

The Risks of Metadata and Hidden Information

Analysis of Microsoft Office Files from the
Web Sites of the Fortune 100

An Oracle White Paper
Updated March 2007

The Risks of Metadata and Hidden Information

Executive Overview	4
Introduction	4
Methodology.....	5
Terminology	5
Study Format	6
Analysis.....	6
Analysis Results.....	6
Study Findings	7
Audio and Video Paths	7
Author History	8
Comments.....	9
Content Properties.....	11
Custom Properties	11
Database Queries	12
Embedded Objects	13
Fast Save Data.....	15
Hidden Slides.....	16
Hidden Text.....	16
Linked Objects	17
Macros and Code.....	18
Office GUID Property	18
Outlook Properties	19
Presentation Notes	20
Printer Information	21
Routing Slips.....	21
Scenario Comments.....	22
Sensitive Hyperlinks	23
Sensitive Include Fields.....	23
Statistic Properties	24
Summary Properties	25
Template Name.....	25
Tracked Changes.....	26
Usernames.....	27
Versions.....	28
Weak Protections	29

Conclusion..... 31
Determining Your Information Exposure..... 31
Metadata and Hidden Information Inspection Versus Content
Filtering 32
About Oracle Outside In Technology..... 32

The Risks of Metadata and Hidden Information

Oracle performed this study to educate users and organizations about the risks associated with information that is commonly exposed when documents are created and shared.

EXECUTIVE OVERVIEW

What is the scope of information exposure in your organization? How many Microsoft Office files are available on your Web site? On partner extranets? How many documents leave your organization each week as e-mail attachments? What measures are in place to inspect file attachments transferred to third parties via instant messaging clients? What process is in place today to inspect these documents for unintended or harmful information exposure?

Oracle performed this study to educate users and organizations about the risks associated with information that is commonly exposed when documents are created and shared. We perceived a general lack of awareness about this subject and a lack of understanding about the specific type of information that can be accidentally exposed through routine business practices and document workflow. The intent of this study is not to find a “smoking gun” or expose potentially sensitive information by identifying specific organizations or specific pieces of content. The sole intent is to increase awareness by exposing the problem in quantifiable terms and provide recommendations for limiting risk factors.

INTRODUCTION

Metadata is commonly described as data about data, or information that is used to describe other information. For the purpose of this study, metadata refers to a variety of information types found inside Microsoft Office files as part of structures such as properties, which can include author names, document titles, keywords, print and save dates, and a range of other information that describes when a file was created, when it was modified, who authored it, and who reviewed and saved the document. Metadata values can also be automatically added to files by other applications and processes, such as external content management and e-mail systems, which insert information that aids in the tracking of files.

Hidden information, relative to this study, represents data that either is not accessible through the Office application’s interfaces or that can be hidden from view using application settings. Examples include comments and track changes, which can be hidden from view as a default setting, or author history and fast save data, which is simply not accessible from the native application interfaces.

The goal of this study is to describe a broad set of data elements that are integral parts of Word, Excel, and PowerPoint file formats. These elements represent information that could contain sensitive, proprietary, or confidential data that is easily overlooked and therefore prone to accidental exposure when files are shared or distributed.

The goal of this study is not to provide de facto definitions of metadata or hidden information, but to use these terms to describe a broad set of data elements that are integral parts of Word, Excel, and PowerPoint file formats. Most important, these elements represent information that could contain sensitive, proprietary, or confidential data that is easily overlooked and therefore prone to accidental exposure when files are shared or distributed.

There have been a number of articles in the press on this topic, including high-profile ones such as British Prime Minister Tony Blair's dossier on Iraq and the United Nations document that deleted the names of individuals implicated in the assassination of former Lebanese Prime Minister Rafik Hariri, but failed to remove the tracked changes from the file. However, we're not aware of a focused study conducted against a reasonably large corpus of real-world documents that provides a useful perspective on the extent of metadata and hidden information disclosure.

Methodology

We visited the Web sites of the Fortune 100 in search of Office files. The total number of files found was 8,846. Of this number, we identified 8,038 as Office 97 or later file types. The analysis and results reported in this study are based on the processing of these 8,038 identifiable files (18 of which returned error conditions, attributable to the files being corrupted in some manner). Word documents were most common, representing 58.1 percent of the total identifiable files, followed by PowerPoint at 26.6 percent and Excel at 15.3 percent. Twenty-four of the Fortune 100 sites did not contain any Office files that were discovered by our process.

Although the files we analyzed are from publicly available Web sites, we're not suggesting the metadata and hidden information problem is central to files published to the Web. In fact, only a small percentage of documents created within organizations are published to corporate Web sites. The majority of documents that represent vectors for information exposure are shared with business partners, customers, prospects, the press, regulatory bodies, and others as e-mail attachments or through some other form of document distribution other than Web publishing.

Terminology

Throughout this paper we use the terms *file* and *document* interchangeably. The use of the term *document* is not meant to indicate the exclusive description of a word processing file, but also applies to spreadsheets and presentations.

We describe the removal of sensitive information in a number of ways, including "remove," "delete," and "scrub." These terms are used interchangeably to describe a process that results in "clean content": a file free of metadata and hidden information deemed to be sensitive.

STUDY FORMAT

The study is laid out in the following format for each of the 29 categories of metadata and hidden information (“target elements”) we examined:

- **Target Element Name**
- **Description**—Description of the target element
- **Risk**—Description of the risks associated with the target element
- **Study Findings**—Analysis of the results, including examples, if appropriate
- **Recommendation**—Recommendations for limiting risks associated with the target element

Analysis

We used Oracle’s Outside In Clean Content, discussed later, to inspect the files.

The Analysis Results table, which follows, shows the occurrence of specific target elements as a percentage of the total file collection as well as the actual file count for such occurrences. Outside In Clean Content is able to generate a detailed report for each file analyzed, providing information such as the number of occurrences of specific target elements, the value of such data (such as the actual text from track changes or the e-mail address of a user from Outlook properties). Although we do not expose any of the specific values or specific text that identifies an individual or organization, we do provide examples of our findings, based on actual metadata and hidden information that was discovered.

Analysis Results

Target Element	Rate (%)	Files Affected
1. Audio and Video Paths	0.4	36
2. Author History	46.4	3,733
Contains paths	36.7	2,950
Contains network share names	14.4	1,158
3. Comments	2.1	165
4. Content Properties	99.8	8,020
5. Custom Properties	5.5	446
6. Database Queries	0.0	1
7. Embedded Objects	24.8	1,994
8. Encryption	1.1	92

9. Fast Save Data	10.1	813
10. Hidden Cells	3.9	315
11. Hidden Slides	1.9	151
12. Hidden Text	0.9	76
13. Linked Objects	0.1	11
14. Macros and Code	5.1	409
15. Office GUID Property	17.2	1,386
16. Outlook Properties	17.1	1,378
17. Presentation Notes	13.6	1,093
18. Printer Information	30.9	2,480
Contains network share names	18.0	1,447
19. Routing Slips	0.0	0
20. Scenario Comments	0.0	0
21. Sensitive Hyperlinks	0.4	29
22. Sensitive Include Fields	0.3	22
23. Statistic Properties	99.9	8,028
24. Summary Properties	99.3	7,978
25. Template Name	7.4	592
26. Tracked Changes	6.5	521
27. Usernames	98.9	7,950
28. Versions	0.0	4
29. Weak Protections	3.5	278

STUDY FINDINGS

Audio and Video Paths

Description

Microsoft PowerPoint supports linking to audio and video files using the **Insert > Movies and Sounds > Movie from File** and **Insert > Movies and Sounds > Sound from File** command. Using this feature results in storing a potentially sensitive link to a local or network file path.

The storage of an external local or network file path caused by linking to audio and video files exposes an organization to the risk that sensitive information could be contained in the directory hierarchy exposed by the path and the risk that the path information can provide a view into the corporate network topology.

Risk

The storage of an external local or network file path caused by linking to audio and video files exposes an organization to multiple risks. The first risk is that sensitive information could be contained in the directory hierarchy exposed by the path. For example, the directory structure might use a taxonomy that includes information such as a client's name or identifier. The second risk is that the path information can provide a view into the corporate network topology. This opens an organization to a network intrusion risk. Although this risk is mitigated by proper network security, it remains a social engineering threat by providing confidential information to hackers attempting to infiltrate a corporate network. The social engineering risk is elevated when path information is combined with other sensitive data, such as valid usernames, e-mail addresses, and e-mail subject lines. *Applies to PowerPoint 97 and later.*

Study Findings

A very low number of occurrences (36 files) were found in our document set, with only a handful of files that contained network path rather than local path information. However, in two instances, files that included a network path also contained Outlook property information, which includes a user's e-mail display name and e-mail address. The combination of valid corporate infrastructure information and valid employee information represents a security concern.

Recommendation

To protect against the wide distribution of path information, we recommend removing audio and video path elements from presentations that are distributed outside the organization.

Author History

Description

The names of the last 10 authors who saved a document are stored in an area of the document that is inaccessible using Microsoft Word. In Word 97 and Word 2000, this information also contains the paths where the document was saved and can include sensitive user logon or network share information.

Risk

Saving the author history within Word documents poses several risks, including exposure of personal information, local or network paths, and an audit trail of previous revisions. Personal information will typically include the usernames associated with the last 10 revisions of the document. Local or network paths will identify where each revision was saved, opening the risks associated with exposing file paths. The combination of usernames and file paths provides an audit trail of previous revisions that might not be desirable. The risk associated with exposing this information often depends on the type of document and the potential

The saving of the author history within Word documents poses several risks, including exposure of personal information, local or network paths, and an audit trail of previous revisions.

reviewers of the document. For example, documents that could be targets of legal discovery and documents that could be published to the Web pose a higher risk than other documents. *Applies to Word 97 and later.*

Study Findings

Eighty percent of the Word documents analyzed (3,733 documents) contained author history data, represented by usernames. In many instances, it appears organizations have configured Word to hide employee names or usernames from the summary information found in the Properties dialog box, often with a default use of a company name or “Licensed User” as the value in these fields. This attempt to prevent exposure of user identities is defeated by author history information. Users are unable to review author history information through the Word application. For organizations that have policies or a desire to keep usernames and identities protected, author history is a serious threat to such intentions.

Recommendation

The combination of usernames in author history and valid path and share name information is a policy and security concern. Except for unusual circumstances, documents intended for external distribution or publication should not expose author histories and path or share name information, which provide third parties with visibility of organizational infrastructure. We recommend removing author history information from documents, with a very strong recommendation for removal of this data from documents that will be distributed beyond organizational boundaries.

Comments

Description

Microsoft Office supports adding user comments to a document through the **Insert > Comment** command. Comments often contain private or sensitive information.

Risk

Document comments can be used to expand upon or clarify visible content and pose low risk when used in this manner. However, comments are also often used for internal commentary and collaboration. In this form, they can expose sensitive discussions, and if released, can represent a leak of information that was not intended. The severity of the threat is highly dependent on the content of the comments. *Applies to Word, Excel, and PowerPoint 97 and later.*

Comments are often used for internal commentary and collaboration. In this form, they can expose sensitive discussions, and if released, can represent a leak of information that was not intended.

Study Findings

Based on the low rate of occurrence (2.1 percent, or 165 files), it appears this group of users has done a relatively good job ensuring comments are removed before publishing documents to the Web. Our review, however, reveals that in the majority of instances, comments were intended for internal review purposes and overlooked when the file was published or distributed. A few noteworthy examples that range from embarrassing to potentially harmful:

- A computer manufacturer's white paper includes a comment that exposes the shortcomings of a partner's software solution (typos are from the original document): "This is a little harder for me offer support but is based on several things. After talking to XXXX they admit that YYYYYY does not scale very well (or not at all) from 4-way to 8-way servers because MS Terminal Services (NT, and 2000) does not scale."
- A software maker's presentation includes the following reviewer's comments: "These speaker notes suck ..."
- An equipment manufacturer's presentation includes the following reviewer's comments (typos from original document), showing surprise at figures: "whow 450 manufacturers is this a real number ,, incredible if its true—in fact all these numbers are large expect the XXX stuff which is believable. If they are right then fine ..."
- A software-maker's presentation includes comments that suggest the author revise the file to include "legally accepted" names and includes the network location of the acceptable list of fictitious names.

Except for a few cases where comments are used as instructions within forms, it's clear that the comments reviewed for this study were intended for internal review and discussion rather than third-party consumption. Additionally, in most instances, the commenting author's name is included with the comment. This provides visibility to internal reviewers and personnel that organizations might desire to remain transparent to third parties.

Recommendation

All documents should be analyzed for the existence of comments before they're distributed, allowing authors or administrators the opportunity to review comments and determine if they're appropriate for consumption by the receiving party. Documents that are distributed in "final" form for publishing or broad distribution to multiple parties or the public should have comments removed, unless they're specifically authored for the receiving party.

Content Properties

Description

Content properties are established using the **File > Properties > Contents** command. They are document properties that provide a view into some of the content within the document. These properties include titles and headings in Microsoft Word documents; sheet names and named ranges in Excel documents; and fonts used, design templates, and slide titles in PowerPoint documents.

Content properties generally represent little or no risk because they primarily mirror some visible content from the document. An exception to this rule occurs when an Office document is encrypted, but the content properties remain accessible.

Risk

Content properties generally represent little or no risk because they primarily mirror some visible content from the document. An exception to this rule occurs when an Office document is encrypted, but the content properties remain accessible. This hole in the Office encryption feature has been closed in recent

versions. However, patching the application will not address existing documents unless they are loaded and resaved by the updated application. *Applies to Word, Excel, and PowerPoint 97 and later.*

Study Findings

Most of the documents in the study (99.8 percent) contained content properties, indicating that almost all authors or reviewers accept the default values populated by the application. Unless the actual values of these elements are sensitive, such as a proprietary or confidential title or heading of a document, there's typically a low risk factor to this information.

Recommendation

Organizations need to determine whether this information poses a threat to internal policies and practices, and should perform analysis and remediation as appropriate.

Custom Properties

Description

Custom document properties can be created using the **File > Properties > Custom** command. They can include user-defined properties or application-generated properties. Custom properties include Checked by, Client, Date completed, Department, Destination, Disposition, Division, Document number, Editor, Forward to, Group, Owner, Project, Purpose, Received from, Status, and Telephone number.

Risk

The risk associated with custom properties varies according to their use. Custom properties are often used by software applications to associate metadata with a

document. For example, content management systems could use custom properties to assist document categorization and facilitate tracking the document lifecycle. Custom properties are also used by individual users to assist in categorization or carry additional information about the document. Depending on the implementation, this information could range from innocuous to highly sensitive. *Applies to Word, Excel, and PowerPoint 97 and later.*

Custom properties can provide insight into content management systems and internal classification of documents, because many management systems populate custom property fields with metadata used by the system.

Study Findings

As a percentage of documents affected, 5.5 percent is statistically low, representing just 446 files. Custom properties, however, can provide insight into content management systems and internal classification of documents, because many management systems populate custom property fields with metadata used by the system. Independent of Microsoft Outlook properties, which are viewable under the Custom tab in the Properties dialog box, other personally identifiable information can be exposed.

Examples:

- A computer manufacturer makes a case study authored by a software partner available, and the custom property information includes the name of an internal author from the software company who is not otherwise made known in the document and whose name is different from the author name that appears in the Summary properties. Further, additional values for the custom property fields include an “XX Confidential” entry under a security classification field.
- A high-tech manufacturing company’s PowerPoint presentation contains the e-mail address of a person from a university—neither of which (the person or the university) are provided with attribution in the presentation and are different from other authors and organizations identified in other metadata fields.

Recommendation

We recommend removing custom property information from documents that are published to the Web or have the potential of being widely distributed. Organizations concerned with exposing unnecessary information to third parties should remove this information.

Database Queries

Description

Microsoft Office supports powerful connectivity to databases that results in database connection and query information being stored in Office documents. This information can include a path or URL to a database server, the database username, database password, and SQL query strings—all of which can be highly sensitive information.

The use of database queries to import data into Excel creates the potential that unauthorized users will be able to independently query a sensitive database at will.

Risk

The use of database queries to import data into Excel is a powerful feature that comes with serious security risks. Specifically, this feature creates the potential that unauthorized users will be able to independently query a sensitive database at will. To allow the query to be updated, whether user initiated or automatic, the document retains the database query parameters. This information can include a file path or URL reference to the database server, SQL query strings that identify the requested data, and the password required to access the database.

A file path to the database server opens all the security threats associated with exposing file paths. SQL query strings can be used to deduce the structure of the database. Storing the database password in the Office document is an option the user can choose when creating the query. This option is often activated to avoid having to re-enter the password each time the data is updated. This information opens an organization to SQL injection attacks.

Proper network security could prevent any external access to the database server, but this provides little peace of mind in the event of a network security breach. Internal access, however, could represent an even-greater threat because the recipients of the sensitive information are likely behind the firewall but possibly prohibited from accessing the database.

Consider an example where the finance department distributes a spreadsheet that at face value simply includes a list of employees by department, but buried within the underlying query lies all the information required to access an employee database filled with confidential data. Extreme caution should be used when releasing spreadsheets that contain database queries. *Applies to Word and Excel 97 and later.*

Study Findings

Only one document in the survey contained a database connection—in this case, a computer manufacturer’s parts spreadsheet that pulled data from a price list that appeared to reside on a local, rather than network, drive.

Recommendation

Files that are distributed across group boundaries should be analyzed for database queries, giving authors or administrators the chance to determine the appropriate action. We recommend removing this information from any files published to the Web or distributed outside the organization.

Embedded Objects

Description

The Microsoft Office embedded object feature (**Insert > Object...**) allows an object to be created in a document (the “source file”) and then inserted into another document (the “destination file”), with a different application. The

resulting object data can then contain any of the hidden and sensitive data issues found in the source or native, application.

Any security threat that has been identified in documents created by an application can also manifest itself when that application serves an embedding. An additional security concern has been found to exist when using embeddings within documents that have been encrypted using Office security options. Surprisingly, embedded objects are not encrypted with the destination file.

Risk

Office applications leverage embeddings to seamlessly work with each other as well as other applications to create compound documents; for example, including a spreadsheet table in a Word document or a chart in a presentation is both common and useful. For any application to allow an embedding to be edited in its native application, the destination file must include a complete copy of the application data associated with the object. This data is in addition to the graphic rendition of the object that is used for display and printing. It is in this data that security risks can be found.

Any security threat that has been identified in documents created by an application can also manifest itself when content from one application is embedded into content from another (i.e. a native Excel chart presented in a Word document). An additional security concern has been found to exist when using embeddings within documents that have been encrypted using Office security options.

Surprisingly, embedded objects are not encrypted with the destination file. For example, if an Excel chart is added to an encrypted Word document, the chart and the entire supporting spreadsheet will be left unencrypted within the Word document. Scrubbing embeddings will remove the ability to make further edits to the embedding while maintaining the most recent graphic rendition of the object. *Applies to Word, Excel, and PowerPoint 97 and later.*

Study Findings

Almost a quarter of the files analyzed contained embedded objects (1,994 documents). With most embeddings, and in particular, Word, Excel, and PowerPoint data, the metadata and hidden information risks are potentially doubled, because the full underlying file can contain the same type of sensitive information as the host file.

Recommendation

Except for cases where recipients of documents will have a definitive need to modify or edit the embedded object, we recommend removing such information (while leaving the graphic rendition of the object's values intact) in all files that are distributed beyond organizational boundaries. For certain organizations, it can be desirable to perform the same removal process for interdepartmental distribution of files.

Fast Save Data

When fast save is activated, deleted text and data can remain in the file even though it is no longer visible or accessible from within the application.

The fast save feature of Word and PowerPoint attaches changes to the end of the existing document rather than completely rewriting the modified document. Unfortunately, this leaves deleted text and data in the document long after removal by the user. This creates the risk of exposing the previous state of a document to recipients. A second risk is that the fast save feature can be used to transfer confidential information through documents in a way that will circumvent most content filtering technologies.

Description

The fast save feature in Microsoft Word and PowerPoint is set using the **Tools > Options > Save > Allow fast saves** command. When fast save is activated, deleted text and data can remain in the file even though it is no longer visible or accessible from within the application.

The fast save feature is enabled by default in Word 97, enabled by default in Word 2000 if it was upgraded from Word 97, and disabled by default in Word 2000 and later. The user can manually enable fast save in all versions of Word.

The fast save feature is enabled by default in all versions of PowerPoint and results in many versions of modified slides remaining in the file.

Risk

The fast save feature of Word and PowerPoint is designed to decrease the time required to save a document to disk. This is accomplished by attaching changes to the end of the existing document rather than completely rewriting the modified document. Unfortunately, this results in leaving deleted text and data in the document long after it was apparently removed by the user, thus, creating the risk of exposing the previous state of a document to recipients. A second risk is that this feature of Office can be used to transfer confidential information through documents in a way that will circumvent most content filtering technologies.

The occurrence of this feature in Word documents is low because the fast save option was turned off by default with the release of Office 2000, though upgrading Office in place can maintain the state of this option. This risk remains a threat in existing, pre-Office 2000 Word documents.

This feature is still turned off by default as of the current release of PowerPoint. As a result, it is common for PowerPoint documents to include multiple prior versions. This is particularly troublesome when considering the frequency with which existing presentations are modified for a slightly different audience. Imagine the risk of distributing a sales presentation to one prospect that was given earlier to another prospect, knowing that the prior version is buried somewhere in the file.

Applies to Word and PowerPoint 97 and later.

Study Findings

Just more than 10 percent of the files (813 documents) contain fast save data. Of these affected files, 31 were Word documents and the rest were PowerPoint documents (representing 33 percent of all PowerPoint files analyzed). We did not attempt to review every bit of fast save data, nor did we perform a detailed comparison of the fast save information to the current version of information in the documents. The noteworthy finding here is the large numbers of files—one out of every three PowerPoint presentations—that contain information authors aren't aware of and that they can't review.

Recommendation

We highly recommend removing fast save data from all files that are shared with third parties or the public. Authors and administrators can't review fast save information using Word or PowerPoint, and the opportunity for accidental disclosure of information is great.

Hidden Slides

Hidden slides are often used to tailor a presentation to a particular audience or to adjust a presentation to meet a required time allotment. In some cases, the hidden slides can contain data not intended for the target audience, creating a risk of leaking sensitive information.

Description

The Microsoft PowerPoint hidden slide feature (**Slide Show > Hide Slide**) allows individual slides to be hidden during the slide show and printing of the presentation. Hidden slides can contain information that is not intended for general release.

Risk

Hidden slides are often used to tailor a presentation to a particular audience or to adjust a presentation to meet a required time allotment. In many cases, exposing the hidden slides does not represent any type of privacy or security concern. In some cases, however, the hidden slides can contain data not intended for the target audience, creating the risk of leaking sensitive information. *Applies to PowerPoint 97 and later.*

Study Findings

As a percentage, the incidence of hidden slides in the survey sample is low, at 1.9 percent (representing 151 PowerPoint presentations or approximately 6.5 percent of all presentations analyzed). The question remains whether any of the hidden slides that have been made publicly available contain information that wasn't intended for such broad access.

Recommendation

Any presentation that contains hidden slides should be reviewed before distribution to determine whether the hidden slide(s) should be removed.

Hidden Text

Description

Text that has been intentionally hidden (**Format > Font... > Font > Hidden**) by the user can contain sensitive information that should be reviewed or removed before distributing the document.

Risk

The use of hidden text exposes the author to unintended information disclosure. Hidden text can be used for internal commentary, temporary display, to suppress printing of certain text, or as a method of deleting text so that it can be later

retrieved if desired. It is less common to find hidden text that provides intended useful content because this is usually done with comments. Releasing documents that contain hidden text to third parties is considered a high security risk when not first reviewed by the author. *Applies to Word 97 and later.*

Study Findings

Less than 1 percent of the documents analyzed contained hidden text, totaling just 76 documents. Our review of a few document samples did not reveal information that we'd consider sensitive or proprietary.

Recommendation

Analysis for hidden text should be performed as a default precaution before documents are made available for third-party consumption or distribution. Organizations should determine the appropriate action when this target element is found, but alerting the author of the existence of such hidden information is a reasonable baseline action.

Linked Objects

Description

The Microsoft Office linked object feature (**Insert > Object...**) allows linking to an external file, or source file, that is managed and rendered by another application. These links can expose local and network path information.

Risk

Office applications enable the primary document, or destination file, to include references to the source file (external document) that are then rendered directly into the destination file. This feature stores a file path or URL to the source file within the destination file. This is done to allow automatic updates to the destination file that incorporate changes to the linked document and to allow direct authoring of the source file within the destination file framework. The existence of path information that supports this feature opens an organization to network intrusion and social engineering risks. Removing the link information can be done without affecting the most recent rendering of the linked object. *Applies to Word, Excel, and PowerPoint 97 and later.*

Office applications enable the primary document to include references to the source file that are then rendered directly into the destination file. This feature stores a file path or URL to the source file within the destination file. The existence of path information that supports this feature opens an organization to network intrusion and social engineering risks.

Study Findings

Only 11 documents contained linked objects, with 10 of them belonging to the same organization.

Recommendation

Organizations should determine their own policy regarding this target element. But for files that are published to the Web, or are otherwise made available for informational purposes (rather than requiring modification or additional editing),

removing linked objects will preserve the visible data represented by the object without disclosing potentially sensitive path or URL details.

Macros and Code

Description

Microsoft Office includes support for Microsoft Visual Basic and can be used to create everything from simple macros to data entry forms to complete applications. Visual Basic can also be used to create macro viruses that travel with documents.

Risk

The risk associated with macros and code being present within inbound documents is a well-known virus threat. The risk associated with outbound documents includes the unintended redistribution of viruses and the potential disclosure of sensitive information contained within an otherwise valid macro. Information disclosure can come in the form of usernames, code comments, and potentially confidential approaches to programmatically accessing corporate resources. Macros and code are often used to support the document creation process but are not intended or desired in the final version of the document. *Applies to Word, Excel, and PowerPoint 97 and later.*

The risk associated with macros and code being present within inbound documents is a well-known virus threat. The risk associated with outbound documents includes the unintended redistribution of viruses and the potential disclosure of sensitive information contained within an otherwise valid macro.

Study Findings

A reasonable percentage of documents (5.9 percent), totaling 409 files, contained macros or code.

Recommendation

Determining the risk associated with releasing documents that contain macros and code typically requires a user review. When macros or code is essential to the function of the document, such code must obviously be left intact. When documents are to be used in a read-only scenario, macros should be removed.

Office GUID Property

Description

The Microsoft Office GUID property is a document property created by versions of Office before the release of Office 2000. This globally unique identifier (GUID) can be used to identify the computer from which the document originated.

Risk

Documents containing the Office GUID property threaten an organization or individual with the risk of losing anonymity. The Office GUID property can be used to uniquely identify the machine on which a document originated. It can also be used to determine if multiple documents originated on the same machine. This property is no longer stored in Office documents as of the release of Office 2000

and is consequently now considered a low risk element. Archived documents and documents created with older versions of Office are still at risk of this disclosure. *Applies to Word, Excel, and PowerPoint 97 and later.*

Study Findings

Of the files analyzed, 17.2 percent (1,386) contained the GUID property. Although we don't place a high risk on this data element, there are few scenarios where it benefits an individual or organization to share files with this identifier.

Recommendation

Unless an organization has specific use cases for the GUID property, removing this identifier from documents is considered a reasonable precaution for individuals or organizations that desire anonymity.

Outlook Properties

Description

The Microsoft Outlook properties are custom document properties that can be added by Outlook to Office documents when they are sent as attachments. These properties include the sender, e-mail address, subject of the e-mail, and routing identifiers associated with the attachment.

Risk

The Outlook practice of adding e-mail metadata properties to Office attachments can result in unintended and sensitive information disclosure. The property metadata can include the sender's e-mail address, e-mail display name, routing identifiers, and the subject line of the e-mail to which the document was attached. Disclosing this information to the recipient of the e-mail does not represent a direct threat because the recipient receives most of this information from the e-mail headers by default. However, inserting this information into the attached documents without any user intervention or awareness allows this information to continue to travel with the document well beyond the initial e-mail recipient. If the document is subsequently published to the Web, it will publicly expose a valid e-mail address, the associated user display name, and a valid related e-mail subject line. The dangers of this release of information can range from simple embarrassment to confidential leaks and, at minimum, present spammers with additional opportunities. *Applies to Word, Excel, and PowerPoint 97 and later.*

Study Findings

A significant number of files in the study (1,378), representing 17.1 percent of the total, contained Outlook properties. This should be an area of real concern for any organization that does not want personally identifiable employee information exposed.

The Outlook practice of adding e-mail metadata properties to Office attachments can result in unintended and sensitive information disclosure. The property metadata can include the sender's e-mail address, e-mail display name, routing identifiers, and the subject line of the e-mail to which the document was attached.

There are many examples where an organization has configured Office to show a default username and company name in the Summary field of the Properties dialog box, or a universal identifier such as “Licensed User,” or a numeric or alphanumeric code rather than an employee name. If the purpose of this configuration is to hide personally identifiable employee information, it is defeated by Outlook Properties.

In addition to exposing employee identification and contact information, we found multiple instances of network path or share information in the same file. Thus, hackers or social engineers now have a glimpse of an organization’s infrastructure as well as valid employee e-mail addresses, e-mail display names, and even the subject line of the e-mail message that contained the Office file attachment.

Recommendation

We highly recommend removing Outlook properties from all documents. Except for instances where employees are sharing files via e-mail with partners, customers, or other third parties, where there’s no chance for further distribution of the file, exposing this type of employee information to competitors, recruiters, spammers, hackers, and social engineers is a risk that should be avoided.

Presentation Notes

Description

The Microsoft PowerPoint notes feature allows notes to be associated with each slide. Notes can contain general content or internal commentary that should be reviewed or removed before distributing a presentation.

Risk

Presentation notes, also referred to as speaker notes, are commonly used to document specific points the speaker would like to make during the presentation. In most cases, these notes represent useful content that can be safely shared with any recipient of the presentation. Often, however, these notes are written in a style that is targeted at the speaker alone and are not to be directly shared with the audience. In other cases, the notes are used to facilitate collaboration between multiple authors or reviewers working on the presentation. Distributing or publishing a presentation that includes speaker notes carries the risk of disclosing unintended or even confidential information. *Applies to PowerPoint 97 and later.*

Study Findings

Presentation notes were found in 1,093 files, representing nearly 48 percent of all presentations and 13.6 percent of all files. We did not attempt to review the full content of all these notes, but the majority of our sampling indicates that most notes are built for the benefit of the speaker as they address an audience, and often contain positioning statements or other comments that are not intended for direct consumption by the audience. Although we did not find any overtly sensitive

In most cases, presentation notes represent useful content that can be safely shared with any recipient of the presentation document. Often, however, these notes are written in a style that is targeted at the speaker alone and are not meant to be directly shared with the audience.

information, authors should consider whether these types of notes should be distributed as part of a presentation that will be fully discoverable by recipients.

Recommendation

Presentations should be inspected for presentation notes prior to distribution, allowing authors or administrators to determine whether such notes are appropriate for third-party exposure.

Printer Information

Description

Printer setup information is often stored within a Microsoft Word or Excel document. For network printers, this information can include potentially sensitive network information and less-sensitive printer model names.

Risk

The release of documents that include printer setup information carries the risk of disclosing sensitive file path information. This information can also include the model of the printer in the form of a text name. The model name represents little or no concern to most users, though it can be used in digital forensics to narrow down the origin of a document. Printer location information is stored in the form of a file path. This carries the typical risks associated with file path exposure, including network intrusion and social engineering concerns. *Applies to Word and Excel 97 and later.*

The release of documents that include printer setup information carries the risk of disclosing sensitive file path information.

Study Findings

Printer information is present in 30.9 percent of the files examined (2,480 files). More than half of these files also contained network share names (1,447 files), exposing potentially sensitive infrastructure information to unknown parties.

Recommendation

The primary risk centers on the exposure of network share names. We recommend removing printer information from files that are distributed outside the organization.

Routing Slips

Description

The e-mail routing feature of Microsoft Office (**File > Send To > Routing Recipient**) stores the e-mail addresses and usernames of recipients in the document.

E-mail routing slips are introduced into documents that enable the document routing feature. Each routing slip can contain the e-mail display name and e-mail address of the originator and all recipients of the routed document. This information will remain in the document after it has been routed and can expose an organization to the release of sensitive information.

Risk

E-mail routing slips are introduced into documents that enable the document routing feature. Each routing slip can contain the e-mail display name and e-mail address of the originator, or sender, and all recipients of the routed document. The routing slip can also contain the subject line, message body, and the date and time stamp of the routing e-mail. This information will remain in the document after it has been routed and can expose an organization to the release of sensitive information. This exposure can be of particular concern with documents that are a target of legal discovery and documents that are made available to the public via electronic distribution or publication. *Applies to Word and Excel 97 and later.*

Study Findings

No documents were found that contained routing slips.

Recommendation

We recommend removing routing slip information from documents that are published or distributed for consumption outside the organization.

Scenario Comments

Description

Microsoft Excel supports associating user comments with the scenario feature (**Tools > Scenario... > Comment**). Scenario comments can include sensitive information and often include hidden author information in addition to the comment.

Risk

The use of scenario comments, similar to document comments, carries the risk of unintended information disclosure. The comments will often include a username and date and time stamp in addition to the comment text. The scenario feature provides a powerful mechanism to quickly analyze multiple models within a spreadsheet. Scenario comments are considered a low risk in terms of unintended information disclosure but do carry some risk because they will not be obvious to the author when reviewing the visible content. *Applies to Excel 97 and above.*

Study Findings

No documents were found that contained scenario comments.

Recommendation

We recommend that spreadsheets distributed outside an organization be analyzed for scenario comments, with an alert to authors or administrators suggesting a review of the data.

Sensitive Hyperlinks

Sensitive hyperlinks are hyperlinks to a resource located on a local or network drive. Releasing network resource names can subject an organization to network security risks through direct intrusion attempts and through social engineering attacks.

Description

The Microsoft Office hyperlink feature (**Insert > Hyperlink**) allows the creation of links to various locations. Two of the possibilities, fully qualified local paths and network paths, can provide unwanted insight into an organization's internal structure. Web links are not treated as sensitive.

Risk

Sensitive hyperlinks are hyperlinks to a resource located on a local or network drive. As such, they carry the risks associated with exposing path information. This includes the release of confidential network topology information and sensitive directory naming conventions. Releasing network resource names can subject an organization to network security risks through direct intrusion attempts and through social engineering attacks. *Applies to Word and Excel 97 and later.*

Study Findings

This target element is nearly nonexistent in the files analyzed, with just 29 files affected. Though the number of incidents in this sample is low, valid path information is exposed to third parties.

Recommendation

Files should be analyzed for sensitive hyperlinks and require review before being made available to third parties.

Sensitive Include Fields

Description

The Microsoft Word include field feature provides non-OLE-based linking to external files (**Insert > Field > IncludeText and Insert > Field > IncludePicture**). These fields can contain fully qualified local paths or network paths.

Risk

Sensitive include fields carry the risk of exposing sensitive local and network file paths, which can provide insight into an organization's internal network structure. The release of path information carries the risks of network intrusion, sensitive information exposure, and social engineering threats. *Applies to Word 97 and later.*

Study Findings

Twenty-two files were found to contain sensitive include fields, indicating the preferred method of inserting external objects and data inside Word documents is via OLE objects.

Recommendation

We recommend that documents that are likely to be distributed to third parties have these data elements scrubbed to avoid exposure of path information.

Statistic Properties

Description

Statistic properties (**File > Properties > Statistics**) are document properties that include Created, Modified, Accessed, Printed, Last saved by, Revision number, Total editing time, Pages, Paragraphs, Hidden Slides, Multimedia clips, and Presentation format. Additional application-maintained properties in this category include Application name, Hyperlinks changed flag, Links up to date flag, and Scale flag. Some or all of these properties should be reviewed or removed before document distribution.

Risk

Statistic properties are document properties that track editing details about the document, such as the amount of time spent editing the document; the number of paragraphs and pages in the document; and when the document was created, last modified, or accessed. Releasing most of this information with the document raises little or no security concerns but is made available for review due to its nature as metadata.

The various date and time stamp statistics might expose a level of undesirable tracking information in extremely security-conscious environments, or in environments where such information can be correlated to time and billing, or raise concern about a document's creation and revision dates. Consider the scenario whereby an author is contracted to produce a document for a client, and the client discovers that the ensuing document was actually created before the parties' relationship. *Applies to Word, Excel, and PowerPoint 97 and later.*

Study Findings

Statistically (99.9 percent), all the inspected documents contained some form of these properties.

Recommendation

Organizations with unique security concerns or information-exposure policies should remove this information from files that are distributed to third parties or made available for public review.

Statistic properties track editing details about the document. Occasionally, releasing metadata like the date and time stamp or time spent editing a document can create concerns if it is correlated to time and billing.

Summary Properties

Description

Summary properties (**File > Properties > Summary**) are document properties that include Title, Subject, Author, Manager, Company, Category, Keywords, Comment, Hyperlink Base, Template, and Preview Picture. Some or all of these properties should be reviewed or removed before document distribution.

Risk

Summary properties include a collection of metadata that summarizes the document, along with attributes of the author or environment of the document. This data is considered a low risk security element for most users. However, one should consider whether properties such as Author, Category, Keywords, and Comment need to be exposed when releasing a document to wider distribution.

A second risk is that encrypted Microsoft Office documents created before version 2003 have unencrypted document properties, partially exposing some information about a document believed to be password protected. A third risk is the potential for incorrect or outdated information to remain in these summary fields when documents are repurposed or continually revised from an original template. In this scenario, it's common to find document titles and author names that are not current. *Applies to Word, Excel, and PowerPoint 97 and later.*

Summary properties include a collection of metadata that summarizes the document, along with attributes of the author or environment of the document. Encrypted Microsoft Office documents created before version 2003 have unencrypted document properties, partially exposing some information about a document believed to be password-protected. Another risk is the potential for incorrect or outdated information to remain in these summary fields when documents are repurposed or continually revised from an original template.

Study Findings

Like statistic properties, summary properties appeared in nearly every document in the study (7,978 files) due to the default behavior of the Office applications. The most common data found in these metadata fields: document title, author name, template name (for Word files), and the company name. Response to this type of information being exposed will depend on the policies of individual organizations, and companies in certain markets can have a more-critical need to suppress personally identifiable information or company-identifiable information.

Recommendation

For documents intended for internal organizational distribution, it's often useful to be able to identify the author or use other summary information to ease collaboration and knowledge sharing. For documents intended for wider distribution, including multiparty or public access, we recommend removing personally identifiable information unless an organization specifically sees value in releasing such detail.

Template Name

Description

If a template other than Normal.dot is used, the document will contain a full path to the template file. This can expose local path or network share information.

Risk

Use of templates other than Normal.dot will result in exposure of a fully qualified local or network path to the template. This element can carry all the risks associated with exposing file paths, including network intrusion and social engineering attacks, as well as revealing confidential naming conventions. *Applies to Word 97 and later.*

Study Findings

Just more than 11 percent of the Microsoft Word documents analyzed (7.4 percent of the total file count, or 592 files) contained template names. A sampling of these files did not reveal naming conventions that we deemed sensitive or noteworthy, and most paths were local, similar to this example: C:\Program Files\Microsoft Office\Templates\

Recommendation

Due to the potential for exposing paths or sensitive naming conventions, and the lack of benefits provided by exposing this information, we recommend removing this element from documents that are likely to be shared outside the organization.

Tracked Changes

Description

The track changes feature of Microsoft Office tracks insertions, deletions, and formatting changes made to the document. Such changes contain deleted text and author and date information that can be unintentionally left in the document upon distribution.

Risk

Tracking changes in documents is a powerful feature that enhances the collaboration process by providing valuable change history. It can be useful for individual authoring and indispensable when multiple authors and reviewers are involved. But a very high information disclosure risk comes with this power. Documents often reach points in their lifecycle where tracked changes should either be accepted or rejected, and a clean version of the document should be saved. This is required when it is no longer desirable to share the history of deletions and additions with the next group of recipients of the document. Many organizations have experienced the fallout associated with releasing a document with track changes still enabled. The results can range from embarrassing to adversely affecting business, and depending on the sensitivity of the content, can even be used to support evidence discovery for litigation. *Applies to Word and Excel 97 and later.*

Study Findings

We were surprised to find that 6.5 percent of the files analyzed (521 documents) contained track changes considering their availability on public-facing Web sites.

Many organizations have experienced the fallout of releasing a document with change tracking still enabled. The results can range from embarrassing to adversely affecting business, and depending on the sensitivity of the content, can even be used to support evidence discovery for litigation.

We did not review all the thousands of deletions and insertions to determine if there were damaging instances of information exposure, such as deleted information that exposes other companies or persons, modifications to business terms, or contractual concessions. We did find examples of track change deletions and insertions that span a broad range of file types, including financial reports, agreements, press releases, and white papers.

Recommendation

Sharing documents with track changes is a critical requirement for many organizations that depend on this feature to negotiate agreements, review documents, and enable multiparty collaboration. We do recommend, however, that files are inspected for the existence of track changes, and authors or administrators are given control to approve the release of such files or given the chance to accept, reject, or simply scrub track change information. For many organizations, we recommend implementing policies that allow certain users or groups of users, such as a legal department, to distribute files that contain track changes, while other users or groups of users, such as sales administration, be allowed to distribute such files only with management approval. Additionally, we recommend a default process whereby documents published to the Web or other wide distribution points, which should be in a final form, have tracked changes removed automatically.

Username

Description

A number of features in Microsoft Office cause usernames to be saved in the document, including the document properties Author and Last saved by, document routing recipients, Word comment and tracked change authors, Excel scenario authors, file-sharing participants, and the last user to edit an Excel document or view a PowerPoint document.

Risk

The existence of usernames in documents represents a potential privacy breach and can also create an unintended audit trail of authors. Usernames can be carried with comments, tracked changes, e-mail routing information, document properties, and author history, to name a few. Keeping track of the users involved in the document creation process provides useful information and is often not considered an information disclosure risk. However, usernames are a form of personal information and there are many scenarios where releasing that information is not desirable.

When a document is going to be shared with a larger audience, such as published to the Web, it's worth considering whether usernames represent an undesired release of personal information. Even documents that are shared with only a small group through e-mail can unexpectedly disclose the names of users who have accessed the

The existence of usernames in documents represents a potential privacy breach and can also create an unintended audit trail of authors. Usernames can be carried with comments, change tracking, e-mail routing information, document properties, and author history, to name a few.

document at some point in its history. This risk can be classified as very serious for scenarios where there are regulatory mandates that identify the unprotected release of personal information as illegal. *Applies to Word, Excel, and PowerPoint 97 and later.*

Study Findings

Nearly every document in the survey contained usernames (98.9 percent, representing 7,950 files). A caveat that needs to be mentioned is that we did not apply any natural language processing or entity/name identification process to our analysis—thus a username of “Licensed User” is counted as a positive incident in our analysis. Extensive sampling of the test files show, however, the majority of documents reviewed contain proper names of users rather than generic identifiers.

An abnormal, but noteworthy, example of usernames:

- An automaker’s press kit document, authored by a PR/design firm, contains more than 500 valid (first/last name) usernames. It appears these entries found their way into the document as a bug or error that copied contact information from the author’s address book. In any case, the information is in the document and exposed to anyone with access to the file.

Recommendation

For files that will be distributed outside the organization, and which have potential for subsequent distribution, we recommend removing usernames as a precaution against exposing personally identifiable employee information.

Versions

Description

The versioning feature (**File > Versions**) in Microsoft Word allows multiple historical versions of a document to be saved within a single file. Versioning is useful during document creation but potentially sensitive once a document is released.

Risk

The versioning feature of Word carries with it a high risk of unintended information disclosure. This feature allows the author to archive the current state of a document into the file so that it can be extracted at a later time if required. Users who rely upon this feature risk accidentally releasing older versions of the document that are not intended to be viewed by the recipient. The severity of this threat is heavily dependent on the sensitivity of the document content. *Applies to Word 97 and later.*

This versioning feature allows the author to archive the current state of a document into the file so that it can be extracted at a later time if required. Users who rely upon this feature risk accidentally releasing older versions of the document that are not intended to be viewed by the recipient.

Study Findings

With only four files containing version data found in the survey, either organizations are very careful with managing documents that contain versions or it's an indication this is a little-used feature. We tend to believe it's the latter.

Recommendation

Documents that are shared across organizational boundaries or are passed between internal groups where sensitive information can be exposed (such as HR documents distributed to a companywide audience) should be checked for versions, enabling the author or administrator to review the content to determine the appropriate action. Documents that are published to the Web or are designated for broad distribution in a final form should have version information removed.

Weak Protections

Description

Weak protections are features of an application that appear to provide a strong level of protection against specific user actions on the document but, in fact, can be easily removed from the file without access to a password.

The Microsoft Word protection features (**Tools > Options... > Security > Password to modify**) and (**Tools > Protect Document... > Password**) are weak protections because they do not result in encrypting the file and are easily circumvented with minor changes to the underlying file.

The Excel protection features (**Tools > Options... > Security > Password to modify**) and (**Tools > Protection > Protect Sheet... > Password to unprotect sheet**) are weak protections because they do not result in encrypting the file and are easily circumvented with minor changes to the underlying file.

Risk

Weak protections carry the risk of leading the user to believe that controls placed on the document are safely protected when they are not. The weakness lies in the fact that because the document is not encrypted, the protection can be easily disabled by hacking the file to overwrite or clear the protection commands. Because these features do not attempt to modify the viewing of a document, they don't pose any direct information disclosure threats. However, if the protection is removed, the user will have access to more features that could indirectly expose additional information.

An example of this risk occurs when assuming that a spreadsheet that includes sheet protection will effectively prevent recipients from examining hidden cells. Once sheet protection is removed, the user will then be able to unhide the cells and expose potentially sensitive information. Another risk example is the scenario in which an author password protects a document from modification, believing that it is a read-only file. When this password can be easily found as a clear text element in

Weak protections carry the risk of leading the user to believe that controls placed on the document are safely protected when they are not. The weakness lies in the fact that because the document is not encrypted, the protection can be easily disabled by hacking the file to overwrite or clear the protection commands.

the file, third parties can perform modifications that might not be noticed or inspected by the original author. *Applies to Word and Excel 97 and later.*

Study Findings

Statistically, a low percentage of files contained weak protections (3.5 percent, totaling 278 files), but we identified some noteworthy instances:

- A Word document hosted on a materials company site, authored by a partner organization, is password protected against modifications, but the password was stored using weak protections and is thus available in clear text through our analysis report (the password works).
- A telecommunications company's Word document containing ordering and preordering technical details is password protected against modifications, but the password was stored using weak protections and is thus available in clear text through our analysis report (the password works).
- A computer maker hosts a product data sheet (Word file) that is password protected against modifications, but the password was stored using weak protections and is thus available in clear text through our analysis report (the password works).
- A telecommunications company has four exhibits to a billing process document, which is password-protected against modifications, but the password was stored using weak protections and is thus available in clear text through our analysis report (the password works). The same password is used for all four exhibits.

In all, 23 files were found with passwords that disabled protection against modification of the documents. One company had 12 documents that represented various schedules to a base file, all which shared the same password using weak protections.

Recommendation

Files should be analyzed for weak protections, and authors or administrators should be made aware of the risks associated with these protections before distributing files to third parties. Similarly, authors should carefully inspect files that have been reviewed by or shared with third parties to ensure modifications haven't been made.

CONCLUSION

The results of this study clearly indicate that the issue of metadata and hidden information exposure is very real. The occurrence of this information within documents published to the Web for broad third-party consumption by organizations with large IT resources raises the question of how much sensitive information leaks from organizations every day during the course of normal business.

The ultimate question is very simple: Why would any organization expose itself to these risks if there are solutions on the market that address the problem?

Determining Your Information Exposure

What's the scope of information exposure in your organization? As a rudimentary test to determine your current level of exposure, take the volume of documents your employees share with third parties or the public and apply the same statistics that have been identified in this study—and you'll start to recognize the risk potential within your organization.

Granted, there are scenarios where some of the target elements described in this study are reasonable to share with known parties—perhaps even a requirement of your business process (such as exchanging legal documents with track changes during contract negotiations with a partner). However, there are more scenarios where it's hard to justify when it ever makes sense to expose certain types of information (such as fast save data, network path, or share information).

Each organization has its own security and information management policies. At a very high level, your organization must ask itself a number of questions to determine if metadata and hidden information disclosure are relevant to such policies.

- Is it acceptable to distribute documents containing employee identifiable information described in this study?
- Is it acceptable for documents that contain comments and tracked changes to be distributed without a review process or automated inspection?
- Is it possible that documents that contain hidden text, deleted text, obsolete text, and fast save data represent information that should not be exposed to third parties?
- Do your security policies prohibit publishing IT infrastructure information, such as server names, paths, and database names?
- Are your employees aware of the shortcomings of the document-protection features (“weak protections”) in the Microsoft Office applications?
- What processes are in place to inspect proposals, legal agreements, presentations, white papers, and other employee-modifiable documents that are shared with prospects, partners, and customers?

- If you currently employ content filtering technologies associated with specific compliance regulations, does this technology also inspect metadata and hidden information?

The ultimate question is very simple: Why would any organization expose itself to these risks if there are solutions on the market that address the problem?

Metadata and Hidden Information Inspection Versus Content Filtering

With a growing focus on content security, driven primarily by compliance regulations, many organizations have implemented solutions that inspect content—particularly that which crosses organizational boundaries—for specific keywords, regular expressions, and text strings. Additionally, many organizations in vertical markets are implementing compliance-specific lexicons to inspect content.

The most common implementation is for outbound e-mail messages, and to a lesser degree, e-mail attachments, to be inspected. Although this approach is appropriate for identifying “known” content violations (such as identification of a keyword that’s on a watch list), it doesn’t address the “unknown” content that is a threat due to where it exists in a document rather than the specific value of the information itself. For example, the text in a hidden comment might not contain words that trigger a security or policy alert, but which, if exposed to a third party, could be damaging. This example can be applied to a broad set of the data elements identified in this study—from usernames and e-mail addresses, to internal server names and path information, to obsolete and hidden text.

Perhaps an appropriate analogy is the importance of antivirus software being able to detect both known and unknown viruses: In most instances, there’s a known pattern that is identifiable, but of equal importance is the ability to spot a behavior or pattern that is suspect. Finally, most content filtering technology was developed to inspect “visible” information and does not address hidden information such as fast save data and author history details.

For comprehensive content and information security, both content filtering and metadata/hidden information inspection should be deployed.

About Oracle Outside In Technology

Outside In Clean Content was used to inspect all the files in this study.

Outside In Technology provides software developers with a comprehensive solution to access, transform, and control the contents of more than 400 unstructured file formats. From the latest office suites, such as Microsoft Office 2007, to specialty formats and legacy files, Outside In Technology provides software developers with the tools to transform unstructured files into controllable information.

Outside In Technology is a family of software development kits (SDKs) that provide a wide range of content solutions. Each SDK is optimized to solve a particular problem, but they are highly flexible and interoperable, generally sharing

Oracle Outside In Clean Content gives software developers the ability to identify and remove sensitive, confidential, or proprietary metadata and hidden information from Microsoft Office documents (versions 97-2007).

the same code base and naming convention. Developers can quickly implement any combination of the SDKs to provide exactly the right functionality in their application while minimizing integration effort and code footprint. The SDKs offer a wide range of options to give the developer programmatic control of their workflow and output.

Outside In Clean Content gives software developers the ability to identify and remove sensitive, confidential, or proprietary metadata and hidden information from Microsoft Office documents (versions 97-2007). As files are created, revised, reviewed, distributed, and reused, they are commonly appended with more than two dozen data elements that, if unintentionally exposed, can lead to corporate embarrassment, breach information disclosure policies, and open new security holes. Oracle's Outside In Clean Content can eliminate this risk, thus enhancing solutions that manage, store, and present content.



The Risks of Metadata and Hidden Information
Updated March 2007

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

Copyright © 2007, Oracle. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission. Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.