clearswift
by HelpSystems

KOFAX

DATASHEET *(Cybersecurity)*

Clearswift Optical Character Recognition: Combating Data Loss in Images

With billions of records lost through data breaches across 2017 and 2018, organizations need to ensure that they can thoroughly scan all content and files passing in and out of a network through email. Today, this means more than just text.

Business Problem

Does your organization regularly convert Excel or Word documents to PDF file format? Do you have one of those multi-purpose photocopier / printer / scanners whereby hard documents are scanned, converted to PDF and stored or shared? Perhaps posted documents may need to have an electronic copy stored or shared so documents are scanned, converted to a PDF and emailed automatically within seconds. This means when a document is scanned, a picture is created for each page and the pictures are stored in the PDF.

Most organizations today use either or both of these information processing methods as standard day-to-day practice. However, from a DLP perspective, PDF documents could be sent around the organization and outside it, with relative impunity as traditional DLP solutions cannot detect the sensitive information contained within these files; the text, inside the images, inside the PDF.

PRODUCT SUMMARY

Products

OCR is an optional, priced module for the following products:

- Clearswift Secure Email Gateway (SEG)
- Clearswift ARgon for Email
- Clearswift Secure Exchange Gateway (SXG)
- Clearswift Secure Web Gateway (SWG)

Support

Clearswift provides 24x7 global support as standard, with additional options for premium support.

This data loss risk doesn't just apply to PDF's, it applies to all types of image formats including screenshots or images (eg. JPG, BMP, GIF, PNG and TIFF) embedded in other files such as Microsoft Office.

How can we scan for sensitive content?

Optical Character Recognition (OCR) is the process of detecting and extracting text from an image file, an image embedded within an electronic document, or a scan of a document.

The OCR process examines the text image and creates computer-editable text from scanning the tiny dots (pixels) that together form a picture of text. So as the OCR engine scans the pixels it builds up what it believes to be a letter, see Figure 1: Pixels become letters become text. This is then matched through a series of pattern and alphabet matching to a corresponding letter.

Figure 1: Pixels become letters become text

The letters are then combined by looking at where the spaces, punctuation and end of a line occurs, and the words checked against language dictionaries to detect the appropriate words. The extracted text is then processed by Clearswift's Deep Content Inspection (DCI) and policy engine to determine if there is sensitive information displayed in the image. With our DCI engine recursing by default to 50 levels deep, the image can be embedded in an Excel spreadsheet, which is embedded in a Word document, which is then shared via a ZIP archive attached to an email. If found, then the extracted text is processed by the Clearswift Adaptive Data Loss Prevention (ADLP) functionality. For example, the attachment with the image containing sensitive data, is blocked.

A further enhancement to OCR analysis enables redaction of text in images, removing only the information which breaks policy by drawing a black box across the words. Clearswift will detect the image, analyse it and redact any sensitive information, allowing the 'safe' file to continue to the recipient, even if the image has been embedded in a Word document or in a Zip file.

Deployment

Our Clearswift Secure Email Gateway (SEG), Secure Exchange Gateway (SXG), ARgon for Email, and Secure Web Gateway (SWG) security solutions have a cost option for OCR to mitigate the risk of data loss through images. It supports multiple languages, enabling it to be easily used by global organizations who operate using more than one language.

The use of language specific dictionaries reduces the number of false positives and increases the recognition rate. The architecture behind SEG has always been scalable, with multiple instances being able to be peered together for both scalability and availability.

The introduction of the OCR option makes use of the scalability, with more instances being able to be added so that even though more processing is happening, the overall throughput of the system can be maintained.

Features

Designed to scale to enterprise deployments, the system provides:

- Granular policy control – allows OCR usage by direction and by sender / recipient combination
- Can scan embedded images, images attached to messages, images in compressed attachments and embedded images in common "Office files" such as MS Office, PDF and LibreOffice
- Supports over 20 different image file formats such as JPEG, PNG, BMP, GIF, and TIFF
- High performance, images scanned <1s (sample JPEG 300Kb)
- Support for 48 language dictionaries (including Chinese, Japanese and Cyrillic)
- Minimum image size is 16x16 pixels; Maximum 8,400 x 8,400 pixels
- Maximum angle of skew 15°; 180° rotation supported
- Supports 8pt text
- Supports most color combinations text / background (sufficient color separation is required)

Cost

Optical Character Recognition (OCR) is a cost option for the Clearswift Secure Email Gateway, Secure Exchange Gateway, ARgon for Email, and Secure Web Gateway (SWG) products, and available for the Secure ICAP Gateway later this year. The cost is dependent upon your installation and throughput processing requirements.

Contact us today for more information.

clearswift
by HelpSystems

www.clearswift.com

About HelpSystems

HelpSystems is a people-first software company focused on helping exceptional organizations Build a Better IT™. Our holistic suite of security and automation solutions create a simpler, smarter, and more powerful IT. With customers in over 100 countries and across all industries, organizations everywhere trust HelpSystems to provide peace of mind. Learn more at www.helpsystems.com.