Knowvation DX™

Making PDFs Safe
Knowvation DX PDF Sanitizer™

EXECUTIVE SUMMARY

The Adobe Portable Document Format (PDF) file format has become ubiquitous throughout industry, government and academia. **While the PDF is an efficient format for transferring files, many users do not realize its potential to transmit sensitive information that is hidden within the PDF.** This attribute of a PDF is particularly important since the PDF format is increasingly used for transmitting documents that have been declassified ahead of release to the public. This whitepaper identifies all the potential hiding places within a PDF, the large vulnerability difference between PDFs originating from secure sources versus non-secure sources, and the methodology used by Knowvation DX to eliminate hidden data.

The PDF format is used for viewing files that have been created in different applications (e.g., word processing, graphics, publishing, scanner, and Optical Character Recognition [OCR]). The PDF reader is made available for free, and allows for easy distribution of documents over computer networks and the Internet. The PDF format is popular since it eliminates the need for possessing the application that was used to create the document when read-only display is all that is required. It also creates a copy which cannot be easily changed and preserves the original content and format.

Despite the widespread use of PDF documents, users who distribute these files may be unaware of the possibility that the files might contain hidden data. **Data can be surreptitiously hidden in many places within a PDF** that are not obvious to the casual PDF user; certain hiding places are not even well known to more experienced PDF aficionados. In addition, there is a need to consider non-visible object data and metadata that is included serendipitously as a normal action of the application creating the document, but not consciously added by the user. There may not be nefarious motives behind the inclusion of this hidden data, but its presence may have unanticipated consequences if it is not removed as part of the document declassification process.

This paper explores the **four general hiding categories** and the myriad of discrete hiding places within each general category. The four hiding categories that are more fully explored in this paper are:

1) **Hidden text** in the OCR Text Layer;

2) **Optional objects** within the PDF that provide information about the PDF's content;

3) **Comments** within the Internal Structural Layer**;** and

4) **Steganographic data** within the PDF Image Layer.

PDFs that are created by digitizing hardcopy media in a secure environment by trusted sources only have only one hidden category where information can be transmitted that is not visible when reviewing the image online -- the hidden text (item 1, above). A "secure environment by trusted sources" is defined as a government facility or contractor SCIF staffed with cleared personnel. PDFs that are created in a non-secure environment by unknown sources may contain many categories where information can be hidden, including optional objects, comments, and steganographic data (items 2, 3, and 4, above). Therefore, it is much easier to eliminate sensitive information from a file created in a secure environment than one created in a non-secure environment.

***PDF redaction* is the process of removing sensitive visible image layer text and hidden OCR text from a PDF document**; redaction is the only process required if the PDF created from hardcopy media was provided by a trusted source. Once the sensitive material has been removed, the document may be distributed in declassified form to a broader audience. ***PDF sanitization* is a follow-on process to redaction. It removes additional forms of hidden content that must be addressed.**

In instances where PDF documents were not created by a trusted source, redaction is not sufficient to ensure that all sensitive content has been removed. Sanitization addresses hidden content in PDFs created from digitized formats (e.g., jpeg, tiff, pdf) or from "born-digital" formats (e.g., Microsoft Office, FrameMaker, computer-aided design software). This hidden content includes optional PDF objects, internal PDF structural data, and steganographic data within the pixels of images.

The Knowvation DX software application suite from Progressive Technology Federal Systems (PTFS) addresses all hidden information. Once the words, concepts and images that are sensitive have been defined, the **Knowvation DX Redaction Editor** combines Knowvation's standard content management capabilities with advanced search techniques such as Variable Adaptive Pattern Recognition (VAPRTM) to identify the visible text that is sensitive. It can also redact sensitive images within manually drawn redaction zones.

The **Knowvation DX PDF Sanitizer** operates in tandem with PDF Redaction to produce declassified PDF documents. The Knowvation DX PDF Sanitizer was designed to address all possible combinations of hidden content in a PDF file. The product is currently at Version 1.2, and removes hidden information such as metadata, bookmarks and PDF notes. The next version will remove all remaining hidden content as well as steganographic data. The Table below summarizes the redaction and sanitization processes.

Table 1 – Redaction and Sanitization Processes

PDF Source	Process	Sub-process	Explanations	Knowvation Services
Trusted and Secure Digitization Process	Redaction	Redact Visible Text	Redact Dirty Words / Concepts / Figures	Done in production by Knowvation DX Redaction Editor
		Redact Hidden Text Layer		
▪ Non-Secure Source - Digitized Hardcopy ▪ All Sources - Born-Digital Formats (MS Office, CAD, etc.)	Sanitization	Sanitize Optional Objects	Remove optional PDF objects that may contain sensitive text	PDF Sanitizer™ 1.2 removes metadata, bookmarks and comments. Complete requirements and solution are defined in the Knowvation Roadmap for future releases.
		Sanitize Internal Hidden Data	Scrub text that is hidden in the internal structure	
		Sanitize Steganographic Data	Disrupt pixels to disable decoding of messages hidden in images	

1. WHY THE PDF FORMAT IS USED

PDF is a de facto standard for sharing information electronically. It is used for converting files that have been created in different applications (e.g., word processing, graphics, publishing) into a format that can be viewed by the Adobe Reader software that is made available for free.

PDF provides excellent fidelity and portability, and allows for easy distribution of documents. The PDF format is used for file exchange via email, for publishing documents on the web, and for interactive content like forms and multimedia. The PDF files retain the appearance of the original content across varying types of hardware and software that will be used to view the files.

The PDF format has been used as a "safe" format for mass distribution to a wide audience via email or posting to public websites. However, the robustness and complexity of PDF formats allow for a wide variety of content types, which increases the likelihood that sensitive data may be unintentionally retained in a file. In this regard, NSA has published information on methods to safely produce PDF files from MS Word. Understanding the inherent risks requires at least a basic understanding of the PDF format. **Even though the format is "open", understanding the structure and the associated risks is not trivial.**

The PDF format has become increasingly popular for physical to digital document conversion projects as well as declassification and FOIA projects. This format provides the exact appearance of the original document but is also searchable (when OCR'd) and can be indexed as a single document or when combined with thousands of documents.

2. PDF HIDING PLACES

Virtually any PDF document may include hidden content that cannot be detected with visual inspection.

2.1 Where Can Information Be Hidden in PDFs?

PDF software provides a set of features that supports a broad range of editing and display capabilities. These capabilities are implemented in multiple structural layers within PDF files, and each layer may contain hidden content. Given the pervasive use of PDF, it is important to address all file layers and content types that may contain hidden data, in order to reduce the risks of accidentally disclosing sensitive information.

Most users familiar with PDF think of it as a collection of pages that generally look the same regardless of operating system or viewer application. However, the internal structure of the PDF file is much more than that. There are three structural components of a PDF:

1. **Visible Image Layer;**
2. **Hidden Text Layer; and**
3. **Internal Structure.**

Content can be hidden in any of these components.

2.2 What Kind of Content Can Be Hidden in the PDF?

Within the PDF file structure, there are four types of hidden content:

1. **Hidden text** in the Visible Image and Hidden Text Layers;
2. **Optional objects** in the Internal Structure that provide information about the PDF content;
3. **Comments** within the Internal Structure; and
4. **Steganographic data** within the Visible Image Layer.

Each is discussed below.

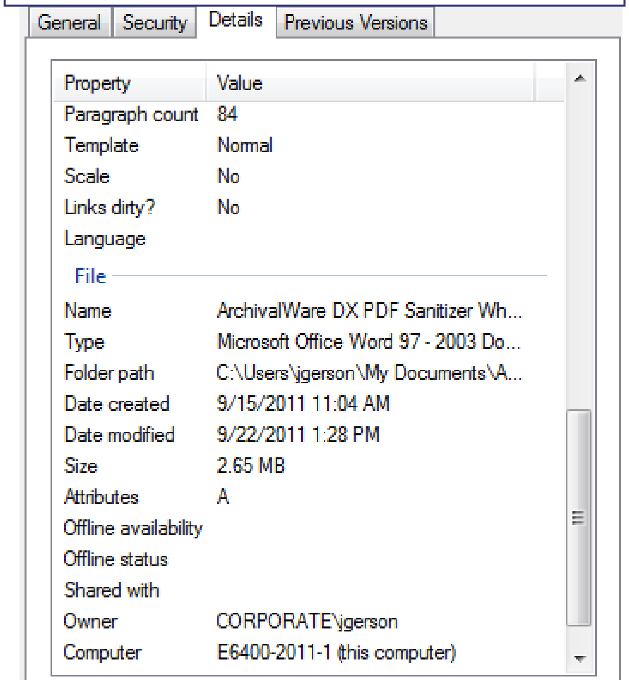
2.2.1 Hidden Text

Text may be present in both the visible text (the searchable part) and in the image of that text (the display and print part) in a PDF document. Hidden text is present in searchable-image PDF files that were **created from page images which have been converted to text using Optical Character Recognition** (OCR).

2.2.2 Optional Objects

The structure of a PDF document provides a general-purpose mechanism for maintaining a binary representation of a set of generic objects. These **"optional" objects are the building blocks for what the user ultimately perceives as a set of pages.**

Figure 2.2.2-1: Data Created by MS Word



Property	Value
Paragraph count	84
Template	Normal
Scale	No
Links dirty?	No
Language	
File	
Name	ArchivalWare DX PDF Sanitizer Wh...
Type	Microsoft Office Word 97 - 2003 Do...
Folder path	C:\Users\jgerson\My Documents\A...
Date created	9/15/2011 11:04 AM
Date modified	9/22/2011 1:28 PM
Size	2.65 MB
Attributes	A
Offline availability	
Offline status	
Shared with	
Owner	CORPORATE\jgerson
Computer	E6400-2011-1 (this computer)

In the vast majority of PDF files, optional objects are used in a manner that conforms to the PDF file format specification, and present no special exposure to carrying sensitive information, concealed or otherwise. However, **due to the general-purpose nature of this object system, it is possible for sensitive information to be concealed in the PDF file** in a manner that is not identifiable through visual inspection by an end user. In fact, **non-visible data may be included as a normal action of the application creating the document, and not consciously added by the user.** For example, when an image is cropped in MS Word, the user only sees the cropped image, but the file retains the entire image and may port the entire image to PDF along with the cropping indicators. MS Word also captures metadata from the IT system it resides on (see Figure 2.2.2-1), and that metadata may be classified at a level above the information in the document. Whether or not hidden data is included consciously, a PDF file runs the risk of serving as the electronic document equivalent of a Trojan horse, appearing from the outside as a benign document, but carrying potentially dangerous material inside. The PDF may also act as a reverse Trojan horse where a released document unwittingly transports sensitive data wrapped in releasable material in a way that is not obvious to the casual reader or recipient.

The following set of **more than thirty objects and artifacts can be included in a PDF file** (Please see Attachment 1 for descriptions of these items):

- Document Metadata
- Object Metadata
- Embedded Content and File Attachments
- Annotations and Comments
- Form Fields and Form Data
- Hidden Text – Not Searchable Image
- Hidden Text – Searchable Image (optional)
- Hidden Layers
- Bookmarks
- Embedded Search Index
- Digital Signatures
- Adobe Reader Extensions
- Article Threads
- Links, Actions and Scripts
- Watermarks
- Stamps
- Bates Numbers
- Typewriter Text
- Sticky Notes
- Highlighted, Crossed Out, or Underlined Text
- Buttons
- Multimedia (Video, Sounds, SWF, 3D objects)
- Drawing Markups
- Markups Added by Third-Party Plug-ins
- Embedded Print Settings

- Overlapping Objects, Obscured Text and Images
- Embedded Images with Reduced Display Dimensions
- Embedded Thumbnail Images
- Alternate Images
- PDF Opening View
- Automatic Page Advancement and Page Transitions
- Deleted Hidden Page and Image Content

It is desirable to remove the optional components within PDF documents because they may contain sensitive hidden content, or because they are otherwise deemed to be inappropriate for declassified publication.

2.2.3 Comments

Comments are the third major type of potentially sensitive hidden internal data in a PDF file. **PDF comments are lines of text, preceded by a "%" character, that can optionally exist within the PDF file, but that have no representation within the PDF object mechanism**. Comments can be used to provide supplemental information about how components of the file are structured, or about where the data was generated. They are used by publishers of electronic documents to provide a linkage between a generated PDF file and some source document. For example, a PDF file produced from Adobe Illustrator may contain PDF comments that are ignored by Acrobat, but that can be used by Illustrator in the event that the PDF file is re-imported into Illustrator. These **comments are typically benign, but because they are not displayed within Acrobat, they can contain data that is difficult to review**.

Comments can be used to embed sensitive information into the PDF file that is completely undetectable to everyone except a knowledgeable expert equipped with a binary file editor.

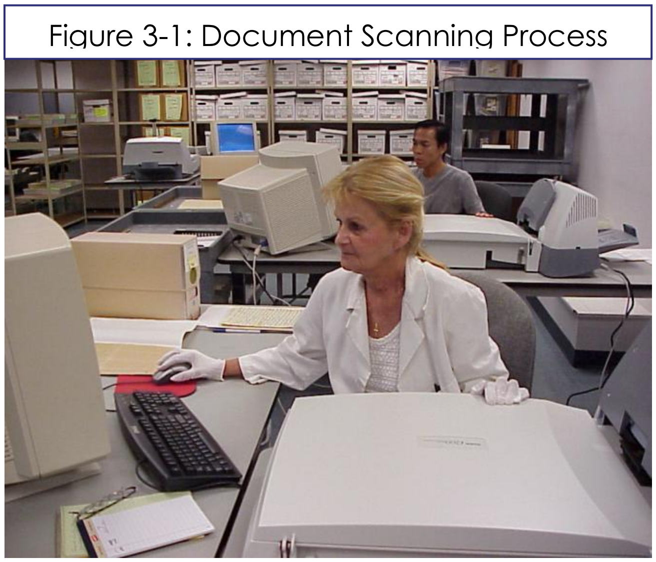
2.2.4 Steganographic Data

Steganographic data is content that may be hidden in the internal structure within raster images that are, or become part of a PDF file. Such data or messages are invisible to the naked eye, yet they can be decoded by a user that has the appropriate software.

Steganographic content may take several forms. For example, it may be image data that is obscure to the normal viewer, typically as a watermark or mask. It may also take the form of binary data that is communicated through the source image's pixel-level data. These hidden messages are very difficult to detect in all their forms.

3. PDF'S FROM SECURE VERSUS NON-SECURE SOURCES

Figure 3-1: Document Scanning Process



PDF files may be produced from hardcopy documents or from electronic files. Hardcopy documents can be converted to PDF via scanning (See Figure 3-1). Electronic files in "born-digital" format are usually converted into PDF using commercial conversion software programs (known as "distillers") that accept source formats such as Postscript, MS Word or Computer-Aided Design format, and produce output to PDF. Note that these electronic files may have originated as hardcopy documents that were converted to text files via OCR. Any application that can send output to a printer can interface with Acrobat's print driver software to generate a PDF file.

There are many **potential sources for data in a PDF document**. Some examples are:

- **Adobe Distiller and the PDFMaker tools** (two of the common applications used to convert an MS Word file to a PDF) **translate much of the layering complexity from one format to the next**.
- **Images placed on top of text in an MS Word document** will continue to hide (but not prevent access to) the underlying text within the resulting PDF file.
- **Metadata captured by MS Word** which may be classified at a level above the information in the document.
- **Non-visible data is included as a normal action of the application creating the document**, but not consciously added by the user.
- **Acrobat Professional and Adobe LiveCycle Designer can generate PDF output** without requiring input from any external source document.

Metadata and non-visible data created by applications are typically present in PDFs without any malicious intent. However, the presence of this data may have unanticipated consequences if the document is distributed without restriction. As a result of these many sources of data in PDFs, a user receiving a PDF document and wishing to remove sensitive data before sharing the document may need a sanitization procedure to fully determine whether sensitive data remains in the file.

PDFs that are created by digitizing media in a secure environment by trusted sources have only one place where information can be transmitted that is not visible when reviewing the image online -- the Hidden Text Layer of the PDF. PDFs that are created in a non-secure environment by unknown sources may be more complex and have many places where information can be hidden, including optional objects, comments, and steganographic data. It is easier to eliminate undesirable information from a file created in a secure environment than one created in a non-secure environment. Therefore, PDFs from a non-secure source require more comprehensive processes to find and remove hidden text that may contain sensitive information.

4. REDACTION AND SANITIZATION

PDF redaction and sanitization are the processes for removing visible text and hidden data that is sensitive from a PDF document so that the document may be distributed to a broader audience.

Content may need to be removed from a PDF document due to security, privacy, confidentiality or other legal considerations. For example, security and privacy considerations may make it necessary to remove the name of the author, specific content, or comments that were embedded in the file when the PDF was created or subsequently edited. Redaction and sanitization both pertain to the removal of visible and hidden content that is not intended for publication and distribution to a broad audience. This includes allowing the PDFs to move from a higher level security domain (i.e., JWICS) to a lower level (i.e., SIPRNet).

We use the term PDF "*redaction*" to refer to the process of removing text from the image layer and OCR or hidden text layer of a PDF document that was provided by a trusted source. In this case, electronic documents were created through digitization (scanning) of hard copy paper or film media by PTFS or another trusted source that had complete control over how the PDF was made and what information was placed into the PDF structure. As a result, only image data and text layer data created via OCR need to be addressed for the removal of unwanted material.

We use the term PDF *"sanitization"* to refer to the follow-on process of removing additional hidden content from a PDF document that was provided by a non-secure source. Here, electronic documents were not created through digitization performed by a trusted source. As a result, redaction is not sufficient to ensure that all sensitive content has been removed. Sanitization additionally addresses hidden content in PDF files created from "born-digital" formats (e.g., Microsoft Office, FrameMaker, CAD software).

4.1 REDACTION

Redaction begins with the identification of sensitive or "dirty" words, concepts or images that need to be removed. Once the visible dirty words or zones of dirty content are found and approved for removal, they are removed and replaced with black, white, or colored rectangles and noted by an Exemption Code stamp. **The sensitive information is removed from both the visible image layer as well as the hidden OCR layer**.

4.2 Sanitization

Three forms of sanitization are needed – removing optional objects, scrubbing internal data, and disrupting steganographic data. Implementing these processes make PDF documents virtually inviolable and enhances their security when shared to lower level domains.

4.2.1 Removing Optional Objects

In order to remove unwanted optional objects, the objects contained within the file must be examined; the **objects defined by the PDF standard must be validated; and the outliers that are to be expunged must be identified**. The validation process can produce a report of the exceptions that it encounters within the file.

The process of validating internal objects for an input PDF file produces a list of non-standard objects, including those outside of the published PDF standard. It is then necessary to ensure that potentially sensitive information contained within those non-standard objects is removed from the PDF file before the file proceeds to the mark-up and redaction processes.

The **process for expunging this sensitive information using Knowvation DX PDF Sanitizer** is a two step process. In the first step, the **objects' references within the PDF file are removed**. In the second step, **a new PDF file is created in place of the original**, utilizing a technique known in PDF as a "referenced links saved" operation. This operation effectively rewrites the document to the binary file system from scratch, including only those objects for which a current reference within the PDF file exists. All objects that are no longer referenced within the PDF file are eliminated, which ensures that the scrubbed objects will have no binary representation. This means that any sensitive information that may have been removed through scrubbing can no longer be reconstructed, even by knowledgeable PDF experts using the most sophisticated PDF toolkits and binary file editors.

4.2.2 Scrubbing Internal Data

The "referenced links saved" operation to remove optional objects by creating a new file will also scrub PDF comments, which are one type of internal data that may contain sensitive information in a PDF file.

4.2.3 Disrupting Steganographic Data

Steganographic data or messages are very difficult to detect in all their forms. Full detection would require pixel-by-pixel analysis of each image found in the PDF file, using a definition file of known steganographic data encryption methods.

Fortunately, measures can be taken to insure that any hidden steganographic data cannot be decoded. **Decoding steganographic data can be disabled by making**

slight modifications to each image found in the PDF file, while still leaving the image in nearly the original display quality.

There's only one solution capable of finding and removing an enterprise's data hidden in a PDF file and allowing the capability to redact or remove that information... Knowvation DX with PDF Sanitizer™.

5. KNOWVATION

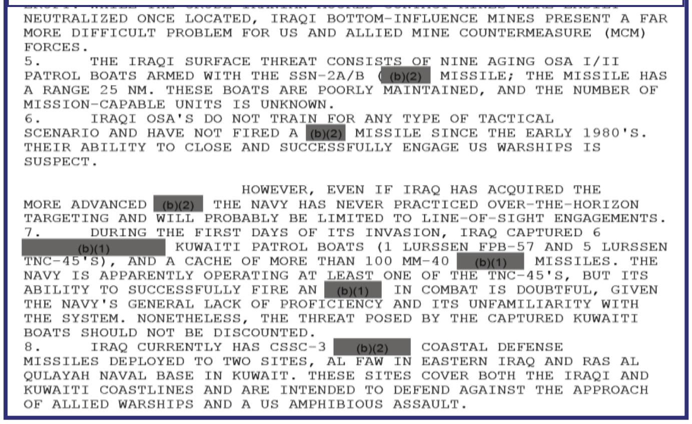
The Knowvation suite of software applications from Progressive Technology Federal Systems (PTFS) addresses both redaction and sanitization.

5.1 Knowvation DX Redaction Editor

The Knowvation DX Redaction Editor removes visible text that is sensitive. It combines Knowvation's standard content management capabilities with advanced search techniques such as Variable Adaptive Pattern Recognition (VAPR™). The basic process is as follows:

- A cognizant person identifies the sensitive or "dirty" words, concepts or images that need to be removed.
- Using an automated full text search and other techniques, the Knowvation DX Redaction Editor "blacks-out" sensitive text and also removes sensitive content within manually drawn redaction zones.
- Optionally, sensitive content may be noted by an Exemption Code stamp (see Figure).
- Areas identified for removal are approved through a semi-manual review.

Figure 3.1-1: Sensitive Text Replaced by an Exemption Code



```
NEUTRALIZED ONCE LOCATED, IRAQI BOTTOM-INFLUENCE MINES PRESENT A FAR
MORE DIFFICULT PROBLEM FOR US AND ALLIED MINE COUNTERMEASURE (MCM)
FORCES.
5.      THE IRAQI SURFACE THREAT CONSISTS OF NINE AGING OSA I/II
PATROL BOATS ARMED WITH THE SSN-2A/B ((b)(2) MISSILE; THE MISSILE HAS
A RANGE 25 NM. THESE BOATS ARE POORLY MAINTAINED, AND THE NUMBER OF
MISSION-CAPABLE UNITS IS UNKNOWN.
6.      IRAQI OSA'S DO NOT TRAIN FOR ANY TYPE OF TACTICAL
SCENARIO AND HAVE NOT FIRED A (b)(2) MISSILE SINCE THE EARLY 1980'S.
THEIR ABILITY TO CLOSE AND SUCCESSFULLY ENGAGE US WARSHIPS IS
SUSPECT.

                         HOWEVER, EVEN IF IRAQ HAS ACQUIRED THE
MORE ADVANCED (b)(2) THE NAVY HAS NEVER PRACTICED OVER-THE-HORIZON
TARGETING AND WILL PROBABLY BE LIMITED TO LINE-OF-SIGHT ENGAGEMENTS.
7.      DURING THE FIRST DAYS OF ITS INVASION, IRAQ CAPTURED 6
(b)(1) KUWAITI PATROL BOATS (1 LURSSEN FPB-57 AND 5 LURSSEN
TNC-45'S), AND A CACHE OF MORE THAN 100 MM-40 (b)(1) MISSILES. THE
NAVY IS APPARENTLY OPERATING AT LEAST ONE OF THE TNC-45'S, BUT ITS
ABILITY TO SUCCESSFULLY FIRE AN (b)(1) IN COMBAT IS DOUBTFUL, GIVEN
THE NAVY'S GENERAL LACK OF PROFICIENCY AND ITS UNFAMILIARITY WITH
THE SYSTEM. NONETHELESS, THE THREAT POSED BY THE CAPTURED KUWAITI
BOATS SHOULD NOT BE DISCOUNTED.
8.      IRAQ CURRENTLY HAS CSSC-3 (b)(2) COASTAL DEFENSE
MISSILES DEPLOYED TO TWO SITES, AL FAW IN EASTERN IRAQ AND RAS AL
QULAYAH NAVAL BASE IN KUWAIT. THESE SITES COVER BOTH THE IRAQI AND
KUWAITI COASTLINES AND ARE INTENDED TO DEFEND AGAINST THE APPROACH
OF ALLIED WARSHIPS AND A US AMPHIBIOUS ASSAULT.
```

As discussed previously, searchable-image PDF files may have sensitive text in both the hidden text layer (from the OCR process) and in the internal image layer of that text. When sensitive hidden text is found, the Knowvation DX Redaction Editor removes sensitive content. **When a sensitive image is found, the Redaction Editor removes the sensitive portion of the image and replaces it with a fill color (white by default) and an exemption code stamp**. Because PDF supports the layering of images on a page, there may be more than one image that is used to display and print text. Knowvation DX reliably removes the redacted portion of the image in every image layer used, regardless of the layering order. The image of the sensitive text in the redacted PDF file can in no event be recovered by extracting the images from the PDF and manipulating them.

5.2 Sanitization

The Knowvation DX PDF SanitizerTM currently runs as a batch process to perform sanitization, but it will become tightly integrated into the Knowvation DX application ingestion process in the next version (v1.3). The **PDF SanitizerTM supports removal of optional PDF objects, scrubbing of hidden data and the disruption of steganographic data.**

5.2.1 Removing Optional Objects

As discussed in Section 2.2, PDF may include many types of objects within a PDF file. The PDF file format specification ascribes special properties to certain of these types of objects. These properties are used to distinguish between the various types of objects that are defined by PDF. For each of these types of objects, the PDF file format specification specifies mandatory and optional properties. For a given type of object, properties that are outside of this set of mandatory and optional properties can be considered a potential source for carrying sensitive information that would not be visible to the end user.

Once the valid set of internal optional objects for an input PDF file has been determined, the **PDF SanitizerTM removes sensitive information contained within those non-standard objects from the PDF file then proceeds to the mark-up and redaction processes**. **First, the objects' references within the PDF file are removed; second, the "referenced links saved" operation is executed to create a new PDF file in place of the original.** This operation effectively rewrites the document to the binary file system from scratch, including only those objects for which a current reference within the PDF file exists. All objects that are no longer referenced within the PDF file are eliminated, which ensures that the scrubbed objects will have no binary representation. This means that any sensitive information that may have been removed through scrubbing can no longer be reconstructed, even by knowledgeable PDF experts using the most sophisticated PDF toolkits and binary file editors. The sensitive information will be irrevocably removed from the file.

5.2.2 Scrubbing Internal Data

The internal layers within PDF documents are one area where sensitive data may be hidden. Sensitive data within internal layers is removed by the Knowvation Redaction Editor. The other major form of internal data is comments. **Comments can be used to embed sensitive information into the PDF file that is completely undetectable to everyone except a knowledgeable PDF expert equipped with a binary file editor.** This exposure makes it necessary to include scrubbing of these comments as a mandatory step in the general mechanism for expunging sensitive content from the PDF file. **The "referenced links saved" operation to create a new file, discussed above, will also scrub PDF comments.**

5.2.3 Disrupting Steganographic Data Hidden in Image Objects

While this third category of hidden data would be extremely unlikely, a planned enhancement to the Knowvation PDF sanitizer will deal with this area. Although it may not be possible to detect all steganographic data that is hidden within raster images, it is possible to prevent the decoding of that data. **Version 2.0 of the Knowvation DX PDF SanitizerTM will include "Disrupter" programs that modify the images within a PDF such that decoding any data or messages that are steganographically encoded is disabled**, while still leaving the image in nearly the original display quality. The Disrupter will recurse through the source code of a PDF file, finding and modifying all raster images found within the file. The Disrupter will not attempt to detect steganographically hidden data in raster image objects (a virtually impossible task). Rather the intent is to disrupt by modifying every raster image object in the PDF file.

There are a variety of **steganographic techniques** employed for digital imagery. The two most common ones, and the technical approaches **used by the Disruptor** to defeat them, are:

1. Image data is communicated through a source image in a way that is obscure to the normal viewer, typically as a watermark or mask - the **Disrupter will essentially punch holes in the source image, by replacing source pixel data with alternate source pixel data**. For example, it might replace every 10th pixel. This has the effect of altering the hidden image. If the obscured image is encoded in a binary fashion, and woven through the source image, then pixel substitution will break that encoding and corrupt the obscured image.
2. Binary data is communicated through the source image's pixel-level data, typically in the least-significant bits, so as to be unrecognized by the naked eye **- the Disrupter will alter the low-order bits for the color components associated with the source image pixel data**. The nature of the alterations must be subtle and slightly randomized so as to not be reversible. For example, in a 24-bit RGB encoding, there are 3 bytes of data per pixel, one byte for each color component. Steganographic techniques can use the low-order bit in each byte to express one bit of data. For example an ASCII string, with 8 byte character encoding can be spread across the RGB components of a pixel at a rate of 3 bits of ASCII data per pixel. Because the substitution is made in the low-order bit, the steganographic content is not disruptive to any but the most discerning eye. To defeat this, the Disrupter will go through the low-order bits of each color component, and add or remove one on a randomized basis, effectively scrambling the steganographic content, rather like moving a magnet over a hard disk.

It should be noted that each of these two approaches bolsters the other, such that a solution that employs both will serve to more effectively defeat even more steganographic methods.

5.2.4 Knowvation DX PDF SANITIZER™ ROADMAP

Knowvation DX PDF Sanitizer™ was initially built to address removal of basic non-visual PDF information such as document metadata, annotations and comments. PDF Sanitizer™ is currently on version 1.2, which provides for sanitation of common PDF file artifacts. Version 1.4 will address a host of additional capabilities and will work in tandem with visible and hidden text redaction. Table 4-1 provides the roadmap for the Knowvation DX PDF Sanitizer™.

Table 4-1: Knowvation DX PDF Sanitizer™ Road Map

Optional Object	PDF Sanitizer™ 1.2	PDF Sanitizer™ 1.4
Document Metadata	✓	
Object Metadata		✓
Embedded Content and File Attachments	✓ (file attachments)	✓ (embedded content)
Annotations and Comments	✓ (visual removal)	✓ (full expunge)
Form Fields		✓
Hidden Text – Not Searchable Image		✓
Hidden Text – Searchable Image		✓
Hidden Layers		✓
Bookmarks		✓
Embedded Search Index		✓
Digital Signatures		✓
Adobe Reader Extensions		✓
Article Threads		✓
Links, Actions and JavaScripts	✓ (initial removal)	✓ (full expunge)
Watermarks		✓
Stamps		✓
Bates Numbers		✓
Typewriter Text		✓
Sticky Notes		✓
Highlighted, Crossed Out, or Underlined Text		✓
Buttons		✓
Multimedia (Video, Sounds, SWF, 3D objects)		✓
Drawing Markups		✓
Document Markups – Third Party		✓
Embedded Print Settings	✓ (initial removal)	✓ (full expunge)
Overlapping Objects, obscured text and images		✓
Embedded Images with Reduced Display Dimensions		✓
Embedded Thumbnail Images		✓
Alternate Images		✓
PDF Opening View		✓
Automatic Page Advancement and Page Transitions	✓	
Deleted Hidden Page and Image Content		✓
Steganographic Data Removal		✓

The design of the Knowvation DX PDF Sanitizer™ is inclusive of all the functions recommended by NSA and implemented in Adobe Acrobat. Our design then goes considerably further and deeper. The product will support removal of optional PDF objects, scrubbing of hidden data and the disruption of steganographic data. When complete, Knowvation DX PDF Sanitizer™ will be the most robust PDF Sanitizer on the market.

5.3 About PTFS

PTFS is a leading Content and Knowledge Management solution provider. Founded in 1995, PTFS has focused on developing enterprise content management solutions for Federal, state, and local government organizations as well as commercial entities. Staffed with knowledge management experts, PTFS understands the difficulties of managing content across an enterprise. Knowvation has been in production for over 15 years, and it's that experience that makes PTFS an industry leading ECM Solution Provider.

5.4 Primary References

- ***Acrobat X Pro Online Help: Removing sensitive content*** **-** http://help.adobe.com/en_US/acrobat/pro/using/WS4E397D8A-B438-4b93-BB5F-E3161811C9C0.w.html
- ***Acrobat X Pro Video on Removing Hidden Information*** **-** http://acrobatusers.com/auc/content/tutorials/acrobat_x/removing-hidden-information.php
- ***Hidden Data and Metadata in Adobe PDF Files: Publication Risks and Countermeasures*** - Enterprise Applications Division of the Systems and Network Analysis Center (SNAC) Information, Assurance Directorate, National Security Agency http://www.nsa.gov/ia/_files/app/pdf_risks.pdf
- ***Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word to PDF*** - Architectures and Applications Division of the Systems and Network Attack Center (SNAC), *Information Assurance Directorate,* National Security Agency http://www.fas.org/sgp/othergov/dod/nsa-redact.pdf
- ***NSA/CSS Storage Device Declassification Manual*** (Supersedes NSA/CSS Manual 1302, dated 10 November 2000.) National Security Agency – http://www.nsa.gov/ia/_files/government/MDG/NSA_CSS_Storage_Device_Declassification_Manual.pdf
- ***Sanitization (classified information)*** From Wikipedia - http://en.wikipedia.org/wiki/Sanitization_%28classified_information%29

ATTACHMENT 1 – TYPES OF CONTENT THAT CAN BE HIDDEN IN A PDF

Objects and Artifacts	Description
Document Metadata	Metadata includes information about the document and its contents, such as the author's name, keywords, and copyright information, used by search utilities. Document Metadata is stored in XMP and in the Info Dictionary. PDF Sanitizer can delete all document metadata, or all except for specific schemas and properties that are specified in advance. For example, retain the PDF/A Schema, retain the Title property value, but remove all other document metadata.
Object Metadata	Metadata that includes information about the object it is attached to, such as IPTC Metadata for an image object, or the information dictionary that is attached to an Article Thread.
Embedded Content and File Attachments	PDF files may contain a significant variety of different content types. These may either be embedded or attached. Embedded content generally appears and runs as part of a page in the file, even though third party applications might provide the functionality necessary to support the content. Attached files generally require that users open the file with an external application, and are not displayed as part of the PDF file.
Annotations and Comments	This item includes all comments that were added to the PDF using the comment and markup tools, including files attached as comments. To view comments in Acrobat, choose the Comments pane.
Form Fields – Stored Interactive Form Data	This item includes form fields (including signature fields), and all actions and calculations associated with form fields. All form fields are flattened and can no longer be filled out, edited, or signed.
Hidden Text – Not Searchable Image	This item indicates text in the PDF that is either transparent, covered up by other content, or the same color as the background. Does not include searchable text created by an OCR engine from the single image of the page. Hidden Text will often not be viewable in Acrobat.
Hidden Text – Searchable Image (optional)	This item indicates text that is created by an OCR engine from the single image of the page. To view Searchable Image Hidden Text in Acrobat, copy the text on the page and paste it into a text editor (such as Word). Searchable Image Hidden Text is not removed by default because Knowvation DX redaction will remove the sensitive areas of such text.
Hidden Layers	PDFs can contain multiple layers that can be shown or hidden. Removing hidden layers flattens remaining layers into a single layer. To view layers in Acrobat, choose View > Show/Hide > Navigation Panes > Layers.
Bookmarks	Bookmarks are links with representational text that open specific pages in the PDF. To view bookmarks in Acrobat, choose View > Show/Hide > Navigation Panes > Bookmarks. PDF Sanitizer can delete all bookmarks, or just some according to certain criteria. For example, delete bookmarks that open Article Threads or are Web Links, but retain the others.
Embedded Search Index	An embedded search index speeds up searches in the file. In Acrobat, to determine if the PDF contains a search index, choose View > Tools > Document Processing > Manage Embedded Index. Removing indexes decreases file size but increases search time for the PDF.
Digital Signatures	A *digital signature*, like a conventional handwritten signature, identifies the person signing a document. Unlike a handwritten signature, a digital signature is difficult to forge because it contains encrypted information that is unique to the signer.

Objects and Artifacts	Description
Adobe Reader Extensions	Adobe Reader Extensions allow the free Reader to perform functions otherwise only available in Adobe's Acrobat desktop software. Acrobat Standard and Acrobat Pro include technology that can enable PDF documents for Reader Extensions by using a digital credential. This credential is located within the Software ("Key") that is embedded in the PDF file.
Article Threads	In PDFs, *articles* are optional electronic threads that the PDF author may define within that PDF. Articles lead readers through the PDF content, jumping over pages or areas of the page that are not included in the article, in the same way that you might skim through a traditional newspaper or magazine, following one specific story and ignoring the rest. When you read an article, the page view may zoom in or out so that the current part of the article fills the screen.
Watermarks	A *watermark* is text or an image that appears either in front of or behind existing document content, like a stamp. For example, you could apply a "Confidential" watermark to pages with sensitive information. You can add multiple watermarks to one or more PDFs, but each watermark must be added separately. You can specify the page or range of pages on which each watermark appears. Unlike a stamp, a watermark is integrated into PDF pages as a fixed element. A stamp is a type of PDF comment, which others reading the PDF can open to display a text annotation, move, change, or delete.
Stamps	Stamp on a PDF page are applied in much the same way that a rubber stamp is used to stamp a paper document. A stamp can be predefined or custom. Dynamic stamps indicate name, date, and time information on the stamp.
Bates Numbers	*Bates numbering* is a method of indexing legal documents for easy identification and retrieval. Each page of each document is assigned a unique Bates number that also indicates its relationship to other Bates-numbered documents. Bates numbers appear as headers or footers on the pages of each PDF in the batch. The Bates identifier is referred to as a number, but it can include an alphanumeric prefix and suffix.
Typewriter Text	With the Typewriter tool you can type on top of any PDF document, even one you created from a scanner. This allows you to easily fill out paper forms on your computer and archive the results electronically, or send the completed form via email. Typewriter text may contain sensitive information.
Sticky Notes	The most common type of comment is the sticky note. A sticky note has a note icon that appears on the page and a pop-up note for your text message. You can add a sticky note anywhere on the page or in the document area. Sticky Notes may contain sensitive text.
Highlighted, Crossed Out, or Underlined Text	The Highlight Text tool, Cross-Out Text tool, and the Underline Text tool can be used to add comments by themselves or in conjunction with notes. As these are a form of comments added in Acrobat after a PDF is created, they may be undesirable in a declassified document.
Buttons	Buttons are most commonly associated with forms, but you can add them to any document. Buttons can open a file, play a sound or movie clip, submit data to a web server, and much more.
Multimedia (Video, Sounds, SWF, 3D Objects)	Adding video, sound, and interactive content transforms PDFs into multidimensional communication tools that increase interest and engagement in your documents.

Objects and Artifacts	Description
Drawing Markups	Drawing Markups are graphic overlays. The Rectangle tool □, the Oval tool ○, the Arrow tool ↗, and the Line tool ╱ are simple shapes. The Cloud tool ☁ and Polygon tool ⬠ create closed shapes with multiple segments. The Polygon Line tool ⌒ creates open shapes with multiple segments. The Pencil tool ✎ creates free-form drawings. These markups may be inappropriate in a declassified document.
Document Markups added by Third-Party Plug-ins and Applications	Document markups created by Acrobat are the de-facto standard for PDF annotations, but third-party applications and Acrobat plug-ins provide their own sets of markups. Usually these can be displayed in Acrobat with the assistance of a plug-in supplied by the third-party developer. These markups are inappropriate for broad public distribution of declassified documents, where the reader will not have the required plug-in for viewing.
Embedded Print Settings	Removes embedded print settings, such as page scaling and duplex mode, from the document. These are inappropriate for broad public distribution of declassified documents.
Overlapping Objects, Obscured Text and Images	Text can be obscured in a number of different ways. For instance, white text on a white background (or other text of any color matching the background color) could be hidden, but still be extractable if all the contents are copied and pasted into notepad. The same technique can also extract text that is inadvertently hidden behind images. This can be a tricky issue in document management, because objects copied and pasted from one program into another prior to the PDF conversion process could contain text that is not easy to detect. Images have a similar risk because they may be hidden behind other images.
Embedded Thumbnail Images	Removes embedded page thumbnails. Embedded Thumbnails are sometimes useful for large documents, which can take a long time to draw page thumbnails after you click the Page Thumbnails button. Embedded Thumbnails show reduced images of pages prior to redaction.
Alternate Images	This removes all versions of an image except the one destined for on-screen viewing. Some PDFs include multiple versions of the same image for different purposes, such as low-resolution on-screen viewing and high-resolution printing.
PDF Opening View	The initial view of the PDF depends on how its creator set the document properties. For example, a document may open at a particular page or magnification. Removing the Opening View settings returns the document to default viewing behavior.
Automatic Page Advancement and Page Transitions for Full Screen Mode	In Full Screen mode, only the document appears; the menu bar, toolbars, task panes, and window controls are hidden. A PDF creator can set a PDF to open in Full Screen mode. Full Screen mode is often used for presentations, sometimes with automatic page advancement and transitions.
Deleted Hidden Page and Image Content	PDFs sometimes retain content that has been removed and no longer visible, such as cropped or deleted pages, or deleted images. Deleted (unreferenced) hidden page and image content is hidden data from previous document saves in Acrobat, and may also have been created in authoring application and then transferred (encoded) into the PDF file.