectangles or making images 'unreadable' by reducing their size has also been used for redaction of hardcopy printed materials. This approach is generally not effective for documents distributed in softcopy form.

3. **Metadata and Document Properties**- In addition to the visible content of the document, most office tools (such as Word) contain substantial hidden information about the document. This information is often as sensitive as the original document and its presence in downgraded or sanitized documents has led to the unintentional disclosure of sensitive information in the past.

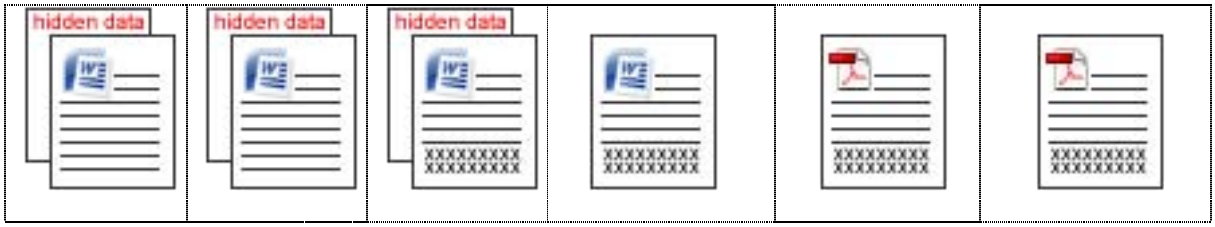
3 Hidden Data and Metadata Removal Tools

Previous versions of Office offered tools for finding and removing hidden data and metadata by download from Microsoft's website. Remove Hidden Data 1.0 (for Office XP) and Remove Hidden Data 1.1 (for Office 2003), hereafter referred to as RHD, allow batch removal of information from Word documents. These tools remove some types of hidden data and metadata, but sensitive information may still be retained. Reliance upon these tools may give the user a false sense of security.

Word 2007 replaced RHD with an integrated 'Inspect Document' capability, which is available without requiring the user to download additional components. While this tool does offer enhanced stability over previous versions, the use of this tool by itself does not replace the manual sanitization procedures outlined in this document. The appropriate use of this tool is as a complementary approach to the steps outlined here. Users should note that while the tool is capable of finding and removing instances of a variety of types of potentially sensitive data, the interface does not display which specific instances are to be deleted, so loss of data may result by running this tool in some cases.

4 Process Overview

An overview of the process is presented here for easy review of the steps. Most of the redaction steps occur in Office, with a final sanitization step included for Acrobat. An overview of the steps is provided here (as shown in Table 1) and more detailed instructions are provided in Section 5.



1. Copy Document (copy.docx)	2. Turn Off Reviewing (copy.docx)	3. Edit Document (copy.docx)	4. Review and Run Doc. Inspector (copy.docx)	5. Verify Settings, Convert (file.pdf)	6. Examine Doc., Save, and Distribute (file.pdf)

Table 1 Overview of the Redaction Process

4.1 Word 2007 Process

1. Open original document. Save a copy of the document with a different name. Retain the original as a backup.
2. Turn off Track changes and remove all reviewing data.
3. Review and delete sensitive content. Delete sensitive text, diagrams, tables, and images. Do not cover up any content- remove it and replace it with innocuous data if necessary to maintain spacing.
4. Check Redacted Content and run 'Document Inspector'.
5. Save the new document. Verify PDF conversion settings. Convert to PDF. Review to ensure redaction was thoroughly performed and that the formatting is acceptable.

4.2 Acrobat 8.1 Process

6. Open PDF Result document in Acrobat Professional 8.1 or higher. Run the 'Examine Document' tool. Select all items for review. Optionally, if bookmarks are to be retained and manually reviewed, uncheck the bookmarks option. Remove any areas identified by the tool. Save the file. The file is ready to be shared.

5 Detailed Process Description

Step 1: Create a Copy of the Document

Open the original document and save it with a different name. The new document is where the redaction will be performed. Retain the original as a backup.

Step 2: Turn Off Reviewing Features and Remove Associated Data

Verify that the track changes feature has been turned off. Select the 'Review' tab, and look in the 'Tracking' area. The display setting should be set to 'Final Showing Markup' as shown in Figure 1 so that all tracking features are displayed for easy review. If another setting is selected, click the down arrow next to it and change it to 'Final Showing Markup'.

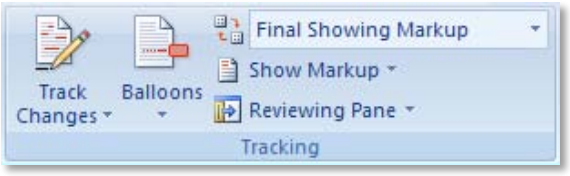


Figure 1 Correct: Tracking Area with Track Changes Off

If 'Track Changes' has been turned on, turn it off. The 'Track Changes' button is highlighted when this feature has been turned on as shown in Figure 2. It returns to the default background color when the feature has been turned off.

Figure 2 Incorrect: Track Changes Turned On!

Navigate through any tracked changes by using the 'Next' and 'Previous' buttons. After reviewing all tracked changes in the document (if any), accept all changes by clicking on the down arrow below the 'Accept' button and selecting 'Accept All Changes in Document' as shown in Figure 3.

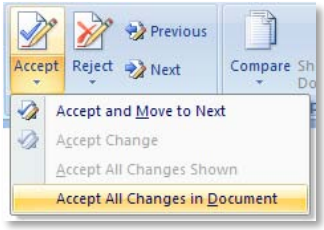


Figure 3 Accept All Changes

Navigate through any comments by using the Next and Previous buttons in the 'Comments' area of the 'Review' tab. Review and remove all comments by using the 'Delete All Comments in Document' as shown in Figure 4.

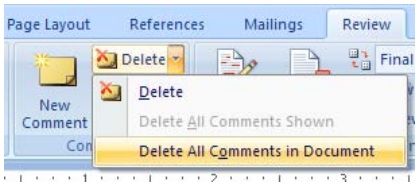


Figure 4 Review and Remove Comments

Step 3: Review and Delete Sensitive Content

Select each chart, diagram, image, or segment of text to be redacted and delete that item. Resizing the item, changing its viewable characteristics such as color, or making the object hidden will NOT effectively remove the data. The content **must** be deleted. If deleting an item changes the format or structure in an undesirable way, replace the item with innocuous content of a size that keeps the formatting correct. If the redacted item is text, replace the text with copies of a single letter such as all A's or all X's (see Example 1). If the redacted item is an image or a chart, replace it with a gray rectangle of the same size (see Example 2).

Example 1: Redacting text. Figure 5 shows a document before redaction. Figure 6 shows the same page after deleting some text. Notice that a paragraph from the next page moves up in the document because deleting the text has shifted the subsequent text.

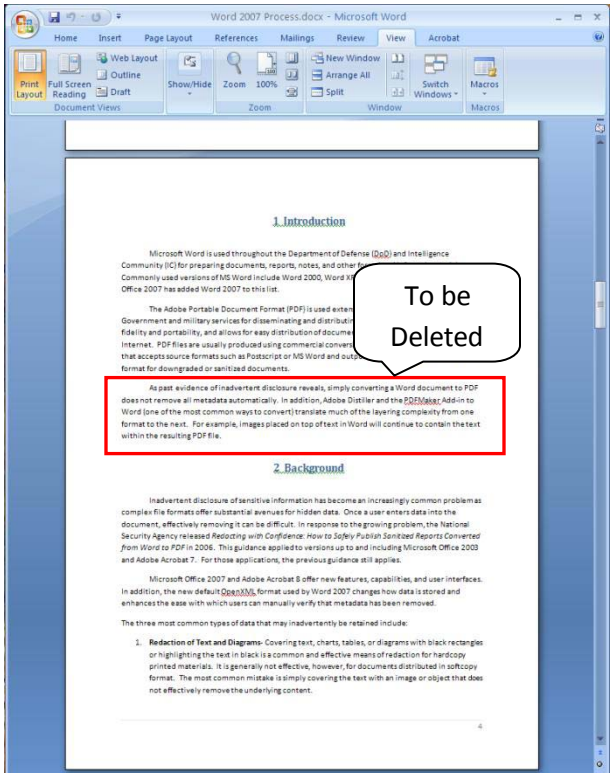


Figure 5 Text View Prior to Deletion

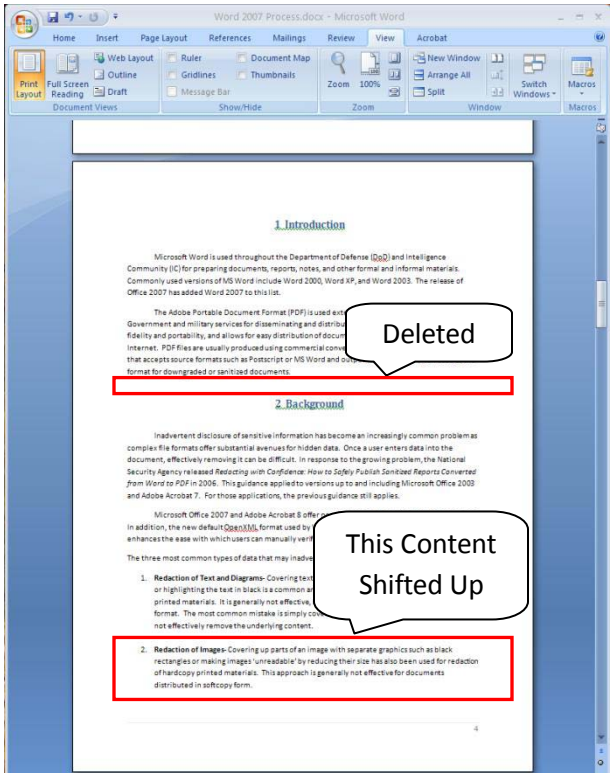
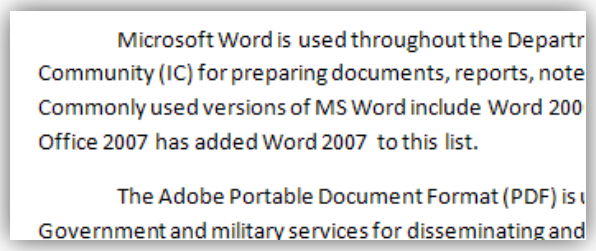
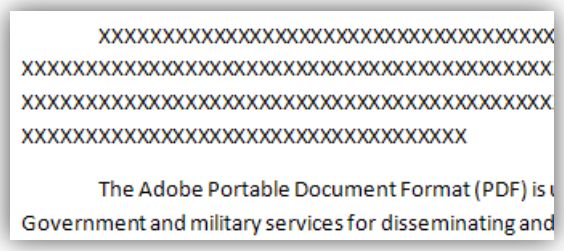


Figure 6 Text View Following Deletion

Because of the shift, subsequent pages may need to be reformatted to ensure figures stay within certain text, page breaks are in the right place, and other such formatting issues do not adversely affect the appearance of the document. This could be time consuming for a large document. If formatting changes are a concern, replace the text with innocuous content of the same size as shown in Figure 7.



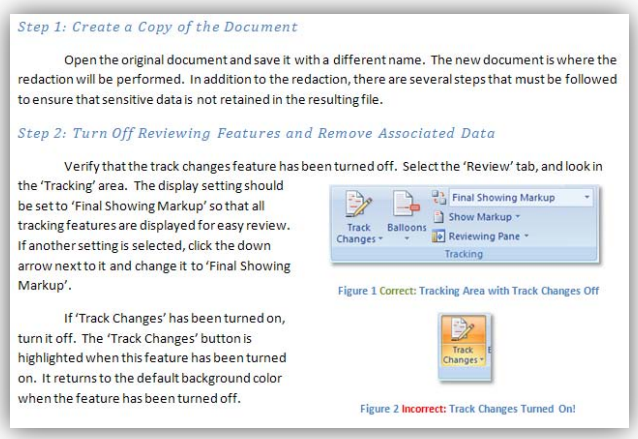
Before



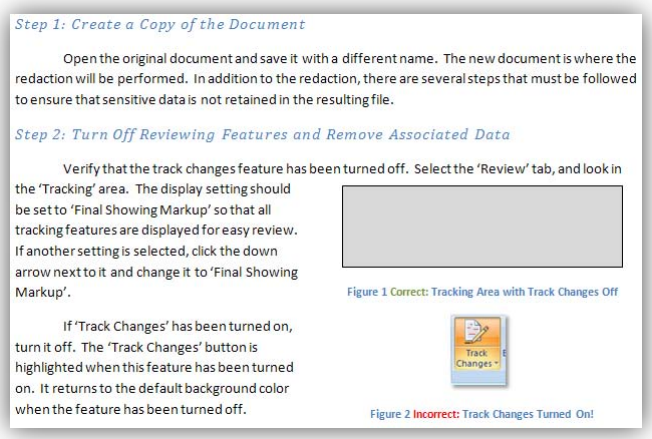
After

Figure 7 Redacted Text Where Formatting and Placement have Been Preserved

Example 2: Redacting a chart, image, or other object. If an image is deleted from a document, the text will shift to fill the void and affect the appearance of the document. To preserve formatting, first delete the image to be sanitized and then insert a grey rectangle that is the same size as the image into the vacant space as shown in Figure 8.



Before



After

Figure 8 Redacted Image or Object Where Formatting and Placement have Been Preserved

The easiest way to ensure that the AutoShape effectively fills the space left by the removal of the previous content is to replicate the size and positioning. To determine the size of the existing object, right click on it before removal and select 'size'. The size will then be displayed. Make a note of the height and width. Alternatively, for some shapes, images, or object types, you may need to right click on the object and select 'Format XXXX' in order to view the size (where XXXX is the content type). To view the positioning, right click on the item and select 'Text Wrapping'. Make a note of the current wrapping type. As with size, some types of objects may require that you use an alternative selection of menu options. If 'Text Wrapping' is not an option, select 'Format XXXX' and select the layout tab to view the current setting.

AutoShapes can be inserted by selecting the 'Insert' tab, and then clicking on 'Shapes' within the 'Illustrations' section as shown in Figure 9. Select the shape that most closely resembles the original image or object. After inserting the shape into the document, right click on the shape and select 'Format AutoShape'. Adjust the size and positioning to match the replaced object by selecting the 'Size' and 'Layout' tabs and adjusting the values accordingly. In the 'Colors' tab, select a grey fill color.

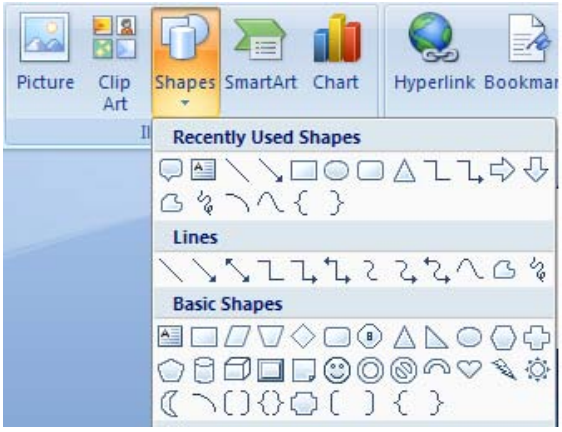


Figure 9 AutoShapes

Step 4: Check Redacted Content and Run Document Inspector

Carefully read over the document to ensure that all material to be redacted has been deleted and, if necessary, replaced with innocuous filler. Check all headers, footers, captions, section titles, footnotes, endnotes, labels, etc. Verify that all comments have been deleted. Ensure that the Print Layout view is selected because some images and objects in the headers and footers may not be visible in the Web Layout, Outline, or Draft views. To view and change the current view, go to 'View' and select 'Print Layout' as shown in Figure 10.

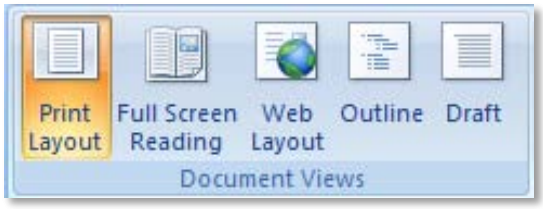


Figure 10 Document Views

To easily flip through the document, use Word's 'Browse By' feature for the different objects. The user interface is shown in Figure 11. By default, the arrows at the bottom of the right scroll bar browse through the document by page when clicked. To customize this behavior, click on the dot at the bottom of the scroll bar that is between the double arrows. This brings up a toolbar with the 'Browse By' options including 'Browse by Comment', 'Browse by Footnote', 'Browse by Graphic', etc. Select one of the options. Now use the double arrows above and below the dot to flip through the document to each instance of that type of object. This is a quick way to make sure all comments are deleted, all images have been redacted correctly, all footnotes have been checked, etc.

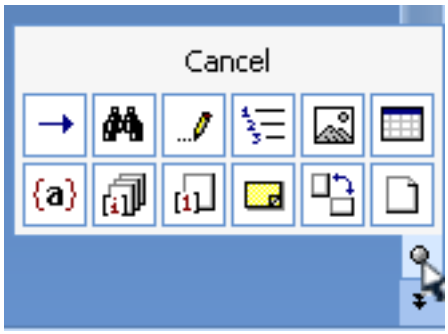


Figure 11 'Browse By' Interface

Next, inspect the document for residual hidden data by using the integrated inspection tool. To run the document inspector, click on the Office icon in the upper left corner of the application. Then navigate to 'Prepare' and select 'Inspect Document' as shown in Figure 12. Ensure all boxes are checked and run the inspection by clicking 'Inspect' as shown in Figure 13.

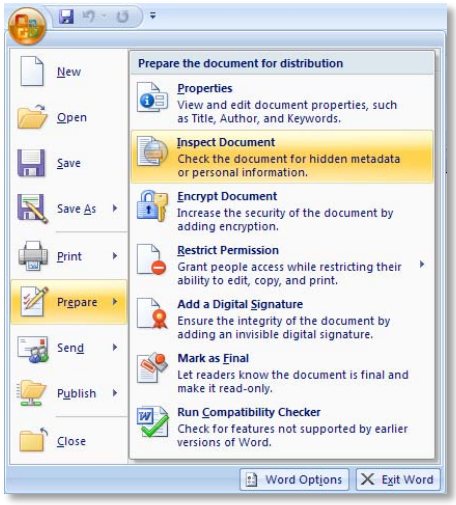


Figure 12 Select 'Inspect Document'

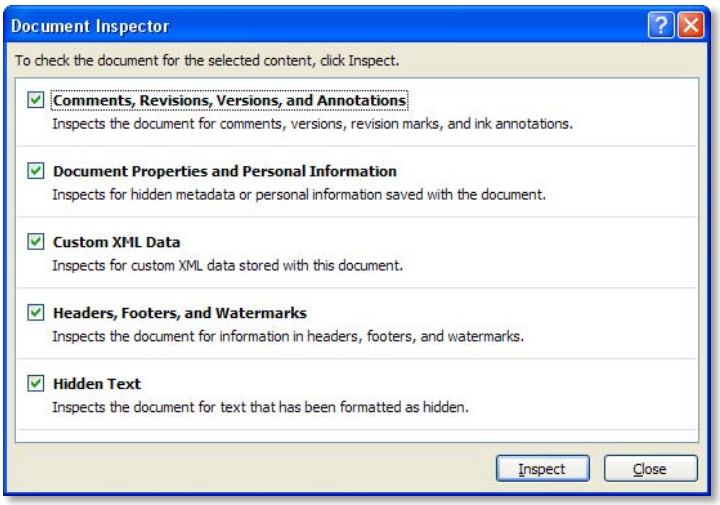


Figure 13 Check all options and Inspect

If any content areas are identified for removal, select remove and re-inspect the content. Verify that the undesired content does not reappear in the subsequent inspection. Headers and footers may need to be manually verified because any content in these areas may trigger the inspection notification, whether the content is sensitive or not.

Step 5: Verify Acrobat Conversion Settings and Convert

For Office 2007 integration (including the availability of the Adobe Acrobat toolbar inside the Word user interface), Adobe Acrobat 8.1 or higher must be used. With previous versions of Acrobat, the settings may not be as easy to modify and the user may not have the option to granularly specify the conversion setting preferences. The 'Preferences' and 'Create PDF' options appear in the 'Create Adobe PDF' section of the 'Acrobat' tab as shown in Figure 14. If the tab does not appear, verify that your Acrobat version is at least 8.1 and follow the instructions on Adobe's site to fix this issue.[1]

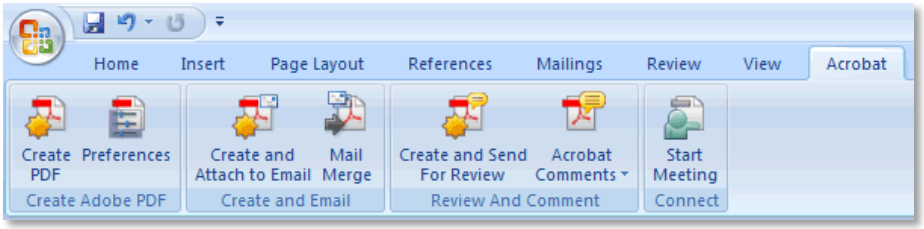


Figure 14 Word Menu Ribbon for Adobe Acrobat Features

Click the 'Preferences' icon. The 'Settings' options are presented as shown in Figure 15. Configure the options as shown, optionally also selecting 'Add Links to Adobe PDF' if you wish to retain any links from the source document. It is imperative that 'Convert Document Information' is **unchecked**. Also uncheck 'Create PDF/A-1a:2005 compliant file' and uncheck 'attach source file to Adobe PDF'. Next, select the 'Word' tab and configure the settings as shown in Figure 16. Uncheck 'Convert displayed comments to notes in Adobe PDF' and uncheck 'Enable advanced tagging'. After selecting the appropriate options, select 'OK' to save them.

[1] http://kb.adobe.com/selfservice/viewContent.do?externalId=kb401734

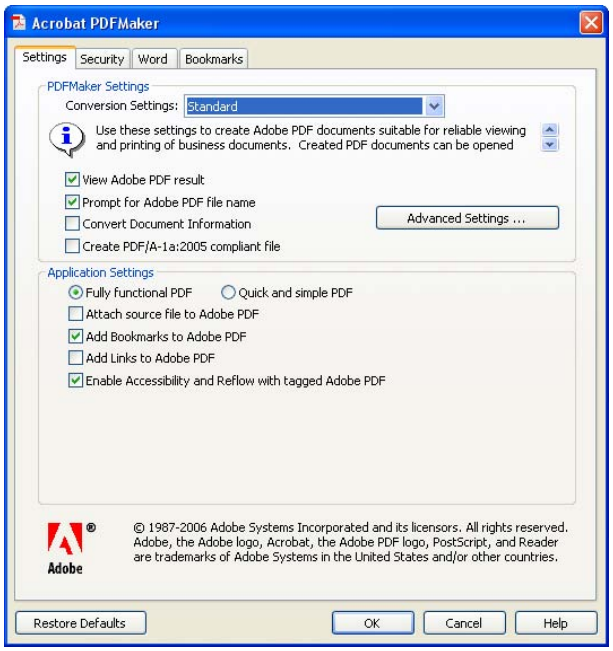


Figure 15 Basic Conversion Settings

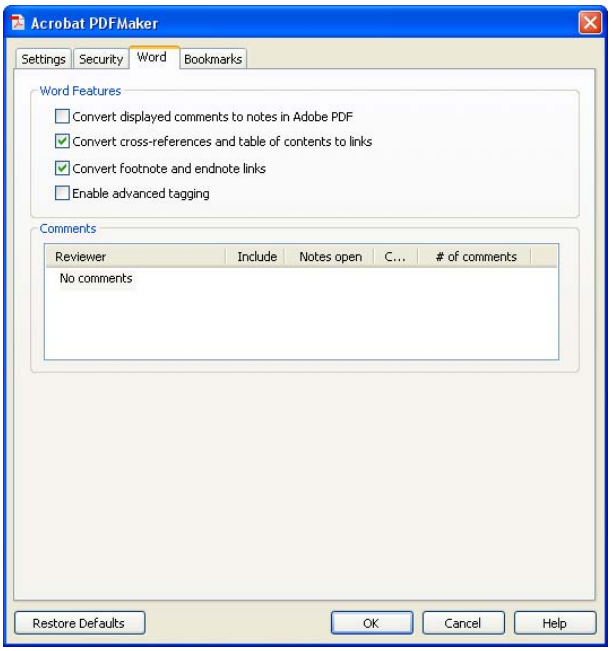


Figure 16 Word-Specific Conversion Settings

Convert the document to PDF by selecting the 'Create PDF' option from the 'Acrobat' tab within Word 2007. Select a descriptive name and save the file.

Step 6: Run 'Examine Document' Tool, Save, and Distribute

Open the resulting PDF file in Adobe Acrobat Professional 8.1 or higher. Select the 'Document' menu, and then select 'Examine Document' as shown in Figure 17.

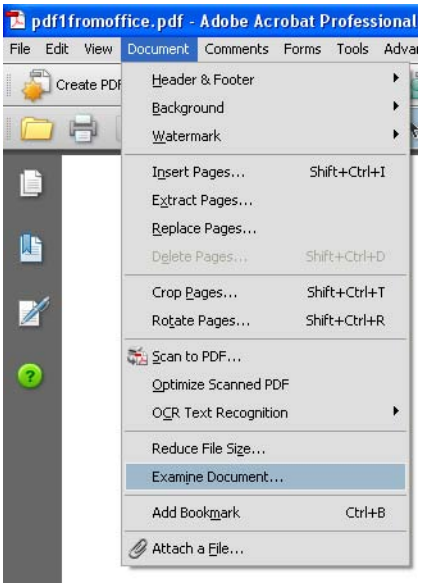


Figure 17 Document Menu

Check all the available options and examine the document. A sample set of results is shown in Figure 18. If hidden text was found, use the 'Preview' link to view any instances of hidden text within the document. This may include text that matches the background color or that was hidden behind an image or object. You should already have verified that such content has been removed before converting to PDF- this is a sanity check! Select 'Remove all checked items' (optionally unchecking 'Bookmarks' if you intend to manually review them) and then re-run the tool to verify that any items found have been corrected. Save the document. This document is the final distributable file.

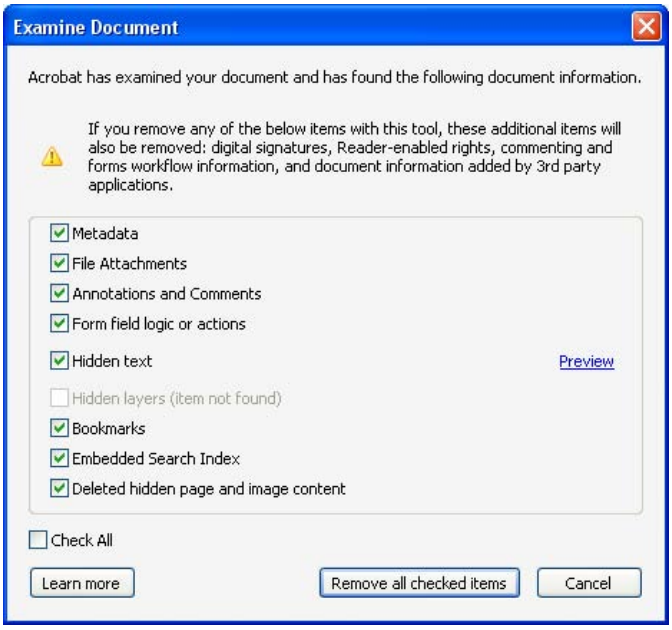


Figure 18 'Examine Document' Interface

6 Conclusion

The new features, capabilities, and interface (including the integrated inspection and sanitization tools) change the way that users interact with Word. The redaction process presented in this document should help reduce the likelihood that hidden data, metadata, and sensitive content are hidden in the resulting file.

The use of Word's 'Inspect Document' feature and Adobe's 'Examine Document' tool are not intended to replace the redaction process, but do supplement the procedure by facilitating additional inspection and sanitization that may catch hidden data that could otherwise remain. Some of the specific benefits that may be provided include the ability to detect and remove:

- Residual Metadata
- Reviewing and commenting data
- Custom XML
- Text that matches background color or is hidden behind an object

More information about the tools is available on the respective developers' websites at http://office.microsoft.com/en-us/help/HA100375931033.aspx and http://help.adobe.com/en_US/Acrobat/8.0/Professional/help.html?content=WS7E9FA147-10E3-4391-9CB6-6E44FBDA8856.html.

References

1. *Redacting with Confidence: How to Safely Publish Sanitized Reports Converted from Word to PDF http://www.nsa.gov/ia/_files/vtechrep/I333-TR-015R-2005.pdf*