White Paper

The Case For Content Security

Introduction

Secure Content conveys the message intended by the author while maintaining its original context without revealing sensitive meta data or content to the reader. The need for Content Security is driven by two factors: the rise of computer–based tools for the creation and production of content, and high speed networks which short-circuits the traditional publishing process by enabling us to share the our work instantly with a global audience. Also, It's now easy to pull digital content out of context and re-use it in ways the author never intended. Indeed, the author may not even be aware of all the content that is hidden in the document's file format. While most of our discussion will focus on the traditional notion of documents that is not the limit of Content Security. Digital content can be text, numbers, E-mail, photographs, video—anything that people can use computers to produce or modify.

The goal of content security, therefore, is to guaranty that the integrity of content remains consistent with the author's original intent, and the document itself does not reveal sensitive details about the originating organization. Most businesses today routinely shred paper documents and destroy digital files at their end of life, but have little or no policies for maintaining the security of content in digital documents while they are in the active phases of their lifecycle. The result can be at least damage to the organization's image or, at worst, violation of the law.

The Content Firewall

All content has a lifecycle (see Gartner Research Note *The GartnerGroup Framework for Content Management* COM-10-1618). It goes through creation, modification, publishing, distribution archiving and destruction. The finished version of any piece of content should be separated from the process of creation and modification by a procedural "firewall." In many cases this is pretty cut and dried due to the nature of the medium. You can't see the raw footage or editing process of a movie on a DVD or see the discarded drafts in a magazine article. In other cases it is not so clear-cut because the "working copy" of some content does not necessarily go through a formal "publishing" process before it is distributed to the world. This is particularly true for government and enterprises that are outside the sphere of professional publishing. Content security imposes this firewall to prevent in-process or sensitive information from becoming public.

There is a fundamental difference between what is appropriate content for internal versus external communications. The process of creating a final publishable version of content on most topics is usually messy. Documents also undergo multiple revisions based on input from many internal stakeholders. There can be a lot of back and forth in this process. Not all opinions or strategies will reflect well on the enterprise to outside observers. Best practices dictate that the details of this process should not be visible to the consumer of the published document, but unless you take measures to secure your digital content it can easily be compromised.

The process of creating documents has become easier than ever. Advances in word processing tools like Microsoft Word enabled levels of collaboration and revision tracking that were once only possible in high-end professional editorial systems used in newsrooms and other professional publishing environments. To enable these features word processing files often contain extra content (technically metadata) that is not visible to the author when the document is opened in the application. It's easy for a hacker to retrieve this "invisible" content, often with embarrassing results for the enterprise. If the final version of the document is printed and distributed on paper it isn't a problem because the invisible content is lost. However, due to the popularity of E-mail and the Internet documents are increasingly distributed in digital form, which means any invisible content is still there for those who want to find it. The problem of hidden content isn't restricted to word processing documents, spreadsheets and presentations can also contain hidden content.

Constructing the content firewall requires replicating the processes used in a paper-based workflow. Instead of sending a document to a printer to create a paper document, you must create an alternative digital version without the hidden content contained in the working copy. While this seems like a sensible idea, it is not part of the routine document creation process in most organizations. Ad Hoc workflow is the norm and standards are not often enforced, which leaves the enterprise open to attacks on its content.

The foundation of the content firewall is using Adobe's Portable Document Format (PDF) for distributing finished digital documents. PDF is the equivalent of a digital copy of the printed page and, when used properly, goes a long way to securing content. Adobe publishes the PDF Reference which has enabled a growing number of firms provide third-party tools for producing and manipulating PDF files. PDF has many built-in features like support for encryption, digital signatures and access control that can be used in securing content. The Adobe Acrobat® program is extensible via third-party plug-ins and external applications can process PDF to further enhance content security.

Compromised Content Security

It's not hard to discover the result of poor content security. It's seems like every few months there is a disclosure that a document containing information which shouldn't be there has made its way into the public sphere. How do they get there? Lax security comes from many sources. Some errors are the result of the mishandled redaction, removing sensitive content, of government or legal documents. Other disclosures are the result of hidden text or metadata left in working documents in native format due to a misunderstanding of the corporate publishing process.

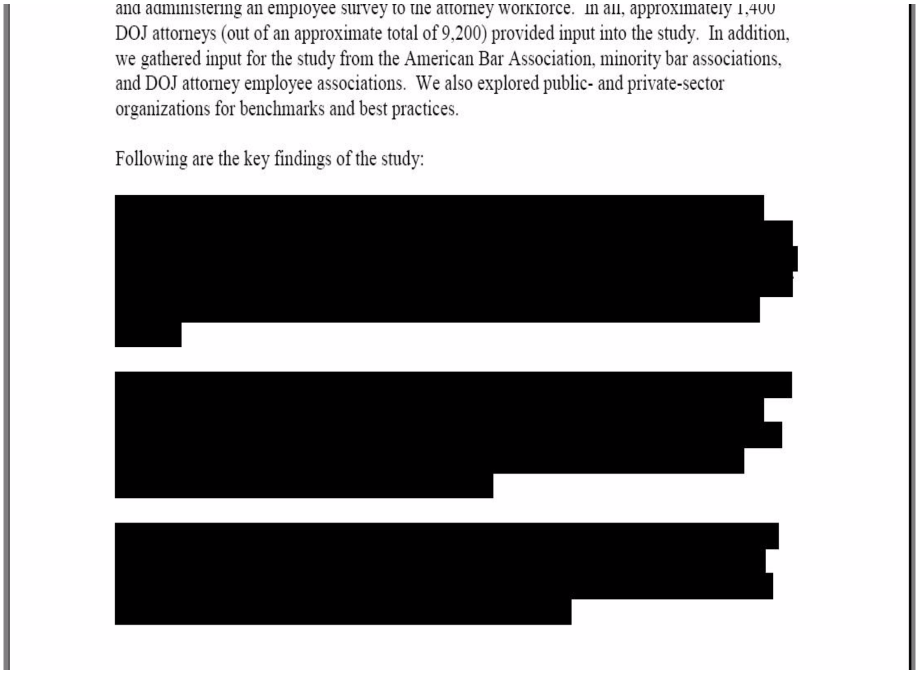
Lessons of FOIA: The Freedom of Information Act (FOIA) allows Americans access to the inner workings of their government minus the bits that might compromise national security, privacy or confidentiality. FOIA is not the only piece of legislation that impacts content; the Privacy Act and the Health Insurance Privacy Assurance Act (HIPAA) also regulate the content security of certain documents.

FOIA actually back to 1960's but has undergone periodic revisions. The process is fairly straightforward. Members of the public could submit a request for documents that cover a specific topic, say the impact of mercury pollution at Air Force bases. Once located, the requested documents would be redacted by blacking out sensitive text with magic markers or actually cutting out excerpts with X-Acto knives. Copies of the redacted documents would then be delivered to the requesting party. As with most things associated with the government, the process includes extensive reporting requirements that further complicate the operation.

In the mid-1990's the Internet was ascending and "Information Superhighway" was hailed as the genesis of next industrial revolution. With this as a backdrop the notion of electronic FOIA or eFOIA was born. Part of eFOIA was establishing electronic reading rooms, searchable repositories of popular FOIA-requested documents in digital form available at each agency's Web site. Documents in these reading rooms are either in Tiff or PDF format.

Once the documents are posted they are available to those who want to read them *and* those who want to hack into the file's internal structures to discover any hidden content that might still be there. If the document has been properly redacted there will be nothing to find, but many times that is not the case.

In the case of a study done by a private contractor for the Department of Justice on their diversity practices we can see the results.



and administering an employee survey to the attorney workforce. In all, approximately 1,400 DOJ attorneys (out of an approximate total of 9,200) provided input into the study. In addition, we gathered input for the study from the American Bar Association, minority bar associations, and DOJ attorney employee associations. We also explored public- and private-sector organizations for benchmarks and best practices.

Following are the key findings of the study:

[redacted]

[redacted]

[redacted]

Figure 1. A PDF file that appears to be heavily redacted, but is it?

This document appears to be redacted, but hidden content is not secure content. A few clicks later we find…

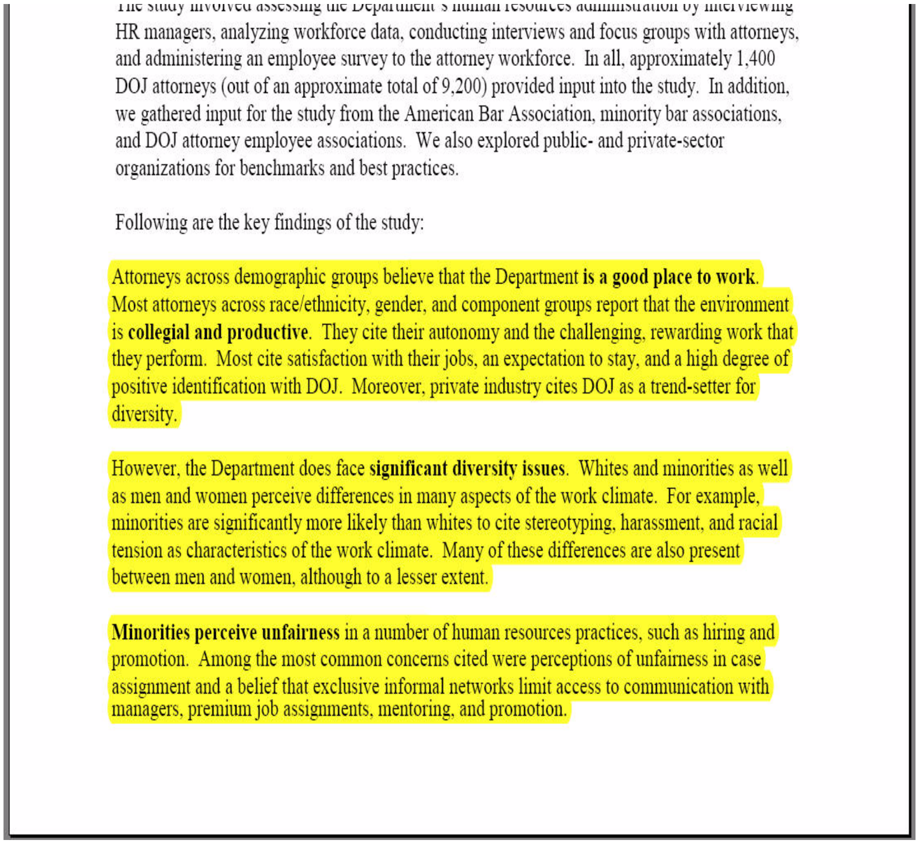


Figure 2, When content is only hidden it can be uncovered and it is not truly redacted.

The study involved assessing the Department's human resources administration by interviewing HR managers, analyzing workforce data, conducting interviews and focus groups with attorneys, and administering an employee survey to the attorney workforce. In all, approximately 1,400 DOJ attorneys (out of an approximate total of 9,200) provided input into the study. In addition, we gathered input for the study from the American Bar Association, minority bar associations, and DOJ attorney employee associations. We also explored public- and private-sector organizations for benchmarks and best practices.

Following are the key findings of the study:

Attorneys across demographic groups believe that the Department **is a good place to work**. Most attorneys across race/ethnicity, gender, and component groups report that the environment is **collegial and productive**. They cite their autonomy and the challenging, rewarding work that they perform. Most cite satisfaction with their jobs, an expectation to stay, and a high degree of positive identification with DOJ. Moreover, private industry cites DOJ as a trend-setter for diversity.

However, the Department does face **significant diversity issues**. Whites and minorities as well as men and women perceive differences in many aspects of the work climate. For example, minorities are significantly more likely than whites to cite stereotyping, harassment, and racial tension as characteristics of the work climate. Many of these differences are also present between men and women, although to a lesser extent.

Minorities perceive unfairness in a number of human resources practices, such as hiring and promotion. Among the most common concerns cited were perceptions of unfairness in case assignment and a belief that exclusive informal networks limit access to communication with managers, premium job assignments, mentoring, and promotion.

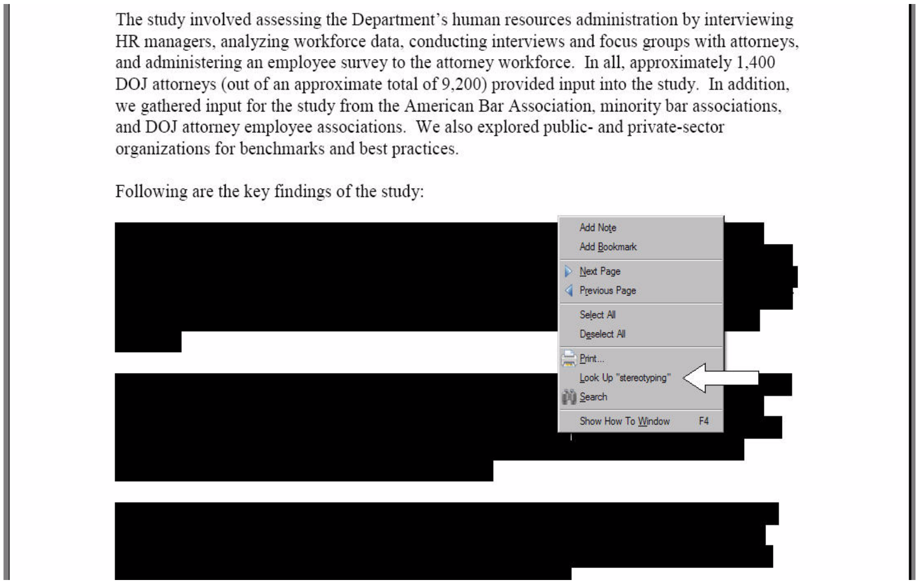
This could have been avoided by properly redacting document using one of Appligent's Redax® family of products. Redax does not hide content it permanently deletes it from the data stream. Leaving content in the document invites discovery later.

Also, consider the case of feature creep. The tool originally used to hide content might change over time. New features may be added that drastically change the functionality of the application in ways you have not considered. Adobe has steadily added features to Acrobat as its vision for the product has changed over time. For example, Adobe introduced spell-checking features for the first time in Acrobat 6. Now it's as simple as placing the cursor on a blacked-out word and right clicking to discover the hidden text.

Likewise, relying on the people viewing your content only having access to limited feature set of the free Adobe Reader is not a viable content security plan. Those most interested in cracking improperly redacted documents invest in the latest tools and plug-ins. So, assuming that your content will be secure because most end users only have Adobe Reader or have not upgraded from

Acrobat 4 or 5 is definitely not a wise idea. Bottom line: *Content security must reside in the file, not in perceived limitations of the viewing software.*

Lets take another look at the original "redacted" document using Acrobat 6.01 Professional.



The study involved assessing the Department's human resources administration by interviewing HR managers, analyzing workforce data, conducting interviews and focus groups with attorneys, and administering an employee survey to the attorney workforce. In all, approximately 1,400 DOJ attorneys (out of an approximate total of 9,200) provided input into the study. In addition, we gathered input for the study from the American Bar Association, minority bar associations, and DOJ attorney employee associations. We also explored public- and private-sector organizations for benchmarks and best practices.

Following are the key findings of the study:

[redacted]

[redacted]

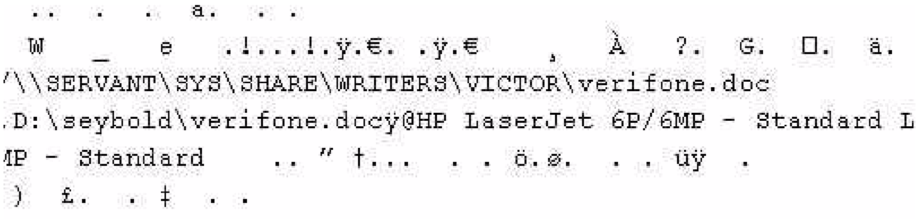
[redacted]

Figure 3, Hidden content can be exposed when new features are added to the viewer. This acts as an advertisement for hacking opportunities.

Going Native: Native Office file formats are a notoriously insecure method for distributing content. The possibility of infection by Macro viruses makes them extremely unattractive to many potential recipients, and the potential bad publicity for the sender should be enough by itself to end this practice. Since many security-oriented managers block these file formats at the firewall which means it won't reach the intended recipient.

While this is the most obvious problem with using native Office documents the Content Security issues can be more damaging. It is much harder to maintain the concept of a content firewall when you rely on the working file formats as a distribution media. Users of any application may not be aware of all of its features or consider the impact of releasing unsecured content to those outside the enterprise.

There is plenty of room for hidden content disguised as features accessed directly from the user interface or buried deep in the file format. Damaging information can be revealed via metadata that expose the internal computing infrastructure. Details like workstation and server naming conventions that



```
  ..   .   .  a.   . .
 W   _   e   .!...!.ÿ.€. .ÿ.€     ,   À   ?.  G.  □.  ä.
'\\SERVANT\SYS\SHARE\WRITERS\VICTOR\verifone.doc
.D:\seybold\verifone.docÿ@HP LaserJet 6P/6MP - Standard L
MP - Standard     .. ″ †...   . . ö.ø.   . . üÿ  .
 )  £.  . ‡   . .
```

Figure 4, This native file reveals the name of the file server, "servant" as well its directory structure SYS\SHARE\ etc. Useful information for the wily hacker.

could prove helpful to someone trying to hack into your system are easily extracted from the file. There are also issues with application features tied to the content. Microsoft Word's Track Changes feature allows anyone to see the user ID's of who worked on a document, what they changed and the timeline of the edits. Even with Track changes turned off Word tracks the ID's of the last 10 authors of the document. This type of hack caused great embarrassment to the UK government of Tony Blair when one of the documents (distributed in MS Word format) used to support the case for WMD's in the run-up to the Iraq war was exposed as containing the recycled work of American graduate students.

Spreadsheets contain formulae that can reveal details of the financial models of your business. The speaker's notes view of a PowerPoint presentation does not go away when you send a copy to each member of the audience that requested it. In some environments the presentations are prepared by a centralized group and distributed to field personnel who may not even look at the Speaker's Notes view.

Alternate Solutions

Let's examine a few of the sub-optimal solutions that are often suggested for maintaining content security before wrapping up.

Lock it away: While some might advocate the mass voluntary adoption of complex Digital Rights Management (DRM) systems this is not the best approach and is unlikely to happen. Many end users are loath to consider that option as the technology provides little benefit to them. Further much published content's purpose is to reach the largest possible audience. Anything that gets in the way of that goal decreases the usefulness of the digital distribution channel. Also DRM alone does not remove the extra content from the data stream, it merely restricts who can see it. As we have shown hidden content is not secure content. If the DRM algorithms are cracked in the future—a likely event, anyone will be able to access the sensitive or harmful information.

Stick with paper: Another option is to consider the printed copy of the content to be the final published form. While this seems to solve the hidden content problem, it actually creates a workflow nightmare latter in the lifecycle. The document must be scanned and converted back into digital form for electronic distribution and archiving. Further, if the text is to be searchable the document must be processed by an Optical Character Recognition application and the results put through a quality control process to assure accuracy. Finally, any metadata describing the document must be re-entered. If the electronically archived document is subject to redaction you are back where you started! The overhead attached to this torturous workflow begs the question: Why not just do it the right way from the beginning?

Best Practices

The first line of defense is to make the content creators in your organization aware of the potential content security problem and create a simple set of procedures that implements the content firewall. This will assure that only the appropriate content is released and the files are scrubbed of any information that is irrelevant to the document's context.

In order to combat the problem organizations need to implement best practices for publishing content.

- Always release documents in a publishing-oriented format like PDF rather than native production-oriented formats like Microsoft Word or Excel (.doc or .xsl).
- Establish a workflow for the release of information outside the enterprise. Make sure the meta data associated with the content matches your corporate policy. Account for both digital and paper final formats.
- Permanently delete all information that should not be released with the document. Never assume that hidden content is secure.
- Use the content security features built into PDF to control copying and printing of the document if its appropriate for your application.
- Label all working documents that have not been checked and purged of unwanted content as “For internal use only” or “Corporate Confidential.”
- Institute training so all staff members understand the importance of content security and the procedures for sharing documents.
- Consider all phases of the content lifecycle for digital documents, including archiving. PDF is gaining wide acceptance as the way to go for archiving digital documents. It has recently been approved by the National Archives (NARA) as a standard archiving format for storing digital documents. The PDF/A working group has developed a specification for assuring a files usability into the distant future.
- If you are releasing redacted documents do a final check to make sure the redaction has been done properly and all of the sensitive content has been permanently removed from the file. Appligent's Redax includes a utility to display the file's data stream to verify the status of the content.

Conclusion

Securing Content is important regardless of your enterprise's mission. It greatly reduces the risk of embarrassment, financial loss or worse from content related disclosures. Remember the goal of content security is not to hide information, it's to assure full but safe disclosure of only the relevant content without providing compromising information about its creation or your technology infrastructure. Additionally, secure content must guaranty that redacted information is not accessible by anyone with tools to interrogate the internal structures of the digital document.

PDF provides the most convenient method for distributing secure content due to the ubiquitous availability of the free Reader software, support for multiple operating platforms and easy integration with the Internet. It's plug-in architecture enables Appligent Redax to be seamlessly integrated into a content security workflow. Since PDF also includes support for sophisticated file security techniques it allows you to control access privileges so you can specify whether a reader can print or cut and paste content. Releasing the final version of content as PDF documents should be the cornerstone of any content firewall.

In the end the content firewall is only as good as the people who build it. This white paper has shown a few examples of poorly secured PDF. Just relying a file format buzzword will not make your content secure. Making content security a central part of the document lifecyle and workflow means giving people the right tools, training and desire to achieve it. True content security will only happen when your organization empowers it.

Victor Votsch

Mr. Votsch is Appligent's Director of Corporate Communication and has over 20 years experience working with and commenting on digital content, including stints as a Gartner Research Director, Seybold Senior Editor and Founding Managing Editor of XML.com.