Proper Redaction

The Importance of Complete Sanitization in PDF Documents

Redaction refers to the process of removing sensitive or classified information from a document prior to its publication. Redaction is used to either remove personal information identifiers, such as social security numbers, names, addresses, or phone numbers, or to limit access to sensitive data prior to releasing the document to the targeted end user/viewer of the document.

Redaction workflows have 2 common steps:

1. The first step is to identify the areas for redaction by searching for the text, string of words and terms, specific images, or areas of the document that you want to redact, and subsequently marking the identified areas that should be redacted with redaction annotations. This allows for a manual review process to occur before the content is redacted.
2. The second step is to apply the redactions, which will complete the removal of the identified sensitive data and replace it with a user-selected redaction annotation appearance. This can be in the form of a black bar, or any other indications that the specific area of the document has been redacted.

There are some important things you should be aware of when redacting your documents, and some potential pitfalls that could happen if you use the wrong tool to redact your sensitive data. Some common issues faced by users include:

- **Incomplete redaction –** be cautious of companies that claim that their tools can remove sensitive data from a document, when all that is done is the placement of black bars over the text. Simply applying black highlighting to block out text in a PDF document is not considered redaction. Full and complete redaction ensures that all underlying text and data in the redacted document is completely removed. You should also consider implementing sanitization which removes all metadata, attached content, comments, and other data elements that might be associated with the sensitive data you are trying to remove from the document.

- **Risks and costs of improperly redacting documents -** these are too high to trust and rely on the wrong tool. Strictly enforced rules apply to documents that contain either classified or sensitive personal information and identifiers, this is especially true in healthcare (Health Insurance Portability and Accountability Act - a U.S. legislation that provides data privacy and security provisions for safeguarding medical information), government (U.S. Federal Privacy Act), and other areas such as the legal and financial sectors. You risk exposure to potential litigation and fines if sensitive information is not fully removed before being released.
- **Potential risk of corruption –** using the wrong tool can potentially distort the text layout in the PDF document, leading to the removal or modification of content that was not meant to be redacted.

Incomplete redaction and the associated risks

Organizations may think that they are implementing redaction by simply drawing black rectangular annotations over the content that they want to redact. However, it is important to note that drawing annotations on top of the content does not actually remove the underlying information from the page, even though the content is no longer visible when viewed or printed. Thus,

- Sensitive content is still searchable
- Even flattening the annotations (which is the process of making the annotations part of the page content rather than content that is placed on top of the page) does not remove the content - flattening simply covers up the sensitive content
- Exposure to potential litigation if the sensitive information was not fully removed before being released

Proper Redaction with Datalogics PDF Java Toolkit

Organizations who use PDF Java Toolkit to implement redaction are guaranteed to achieve full and complete removal of sensitive information. Rather than simply drawing a black rectangular annotation on top of sensitive content, the PDF Java Toolkit does completely remove all sensitive content when used to apply the redaction annotations - ensuring that content under the redaction annotations is fully removed so that it is no longer in the document.

Additionally, the PDF Java Toolkit can accommodate overlay text in the redaction annotation to specify the type of information that was removed, or codes that refer to the content type such as "name" or "address". This is especially relevant to government agencies as it allows them to meet the specifications of the Freedom of Information Act and the U.S. Privacy Act of 1974 which requires that these agencies specify the exact exemption for each redaction in disclosed public documents.

The Datalogics PDF Java Toolkit can provide a streamlined and flexible redaction process. Applications can be developed to facilitate batch redaction, thereby automating the redaction of large collections of documents. For example, an application can allow users to specify a location on the page to redact, and then batch process the redactions all at once. Because of the flexibility of PDF Java Toolkit, the user can choose to create different redaction workflows to redact a set of PDF documents that have fields in alternate locations on the page. This provides limitless flexibility for applications across the organization as it is not limited to one specific document layout.

Another benefit is the additional level of sanitization that can be applied after the redaction process, which ensures that any non-visible elements in a PDF document - potentially containing sensitive data - is completely removed. Examples of non-visible data elements contained in a PDF (which are often overlooked) include:

- Metadata - who created it, how it was created, location, and time created
- Attached content
- Comments from a review process
- Previous versions - these occur because of incremental saves and could be restored
- Embedded search indexes - a search index could potentially have references to sensitive content in the document
- Obscured text - this refers to text that could have been partially covered by an image or text that might be off the page during the creation of the document
- Links - this prevents someone from finding references that might exist elsewhere internally

About PDF Java Toolkit

Datalogics PDF Java Toolkit is a native Java library that provides high-level APIs for automating PDF workflows like processing PDF forms, verifying digital signatures, and extracting text. It also offers low-level APIs for working directly with the structure of the PDF for those times when needed. While written with Java developers in mind, Datalogics PDF Java Toolkit can be used with any Java Virtual Machine (JVM) language.

Datalogics and the Datalogics logo are registered trademarks of Datalogics Incorporated. Adobe and the Adobe logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. All other trademarks are the property of their respective owners. ©2017 Datalogics, Inc.