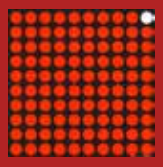

BRIS201502-PII

Boosting PII Detection and Protection in "Unstructured" Content

Summary

Regulatory mandates from countries around the globe provide privacy protection for personal information. Detecting and protecting PII is particularly challenging when it is contained in "unstructured" content, i.e., in the documents and user-generated files contained on file shares, personal computing devices, and content management systems. This paper discusses how visual classification can increase an organization's ability to first detect PII in unstructured content and then provide more options for protecting it.

Contents

Boosting PII Detection and Protection in Unstructured" Content

Introduction

Because of the significant reputational and financial consequences of failing to protect content containing personally identifiable information ("PII"), corporations and governmental agencies have made it a major goal to identify and protect such content. Privacy expectations arise from a number of laws in different jurisdictions and are sometimes referred to by various acronyms such as HIPAA or PCI, but we will refer to them collectively in this posting as "PII."

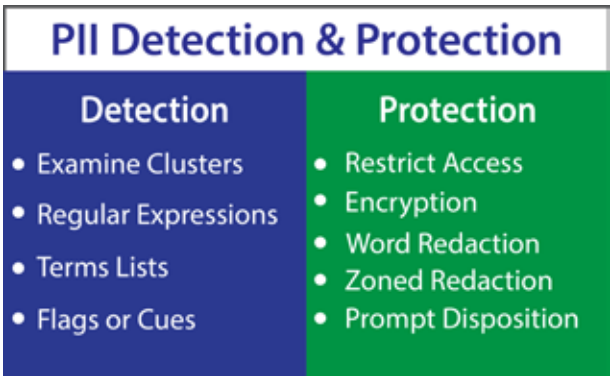

Figure 1. PII Detection & Protection Overview

The Challenge of Unstructured Content

One of the most challenging aspects to identifying and protecting PII is how to deal with "unstructured" content, i.e., with documents or files on file shares, personal computing devices, and content management systems. These files can be generated within and outside the organization using many applications, can be converted to multiple file formats (most commonly to PDF), and seemingly have unlimited form and content.

By contrast, structured data like those in databases and support systems have defined fields in tables that have defined relationships with each other. To protect social security numbers in a database, you control access to the field for social security numbers. With documents things aren't that simple.

Pattern Searching for PII

For example, most PII detection systems provide the ability to look for social security numbers using an expression like "NNN-NN-NNNN" where N is a digit between 0 and 9. For that to work there has to be accurate text associated with the documents being searched, but in many corporate collections a sizable proportion of documents does not have associated text.

Documents without text can be created when certain software applications save files to PDF but do not automatically include editable or searchable text. Further, scanned TIF or PDF documents may not have accurate text or any text. The documents display fine on screen or when printed, but the numbers and letters are just pixels to the system, it doesn't know which characters were intended.

Social security numbers in such documents are invisible for purposes of looking for "NNN-NN-NNNN." Even where there is text, the text may be inaccurate. For example, if a zero ("0") is actually a capital "oh" ("O"), or what should be a one ("1) is a lower case "el" ("l) or a capital "eye" ("I"), the search algorithm will fail. And of course, handwritten content will not be recognized at all.

Lists of PII Terms

Even where there is perfect text, many times PII does not occur in predictable text patterns, for example, first and last names, addresses, and account passwords come in a virtually infinite number of combinations. Some PII elements like medical diagnoses could be contained in a long list of possible terms, but those would again depend on the presence of accurate text, and just because a term was a diagnosis would not mean it was used in a setting where it would be PII.

Using Visual Classification to Identify PII Document Types

Visual document classification provides a whole new approach to meeting the challenge of detecting and protecting PII in "unstructured" content.

A key point is that PII does not occur at random across document types or within documents. It's like gold, it occurs in veins or in ore, and visual classification is very useful in identifying the PII veins and ore present in the mountains of unstructured content maintained by large organizations.

Visual classification works by clustering all documents based on their visual similarity so that the document types that normally contain PII can be identified. This approach ignores whether text is available and focuses on appearance, thereby normalizing documents regardless of the type of file in which the content was stored.

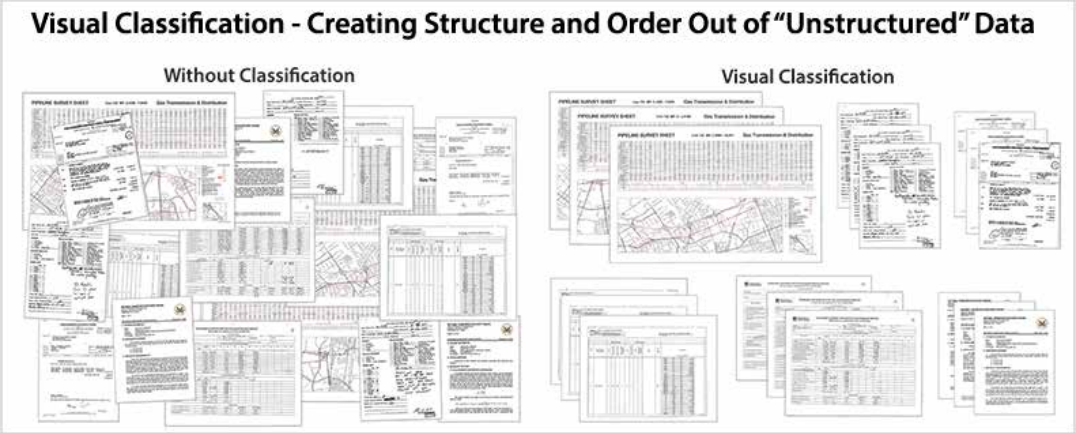

Figure 2. Visual Classification Automatically Clusters Electronic Documents and Scanned Documents Based on Visual Appearance

Reviewing Visual Clusters for Multiple Information Governance Purposes

Typically the number of visually-similar clusters is less than one percent of the total number of documents, and the clusters can be viewed starting with the largest clusters first. Within a few days, the PII detection team can:

- Review clusters representing well over 99% of all the documents in an organization,
- Eliminate clusters that do not have ongoing business or legal value,
- Tag those remaining clusters that contain PII,
- Assign document-type name labels to them, and
- Identify the PII attributes or data elements present in each document type.

Subsequent reviews only have to examine new clusters of visually-similar documents that have formed since the time of the last review. Decisions made about existing clusters are simply applied to documents

that are later added to the cluster.

Whether or not those documents had associated text values, after the review of visually-similar clusters, the organization can now decide what types of protection is warranted for each type of document:

- What level of storage is indicated, e.g., should some of the clusters be on encrypted servers?
- Which people or job functions should be able to see specific document types?
- What retention period should apply?

Visual Classification and Text-Based Approaches Are Complementary

Visual classification is not used to the exclusion of text-based approaches. In fact text-based pattern and term searching techniques can be used in conjunction with visual classification to provide the most comprehensive detection and protection options available.

After visual clusters are formed, searches can be made for patterns like social security numbers or for lists of potential PII like medical diagnoses. The results are then viewed arranged by visual cluster to determine whether some of the clusters that were not originally tagged as regularly having PII ought to included in the PII category of clusters.

Note that even if not all documents in a cluster have associated text, the ones that have text can be identified as having PII and this can result in all the documents in the cluster receiving the additional PII protection they warrant.

Detecting PII Flags or Cues

Text search can also locate words that often serve as flags or cues for PII. For example, the terms "SSAN" or "SS#" or "Social Security Number" will often serve as flags that the information close by includes social security numbers. If documents cannot be sorted or arranged by visual cluster or document type it could be very burdensome to review the results of such a search because there can be so many hits. However, when the results can be reviewed by cluster or document type, attention can be focused only on those clusters that have not already been designated as containing PII.

Protective Measures

Once documents or clusters have been identified as having PII they can be afforded the appropriate level of protection. These include:

Encrypted Storage

Encrypting data helps protect it and lowers regulatory risks in the event of a data breach. However, organizations may not want to encrypt everything they have. Visual classification greatly reduces the content that is kept because much content can be disposed of, and then only a portion of what is retained may warrant encryption.

Restricted Access

Having consistent, reliable document type classification of all stored content permits organizations to restrict employees' access to only those documents they need to perform their jobs.

Prompt Disposition

Without consistent document classification, many organizations end up keeping everything either forever or for the longest retention period associated with any of the documents in a collection. Consistent, reliable classification permits granular retention schedules that can be readily applied, considerably reducing the volume of content at risk.

Three other protective options ought to be considered:

Text Redaction

BeyondRecognition's visual classification system is based on cataloging the graphical elements on all pages. As part of that process it content-enables image-only documents to provide searchable text. Whether it provided the searchable text or the text was already present when the documents were processed, BR knows the page coordinates for the text values associated with the pages. It uses those coordinates to perform high-speed, highly-accurate redactions using expressions or word lists, on the order of 700,000 redactions per CPU per hour.

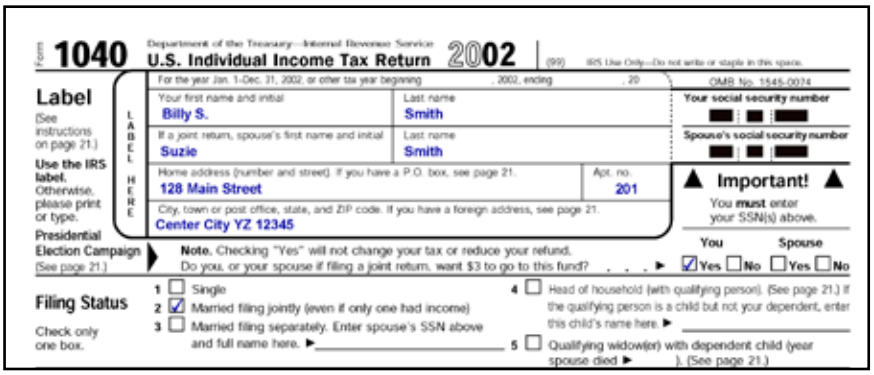

Figure 3. Example of Text-Based Redaction

In this example, the social security numbers were redacted based on the presence of character strings of "NNN NN NNNN" where N is a digit between 0 and 9.

This industrial-grade redaction capability means that when producing or turning over documents to third parties, the PII can be simply removed. It also provides the organization with the option to work with redacted copies of documents where circumstances warrant. The obvious benefit is that people can't steal or inadvertently disclose PII that isn't there.

Zoned Redaction

Many times some forms may be completed with handwriting that is not susceptible to text or word-based redaction. Other times, the words in a part of a document are too variable to be able to specify what patterns or words will be used. And as already discussed, some documents do not have accurate, reliable text.

In these circumstances, BeyondRecognition can also provide zoned redaction where all of the content that falls within certain page coordinates will be redacted.

Figure 4. Example of Zoned Redaction

In this example, IRS Form 1040s are placed in the same visually-similar cluster and the social security number entry area was zoned for redaction.

Logging Redactions

Regardless of which type of redaction is used, BR provides a detailed log of all redactions, including what was redacted and the reason given for the redaction. Both text and image layers in redacted documents are redacted.

Expedited Disposition

Often documents are used to collect information that is then input into a database or decision support system. If organizations knew the information was accurately entered, the documents could be viewed as transitory and be scheduled for immediate or expedited disposition.

Visual classification technology provides an automated way to validate that the information on specific documents were added correctly and to then flag those documents for expedited disposition. As detailed on the BR website, automated attribute extraction can pull specified data elements and format them. These data values could then be checked against the values maintained in databases or decision support systems.

[Continued next page.]

A Final Word

BeyondRecognition is the developer of visual classification technology and that technology is available through the BR network of companies.

Detecting and protecting PII is one of several major document-centric information governance initiatives that are dependent on consistent document classification. Others include records retention and disposition, file share remediation, content migration, silo consolidation and digitization. The good news is that visual classification can serve as the foundation for all of those initiatives, making the most effective use of the time and energy invested in reviewing and classifying an organization's otherwise "unstructured" documents.

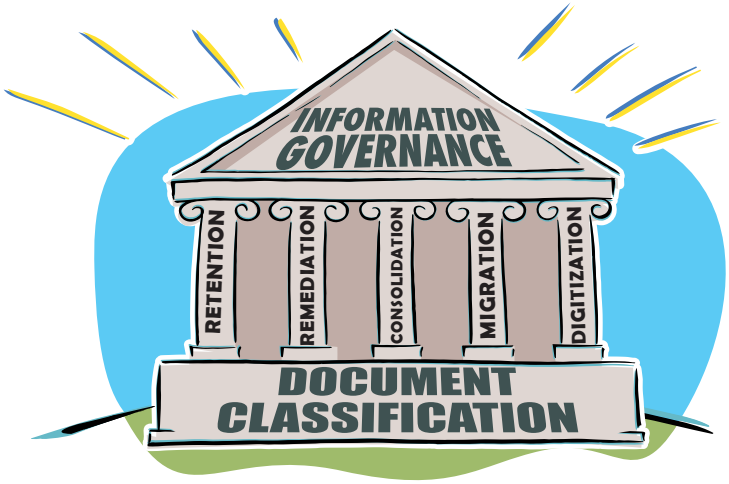

Figure 3. Information Governance Pantheon

Document-centric information governance initiatives all rest on or depend upon consistent document classification. That includes assigning retention periods, remediating file shares, consolidating silos, migrating content, and archive digitization.

More Information

For more information on how BeyondRecognition can help you detect and protect PII in your organization or help you with your other document-centric information governance initiatives, please visit our website at www.BeyondRecognition.net or contact us at IGDoneRight@BeyondRecognition.net.

