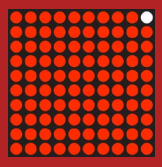


BRTO201501-01 VC

Visual Classification: The Emerging Foundational Technology for Document-Related, Classification-Dependent Information Governance Initiatives

Key Takeaways

"Classification-dependent information governance initiatives" are those initiatives or undertakings whose successful completion depends on the consistent classification of the documents involved. Examples include assigning retention and disposition schedules, file share remediation, content migration, silo consolidation, and determining the appropriate storage location, security level, and access rights for stored content.

Comprehensive document classification technologies will (1) deal with all of an organization's documents, both native electronic and scanned paper documents, and will (2) cover the entire process from collection through evaluation, designating document types, extracting attributes, redacting where indicated, and loading into designated content management system.

Visual classification provides consistent, enterprise-scale classification of all of an organization's documents both electronic and paper, and serves as a knowledge base of decisions made for each cluster regarding retention, document type designations, and attributes extracted. Critical factors leading to increasing adoption:

1) Clustering is based on visual appearance of documents regardless of the type of file containing the documents, thereby normalizing them for comparison purposes. All other technologies are text-dependent and do not address all of an organization's documents.

2) Visual clustering is completely automatic, greatly speeding project launch and completion.

Contact BeyondRecognition at IGDoneRight@BeyondRecognition, LLC or visit www.BeyondRecognition.net

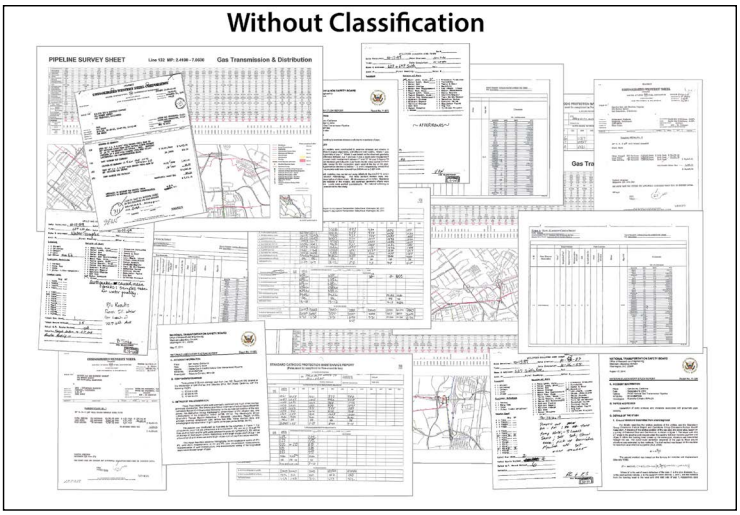
Purpose

This document provides an overview of BeyondRecognition's new visual classification technology for information governance technologists and practitioners who want to advise clients on the current state of the art as well as information governance stakeholders who want to undertake document-related information governance initiatives without the restrictions and limitations of prior technology. These initiatives could include:

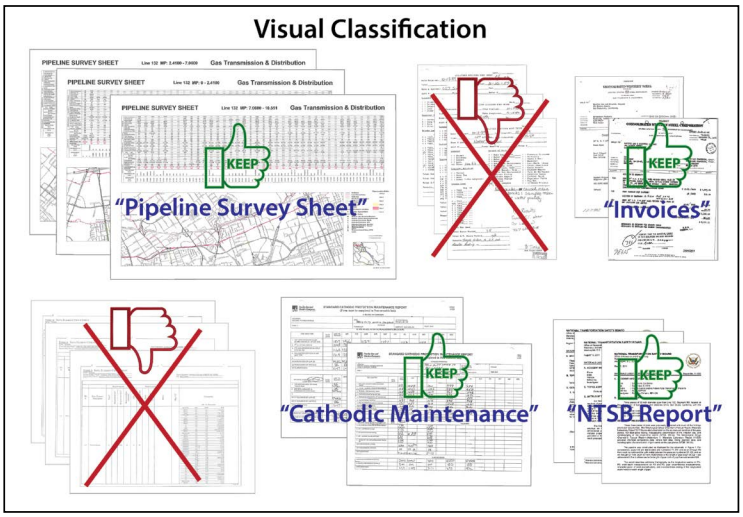
- File share remediation
- Content migration
- Archive digitization
- System decommissioning
- Setting access rights
- Retention and disposition scheduling

High-Level Overview

Without consistent document classification, there is no practical way to differentiate among documents in a collection to determine what ought to be kept, where it ought to be stored, who ought to have access to it, and how long it ought to be kept.



Visual classification automatically clusters visually-similar documents. Clients can examine those clusters to determine which to keep, and what document type to assign to the ones they retain.



BeyondRecognition's document classification technology and process is *comprehensive* in (1) dealing with all of an organization's documents regardless of whether they have associated text or not, and (2) dealing with the whole classification process commencing with collection through processing, evaluation, and loading the results into any designated management system. *Other classification technology is text-*

dependent and fails when there is no text associated with documents or the text is of poor quality.

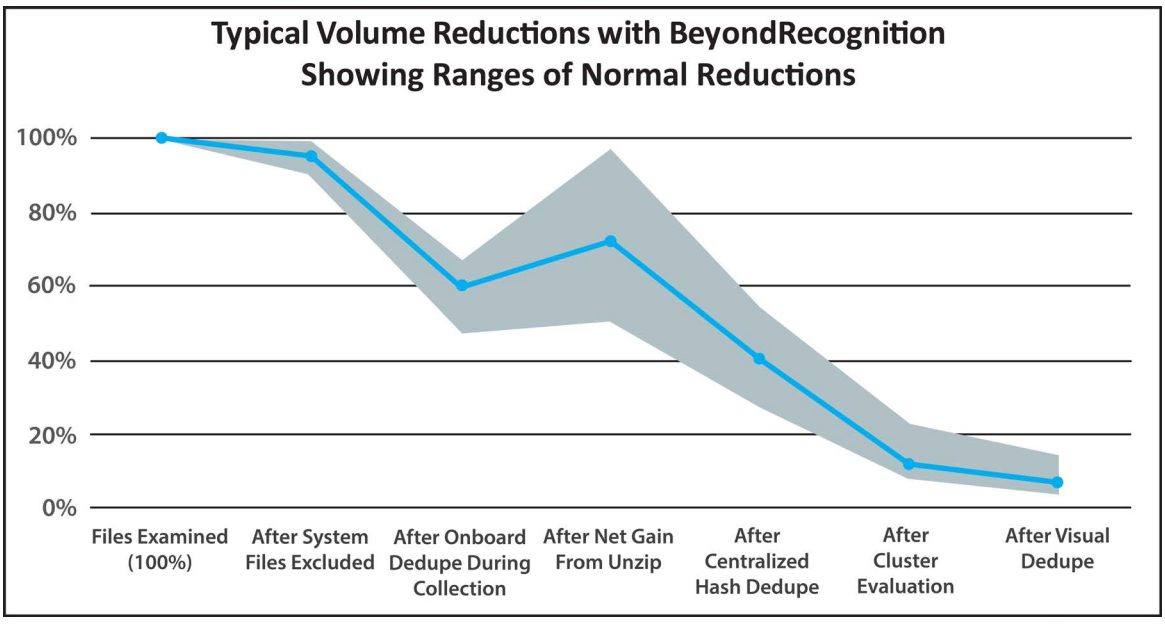
BR's technology and process is hardware scalable so that even collections from the largest enterprises can be processed. Other technologies can have choke points on classification processing which limit their abilities to process enterprise-scale collections. BR can scale by adding computing resources for classifying even the largest collections.

BR's design philosophy is built around making decisions once and then applying them instead of making them repeatedly. We call this *single-instance persistence*. For example, when a client determines that a particular cluster of documents does not have ongoing value and can be disposed of, such decisions will continue to be applied going forward. New decisions are required only when new clusters form.

Finally, BR's process is designed to eliminate the irrelevant or redundant content as early as possible in each step of the process using the minimum personnel and computational resources required. For example, BR employs hash-based deduping on its collector to avoid collecting multiple copies of the same files.

Benefits of Visual Classification

The most directly observable benefit of the BR process is a dramatic reduction in the number of documents or files being maintained by an organization. This is achieved by removing duplicates and by removing documents that have no ongoing legal or operational value to the organization, all accomplished in a completely transparent and defensible manner. Every collection is different, but this graph shows the magnitude of reductions in files typically achieved by BeyondRecognition during the steps outlined in the next section:



There are other significant benefits:

Data-Driven Process Equals Rapid Project Launch. The clusters of visually-similar documents form automatically without having to write rules, or select exemplars. This means a rapid project launch as project managers can immediately start evaluating the document clusters.

Project Timeline Compression. Because more computing resources can be added as needed to achieve desired milestones or completion deadlines, project timelines are greatly compressed compared to alternative approaches – what might take others years can be done in months with BeyondRecognition.

Much Improved End-User Content Management System Experience. End users will find a much improved CMS experience because there will much less unnecessary content and less redundant content in the system, and their access to content can be restricted to what they need to do their jobs. Also the searchability of the content will be much improved because BeyondRecognition:

- Adds a text-layer to image-only documents, and
- Adds field-searchable attributes for documents.

Overview

The major steps or phases in BR processing discussed in more detail below are:

1. COLLECT
- Log all files examined
- Dedupe
- Log but don't collect System Files

2. PROCESS
- Centralized Deduping
- Cluster

3. EVALUATE
- In or Out Decisions
- Designate Document Types
- Attribution

4. EXPORT FROM BR, LOAD IN CLIENT CONTENT MANAGEMENT SYSTEM
- Visual Deduping
- Redaction

Special Cases. Following the description of the major steps, this document will also describe two special cases, email, and scanned paper archives.

Individual Steps

Here is more detail on each of the steps in BR's visual classification processing:

1. COLLECT

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

BeyondRecognition provides USB devices that connect on a client's network to collect data. Prior to use, a file is placed on each device that instructs the Collector what paths or devices to consider. As files are collected they are encrypted and compressed using algorithms that meet the highest security standards.

SHA Hashing. The Collector calculates the SHA hash value for each file it encounters and uses these values as a basis for excluding known system or software related files, and for deduping the files it collects.

Logging. The Collector logs all files it encounters whether or not each file is copied onto the Collector. The log files enable BR to provide accountability for all files examined.

Excluding Known System Files. Files whose SHA hash values match entries on BR's known system file list are logged to record the name, size, location and hash value of the file, but the files are not copied onto the Collector.

Onboard Deduping. The Collector contains a list of hash values of all files collected by BR prior to deployment of the Collector and this list of already collected files is supplemented with the hash values of additional files it evaluates. The Collector's log is updated to add the file name, file size, location, and hash value of all files that are evaluated, but collects only a single instance of non-systems files. Note that some collection systems may collect only single instances but fail to track information about where all the copies were located. Not BR.

Container Files. The Collector does not unzip or decompress container files like *.zip or *.rar files to attempt to dedupe files contained within those container files. The onboard deduping is limited to deduping the overall container files.

2. PROCESS - Centralized Deduping

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

After a Collector has copied single-instances of unique files, it is sent to the centralized processing facility where its contents are copied onto a server. At this point container files are unzipped or decompressed and another SHA deduping process takes place. Duplicates are logged and

unique files are passed onto the next step.

2. PROCESS - Cluster

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

A BR server clusters the single-instance, unique files based on visual similarity. The documents in each cluster look substantially alike and evaluations of all the documents in each cluster can be made by examining one or two documents per cluster.

There are a few key points about this critical step:

Comprehensiveness. While all documents can be represented visually, not all document files have associated text. In some collections, 20-30% or more of the documents may have no text, e.g., image-only PDF or TIF files. BeyondRecognition is the only technology that bases its clustering on the visual representations of documents. BR in essence normalizes the documents so they can be compared regardless of the type of file in which they were located. Other technologies are text-restricted, meaning they examine or evaluate only the text associated with the document files and hence will have limited or no ability to cluster or classify documents that have no text or have poor quality text.

Scalability. BR processing throughput can be increased by adding more computing resources.

Data-driven. The clustering is automatic. The groups are self-forming without direction from operators or consultants. Clients can begin working with the clusters within just hours after the beginning of the clustering process, dramatically shortening project launch timelines.

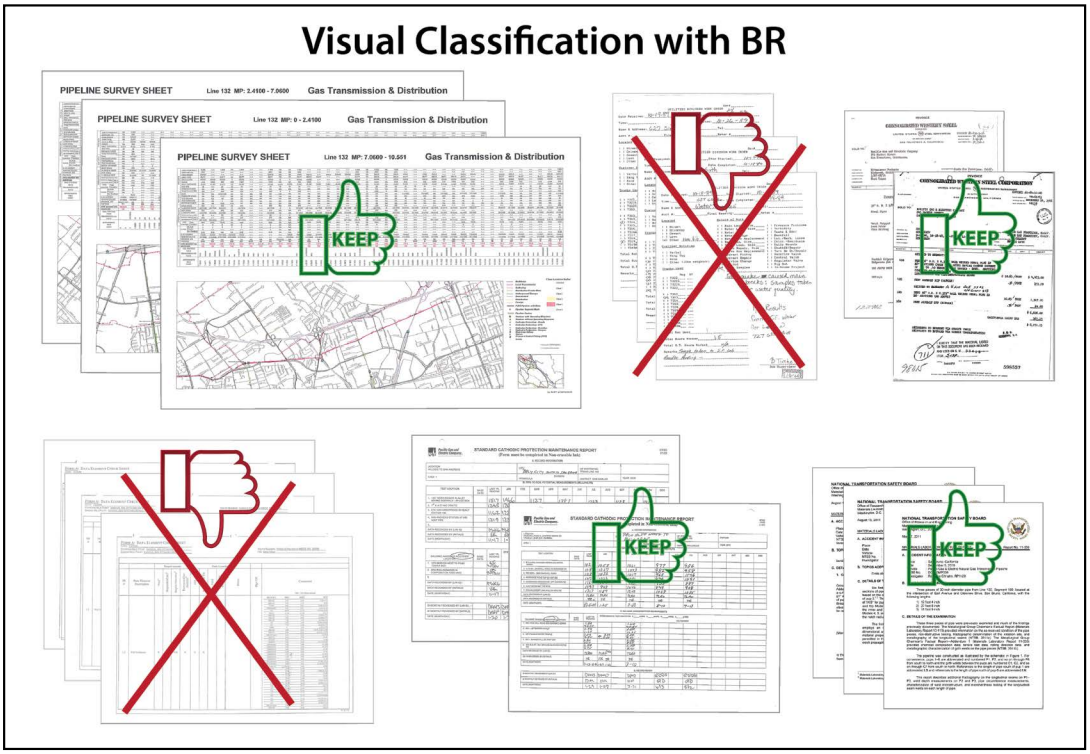
Exportability. The cluster IDs for each cluster can be exported to downstream platforms to enable them to make use of the visual similarities of documents in the same clusters. For example, in ediscovery review, reviewers can be assigned documents from the same clusters making the review faster and far more consistent than would otherwise be the case.

3. EVALUATE - In or Out

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

As clusters are formed by the first documents being processed, client's knowledge workers or subject matter experts begin evaluating them, starting with the clusters with the most documents first. The nature of the evaluation will vary with the purpose of the process. For information governance reviews, the evaluation is whether the documents have ongoing business, regulatory, or legal value and hence ought to be retained, or whether they can be

disposed of. In ediscovery reviews the evaluation will be whether the documents are relevant to a document request or otherwise relevant to the litigation.



Two key points about evaluation:

Persistence. The evaluation decisions are persistent, meaning they will be applied to documents that are subsequently added to the clusters. The persistence characteristic permits clients to begin evaluation as soon as the clusters start forming, i.e., they do not need to wait until all processing has concluded.

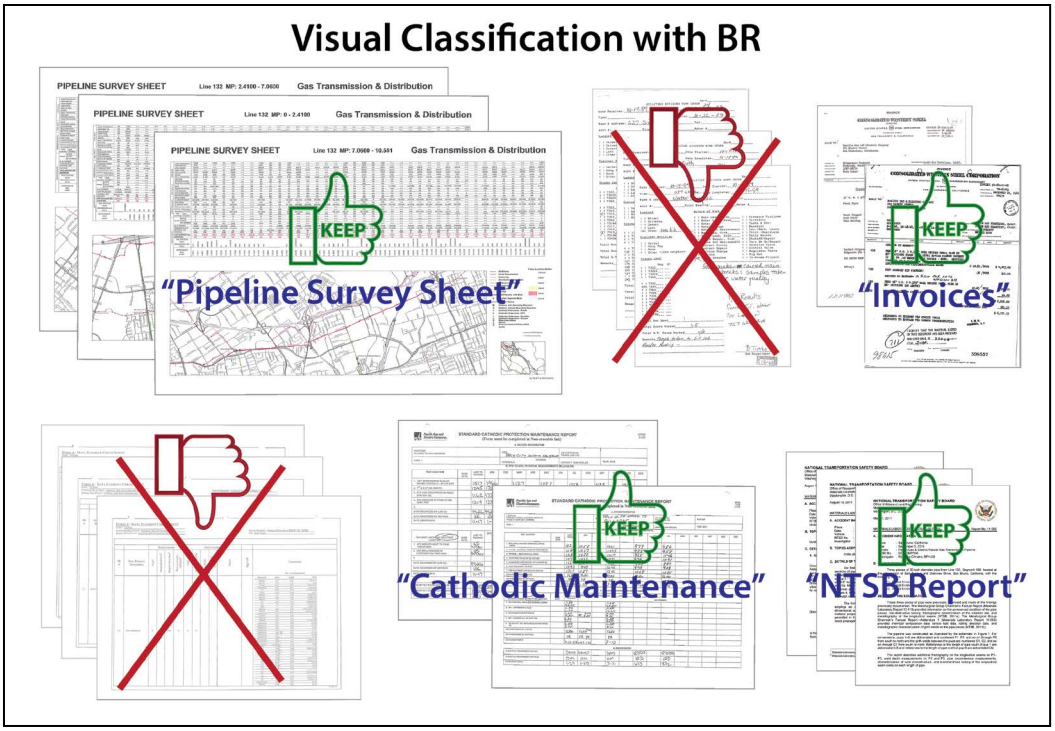
Rolling Intelligence. Decisions made in one information governance initiative can be rolled forward into subsequent decisions. For example, if a cluster is evaluated as a record during a file share remediation initiative, that decision does not have to be made again in a content migration initiative or a paper archive digitization initiative. Projects keep getting easier to do because more and more intelligence is accumulated and rolled forward.

3. EVALUATE - Document Type

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

Document-type labels can be applied to clusters that are being retained. BeyondRecognition provides a user-definable three-level document-type taxonomy or tree for this purpose.

Typically clients use the first level for business unit or function, the second layer for document type, and the third layer for sub-type. Multiple clusters may have the same document-type label, e.g., there may be several clusters that are labeled as "Invoices."



The labeling decisions are persistent and are applied to documents that are subsequently added to the clusters that use that label.

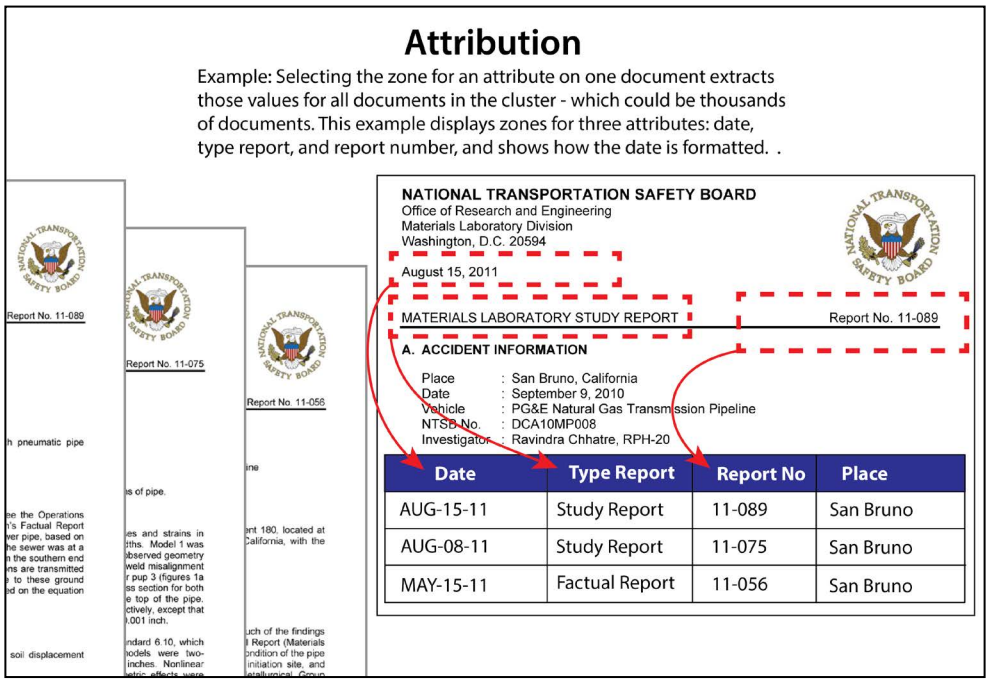
EVALUATE - Find

Because of the unique way that BeyondRecognition indexes the glyphs or graphical elements on each page of each document it can provide fixed and relative positional search operators for finding documents. Fixed positional searches look for items with the specified coordinates. Relative operators look for within a set of coordinates identified relative to other search specifications. Searches can specify ranges for dates, numbers, and text values, and complex searches can be built using the Reverse Polish Notation ("RPN") search logic. Find is very useful when trying to differentiate among documents within a cluster or for identifying documents within the entire collection.

3. EVALUATE - Attribution

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

Attribution involves extracting specific data elements from specific document types, e.g., pulling the API well number from a well log or a borrower's social security number from a loan application. Because of the visual similarity of documents in the same clusters, extracting attributes can be as simple as clicking and dragging to designate what data elements to extract.



BeyondRecognition provides numerous delimiters that help specify what data elements to use to help identify the data elements of interest and filters to provide a variety of ways to format the extracted information.

Where documents are of such poor quality that automated attribution does not yield satisfactory results, BR technology can be used to have data entry specialists key those portions of the documents that contain the desired attributes. In doing this, the portions of the document image being entered are split off from or disassociated from the rest of the document image so the person doing the entry does not see or know the other information on the page or document. For example, the person keying the loan applicant's name would never see the part of the document that has the address or social security number.

4. EXPORT/LOAD to Client Content Management (with Options for Reduping/Repopulating)

1. COLLECT			2. PROCESS			3. EVALUATE			4. EXPORT/LOAD
Hash	Log	Copy	Unzip	Hash	Cluster	In or Out	Doc Type	Attribution	Client Content Mgt

BeyondRecognition can export documents, document-type labels, and attributed data in any format desired by the client for their content management system, including PDF, PDF

with text, TIF, TIF with Text, CSV, PDF with internal metadata, etc. BR can also build load files to virtually any specification.

As part of the final export process, BR can "redupe" or repopulate duplicates to meet unique requirements of the client. For example, in ediscovery cases, sometimes agreement with the opposing counsel may require that documents not be deduped. In that case, BR can recreate the original file sets of all the duplicates of the documents being produced.

EXPORT – Visual Deduping

Many times there are several versions of a document that are visually indistinguishable from one another. For example, if a Word document is also saved as a PDF, and then one of those copies is printed and ultimately scanned to a TIF format, someone looking at the three versions could not identify any differences among the Word, PDF, and TIF versions. These are "visual duplicates." Normal duplicate detection using MD5 of SHA hash values will not identify these as duplicates because they are in different file formats and will not be bit-for-bit duplicates.

BeyondRecognition can add another layer of duplicate content detection using visual deduping to identify unique content. Clients can elect which visual duplicate or duplicates to retain in their ECM system

EXPORT – Redaction

BeyondRecognition can redact PII or other sensitive data using two complementary approaches. The first is based on identifying text strings that represent PII such as social security numbers or email addresses. Because of the way BeyondRecognition processes documents it knows the page coordinates of each glyph or graphical element, and can place redactions that are very precise in obliterating the redacted data without compromising adjacent non-redacted information.

The second approach is to redact zone coordinates on pages within clusters. For example in the cluster for IRS Form 1099 the whole block in which social security numbers are supposed to be written could be redacted even if the entry on some forms was handwritten. Redactions performed with BeyondReview can be done 17 times faster than a manual review.

Special Cases

Email

BR opens email stores such as PST files and identifies each email contained in them, and then dedupes and consolidates them as described below:

Normalized Email Deduping. Email requires additional processing because minor differences in the copies of emails stored by each recipient prevent them from being bit-for-bit duplicates. For example each email copy may have been routed slightly differently and have been sent at

slightly different sent times. To overcome this challenge, BR normalizes emails using the RFC 5322 Internet standard. This converts emails into the standardized format used to transmit them. When selecting which deduped email to use for analysis, BR choses the sender's email so that any BCC information is included.

Consolidating Embedded Email. The second step to consolidate duplicate email information contained in subsequent replies or forwards so that it is represented only once. For example, if a reply incorporates the body of the initiating email and involves the same email addresses, the content of the initiating email is duplicated in the reply and only the reply needs to be evaluated.

Example: Assume Person A emails Person B and C, then B replies all and then C replies all. There are three emails each contained in the email records of three people for a total of nine emails. Normalized email deduping would reduce the nine emails to the three unique emails. Consolidating the embedded emails would result in selecting the last email that included the earlier two emails. In this example applying both steps in sequence would have reduced the emails under consideration from nine to one, an 89% reduction in the number of objects to consider.

Payload Analysis. Email attachments are processed along with other document files, meaning they are clustered and evaluated. The client has the option of applying the evaluation decision made regarding the attachments to the emails that transmitted them. For example, if an attachment is considered to be a record, the emails that transmitted it could also be considered to be records.

Scanned Paper Archives

Scanned paper can present major difficulties in identifying duplicates, identifying logical document boundaries, and obtaining searchable or extractable text. BR handles all three of these challenges exceptionally well:

Visual Duplicates. Digitized paper documents represent significant challenges that are handled uniquely by BeyondRecognition. The first is that although there may be a large proportion of duplicates within the paper collection and overlapping with native electronic files, normal hash-based deduping will not identify them. For example, it is almost impossible to scan the same sheet of paper twice and have the resultant image files have the same hash values especially if the paper is moved in any way. Detecting duplicates with hash values requires that the files be bit-for-bit the same and any slight realignment of the page on the scanner will cause pixels being captured in different orders yielding different hash values.

BR's visual duplicate technology enables it to disregard such inconsequential differences and focus on the appearance of the pages when identifying duplicates, either within the collection of other scanned documents or as compared to native files that may have been used to print the documents in the first place.

Document Unitization. The other challenge has to do with document unitization or page breaks between documents. When documents have yet to be scanned, requiring the scanning crew

to use slip sheets or otherwise indicate document boundaries slows the process significantly and is error prone. On the other hand, many times when digitized images are received the scanning operators did not use any logical boundary indicators, e.g., when documents are scanned during merger and acquisition activities, a whole box of documents are often either all single-page TIF files or are all in one large multi-page image file.

BeyondRecognition has the answer to both of these document unitization challenges. BeyondRecognition can learn what the first pages of documents look like and this knowledge can be applied to determine where the documents boundaries for documents should be placed. This has been found to be more accurate and consistent than manually-assigned document breaks.

Searchable or Extractable Text. As part of its scanned document processing, BeyondRecognition can perform glyph recognition to yield the text associated with the document images. This can be included in whatever the deliverable the client specifies, e.g., PDF with image and text layers.

Conclusion

BeyondRecognition has developed some truly unique technology with functionality never before available for information governance. BR owns the code used in its technology and has an extremely broad ability to develop custom applications.

We would like to show you how well visual classification can work for your document-related information governance initiatives. *Contact us to discuss free visual clustering of your documents on a healthy representative sample of your content.*

For more information on BeyondRecognition, LLC, visit us at www.BeyondRecognition.net or email us at IGDoneRight@BeyondRecognition.net.

