AERSTONELABS

[SIFT: Automated Metadata Tagging]

Jason Winder, Managing Partner, Aerstone Labs

jason.winder@aerstone.com | 301-257-5931 [Unclassified]

jason.h.winder@coe.ic.gov | 578-9976 [Classified]

Tuesday, June 13, 2017

Table Of Contents

1 Introduction

One of the key strategic goals of information management is to reduce the cost of retrieving information. Information should be delivered to the people who need it, where they need, when they need, in a format that they can consume. This is an extremely difficult problem however, noting that only the least valuable information tends to be put in motion easily. The most valuable information tends to gather at the bottom of deep "data gravity wells" that resist information retrieval. The reasons for this are several-fold.

> "Compared to the cost of moving bytes around, everything else is free."
>
> -- Dr. Jim Gray, Microsoft

2 Background

Most organizations maintain a large corpus of valuable information, across multiple security enclaves, that isn't properly tagged with useful metadata – i.e., useful and actionable information about file contents. This practice creates a severe organizational risk that many organizations underestimate, however the heart of this problem is the tension between a desire for data *confidentiality*, against a need for data *availability*.

2.1 Confidentiality

Every organization has valuable information. Some of this information is sensitive, legally or otherwise – including personally identifiable information (PII) like customer social security numbers and credit card information. Organizations also usually maintain sensitive information about their employees, such as salary information, or HIPAA-protected medical information. Organizations may further have strategic and financial information, and intellectual property, which shouldn't be disclosed outside of an explicit need to know. Most organizations err on the side of caution, and implement a series of mechanisms to protect sensitive data from being put in motion – including policy, hardware, and software solutions.

At one end of this spectrum, organizations with a single flat network frequently protect network shares or content repositories with discretionary access controls, and limit access to entire systems (e.g., financial or HR systems) to a small set of staff. At the other end of the spectrum, complex classified multi-level security environments implement air-gapped security enclaves and sophisticated insider threat monitoring to ensure that sensitive information remains protected. In any case, most organizations rely only on *where* a file is saved, versus *what* a file contains, as the core protection against sensitive data in motion.

The underlying assumption with this approach is that rules will be followed, that technology will function correctly, that employees will not be sloppy, and that sensitive information ultimately won't be spilled onto inappropriate enclaves. These assumptions are incorrect, which leads to data breaches and unintentional data disclosures that can have a cataclysmic effect on an organization.

2.2 Availability

Even as organizations strive to protect sensitive data, we should acknowledge that this same information frequently has the most value to an organization: intellectual property, business intelligence, and mission data of various kinds. Attempting to deliver this information proactively (or even reactively) to the correct target audience is frequently thwarted by the same security controls desired to protect this information. And while automated full-text search hopes to solve this problem in aggregate, there are other aggravating factors that make useful information very difficult (or impossible) to find.

First, full-text search doesn't allow sophisticated content targeting. This comment should resonate with anyone who has ever attempted to search for a relatively common word or phrase in a massive content management system. Full-text search is a second-best solution, which hopes to fill the gap left by content creators who don't metadata tag their assets – even while acknowledging that manual tagging off assets simply cannot scale to exabytes of storages.

Second, many digital artifacts are not text searchable. This not only includes pictures, video, or scanned text PDF documents, but also files that contain pattern-based data – e.g., social security numbers, credit card numbers, or IP addresses. And while some OCR solutions exist that can extract text from scanned document files, these same OCR engines return exceptionally poor results with unstructured files (like images stamped with text). Any file with a low signal-to-noise ratio typically causes OCR engines to fail with under 10% accuracy.

3 SIFT™ Solution Overview

SIFT ("Secure Integrated File Transfer") is a browser-based file inspection and metadata tagging solution. It is designed first and foremost to protect an organization from accidentally spilling information onto unauthorized network enclaves. SIFT can be used as a stand-alone data transfer portal, or integrated seamlessly with existing document management systems or high assurance guards. Once configured to search for the kind of data an organization considers sensitive, SIFT can be used to implement data transfer approval workflow, and to optionally metadata-tag documents with any discovered keywords. These keywords may be static terms (like TOP SECRET or TREADSTONE), regular expression-based patterns (like social security numbers, or IP addresses), or specific shapes in pictures or videos (i.e. faces, buildings, tanks, text, etc.). SIFT can also be used by system administrators to scan specific network locations for spilled documents, and ships with a RESTful API that allows it to be implemented in line with document management solutions, including high assurance guards, content management systems like Adobe Experience Manager (AEM) or Microsoft SharePoint, or email gateways. Detailed reporting data provides valuable real-time audit data about document transfers, and can be ingested into agency data visualization tools.

3.1 Technology & Performance

SIFT is built on open-source technology, combined with patent-pending algorithms that specifically enhance OCR fidelity. These open-source technologies include the Google Tesseract OCR engine, as well as the Google Tensor Flow machine learning (ML) platform. SIFT performs well out of the box, with OCR accuracy rates that are as high as 10X better than native Tesseract scanning, but can also be trained to perform even better against predictable asset formats. The underlying OCR/ML engines can also be swapped out for other engines as desired. In terms of throughput, SIFT's processing speed is heavily dependent on server memory and document size/complexity. As a baseline, a 1-page PDF or JPEG file will be scanned in <1 sec on a dual core 8GB system.

3.2 File and Language Support

SIFT ships with natively support for hundreds of languages, as well as support for all the common file formats – including MS Office and Adobe Acrobat, multiple picture formats, MP4 video, zipped archives, plain text documents, and web format files. SIFT performs recursive scanning on zip archives, as well as on OLE-embedded documents in MS Office files. SIFT was designed in a highly modular fashion, and support for additional file types can be added in days (not weeks).

3.3 Platform Support

SIFT has been designed modularly, and can scale to support massively parallel deployments. SIFT has been successfully deployed on on-premises or cloud-hosted Windows or Unix hardware, Apache Spark clusters, the Pivotal Cloud Foundry (PCF) PaaS platform, and within Docker containers. SIFT is currently available in the unclassified AWS cloud, and will soon be available in the C2S/UC2S environments as well. Each component of the SIFT software, including the web front-end application, the ML/OCR application, the message queue, and the configuration database, can all be load-balanced or clustered (as appropriate) to support system load and availability requirements.

4 SIFT™ Use Cases

SIFT offers four core use cases, each of which addresses a different aspect of the data confidentiality and availability problem.

4.1 Metadata Tagging

Non-searchable documents such as pictures, scanned .PDF documents, and video are routinely added to enterprise or web content management systems, without being stamped with useful metadata tags. This leads to a number of limitations, including:

1. **Enterprise Search.** These assets cannot be picked up by enterprise search, which assumes the underlying systems that own and contribute data are performing metadata tagging.
2. **Attribute Based Access Control (ABAC).** Assets that are not correctly metadata tagged with the appropriate CAPCO markings cannot be controlled using ABAC solutions.
3. **Data Loss Prevention.** Untagged assets cannot be prevented in an automated way from exfiltrating the agency network.
4. **Content Targeting.** The ability to target content proactively to data consumers is predicated on metadata tagging. Unsearchable content cannot be served to end users in any automated fashion.
5. **Digital Rights Management (DRM).** Once assets are correctly tagged, DRM solutions can be implemented to control the actions that users may take (forward, print, etc.) against sensitive files.

4.2 OCR/ML Microservice

Many organizations already have a variety of data management and analysis tools in place, including content management systems, and business intelligence toolsets. SIFT's RESTful API can be consumed by these services to add an OCR or ML capability to existing processes. This allows organizations to take complete advantage of the full set of information acquired or generated, whether structured or unstructured.

4.3 DTO Process

Many classified environments have a formal Data Transfer Officer (DTO) program in place, in order to limit and control the flow of information from high to low. In most cases, these programs are highly distributed across the organization, without centralized control or auditing. Moreover, the tools used to support DTO efforts are not capable of analyzing unsearchable assets, cannot handle pattern-based text, and cannot be configured to recognize shapes or images. SIFT allows organizations to centralize their DTO function, to ensure proper reliability, control, and oversight.

4.4 Spill Remediation

It is exceptionally difficult to search and monitor network shares or cloud object storage repositories for spilled data. Without proper metadata tagging, unsearchable files can easily evade OS-level content searches. SIFT supports both manual and scheduled searches for files, across any organizational network location. This capability helps organizations proactively detect and remediate data spillage before financial or legal losses occur.

5 SIFT™ Summary

SIFT's value proposition effectively boils down to two basic use cases:

- **Does your organization have valuable documents that aren't properly metadata-tagged?**
  SIFT turns unsearchable and unprotectable files into highly searchable and protectable assets, by:

  - **ASSESSING** the contents of searchable AND non-searchable assets
  - **STAMPING** files with tags based on explicit keywords AND dynamic data patterns
  - **INTEGRATING** easily with other document analysis and translation solutions

- **Does your organization have different security enclaves, or users with different needs to know?**
  SIFT supports the advanced protection of sensitive files, by:

  - **SERVING** as a centrally-managed cross-enclave transfer process
  - **ENABLING** dissemination control techniques, like ABAC and DRM
  - **INTEGRATING** easily with other content publishing or data transfer solutions
  - **SUPPORTING** the regular scanning of organizational file shares for spilled data

---

Secure By Design

Aerstone Labs develops advanced enterprise security software solutions that solve complex problems for large businesses and institutions. Our commercial focus is on developing patentable techniques that can enhance the confidentiality, availability, or integrity of data in enterprise environments. In many cases, solutions in a particular space tend to converge on each other in both features and functionality. Aerstone Labs delights in rethinking the challenges of the enterprise security process, and identifying core capabilities that existing markets have failed to address. Whether as stand-alone solutions, or enhancements to existing technologies, our technologies provide meaningful enhancements to enterprise security posture — and change the game for other software in our space.

For a free demo or consultation, contact us today: info@aerstonelabs.com