Guide to Managing Unstructured Content

Practical advice on gaining control of unstructured content

By John Martin
Note: Illustrations listed on page x are
(c) Juergen Priewe, obtained from Dreamstime.com
June, 2009. I was walking through one of my client’s mortgage servicing centers when I turned a corner to go down one of the “cube alleys” in the Center and narrowly escaped a head-on collision with a worker rushing to a copy machine with a stack of papers. As we swerved to avoid the collision, he stumbled, bumping a three-foot stack of banker boxes. The stack tumbled and, like a row of dominoes, proceeded to collapse a seemingly endless row of physical files, rubber-band-clad legal files, boxes, and manila folders.

Within seconds a hundred square feet of industrial carpet became a graveyard of random documents, images, and transcripts. This was the raw material needed to complete administrative and processing requirements so borrowers could undertake specific loss mitigation activities that could allow them to retain their homes.

Two hours later I found myself in a meeting of senior executives discussing a quality control review of over 5,000 processed loan modifications. The review indicated a greater than 40% defect rate in the data used for eligibility assessment and completion of required forms. Data to support the factors used to determine borrower remediation was either missing, unverified, or unsubstantiated by a physical source document.

Foreword
It was at that moment that I began to understand the enormity of the problem in our industry — the lack of accurate, verified, classified, archived, persistent and retrievable “digitized” data. We were a “stare and compare” industry and we often got it wrong. We continue to have miles and miles of physical document stores that have been created as evidence of completed business transactions, and we need to convert these polaroid snapshots of information into digital libraries that codify content into a language, vocabulary, and syntax.

June, 2016. This book presents a scientific solution to the biggest challenge we still have in business today — to be able to extract and store data in a manner consistent with the taxonomy the specific business and industry require to turn data into intelligence.

I believe the technology discussed on these pages delivers a new and disruptive ability to apply sophisticated rules engines, artificial intelligence, machine learning, and analytics to unstructured content at a speed never before experienced, particularly in financial services. It also provides the ability to re-constitute peer-to-peer transaction records that can become a foundation for creating block chain ledgers defining the immutable history of what was done where, when, by whom, and with what specific outcomes.

The most important contribution of this work is the brilliant method by which the author has rendered a highly complex, sophisticated technology into a conceptual framework and descriptive narrative that is relevant and understandable to business process owners. This book is an essential text that allows business leaders to contemplate use cases that enable them to complete the critical first step in the journey to process digitization and automation – data extraction, classification, and attribution.

Without the alphabet, dictionary, and library made possible with this new approach and technology, businesses that rely on vast amounts of structured and unstructured data will continue to feel that the source of truth they need to innovate is unavailable and unattainable. It will continue to feel like the experience I had in 2009 — oceans of data locked in forms, files, and image repositories that sometimes spill on the floor and get lost forever.

Henry Santos
Charlotte, NC, October, 2016
Preface

Among the most important responsibilities of information governance professionals are those relating to “unstructured” content – the files and documents stored on file shares, individual computing devices, and content management systems. Information governance professionals have long known what they should do with unstructured content:

- Classify it
- Apply retention schedules
- Dispose of unwanted files
- Avoid duplication
- Control access

Where there’s been less clarity is on exactly how to do those things.

This book addresses that question. It tells specifically how to perform basic information governance tasks on unstructured content. It is intended to be intensely practical and realistic; it is based on decades of experience selecting, analyzing, preserving, and migrating such content and testifying about it.
This book uses a simple RCAV (Rationalize, Classify, Attribute, Validate) model to organize the discussion. This is how the chapters are organized:

I - Introducing the RCAV model for governing unstructured content
II - An overview of visual classification technology
III - An in-depth look at the RCAV model using visual classification
IV - Use cases for visual classification.
V - Selecting file or document classification services and technology
VI - Getting started with visual classification
VII - How to contact us for further information

I welcome questions, comments, or suggestions on this book. Send them to me at john.martin@beyondrecognition.net. I look forward to hearing from you!

John Martin
Houston, TX, October, 2016
Table of Contents

Foreword ................................................................. i
Preface ................................................................. iii
Table of Contents ...................................................... v
Sidebars ............................................................... ix
Figures ................................................................. x
Tables ................................................................. x
Illustrations ........................................................... x
I. The RCAV Model for Managing Unstructured Content .................. 1
   A. Rationalize — How Many Files & Where? ............................ 3
      1. Anticipate Interruptions ............................................. 3
      2. Beware the 260-Character Path/Filename Limit in Windows ... 3
      3. Log All Files Reviewed ............................................. 4
      4. Calculate Hash Values and Record in Log ..................... 4
5. Identify NIST-Listed Software Files ...................................... 4
6. Copy Only Deduped, DeNISTed Files .................................... 5
7. Use Hash Values for File Names .......................................... 5
8. Compress Collected Files .................................................. 5
9. Check for Duplicates in Container Files .................................. 6
10. Anticipate Encrypted Files .............................................. 6
11. Encrypt Collected Files ................................................. 7
12. Use Hash Values to Audit Compartmentalized Storage ................. 7

B. Classify — What Type of Content? ...................................... 8
1. Anticipate Constant Change ............................................ 8
2. Be Aware of Text Dependency Limitations ............................ 9
3. Address Document Unitization issues .................................. 9
4. Anticipate Scale & Granularity ......................................... 10
5. Anticipate “Do-Overs” ................................................ 10
6. Anticipate Multiple Classifications for Multiple Reasons ............ 10
7. Use Stakeholder Collaboration ......................................... 11
8. Anticipate the Awareness Gap .......................................... 11
9. Use Classification Matrix Approach .................................. 11
10. Test Consistency ...................................................... 12
11. Start-Up and Ongoing Maintenance Costs ............................... 12

C. Attribute — Extract What’s Important .................................... 14
1. Identify Specific Desired Attributes for Each Classification .......... 14
2. Examine Data Fields in Control or Retrieval Systems ............... 14
3. Be Aware of Classification-Dependent Limitations ................. 14
4. Be Aware of Text Conversion Dependence ............................ 14
5. Be Aware of Extraction Tool Dependence ............................. 15
6. Consider Non-Textual Attribute Extraction ........................... 15
7. Normalize Extracted Values ............................................ 15
8. Use and then Update External Authority Lists ....................... 16
9. Involve All Key Stakeholders ........................................... 16
10. Build Comprehensive Logs for Full Audit & Presentation .......... 16
11. Consider Full Text as an Attribute ..................................... 16

D. Validate — How Dependable Is it? ....................................... 17
1. Automate the Validation ............................................... 17
2. Validate Deadline Compliance .......................................... 17
3. Compare Before & After Totals ......................................... 17
4. Measure Classification Accuracy ...................................... 18
5. Measure Accuracy of Attribute Extraction .......................... 19
6. Use Ad Hoc Analytics ................................................. 20
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>Double-Check Duplicate Tracking</td>
</tr>
<tr>
<td>8.</td>
<td>Know What Didn't Get Captured or Extracted</td>
</tr>
<tr>
<td>9.</td>
<td>Reconcile Authority Lists</td>
</tr>
<tr>
<td>10.</td>
<td>Compare Results of Using Other Systems or Approaches</td>
</tr>
<tr>
<td>11.</td>
<td>Incorporate User Feedback</td>
</tr>
</tbody>
</table>

**II. Overview of Visual Classification** ........................................ 22

**III. Visual Classification in Depth** ........................................ 25

**A. Rationalize — How Many Files Where** .................................. 26

1. Collectors  | 26 |
2. Logging & Hashing  | 26 |
3. DeNISTing  | 26 |
4. Deduping & Copying  | 26 |
5. Compression & Encryption  | 27 |
6. Capacity  | 27 |
7. 2,600 Character Path Limit  | 27 |
8. Inherited Security  | 27 |
9. UnZipping  | 27 |
10. Rationalization Deliverables  | 27 |

**B. Classify — What's in the Files (Visual Clusters)** .................. 29

1. Visual Representations  | 29 |
2. Global Glyph Catalog  | 29 |
3. Cluster Visually-Similar Files  | 30 |
4. Classify — Assign Document-Type Classification Labels  | 30 |
5. Classification Tree  | 31 |
6. Benefits of Visual Classification Approach  | 32 |
7. Document Unitization of Multi-Document Files  | 33 |
8. Document Unitization of Single-Page Files  | 34 |
9. Learning from Previously Classified Content  | 34 |
10. Classification Deliverables  | 34 |

**C. Attribute — Extract What You Care About** .......................... 35

1. List Attributes  | 35 |
2. Locate Attributes within Clusters  | 35 |
3. Delimiters  | 35 |
4. Formatting  | 36 |
5. Single-Instance Glyph Editing  | 36 |
6. Non-Textual Attribute Extraction  | 36 |
7. Using Existing Authority Lists to Find Potential Attribution Locations  | 36 |
8. Content Enablement — Converting All Character-Related Glyphs  | 37 |
9. Negation to Yield Residual High-Value Terms  | 37 |
10. Benefits of Visual Classification Approach to Attribution ........................................... 37
11. Attribution Deliverables ......................................................................................... 37

D. Validate — Be Confident of Results ........................................................................ 39
1. File Count .................................................................................................................. 39
2. Classification Accuracy .............................................................................................. 39
3. Attribution — Specifying ......................................................................................... 40
4. Attribution — Single-Instance Editing ...................................................................... 40
5. Attribution — Comparison with Authority Lists from Systems of Record .............. 40
6. End-User Feedback & Course Corrections ............................................................... 40

IV. Visual Classification Use Cases .................................................................................. 41
A. Granular Classification & Attribution Use Cases ...................................................... 42
B. “Large Bucket” Classification ................................................................................... 43
C. Special Use Cases ....................................................................................................... 43
1. Scoping Projects ......................................................................................................... 43
2. Email Management ..................................................................................................... 43
3. PII Detection & Redaction/FOIA Responses .............................................................. 45
4. Taxonomy/Term Selection .......................................................................................... 46
5. Fax Servers .................................................................................................................. 46
6. Reclassifying Previously Classified Content ............................................................. 48
7. FAAT — Fully Automated Attribute Tables ............................................................... 48
8. Contract Review w/Duplicate Block Negation ............................................................ 48
9. Public Agency Comment Analysis .............................................................................. 49

D. New Search Functionality .......................................................................................... 49
1. New Data Types and Range Searching: Numbers & Dates ........................................ 49
2. Absolute and Relative Positional Search ..................................................................... 49
3. Graphical Search ......................................................................................................... 50

E. Summaries of Use Cases by Deliverables and Functions Involved ............................ 50

V. Selecting File or Document Classification Systems .................................................... 53
VI. Getting Started: One Bite at a Time .......................................................................... 56
VII. Further Information ................................................................................................. 57
Sidebars

I. The RCAV Model for Managing Unstructured Content ........................................... 1
   SB 1. How to Estimate Volume of Files Not Collected Due to Long Path/Filename ..... 3
   SB 2. How to Test to See if Your System Can Search Files on Long Paths ............... 4
   SB 3. Identifying Duplicate Emails ....................................................................... 5
   SB 4. Problems Hashing *.MSG files .................................................................... 5
   SB 5. Lotus Notes Potential Gotchas ..................................................................... 6
   SB 6. PDF: Versatile Work Horse ......................................................................... 7
   SB 7. Determining How Many Non-Textual Files An Organization Has ................ 9
   SB 8. Limitations of OCR ..................................................................................... 10
   SB 9. Non-Textual Attribution: Signatures ......................................................... 16
   SB 10. False Positives and False Negatives ......................................................... 18
   SB 11. QA for Scanning Projects/Processes ....................................................... 19
   SB 12. Incorporating Validation Measurements in Service Level Agreement (SLA) ... 19

II. Overview of Visual Classification ................................................................. 22

III. Visual Classification in Depth ................................................................. 25
    SB 13. Rationalization Overview .................................................................... 26
    SB 14. BR Collector At-a-Glance ................................................................ 27
    SB 15. Classification Overview .................................................................... 29
    SB 16. Visual Classification At-a-Glance ....................................................... 29
    SB 17. Taxonomies .......................................................................................... 31
    SB 18. Attribution Overview ........................................................................ 35
    SB 19. Visual Attribution At-a-Glance ........................................................... 35
    SB 20. Overview of Validate ......................................................................... 39
    SB 21. Validation At-a-Glance ...................................................................... 39

IV. Visual Classification Use Cases ................................................................. 41
    SB 22. Nine Levels of Glyph Patterns ......................................................... 47

V. Selecting File or Document Classification Systems ....................................... 53

VI. Getting Started: One Bite at a Time ........................................................... 56

VII. Further Information ..................................................................................... 57
Figures

Figure 1. RCAV Model .......................................................... 1
Figure 2. Example of Long File Path ........................................ 4
Figure 3. Mapping Structured & Unstructured Content .......... 15
Figure 4. Extracting Non-Textual Elements ............................... 16
Figure 5. Thumbnail Images Showing Text Not Needed for Classification . 22
Figure 6. Classification Labels Applied to Clusters of Visually-Similar Files .......... 23
Figure 7. Document Attribute Matrix ........................................... 24
Figure 8. Automatically-Formed Clusters of Visually-Similar Files ................. 30
Figure 9. Clusters of Visually-Similar Files with Classification Applied .......... 31
Figure 10. Example of BR Document Specialist Interface ..................... 32
Figure 11. Attribution: Extracting Specified Data Elements ...................... 36
Figure 12. Zonal Redaction of Handwritten SSANs in IRS Form 1040 ................. 45
Figure 13. Text-String Search Redaction of SSANs in IRS Form 1040 ................. 45
Figure 14. Graphical Representation of Where PII Is Found ..................... 46

Tables

Table 1. Document Classification Matrix .................................. 13
Table 2. Example of Impact of Persistent Classification ................. 33
Table 3. Re-Usable Knowledge Gained During Visual Classification Process .... 42
Table 4. Impact of Classification Inheritance on Email Threads ........... 44
Table 5. Example of Table in Unstructured Content ...................... 48
Table 6. Deliverables in Different Visual Classification Use Cases .......... 51
Table 7. Different BR Function in Different Visual Classification Use Cases .... 52
Table 8. Checklist for File Classification Technology ..................... 54

Illustrations

The chrome and/or red sphere illustrations on the following pages are copyright Juergen Priewe, www.dreamstime.com/joggi2002_info: cover & pages i, iii, v, 1, 3, 8, 14, 17, 22, 25, 26, 29, 35, 39, 41, 53, 56, and 57.
There are four basic steps to managing unstructured content. They are:

1. **Rationalize.** This is an inventory and collection process and answers the most basic of questions: "How many files do you have and where are they?"

2. **Classify.** This begins the process of creating useful intelligence from the content. It answers another basic question: "What's in those files?" Accurate classification enables all downstream information governance activities, e.g., assigning useful record retention schedules and performing accurate, uncluttered retrieval.

---

**I. The RCAV Model for Managing Unstructured Content**

![RCAV Model Diagram](image)

*Figure 1. RCAV Model*
3. **Attribute Extraction.** This involves a more granular determination, “What types of information do you want to pull or extract from within each classification?”

4. **Validate.** This involves measuring and confirming that all files have been accounted for and that resulting classifications and attributions are valid. It's basically, "Do you have everything and is it accurate?"

In the following sections we provide tips or things to consider during each of those four steps.
A. Rationalize — How Many Files & Where?

Managing unstructured content begins with inventoring it. Files must be located, counted, and copied for analysis and processing.

Here are suggestions for rationalizing unstructured content:

1. Anticipate Interruptions

Enterprise-scale file inventoring and copying is an ongoing, resource-intensive operation. Anticipate that it will be occasionally interrupted and be sure that your process can resume without having to repeat all the analysis.

2. Beware the 260-Character Path/Filename Limit in Windows

Software for indexing, logging or copying unstructured content files is often subject to the 260-character path/filename limit inherent in most Windows applications. That means that files where the path/filenames exceed that limit will not be examined by that software - they are simply invisible to it. They won't be indexed, logged, or copied.

BR has found cases where 2% of files it collected had been inaccessible to normal Windows applications because of overly long file paths. This can happen in many ways, for example:

- Zipped or compressed folders can contain multiple levels of folders that although not troublesome in their original location become excessive when unzipped or copied onto other, longer paths.

Sidebar 1

How to Estimate Volume of Files Not Collected Due to Long Path/Filename

Windows counts the space used by files on long file paths even if most Windows applications can’t read them. To determine if any of your files are inaccessible due to path/filename restrictions, first determine the space-used metric supplied by Windows (by, e.g., right clicking on a drive and selecting “Properties”) or, if you use cloud storage, obtain the space used number from your cloud provider.

Next have your indexing or collection software create an audit log of all files on that drive, including the size of each file. Total the file sizes and compare that to the “size on disk” metric provided by Windows. The difference can be files hidden because path/filenames are greater than 260 characters.
• Drives or folders can be mapped to other drives or folders resulting in unexpectedly long paths.
• Users sometimes use folder names to store notes about the files in the folder, e.g., “not to be included in the survey responses.” When folders get copied and moved to other locations, the total characters for all folders in the path can easily exceed 260 characters.

3. Log All Files Reviewed

All files examined during the inventorying process should be logged whether or not they are copied. This permits audits of what was examined and can provide support for why any processing decisions were made. Ongoing, accurate logging goes to the very heart of data integrity.

4. Calculate Hash Values and Record in Log

Hashing algorithms have long been used on the Internet to provide a way to determine if emails or files were altered during transmission (see RFC 822 and subsequent standards). If two files have the same hash values they are identical, especially if the more secure SHA hash standard was used.

Hash values should be calculated for each file and stored in the log. This is a critical data integrity requirement. By doing this the files can be rehashed later to confirm whether they have changed in any way. The hash values also enable several other important functions as described in several of the following tips.

5. Identify NIST-Listed Software Files

Information governance concerns itself with the content created by or received by an enterprise. One way to focus on content is to exclude system files distributed by software companies such as executables, help files, and templates. The National Institute of Standards and Technology (“NIST”) publishes a list of known system files and their hash values, and this NIST list can be used to identify such files.

Sidebar 2

How to Test to See if Your System Can Search Files on Long Paths

To see how well your indexing or file collection works you can create a long path, place a file containing a unique word on it, and see if the system can locate it. Here’s how:

Go to Windows Explorer and starting with the root drive (probably C://) create a folder named 12345678901234567890 which is 20 characters long. Then add deeper folders using the same name. Repeat until you have created 12 folders with the same name. Your system may truncate the last folder name. At this point the last folder will use a total of up to 256 characters as follows:

<table>
<thead>
<tr>
<th>Path</th>
<th>Char’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>c://</td>
<td>4</td>
</tr>
<tr>
<td>12 folders @20 characters</td>
<td>240</td>
</tr>
<tr>
<td>12 forward slashes</td>
<td>12</td>
</tr>
<tr>
<td>Total (less truncated char’s)</td>
<td>256</td>
</tr>
</tbody>
</table>

Make a text file that contains a unique text string, e.g., “xiobionic,” and save it with the name 1234567890.txt at the end of path. That will make the total path/filename exceed the 260 character Windows limit.

Now see if your Windows search engine can find the copy of the file that is on the long path. Normally, it won’t.
When such files are located, their presence should be noted on the log, to include the filename, location, hash value, and if the hash matched a NIST list entry.

6. **Copy Only Deduped, DeNISTed Files**

Hash values for content files should be compared to hash values of files already collected, and only new files that do not appear on the NIST list should be collected. Collecting only one instance or copy of each unique file minimizes the number of files to be copied and permits the collection system to copy drives that were larger than those of the collection device itself because it is not copying known system or duplicate content files. Note that the detailed audit log retains all information about where the duplicate copies of all files were located; there is no need to copy bit-for-bit duplicates to track this information.

7. **Use Hash Values for File Names**

Managing millions of files requires that they be stored in a standard fashion, i.e., not kept in their original folder paths which may present numerous difficulties in terms of length of path, inefficient numbers of files in individual folders, and possibly unacceptable characters. Using hash values for names has two significant advantages:

- **Noncollisionable Names.** While duplicate file names don’t present many problems if they are each in unique folders, when files are moved to new folders as happens with standardized storage, there can be “collisions” when two or more files in the same folder have the same name. Basically only one of them will be kept, others with that file name will be essentially lost. This does not happen if hash values are used as names.
- **Self-Authentication.** When the value obtained by hashing a file is used as its name it makes the file self-authenticating because the hash value can be recalculated and compared to the file's name.

8. **Compress Collected Files**

Collected files should be compressed when not being analyzed in order to preserve data storage resources. By deNISTing, deduping, and compressing copied files, collection devices can often collect

---

**Sidebar 3**

**Identifying Duplicate Emails**

Identifying duplicate emails is complicated because different email clients can store emails in different ways and those differences result in different hashes for different clients. To use hashing, the emails should be restored to the form used to transmit them (see RFC 5322), and the hash calculations should not include metadata values expected to be different in different copies, e.g., routing information or date last accessed.

**Sidebar 4**

**Problems Hashing *.MSG files**

Emails are often saved as *.msg files when preserved for e-discovery or other purposes, with hash values calculated and saved so to prove the files hadn't been altered.

One big problem with this approach is that if that copy is later opened in some versions of Outlook, the “date opened” metadata contained within the *.msg file itself may be changed resulting in a new hash value for the whole file.

Here are some solutions to this:

1. Save the *.MSG files only on read-only media so the metadata can’t be updated.
2. Save the emails as PDFs, EML or some other static format.
from drives two-to-three times larger than the storage on the collection devices.

9. **Check for Duplicates in Container Files**

Once copies of files have been centralized for processing, files that contain other files (e.g., *.ZIP or *.RAR files) need to be recursively opened, logging each new file or container file and its associated hash value. The goal is to identify the set of unique, deduped content files copied from the source files.

10. **Anticipate Encrypted Files**

Some files will be encrypted, making it impossible to accurately classify them or search on their contents. Decisions should be made ahead of time about how to process encrypted files as there are several options, each with different costs and benefits. If files are encrypted should be indicated on the log to be treated as possible exceptions, and the log should be updated if encrypted files are decrypted.

Here are ways to handle encrypted files when processing large volumes of data for information governance or e-discovery purposes:

- **Request passwords from users.** This is time-consuming and fraught with difficulties, e.g., when users have resigned, died, or just don't remember the passwords.

- **Brute force attempts.** In this approach programs try virtually all possible passwords until a successful one opens the file. Depending on the length of the pass code this can be a rather quick process or a long and resource-intensive one. If a password for one file is located it can be tried against the other encrypted file to see whether the user re-used that same password.

- **Dictionary.** Software can try all character strings found in the unencrypted files to see if the user had stored the password in one of those files.

- **Key recovery.** Rather than try to find the password that the encrypting software uses to calculate the key that opens the file, programmatically generate all the possible keys.

- **Email.** When the contents of individual emails have been encrypted, it may be possible to find an unencrypted version of the email by looking later in the thread on other custodians’ files.

---

**Sidebar 5**

**Lotus Notes Potential Gotchas**

Lotus Notes content is stored in a Notes databases that have unique characteristics that can cause problems when collecting Notes messages or attachments. The difficulties arise because Notes uses views to permit users to examine and export content, and those views can important limitations:

- **Omitted Emails: Time to build views.** When Notes content is exported to a new Notes container, it takes time for that container to build its views. If users try to export data from the new Notes container before the views are completed, not all intended records may be exported. Symptoms of this problem may include not exporting Sent emails or having missing time periods.

- **Field Truncation.** The view may impose truncation limits on content in the fields in the view, e.g., the CC list may be cut off after a certain limit is reached, or only a certain number of characters may be included in the message body.
11. **Encrypt Collected Files**

Collected files represent potential security risks if they are stolen or unauthorized people gain access to them. To minimize those risks, encrypt the collected files. If passwords are used for both compression and encryption, security can be maximized by using different key lengths for both processes.

12. **Use Hash Values to Audit Compartmentalized Storage**

Many organizations have content so sensitive it is supposed to be only maintained on servers physically disconnected to any other network. One way to find if there's been “leakage” of this content is to calculate the hash values on the isolated or quarantined servers and when the general enterprise content is inventoried and hashed, compare hash values to see if copies of the secure content are being stored in unsecured locations or devices.

---

**Sidebar 6**

**PDF: Versatile Work Horse**

The PDF file format standard was developed by Adobe but later released and published as an ISO standard. PDF can be a workhorse for information governance for several reasons:

**Universal Viewing.** No special software is needed to view the primary content, i.e., users don't need licenses to the software used to initially create the content.

**Embedding.** PDF standards contemplate the embedding of content well beyond what is visible to the casual user. This could include the file classification for the visible content, values extracted from it, translations, text layers, and even the original native electronic file used to create it. Embedding content in a way that can be indexed helps make PDFs more distributable because other content management systems can use the embedded content to make the PDF more searchable and usable.

**Compression.** PDF standards permit the use of multiple compression techniques on individual files so the best compression can be obtained to reduce storage and transmission requirements. While using additional techniques imposes additional computational requirements on the systems compressing the PDFs, they impose no additional resource requirements on PDF viewing software.
Consistent classification is the most critical challenge in managing unstructured content. If you can't classify items, you can't manage them.

It has historically been very difficult to achieve consistent, scalable classification. Here are just a few of the downstream tasks that become much more difficult if not impossible without consistent classification:

- Classification-based retrieval
- Setting records retention periods
- Determining user-level access rights
- Setting department or business unit level access rights
- PII detection
- Setting system security specifications for content

These difficulties also lead to end-user workarounds that defeat many of the reasons for having ECM systems, e.g., users maintaining private but duplicative stashes of content.

Note that file type is not a sufficiently useful file attribute for most information governance purposes. The file is just a container and information governance focuses on content. For example, a PDF could be virtually anything from a spreadsheet to a presentation, Word document, or website. Classification must go beyond file type to accurately label the type of content.

Here are considerations when setting up automated file classification systems:

1. **Anticipate Constant Change**

Changing business needs and changing regulatory requirements results in a constant change in the documents used to perform business functions. Individual forms or templates change 10-15% per year on average, and many new document types are added each year. This constant change imposes the requirement that classification systems be flexible and responsive, ideally alerting administrators when new classifications are required.
2. **Be Aware of Text Dependency Limitations**

Most automated classification systems rely on the presence of accurate textual representations of the files being classified. Classification systems must anticipate multiple problems with a text-based approach:

- **Language.** Systems that seem to work fine with English documents may fail when presented with other languages that were not part of the original training sets or scripted rules. Multiple languages will also cause obvious problems with approaches based on taxonomies. Machine translation of content may not yield the desired classification accuracy.

- **Non-textual files.** Many files have no text associated with them, e.g., files output as PDF or TIF files from user-software or captured as image-only documents by scanning or faxing. This may be a minor issue in some collections but in others non-textual files may account for appreciable percentages of all files. At the very least the percentage of non-text files ought to be measured to help determine what sort of remedial effort may be justified (see sidebar).

- **Poor-quality text files.** Text layers can be created by optical character recognition (“OCR”) software, but the resultant associated text can be riddled with errors, making text-based classification very problematic (see OCR sidebar next page). One area of particular concern is being able to classify all versions of the same document type consistently, e.g., to classify the original Word document with the PDF version and the scanned TIF version.

- **Sentence dependency.** Some auto-classification systems analyze text as presented in sentences and ignore non-sentence text. This causes them to fail to accurately classify documents like check lists, spreadsheets, PowerPoint presentations, and many form-based documents.

- **Numeric Text.** Text analytics and text search systems may ignore numeric text strings and evaluate pages and documents without considering their numeric text.

3. **Address Document Unitization issues**

One of the most basic, and often incorrect, assumption about file classification is that there is just one document per file. Virtually all content management systems or file search systems permit only
one set of fields to manage a file, e.g., there is one document type, one author, one date created field.

Document unitization problems can arise when users assemble multiple documents or files, possibly even those created in different applications, into one PDF. Unitization issue are also common in files created by scanning or faxing paper documents.

The one-document-per-file limitation causes embedded documents to be "lost" in the sense they are not represented in the fielded data that describes individual documents, e.g., Date, Author, Title, etc. Having multiple document files can also cause errors in text analytics systems that depend on comparing the text in individual documents.

At the very least, sampling of the various sources and storage locations should quantify the extent of this issue. The unitization issue impacts all later downstream functions, e.g., assigning retention periods, accurate retrieval, and extracting document attributes.

4. **Anticipate Scale & Granularity**

Some text analytics systems used to place a few hundred thousand files into relatively large “buckets” such as Responsive or Non-Responsive for e-discovery may not provide the granularity and scalability to have hundreds of classifications for millions of files. Anticipate your needs for scale and granularity when comparing competing solutions.

5. **Anticipate “Do-Overs”**

Business environments and missions evolve over time and those changes can affect how files or documents are classified. A well-designed classification system will permit reclassification without completely reprocessing large numbers of files.

6. **Anticipate Multiple Classifications for Multiple Reasons**

There are many reasons to classify unstructured content:

- File share remediation to identify and remove

---

**Sidebar 8**

**Limitations of OCR**

OCR is often used to obtain text from image-only files for use with text-based classification systems. However, several things can cause inaccurate or missing text:

- **Language.** Individual OCR software may not correctly convert the glyphs associated with all the foreign languages that may be encountered in a set of files.
- **Font Size.** OCR may not convert characters with very large or small font sizes. This can make the most important characters and words unavailable for text-based systems.
- **Uni-Dimensional.** OCR does not catalog page placement information for characters even though page coordinates can be useful for classification and attribution. OCRd text basically only has one dimension: length, where words are either in front of or behind other words.
- **Non-Symmetrical DPI for Faxes.** Faxes are often stored in files where the number of dots per inch horizontally is not the same as the DPI vertically, and OCR engines can have difficulty with this non-symmetrical DPI.
- **Sequential Editing.** OCR errors typically must be corrected sequentially with the same errors being repeatedly being edited. Global spell checking this can also introduce other errors.
- **Case Sensitivity for Editing.** The use of spell checking to correct OCR text may not permit the case of the letters to be considered, e.g., cat and CAT will be treated alike.
- **Partial Text.** When files contain some text, OCR engines may not attempt to convert images to text. This can cause OCR to miss text from images embedded in documents, and miss text when textual headers, footers, or legends were added to image-only PDFs.
- **Non-Textual Glyphs.** Often there are important non-textual characters or glyphs that do not get converted to characters by OCR, leaving them invisible for text analytics or text-based retrieval, e.g., logos, map symbols.
unneeded files

- Responsiveness/Relevance in discovery
- Setting retention schedules
- Setting access rights based on the work section or job classification of individual workers
- Setting storage security requirements
- PII detection/protection
- To indicate need for specific image enhancements

Classification systems ought to permit multiple looks at the same content, otherwise the organization may have to pay for and support multiple classification systems for multiple purposes.

7. **Use Stakeholder Collaboration**

Multiple heads are better than one for capturing enterprise knowledge about the business reasons that different documents types are created and used and what the regulatory requirements are that govern their use and disposition. Classification work flows that involve collaborative classification schemes will avoid many downstream problems later on.

For example, if subject matter experts from the business unit, finance, and records management can all consider the same documents at the same time they can help ensure that the most useful classifications for the whole enterprise are applied.

This approach also pays huge dividends in the next step of Attribution.

8. **Anticipate the Awareness Gap**

While using collaborative teams will help ensure that the collective wisdom of the enterprise is brought to bear on the classification task, there will still be document types of which team members are simply unaware. They may do a good job classifying files or documents presented to them, but how do you know that substantially all files being classified have been considered?

9. **Use Classification Matrix Approach**

A typical file classification scheme has the business unit or function as the top level with each unit or division handling for lower level classifications. This can lead to overly complex and possibly inconsistent labels for files that are for all purposes the same types.

For example, Finance might call one document type “Authorization for Expenditure,” Exploration & Production could call the same thing “AFE,” while Engineering might call it “Expenditure Authorization.” These different labels for the same type of document cause clutter and confusion later downstream.

An alternative approach is to build a matrix with the business units or functions listed in the left column and document types listed as column headers with Xs or check marks within cells to indicate if that document type is used in the unit or function. The advantage of this approach is that it minimizes the number of classifications and helps assure the use of consistent classification labels for the same documents. See the example in Table 1.
10. Test Consistency

The top three criteria for classification are consistency, consistency, and consistency. Test and measure consistency. Do you get the same results when you reprocess the same content? When you do text searches do you find the same type of content has been assigned two or more classifications? Consider having an alternative classification technology re-classify already classified documents and determine which system performs more accurately.

11. Start-Up and Ongoing Maintenance Costs

Different classification technologies involve different costs of consultants, in-house staff, computing resources, and per file or per gigabyte licensing fees. When evaluating competing alternatives, ask for explicit maximum costs and agreed accuracy levels to be included in a service level agreement (“SLA”).
<table>
<thead>
<tr>
<th>#</th>
<th>Top-Level Doc Type</th>
<th>Finance</th>
<th>HR-IT-Legal</th>
<th>Operations</th>
<th>Subsurface</th>
<th>Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AFE</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>Analysis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>Assessment</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>Audit</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>Basis of Design</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>Blueprint</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>Book</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>Budget/Forecast</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td>Bus Case Proposal</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>Calculation/Analysis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td>Catalogue</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Certificate</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>13</td>
<td>Checklist</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>14</td>
<td>Claim</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>15</td>
<td>Conservation Inf.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>16</td>
<td>Contract/Agreement</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Cost Estimate</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>18</td>
<td>Cost Time Resource</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>19</td>
<td>Data</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td>Data Sheet</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>21</td>
<td>Definitions</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>22</td>
<td>Details</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>23</td>
<td>Diagram</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>24</td>
<td>Drawing</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>25</td>
<td>Drill Well on Paper</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>26</td>
<td>Evaluation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>27</td>
<td>Emergency Response</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Form</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Guidance Docs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>30</td>
<td>Handbook</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>31</td>
<td>HAZID</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>32</td>
<td>HAZOP</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>33</td>
<td>Indexes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>34</td>
<td>Instruction</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>35</td>
<td>Interpretation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>36</td>
<td>Legends &amp; Symbols</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>37</td>
<td>Licenses</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>38</td>
<td>Logs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>39</td>
<td>Mgt of Change Doc.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>40</td>
<td>Manuals</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>41</td>
<td>Maps/Charts</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>42</td>
<td>Material SDS</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>43</td>
<td>Minutes of Meetings</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>44</td>
<td>Organization Chart</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>45</td>
<td>Papers</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>46</td>
<td>Permits</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>47</td>
<td>Philosophies</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>48</td>
<td>Photographs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>49</td>
<td>Plan</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>50</td>
<td>Policy Documents</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>51</td>
<td>Presentation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>52</td>
<td>Press Release</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>53</td>
<td>Procedures</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>54</td>
<td>Programs/Plans</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>55</td>
<td>Query</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>56</td>
<td>Questionnaire</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>57</td>
<td>Quote</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>58</td>
<td>Recommendation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>59</td>
<td>Register/Schedule</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 1. Document Classification Matrix
C. Attribute — Extract What’s Important

Document attributes are things that are evident on the face of the document. Attribution involves identifying which of those data elements are significant from a business, legal, or regulatory purpose in each of the classifications, and extracting them so they can be used for retrieval, reporting, or as part of a decision management tool. They can be used to populate or to validate structured database entries.

Here are considerations when attributing files:

1. Identify Specific Desired Attributes for Each Classification.

For each classification, the organization should identify the types of data elements or attributes it wants to extract from members of that classification. This list becomes a checklist to consider when designing or performing attribution.

2. Examine Data Fields in Control or Retrieval Systems

If information from unstructured content is used to populate or to audit a process control system (e.g., the name of a loan applicant is used in a mortgage loan management system), those data elements should be considered for inclusion on the attribute list as automated extraction can speed input and auditing of such data. See Figure 3.

3. Be Aware of Classification-Dependent Limitations

The ability to extract desired attributes depend often on the consistency with which initial classification is performed. If classes are consistent and files within classes are very much alike it will be much easier to extract data values. On the other hand, inconsistent or overlapping classifications will lead to poor quality attribution. Keep those limitations in mind when considering what to try to extract.

4. Be Aware of Text Conversion Dependence

Systems that use text conversion technology to provide textual values from image-only files will
be subject to the limits of that technology. If the underlying OCR system won't recognize characters with font size greater than 24, then any text of larger size won't be available for auto-extraction, and the same will be true for minimum font size limitations. As discussed earlier, text dependence is also a significant issue for non-textual files, poor quality text files, and foreign language files.

See Section 1.B.2. Be Aware of Text-Dependency Limitations, and Sidebar 8, Limitations of OCR.

5. **Be Aware of Extraction Tool Dependence**

The ability to extract desired attributes is dependent on the tools available for the extraction, e.g., can the extraction specifications include absolute page coordinate zones or positional specifications relative to other zones? There is no point in specifying attributes that the available tool set cannot identify.

6. **Consider Non-Textual Attribute Extraction**

Although people usually associate attribution with text extraction, be aware that non-textual elements, such as signatures or logos, can also be extracted and stored in a properly designed system.

See Sidebar 9, Non-Textual Attribution: Signatures.

7. **Normalize Extracted Values**

When designing attribution, provide a mechanism to normalize content so terms are stored in a consistent format. For example, in the Oil & Gas industry, the Well Number is a key data element, but it can occur in many slight variations, some with spaces, some with dashes, some with spelling variations. Normalizing such data elements maximizes the value of the attribution for reporting and retrieval purposes.
8. **Use and then Update External Authority Lists**

Often management systems maintain authority lists that ideally represent all known values that can appear in a particular field, e.g., a list of the American Petroleum Institute Well Numbers from an oil & gas well tracking system that tracks active oil and gas wells. The authority list can help find in which document types those terms occur, and this can guide attribution guidelines. If references to the terms are essentially random within a type, they can be ignored. If they are consistent or largely consistent, they can be considered for inclusion.

Comprehensive attribute extraction typically results in identifying a significant additional number of values for the authority list, and those will need to be reconciled.

9. **Involve All Key Stakeholders**

Have all significant stakeholders involved in deciding what to extract, how to label each item, and how to format it. Ideally this collaboration would take place in the same room looking at examples of the file types at the same time.

10. **Build Comprehensive Logs for Full Audit & Presentation**

Logging of all extracted values permits the full auditing and authentication of those values. The system should know the document, page, and page coordinates for each extracted value. This not only greatly expedites auditing any control system to see if the supporting documentation confirms the source of the entered data, but it can also be used for on-screen presentations to minimize the “stare and compare” time otherwise spent comparing data from different documents.

11. **Consider Full Text as an Attribute**

One of the attributes of a file may be the text associated with the entire document without trying to identify specific attributes or field values in the full text. This is sometimes called “content enablement,” and it should be considered as a supplement to extracting field values. While full text can be useful, it will not provide for the precision available by searching specific fields in a database and will not provide the level of report formatting and sorting available with attributed data values. When extracting full text, always measure the quality at the character and word level to give a sense of how reliable it will be when used for retrieval or other analysis.
D. Validate — How Dependable Is it?

Validation involves measuring how accurate and complete the resultant data is. It incorporates questions like, did we account for all the files we encountered doing the inventory? Are the files classified accurately? Were the correct attributes extracted from the files and were they properly formatted and loaded?

Here are tips or hints for this phase:

1. **Automate the Validation**

Whatever the process involved in validation, it should be automated so validation is consistent and results can be compared over time. An automated process should also save the time of the people doing the validation, making it far more likely that they will validate data regularly. One sign of a non-automated process is if someone must use Excel to generate a random list of files to examine or has to use Excel to tally the results of having evaluated files.

2. **Validate Deadline Compliance**

Besides focusing on the results of the classification and attribution, it is also important to validate that the results were delivered *on schedule*. Mechanisms should be in place to document when files were available for classification and when the results were received.

3. **Compare Before & After Totals**

All files initially examined should be accounted for. The accounting should include categories like:

- Files Examined
- Files Uncompressed
- NIST list
- Content Duplicates
- Unique Content Files
- Encrypted/Decrypted
The unique content files should also be accounted for:

- Unencrypted
- Remaining Unencrypted

The number of files retained should be compared to the number loaded into whatever the target search or content management system.

See Sidebar 1 in Chapter I about calculating the volume of files not collected or indexed because the total path and filename exceeds the Windows 260-character limit.

### 4. Measure Classification Accuracy

Consistent, accurate classification is the lynch pin for practically any document-related information governance initiative so it is critical to assess the quality of the classification. This can be done in several ways:

- **Reprocess a Benchmark Sets of Files.** If classification involves writing new rules or scripts that assign classifications, it will be useful to keep a standard set of documents that are reprocessed for each new rule/script change to see if the changes have adversely affected existing classifications. Regardless of how classifications are assigned it will always be wise to periodically reprocess benchmark sets of documents or files to ensure that the system has not gone off track.

- **Peruse Files within Same Classifications.** Another way to perform ad hoc review of classification consistency is to examine sets of documents all assigned the same classification. This need not be a lengthy process as systemic errors will often be obvious.

- **Review of Randomly Selected Documents.** End users can randomly select documents, sort them by assigned document type, and evaluate whether the assigned document types are accurate. Depending on the sample size, this may not adequately assess classifications assigned to relatively few documents as they may not be included in the sample. To combat this, the process used to select the random sample should ensure that at least some minimum number of files are selected from each classification.

- **Review Results of Full-Text Searches for Specific Document Types.** Finding files correctly classified when you search by specifying the class can lead to a false sense of confidence about the quality of classification for all the content. The search results don't show

---

**Sidebar 10**

**False Positives and False Negatives**

- **False Positives**: People searching for the incorrectly applied classification will have the incorrectly classified document in the search results.
- **False Negatives**: People searching for the correct document type will miss finding the misclassified file.
- **Cascading Attribution Errors**: Because attribution is linked to document type, assigning the incorrect document type can cause correct fields or attributes to be missed while extracting incorrect ones.
the files that should have been assigned that classification - it doesn't show the false negatives. One way to gauge the extent of this is to try to retrieve documents within the desired classification without using the document type as a search parameter. Use full text search to try to pull up files in the desired class. You may find files that look like they should have had the desired classification but didn't. Maybe some classifications could be consolidated. One problem with this approach is that it is text-dependent and will not work with documents with text dependency problems (e.g., no text, poor text, non-English text).

- **Search for Sets of Related Document Types.**
  In collections where one would expect to find a several document types relating to the same transaction, event, or item, users can search terms that describe that transaction, event or item and see if all the expected document types are represented. For example, in a home mortgage/loan system the user could search for a specific borrower and see if all the expected document types associated with a loan are present. This again is subject to all the usual limitations of text-based search.

- **Evaluate Files with Residual Classification.** It can be informative to review files or documents that were assigned the “Other” or “Unknown” classification. You may find there are several files not being classified by a text-based algorithm because of text-related issues (no text, poor text, language, lack of sentences). You may also discover new document types that should be included in the overall classification scheme.

### 5. Measure Accuracy of Attribute Extraction

It is vital to measure the accuracy of attribute extraction so the organization can be sure it’s receiving value for the effort and funds expended on attribution and so it does not place unwarranted reliance on the results of attribution.

Here are ways to confirm the accuracy of attributed data:

- **Compare extracted data value to source document.** The best way to confirm accurate extraction is to go to source documents to confirm

---

**Sidebar 11**

**QA for Scanning Projects/Processes**

Here are two ways to evaluate whether all the pages in a set of documents were scanned. One is to record the weight of each container prior to scanning and then estimate the number of pages based on the net weight and compare that to the scan count. This can help disclose when major portions of a container were not scanned.

Another technique is to turn image capture off on the scanners and to feed the container contents a second time to just obtain page counts to use to compare with the number of saved images.

Manual comparisons of images to hard copy content is very expensive and significantly increases labor costs.

---

**Sidebar 12**

**Incorporating Validation Measurements in Service Level Agreement (SLA)**

Validation measures should be included in the Service Level Agreement, including:

- How accuracy is measured
- How files or data values are selected
- Time limit for testing
- How long does the vendor or process have to deliver the results back.
- Consequences of validation failure (do-over, lower price, not paid, etc.)

Internally an organization should:

- Identify who’s responsible for testing
- Detail how to notify the vendor and accounting department of any necessary adjustments
the data value. However, in some systems this will be a time-consuming process.

- **Compare extracted values to trusted authority list.** If an extracted value matches one of the values on a trusted authority list, the organization may be able to assume it was extracted accurately.

- **Confirm missing values.** If the file or document classification was accurate and the attribution list identifies which values are associated with a particular document type, quality assurance efforts should involve examining fields expected to have values but don’t.

- **Focus on the key data elements.** Some data elements will be more critical than others, e.g., the loan number will be a critical value in a home mortgage tracking system. Focus on the critical items.

6. **Use Ad Hoc Analytics**

Reviewing metrics associated with rationalization, classification, and attribution can disclose many unexpected results. For example, document date is a commonly extracted document attribute (this is the date that appears on the face of the document, not the system date), and reviewing graphics showing volume on the Y axis and time on the X-axis can make spikes or gaps readily apparent. Analytics can also make outliers more apparent, e.g., document dates well before or after expected date ranges. People familiar with the content should be able to explain these aberrations, e.g., a spike was caused by acquiring another company, or a gap was caused by a company-wide strike.

7. **Double-Check Duplicate Tracking**

Some software only tracks one location for duplicates. However, there can be many reasons to track back to all the original sources of duplicate files, and consider doing periodic audits to ensure that all copies are tracked.

8. **Know What Didn’t Get Captured or Extracted**

Validation should include looking at what wasn’t captured. For example, if the data value for Borrower’s Name should have been extracted from Loan Applications, the validation should include looking at Loan Applications with no Borrower’s Name listed to determine why attribution failed to provide that data value.

9. **Reconcile Authority Lists**

Most systems can generate authority lists which are usually alphabetically-sorted lists of the unique values that appear in specific fields. Reviewing such lists is often an easy way to spot data values that either shouldn’t be there or should be there in a different format. Someone familiar with the content may also notice that terms that should be included are not on the list. At the beginning of the process, domain experts should examine the new values being extracted from the unstructured content and compare those with the previous authority lists to confirm that the correct values are being extracted.
10. Compare Results of Using Other Systems or Approaches

Other than not processing files that ought to be included, the biggest error that can occur is to have the wrong classifications assigned to files. This causes numerous downstream problems with classification-dependent attribution and loss of user confidence in being able to depend on classifications for retrieval.

One way to validate the accuracy and consistency of overall classifications is to apply the same rules using a different classification technology, or to apply different technology and then compare any differences in classification groupings. It may disclose that both systems have strengths or it may disclose that one is superior across the board.

11. Incorporate User Feedback

The real acid test is whether users find the classifications and attributions to be useful and reliable. If not, all the effort may be for naught. When the outputs are first loaded, users should be encouraged to share what they like or don't like about the results. Are they looking for things they can't find? Are they experiencing clutter in their search results? These are just some questions whose answers should be gathered and analyzed so classification labels and attribute names and formatting can be adjusted going forward.
II. Overview of Visual Classification

Visual classification groups files or documents based on their appearance - without being subject to the multiple limitations and difficulties of text-based classification approaches. To show you how this sort of facial recognition for documents works, you can probably identify which of these thumbnail representations of documents is the email, the agreement, the invoice, and the letter, despite not being able to read any of the text:

Figure 5. Thumbnail Images Showing Text Not Needed for Classification
By focusing on appearance, visual classification can use document properties or characteristics like layout and content placement to help identify similar objects. After files have been accurately grouped and classified, the text characters shown on the document can be used to make the content searchable and to extract specific data values, but text is not needed or used for the initial classification.

Before drilling down into specific features in the next chapter, here’s a high-level overview of how visual classification works:

Visual classification creates visual representations of all content files and then clusters or groups visually similar ones. By basing classification on the appearance of the contents, visual classification normalizes the content regardless of the container that stored or transmitted it. For example, the same content could exist in these four different file types:

- The Word file (*.docx) that was used originally to create the document,
- The PDF version used to distribute it,
- A scanned TIF copy of a printed version, and
- A FAXed copy.

Visual classification groups all four copies in the same cluster regardless of whether text was included in any of them and regardless of whether some of the image-only copies had varying resolutions.

The most significant advantages of visual classification are:

- **No text dependencies.** Visual cluster formation is based on what could be considered a kind of facial recognition or fingerprinting for documents. It has none of the usual text dependencies, i.e., require no text much less perfect text or complete sentences, and it is language agnostic.

- **Automatic cluster formation.** Visual classification clusters form automatically, there are no rules or scripts to write, no taxonomies to create and tune, and no exemplars or seed sets to select. Once clusters are formed, domain experts can apply document-type labels or classifications to them.

- **Consistency.** Visual classification consistently clusters visually similar doc-

![Figure 6. Classification Labels Applied to Clusters of Visually-Similar Files](image)
documents. This consistency enables simpler, faster implementation of other basic downstream
information governance tasks like attribute extraction and retrieval.

With visual classification it is much easier to identify which data elements to extract from the face
of the documents in each cluster than in text-based systems where the clusters are far less consistent
and much more unpredictable.

Once documents are clustered based on their visual similarity, it’s easy to see that seemingly un-
structured content really is structured within clusters. The same types of data elements occur in
generally the same places on the documents within a cluster. The document types and data elements
weave a fabric or structure that is a counterpart to fielded database management systems.

As illustrated in the following document attribute matrix, many document types share common
data elements, and those common data elements can be used to consolidate information about
specific items referenced in multiple document types.

![Document Attribute Matrix](image)

**Figure 7. Document Attribute Matrix**

With this overview in mind, let’s review more of how visual classification works in each of the
primary steps for managing unstructured content: rationalization, classification, attribution, and
validation. Within each step we will talk about the intelligence accumulated that becomes available
for later collections as well as the deliverables that are tied directly to the processed content.
III. Visual Classification in Depth

This chapter presents a closer look at each of the four steps in the RCAV model as applied to visual classification technology developed by BeyondRecognition (“BR”):

- Rationalize
- Classify
- Attribute
- Validate
A. Rationalize — How Many Files Where

To manage unstructured content, you have to rationalize it. You have to figure out how much of it you have and where it is.

1. Collectors

BeyondRecognition (“BR”) accomplishes this with its collectors, which are USB devices that attach to networks. They can be thumb drives for smaller collections or servers with a PB of RAID storage.

2. Logging & Hashing

Collectors are scripted to tell them which drives or folders to examine. They can examine many operating systems including Windows, Unix, and Apple. When examining the designated drives or folders, Collectors log the name and location of each copy of each file it examines, and calculates and records the SHA hash value of each file.

3. DeNISTing

SHA hash values are used to determine if each file is listed on a list of known commercial software executables and other software publisher-related files. BR uses the software reference list published by the National Institute of Standards and Technology (“NIST”) as augmented with lists from other trusted sources of files and hash values. If a file’s hash value appears on this augmented NIST list, that is noted in the log, but the file is not copied.

4. Deduping & Copying

The SHA hash values are also used to determine if another instance of a unique content file has already been copied to the Collector so that only a single instance of files with the same hash value is collected. By deduping and deNISTing, the Collector maximizes its use of its storage capacity by only copying single instances of content files.
5. **Compression & Encryption**

The Collector compresses the files it stores to further increase the data it can collect. As an additional layer of security it also encrypts the stored file using an encryption key that is a different length than the one used for compression.

6. **Capacity**

Because Collectors can be locally attached where they are not limited by network bandwidth, they can collect up to a PB in 7 to 10 days. They can also be configured to collect over networks but not compete for resources during heavy traffic times.

7. **2,600 Character Path Limit**

Unlike standard Windows applications, the Collector can examine folder paths of up to 2,600 characters, ten times the depth of standard Windows applications. BR has encountered instances where up to 2% of the files copied were originally on paths whose length exceeded the Windows limits. Those files would have been invisible or lost to the standard Windows collection devices or software.

8. **Inherited Security**

Collectors use an inherited security model in which the folders and files they can access is controlled by the security settings on the port to which they are attached.

9. **UnZipping**

Collectors do not unzip or uncompress files during the logging and collection phase. However, when data is brought back to the BR server, the server recursively opens all container files, hashing and logging all contained files. Those files are then subjected to an additional round of deduping and deNisting so the classification process will have unique copies of all content files for analysis.

10. **Rationalization Deliverables**

There are two types of deliverables from the rationalization step:

**Intelligence.** The scripts developed to instruct the Collector what drives or folders to inventory and copy can be used again later to collect any newly added files. The list of hash values associated with already collected content files can also be used to streamline future collection efforts.

**Content-Specific Results.** Rationalization produces the following deliverables that are tied directly to the files rationalized:

- **Rationalization Log.** The log includes values for each file:
- Original path/file name location
- SHA hash value
- File Size
- NIST Y/N
- Dupe Y/N
- Parent SHA Hash Value (for unzipped files)

**File Copies.** Single-instance copies of content files named using the file's SHA hash value. The SHA hash value can be used to determine where each copy of the file was located by reference to the Rationalization Log.
B. Classify — What’s in the Files (Visual Clusters)

This is the process used to classify files using visual classification:

1. **Visual Representations**

   After content files have been collected and deduped, BR makes temporary/interim visual representation of each file. If files are encrypted, the fact of the encryption is noted on the log and the encrypted files are handled as instructed by the client, e.g., brute force decryption is attempted, custodians are contacted for passwords, etc. See Section I.A.10, Anticipate Encrypted Files.

2. **Global Glyph Catalog**

   Visual classification examines the visual representations of the files and catalogs where each individual symbol or graphical element is located. A symbol or graphical element could be a letter, number or punctuation mark or it could be a Nike logo or a staple hole. It is essentially a set of contiguous pixels on a contrasting background.

   The glyph catalog is then used to identify various levels of patterns including the page-level pattern used to identify document boundaries and classify documents. It enables several important tasks in visual classification, such as clustering pages and permitting single-instance editing of text values associated with specific glyph clusters. *We will discuss the editing function in more detail in the Attribute and Validate sections.*

---

**Sidebar 15**

Classification Overview

**Setup:** Need:
- Single-instance copies of content files
- Classification Tree

**Output:**
- Classifications assigned to files in visually-similar clusters

**Sidebar 16**

Visual Classification At-a-Glance

- Consistent classification across all content sources
- Objective *automatic* clustering - minimal front-end loading
- NOT text-based, classifies even non-text files
- Language agnostic
- User-definable multi-tier classification tree
- Scalable to 100s of millions of files
3. **Cluster Visually-Similar Files**

Using the catalog of the glyphs or graphical elements contained on each page, visual classification *automatically* clusters visually-similar files. Users can begin reviewing clusters within a day of the beginning of the classification phase.

This is what clusters of files from the pipeline industry might look like with no prior human intervention to define or form the clusters:

**Initial Automatic Visual Clustering —
No Setup Required for Scripts, Exemplars, or Seed Sets**

![Figure 8. Automatically-Formed Clusters of Visually-Similar Files](image)

One fact may be obvious: *Once files have been clustered based on visual similarity, they do in fact have a recognizable structure.* Specific types of information occur in the same places on the pages within the documents in a cluster. This will be exploited during attribute extraction discussed in the next section.

Note also that visual classification places files of different file types in the same clusters. For example, scanned files or faxed files will be placed in the same clusters as files such as Microsoft Word that were the original source for later PDF, TIF, and FAX versions of the same content.

4. **Classify — Assign Document-Type Classification Labels**

Because documents within visually-similar clusters are so alike, all the files in a cluster can be reliably classified by examining just a few. *The designated classifications are “persistent,” meaning they*
will be applied to all files that later fall in that cluster.

This is how the clusters shown above might be classified in a records-management type review:

```
Figure 9. Clusters of Visually-Similar Files with Classification Applied
```

5. Classification Tree

When reviewing visually-similar clusters, subject matter experts can select from the choices available on a classification tree displayed on part of the screen. The classification tree can include up to three levels and is user-definable.

The software interface used for classifying clusters of visually similar files is shown on the next page. The clusters are shown in the upper left part of the screen, sorted with clusters with the largest number of contained files at the top. The main part of the screen shows a document image, and the table at the bottom of the screen has one row for each document or file in the cluster. Changing rows in the cluster area changes the cluster being considered, and changing the row in the document table at the bottom changes the document image being displayed in the main part of the screen.

Double-clicking on the desired classification in the classification tree on the upper right hand part of the screen assigns that document type label to all documents in the cluster.

```
Sidebar 17

Taxonomies

Visual classification can significantly affect the effort expended to develop and maintain taxonomies.

If taxonomies are only used to classify files, the consistent classification from visual classification can eliminate the need for taxonomies. If taxonomies are used in other systems for other purposes, visual classification can provide word frequency lists from just the files in a class to help develop taxonomies more accurately and more efficiently.
```
When files are being classified as part of an ongoing records management or content migration process the classification tree can involve over a hundred classifications, typically with the business unit or function as the top classification, and then the main document type as the second level. Clients can also use a third sub-document type classification level.

Other use cases (see Chapter III), can involve much simpler classifications, e.g., for e-discovery, clusters could be classified as either Responsive, Non-responsive, or Needs Review.

6. Benefits of Visual Classification Approach

Some benefits of visual classification include:

- **Persistence and Multiplier Effect.** File classifications persist as new files are added to clusters. The return on the investment of time in classifying a cluster grows as the clusters grow over time. The following table shows how the effort (about 90 minutes) to classify 73 clusters in the first segment processed directly affected the 19,682 documents in those clusters, but then also classified documents that were later added to those clusters in the next 12 segments, ultimately impacting 327,497 files.

- **Convergence.** With visual classification a point of convergence is reached at which time incoming files substantially all fall into clusters that have already been classified.

- **Easier maintenance.** After convergence classification effort is limited to any new clusters that form, making maintenance a very manageable task.
7. Document Unitization of Multi-Document Files

There can be more than one “document” per file, especially with PDF or TIFF-format files. In some collections there may not be many multi-document files However, in others there may be a sizable percentage of files containing multiple documents.

If the files that contain multiple documents are not identified and the proper document boundaries inserted, usually only the top document is correctly referenced in the content management or process control system. The document type, author, date, title, and other attributes of the second and subsequent documents are lost for search, display and reporting purposes. The combined documents also make text-based analytics or classification systems less accurate even if all the pages have perfect English text in complete sentence structure.

Visual classification can be used to identify the pages that look like other pages that begin documents and to then insert logical document boundaries at those points. The document boundaries inserted using visual classification are more accurate and more consistent than document boundaries inserted by humans performing manual boundary assignment. The outcome is that the embedded documents, the ones previously buried under the top document in a file, are now fully attributed (See next section) and searchable.
Visual classification clients may accept initial document unitization boundaries or detect and insert proper document boundaries in files containing multiple documents.

### 8. Document Unitization of Single-Page Files

The logical opposite of multi-document files are collections comprising single-page files where it takes one or more page files to fully represent an individual document. Visual classification can be used to identify the pages that begin the different document types and then to associate pages into documents. This automated unitization of single-page files is faster and more accurate than manual unitization.

### 9. Learning from Previously Classified Content

Visual classification can examine content that was classified by another system to learn how files have been classified in the past. In this use case, the original classifications are an additional input. Using previous classifications has the added benefit of auditing those initial classifications. Typically, the original classification system is found to have assigned multiple classifications to files that contain the same content.

### 10. Classification Deliverables

These are the outputs from the classification process:

- **Re-usable Intelligence.** These are decisions that can be reused for other collections:
  - Classification tree or matrix that lists approved document types.
  - Classification decisions tied to Cluster IDs.
  - Graphical profile of what files in visually-similar clusters look like.

- **Content-Specific Results.** These are the results of processing the specific content:
  - CSV with hash values in one column and assigned classifications in another, and/or
  - Folders with classification labels for names with each folder holding all files assigned to that classification.
C. Attribute — Extract What You Care About

Attribution involves specifying the data elements you care about on the face of the documents in each cluster and then extracting them so they can populate or audit content management systems or business process support systems.

1. **List Attributes**

Attribution starts by creating a list of the attribute values that should be extracted for each type of document on the classification tree or matrix. Typically, the group of subject matter experts who established and assigned document classifications will also specify which values to extract for each document type.

2. **Locate Attributes within Clusters**

Another, larger group of document specialists will review a few documents in each cluster and identify the data elements to extract by clicking-and-dragging a rectangle around those elements using a graphical user interface, and then specifying what attribute name or label to use.

3. **Delimiters**

There several delimiters that can be used to identify what to extract from within the attribution zones. They include delimiters for:

- Relative size of the font
- Text strings preceding the value
- Text strings within the value
- Number of lines before the value

---

**Sidebar 18**

**Attribution Overview**

**Setup:** Need:
- Classified Groups of Visually-Similar Files
- Attribute List for Classifications

**Output:**
- CSV file w/Attribute Values
- Customized Output

**Sidebar 19**

**Visual Attribution At-a-Glance**

- Attributions set within clusters
- Tools include zoned and text string-based extraction
- Extraction zones can be absolute page coordinates or relative to other zones.
- Each value can be formatted
4. **Formatting**

Users can also specify the formats to use for the extracted values, for example several date formats can be used.

5. **Single-Instance Glyph Editing**

When domain experts are designating the page location from which to extract attributes the system provides immediate feedback on exactly what was extracted. It also uses many techniques to determine if it has associated the correct text value for the extracted values. Users can specify what text values should be associated with individual glyphs. These corrections are then cascaded throughout the collection wherever that glyph occurs. The result of the cascading corrections is that individual corrections can affect hundreds of thousands and sometimes millions of edits without having to correct each occurrence or each word in which the glyph occurs.

6. **Non-Textual Attribute Extraction**

While the normal object of attribution is to extract text values for the attributes, sometimes non-textual objects may be useful, e.g., logos or signatures. BR permits the extraction of such non-textual objects by its graphical search capability made possible by the global glyph catalog that stores the appearance and location of all the individual glyphs.

7. **Using Existing Authority Lists to Find Potential Attribution Locations**

One of the ways to find where desired attributes are within collections is to search the collection to find documents containing terms on authority lists from other systems of record. For example, a Well Tracking system might have the names of all known wells in a specific field. To find which doc-
Document types might be a source for content relating to wells, the whole collection could be searched for those terms.

8. Content Enablement — Converting All Character-Related Glyphs

Although attribution is usually associated with trying to locate what might be thought of as fielded data, one attribute commonly extracted is the “all text” attribute, i.e., text values for all glyphs in a document that have associated text values, whether or not the glyphs are associated with a more specific attribute.

9. Negation to Yield Residual High-Value Terms

Some visually similar clusters contain what are essentially form or template-based documents. Because the global glyph catalog knows the identity and location of all glyphs it can negate or drop off the glyphs that occur in the same place in all the documents in a cluster. Dropping the recurring elements found in the underlying form or template leaves just high-value terms.

Having just high-value terms can make it possible to locate the other document types that may pertain to the same person, incident, or other logical “object.” For example, in a mortgage loan file there will be multiple document types but many will have the name, address, phone number, date of birth, etc. for the mortgage applicant. Negation can provide a way to help associate all the documents pertaining to a particular person or transaction.

10. Benefits of Visual Classification Approach to Attribution

There are two main advantages to attribute extraction with visual classification:

- **Tight Clusters Simplify Attribution.** Grouping visually similar documents results in clusters of documents where the same types of information appear in the same places on the faces of the documents in a cluster. This uniformity within the clusters makes for much easier attribute extraction.

- **Fixed and Relational Page Coordinates.** The glyph catalog knows where every graphical element on a page comes from. This lets users specify the zone from which they would like to extract attributes. The zonal specifications can be either fixed locations on a page or from zones that are relative to some other zone.

11. Attribution Deliverables

These are the outputs from the attribution process:

- **Re-usable Intelligence.** Decisions made during Attribution can be reused for later processing in the same or other collections:
  - Lists of data elements to extract from each classification.
  - Attribute extraction specifications for each cluster ID to include:
    - Page location coordinates
• Delimiters to specify data elements within the zones
• Formats to use for selected data elements.

**Content-Specific Results.** These are the results of processing the specific content:
• CSV files with hash values and each extracted attribute.
• Folders with classification labels for names with each folder holding all files assigned to that classification.
D. Validate — Be Confident of Results

Validation involves measuring and ensuring that all information gathered about unstructured content is correct and complete. BeyondRecognition’s process includes these checks, beginning with the most basic:

1. **File Count**

The BR Collector creates a log of all files encountered and their hash values, whether or not copied. That log is expanded with information from files extracted from container files like (*ZIP or *RAR) on the BR server. The log is then updated during processing to track and confirm these conditions for all files:

- Located beyond 260-character path
- System file as indicated by expanded NIST list
- Unique non-system file
- Duplicate non-system file
- Corrupted
- Encrypted
- Decrypted
- Classified
- Loaded

2. **Classification Accuracy**

Consistency is the most important measure of the quality of classification. The BR interface permits users to select any visual cluster and quickly browse the files placed in that cluster to confirm that they are appropriately labeled.

The same approach can pull up all clusters of visually-similar files associated with a given document type label. For example, there may be several visual clusters for invoices.

---

Sidebar 20

**Overview of Validate**

**Setup:** Need
- Earlier Output
- User Feedback

**Output:**
- Validated Text Values
- Adjusted Classes or Attributes

Sidebar 21

**Validation At-a-Glance**

- Complete audit trail via logs
- File Hashes and Cluster IDs are key data elements tying together data from all four phases.
The visual cluster ID can also be loaded as a metadata value into whatever content management system will manage the unstructured content so the validity of the clustering or label assignment can always be verified.

3. Attribution — Specifying

When the document experts are specifying attribute zones, extraction rules, and formatting specifications, the BR system presents the extracted values in real time in a spreadsheet-type table with rows for documents and columns for attributes. This immediate feedback permits the experts to adjust the zones, rules or specifications in a highly interactive manner.

4. Attribution — Single-Instance Editing

One of the biggest challenges in extracting document attributes is to ensure the accuracy of the text values associated with specific glyphs. BR permits users to change those associations with a process in which editing the association once impacts all occurrences or instances where that glyph occurs. This improves consistency and decreases the effort required to obtain high levels of accuracy.

5. Attribution — Comparison with Authority Lists from Systems of Record

Often the values extracted from unstructured content can be compared to authority lists of values in other systems of record, e.g., the API Well Number in oil & gas documents will often be contained in at least one other management system. Those values can be compared to the extracted values and any differences reconciled. Invariably the extracted values are more comprehensive and authoritative.

6. End-User Feedback & Course Corrections

The acid test of any unstructured content management system is whether it helps end users use the content or it helps managers manage the underlying business process. With visual classification each piece of managed content is tied together with three values:

- **Hash values** that identify unique values and the locations where they occurred,
- **Visual Cluster IDs** that identify which files were visually similar, and
- **Classifications** that identify which Cluster IDs were associated with them, and which attributes were to be extracted from the files in the associated clusters.

All three of those values can be loaded in the final content management system so that as users and managers work with the content they can identify where corrections ought to be made, or additional attributes extracted. Those values provide an audit trail of how/why decisions were made, giving a process transparency unrivaled by other systems which typically only provide the conclusion reached about the classifications assigned or attributes extracted.
IV. Visual Classification Use Cases

Visual classification has many use cases in information governance, information security, and records management. Anytime an organization must manage unstructured content the process must include classification. Without classification, all files must be treated alike and because some files must be retained, that would mean that everything would have to be retained. No other technology does a more consistent or a more scalable job of classifying files.

Visual classification use cases can involve fairly granular classification and attribution, or can involve a relatively small number of classifications with little or no attribution. This chapter describes these types of use cases:

- Fairly granular classification schemes and detailed attribute extraction
- Broad classifications with little if any attribution
- Special use cases including:
  - Determining project scope
  - Email
  - Redaction
  - Taxonomy development
  - Contract review
• Graphical search
• Agency comment response system

Whichever use cases are first implemented, the intelligence developed in implementing them is available to reduce the time and effort required to implement later use cases and ensure consistent classifications across the organization. For example, if file share remediation is the first project, the document classification tree and the attribution lists and zones developed for it could be used for other projects such as paper archive digitization.

This table summarizes the types of intelligence that can be rolled forward for later processing or projects.

### Re-Usable Knowledge Gained During Visual Classification Process
*(Eliminates duplicates on subsequent processing, reduces work required to set up or process other use cases in the organization, and ensures consistency across all enterprise use cases.)*

<table>
<thead>
<tr>
<th>Step</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Step Preparation</td>
</tr>
<tr>
<td>Rationalize</td>
<td>File Location Mapping</td>
</tr>
<tr>
<td>Classify</td>
<td>Classification Tree</td>
</tr>
<tr>
<td>Attribute</td>
<td>Attribute Lists</td>
</tr>
<tr>
<td>Validation</td>
<td>Edited Glyph-Text Pairs</td>
</tr>
</tbody>
</table>

Table 3. Re-Usable Knowledge Gained During Visual Classification Process

Here are descriptions of the multiple use cases:

### A. Granular Classification & Attribution Use Cases

These use cases all involve a fairly detailed or granular classification tree, typically involving over 100 classifications, with attribution used to pull eight or more attributes per classification:

- Records Management
- Content Migration to New ECM or Cloud-Based Content Management System
- System Decommissioning
- Silo Consolidation
- Paper File Digitization
- M&A Due Diligence and Record Assimilation

The outcome of these projects is that files that had been regarded as unstructured are assigned granular document-type designations that permit application of suitable records retention periods, and attributes are extracted to populate or validate decision support systems or content management systems.
Information security is enhanced by the ability to limit access to specific document types to those with a business need for such access. Information retrieval is simplified because of the consistency and dependability of the classifications and attributions.

When implementing these use cases, organizations often identify many files that need not be retained, freeing IT resources and eliminating much clutter from content searches. The scheduled disposition of the remaining content based on event triggers prevents future content build-up and hoarding.

When duplicate or unneeded files are identified, BR can provide “Terminator” software that forensically deletes those files.

**B. “Large Bucket” Classification**

“Large Bucket” use cases involve relatively simple classifications, generally with 10 or fewer rather broad classifications. For example, an e-discovery use may involve just separating responsive from non-responsive documents. These use cases rarely involve significant attribution although the files may be content enabled, i.e., made text searchable.

- Legal Hold (Hold or No Hold)
- E-Discovery (Responsive - Non-Responsive - Privileged - Sensitive)
- Simple Non-Record Remediation (Retain - Dispose)
- Image Enhancement (Enhance - No Enhance)

**C. Special Use Cases**

These use cases involve either using a small subset of the visual classification technology described earlier in this book or using some supplemental processing and technology:

1. **Scoping Projects**

Prior to deciding about individual initiatives, organizations usually want to determine their scope, i.e., to determine how large they are. The BR Collector can crawl specified directories, creating logs that contain the full path/file name of files encountered with their hash values, but without copying the files. This log can then be used to help scope the effort required to copy and process deduped content files.

2. **Email Management**

Visual classification provides an additional powerful tool to help identify emails of interest. For records management purposes, emails “of interest” may those considered to be “records” that must be managed. In e-discovery, emails “of interest” would be those that are responsive to discovery requests and must be produced to the opposing party.

Typically, many tools and techniques are used to identify emails of interest. For example, sender’s domain analysis or time period limits can be used to collect emails of interest for e-discovery, and
collections are often limited to specific custodians.

Visual classification adds to the tool available for analyzing email collections. Those other tools use one or more of the several unique features about email:

- **Duplicates.** Because of the way emails are distributed there are typically large numbers of duplicates of emails within an organization. One of the most fundamental ways to manage emails is to identify and consolidate duplicates.

- **Visible Structure.** Emails are semi-structured with visible structure for sender, recipients, date/time, and subject.

- **Hidden Structure.** Emails also have hidden metadata or structure with values that enable emails to be associated into conversations or threads.

- **Parsable Syntax.** The information about senders and recipients has a formalized syntax that can be parsed to reveal the domain name and user name within the domain for senders and recipients.

- **Attachments.** Emails not only contain information but are also used to transmit information in attached files. Often emails and their attachments must be evaluated together to understand the overall context.

Visual classification expands on the other tools by providing a way to evaluate email attachments. Attachments analyzed by visual classification to help identify which groups of visually-similar files are of interest.

Emails that attached files of interest are deemed of interest as they provide context for the attachments. The threads that contained attachments or emails of interest are also considered of interest. The thread inference causes an evaluation of even emails with no attachments if they were part of a thread that contained attachments of interest.

In this approach, emails are first normalized and then deduped with individual emails being placed in the threads in which they occur. See Sidebar 3 on Identifying Duplicate Emails in Chapter 1.

Using of visual classification is a very efficient way to classify large numbers of email objects. For example, using these assumption, evaluating one visual cluster containing email attachments results in classifying 600 threads with 2400 email instances.

<table>
<thead>
<tr>
<th>Impact of Classification Inheritance on Email Threads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor</strong></td>
</tr>
<tr>
<td># Email Attachments in Visually-Similar Clusters</td>
</tr>
<tr>
<td># Emails to which each Attachment is Attached</td>
</tr>
<tr>
<td># Emails in a Thread</td>
</tr>
<tr>
<td># Duplicates per Email</td>
</tr>
<tr>
<td>Impact Per Attachment Cluster (30 x 5 x 4 x 4)</td>
</tr>
</tbody>
</table>

Table 4. Impact of Classification Inheritance on Email Threads

**Thematic Analysis.** Visual classification can help identify terms that can be used to select other emails of interest not previously identified. To do this visual classification evaluates the frequency of terms in the message attachment groups of interest so users can select themes or search terms to identify other emails.
3. PII Detection & Redaction/FOIA Responses

Without visual classification, the only ways to detect PII or other items needing redaction and to then redact them involve either text string search/replace technology or manual visual review. Visual classification provides an additional way to identify the document types that typically contain PII and to then redact them either using text string searches or by using visual zones based on absolute or relative page coordinates. For example, until now redacting handwritten information involved manual redaction, a slow and expensive process. Visual classification provides a rapid way to identify the classifications of documents that typically contain PII, and to redact them using either text string replacement OR zonal redaction. This is an example of how zonal redaction could redact the social security number portion of a document, even on a handwritten form:

![Zonal Redaction of Handwritten SSANs in IRS Form 1040](image)

Zonal redactions within clusters of visually-similar files is in addition to text search/replace technology that can look for typed text strings:

![Text-String Search Redaction of SSANs in IRS Form 1040](image)

A key point about redaction is that most PII or other content that requires redaction does not occur randomly across all document types. Some of it might appear almost randomly, but most of it occurs where you'd expect it.
Visual classification can also produce detailed logs showing the document, page, and page coordinates redacted, the reason for the redaction, and what was replaced. It can also apply labels or legends to the redacted content.

4. **Taxonomy/Term Selection**

The term “taxonomy” can mean several things. One is a listing of terms appearing within files that help place those files within a classification hierarchy because they are unique to a particular classification or because some combination of those terms is highly correlated with a particular classification. These taxonomies are often expensive to develop and maintain and often do an imperfect job classifying files.

Visual classification can obviate the need for taxonomies for classification because the clustering or grouping of files occurs without resort to the terms occurring in the files. This text independence has many advantages, e.g., See the discussion on text dependence at Section I.B.2., Be Aware of Text Dependency Limitations, and Sidebar 8, Limitations of OCR. Visual classification yields consistent, accurate classifications without the effort required to develop or maintain taxonomies.

To the extent an organization wants to use taxonomies for some other purpose, visual classification can greatly reduce the effort required to develop and maintain taxonomies. It can provide a listing of all the terms that occur within files that fall under any classification and remove words that because of their widespread occurrence within the document population have little value in identifying files within any specific classification.

5. **Fax Servers**

Fax servers can receive high-value documents on an ongoing basis, and managing that content can be challenging for several reasons:

- **Cover Pages.** Many classification systems are based on analyzing the top page of a document. However, with fax systems there may be one or more cover pages that are not helpful in determining what type of documents were transmitted by the cover pages.

- **Time Sensitivity.** Unlike some collections of image-only documents that are primarily just of historical significance, faxed documents are often an integral part of a business process where time is of the essence.
• **Variable Quality and Resolution.** Faxes are sent from a wide variety of image capture systems with many resolutions and a wide range of quality.

• **Non-Symmetrical Resolution.** Text-based classification and routing systems for incoming faxes depend on obtaining the needed text by OCR processing. However, OCR may depend on having symmetrical DPI resolution, e.g., 300 x 300 dpi, or 200 x 200 dpi but faxes can have non-symmetrical resolutions, e.g. 204 x 98.

Visual classification has a rich set of functionalities that enable it to meet the challenges posed by files and documents received by fax servers:

• **Logical document unitization.** Visual classification can determine the differences between one or more cover pages and where the actual documents begin. This helps not only distinguish transmitted documents from cover sheets, but can separate individual pages into multiple documents transmitted at the same time.

• **Speed and Scalability.** Visual classification is text-independent and can scale to classify millions of documents per day, permitting it to rapidly and accurately classify incoming documents.

• **Resolution-Independent.** Visual classification analyzes the relationships among the glyphs on a page and the geometric nature of those relationships remain even if resolutions are different.

• **Text Quality.** Single-instance editing of glyph-text value pairs results in higher quality text than typical character recognition technology.

• **Known Unknowns.** Because of the accuracy of the classifications and the listing of attributes known to be associated with specific document types, BR will know when an attribute value should have been extracted from incoming files, and can alert operators if they are missed. Because it will know from where on the page the value should have been extracted, it can take an operator directly to the page coordinates from where the missing values should be obtained.

### Sidebar 22

**Nine Levels of Glyph Patterns**

Visual classification can focus on different levels of objects on a page, much as a lens can focus on different parts of a picture. These are the nine levels of glyph pattern matching that help accomplish many of the special use cases described in this section:

1. **Individual symbol or character.** Each glyph may be associated with one or more text equivalents. Higher level abstractions or patterns can use either the original graphical representations or their text conversion equivalents.

2. **Word.** A word is a set of glyphs that are contiguous on a line. A “word” may be completely numeric. Words can provide the context within which to determine if ambiguous symbols are either alpha or numeric. For example, in certain fonts a numeric 1 may look like a lower-case alpha l, or a numeric 0 may look like a capital alpha O.

3. **Phrase.** A set of contiguous words. BR can determine which strings contain the most shorter strings.

4. **Line.** A horizontal set of words that goes from edge to edge on a page with no negative space.

5. **Paragraph.** A set of two or more related lines.

6. **Block.** A set of symbols surrounded by negative space, possibly as small as a single symbol.

7. **Table.** A set of blocks arranged in a grid with column heading, row headings, or both.

8. **Page.** What was printed on one side of one physical sheet of paper or it’s electronic equivalent.

9. **Document.** A “file” can contain more than one document, as when multiple files are combined in one PDF. Also a document may span two or more files, as when documents are scanned into TIFs which are each individual pages.
BR’s visual classification can be integrated in virtually any work flow process for ingesting and classifying faxes.

6. **Reclassifying Previously Classified Content**

When content has already been classified visual classification can group those documents based on visual similarity. If the group or cluster of visually-similar documents had been assigned several classifications, it could use the counts of those classifications as a voting engine, assigning all the documents in the visually-similar cluster the document type assigned to the largest number of documents in that cluster.

7. **FAAT — Fully Automated Attribute Tables**

Sometimes the most valuable information in a document is contained in data tables. What many clients want is to extract the table values in a way that permits the client to populate the equivalent spreadsheet or data table with the same row and column headers and data values. FAAT, or Fully-Automated Attributed Tables provides the easiest way to accomplish this. Visual classification recognizes the existence of a table by the hierarchy of glyph patterns it contains (See Sidebar 23, Nine Levels of Glyph Patterns) and can then identify the row headers, column headers, and data value cells, and output row/column header-data pairs, doing both the table identification and the table data extraction without manual intervention.

For example, the data in the table to the right might be output as:

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>12,562</td>
<td>15,894</td>
</tr>
<tr>
<td>Average Age</td>
<td>37</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 5. Example of Table in Unstructured Content

8. **Contract Review w/Duplicate Block Negation**

Organizations with thousands of contracts in place would often like to review the entire collection to determine that all provisions meet current standards and to build a contract administration database or system to track things like contracting parties, contract date, termination dates, venue, prompt payment discounts, etc.

Visual classification can work on a paragraph level using abstractions of the textual values associated with the paragraph to identify where instances of unique paragraphs occur in the collection and to extract attributes or values from them.

The value of this approach is that individual clauses or paragraphs must only be reviewed once and values can be extracted in a way that lets them be used to either load or validate data elements in a contract administration database.
9. Public Agency Comment Analysis

Federal and state agencies contemplating rule changes and other major decisions often must post them and then respond to any comments made on-line or by email or letter. Because special interest groups often suggest specific language for their supporters to use in commenting, there are a high number of duplicate comments received.

Visual classification isolates individual paragraphs and using an abstraction of the text values in the paragraphs, identify where there are duplicates. This permits the agencies to formulate responses to unique paragraphs and use those responses wherever copies or duplicates of the unique paragraphs occurred.

This process can save considerable time and ensure uniformity of responses to the comments.

D. New Search Functionality

The technology that enables visual classification also enables several powerful new ways to search content:

1. New Data Types and Range Searching: Numbers & Dates

Traditional full-text search engines treat all text strings alike, other than they may not permit searching against purely numeric strings. Users who wanted date range searches or searching for numeric values within certain ranges are left to construct difficult searches or to have to rely only on fielded databases that supported such functionality.

BR’s visual classification adds a new range of search functionality for numbers and dates:

- **Numeric-Only Text Strings.** BR’s permits searching for text strings even when they are purely numeric.
- **Automatic Text Equivalents for Numeric Searches.** Searching for “10” also finds “ten.”
- **Date or Numeric Range Searching.** Users can specify desired ranges of dates or numbers even if there is not a specific numeric or date field set up for that collection. For example, the search could specify that search result documents had to contain a date within a specified range.

2. Absolute and Relative Positional Search

Traditional text search applications provide a relatively fixed number of search operators, e.g.,

- **Boolean**
  - And
  - Not
  - Or
  - XOR
- **Relational**
  - Before
• After
• Within x words

• **Wild cards**
  • ? for individual characters
  • * for multiple characters

• **Logical Unit**
  • Sentence
  • Paragraph

Because of the way visual classification builds a catalog of exactly where each glyph or graphical element occurs, it expands on the types of search operators available, to now include:

• **Absolute Page Coordinates.** Users can specify the page coordinates within which a search term must occur to satisfy the search requirements.

• **Relative Page Coordinates.** The page location coordinates can either be fixed or they can be relative to some other search requirements. For example, an absolute page coordinate search could look for the dollar amount over $32,000 in the right one-third of a page and then a relative search could be for the term “Entertainment” to the left of the first term.

### 3. Graphical Search

Users of content search engines or content management systems have long been accustomed to only searching for specific text strings or terms within particular files or documents. However, as part of its cataloging of all the glyphs or graphical elements that occur in files, visual classification permits users to search for even non-textual graphical elements such as logos or stamps.

### E. Summaries of Use Cases by Deliverables and Functions Involved

The tables on the following two pages provide summaries or overviews of some of the principal use cases for visual classification. The first summarizes the types of project-specific data deliverables, and the second summarizes the BR functions used in those use cases.
# Project-Specific Data Deliverables in Different Visual Classification Use Cases

<table>
<thead>
<tr>
<th>Project-Specific Data Deliverables</th>
<th>Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scoping</td>
</tr>
<tr>
<td></td>
<td>Records Mgt.</td>
</tr>
<tr>
<td>Rationalization</td>
<td></td>
</tr>
<tr>
<td>Log of each file encountered with:</td>
<td></td>
</tr>
<tr>
<td>• Location - Path/Filename</td>
<td></td>
</tr>
<tr>
<td>• Hash Value</td>
<td></td>
</tr>
<tr>
<td>• NIST</td>
<td></td>
</tr>
<tr>
<td>• Dupe</td>
<td></td>
</tr>
<tr>
<td>Add to log:</td>
<td></td>
</tr>
<tr>
<td>• Encrypted</td>
<td>X</td>
</tr>
<tr>
<td>• Decrypted (if requested)</td>
<td>X</td>
</tr>
<tr>
<td>• Info on Unzipped Files</td>
<td>X</td>
</tr>
<tr>
<td>Working copy of each file</td>
<td>X</td>
</tr>
<tr>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td>Additional log entries for:</td>
<td></td>
</tr>
<tr>
<td>• Visual cluster ID</td>
<td>X</td>
</tr>
<tr>
<td>• Assigned Classification</td>
<td>X</td>
</tr>
<tr>
<td>Glyph catalog for graphical searches</td>
<td>X</td>
</tr>
<tr>
<td>Attribution</td>
<td></td>
</tr>
<tr>
<td>CSV with attribute names and values</td>
<td>X</td>
</tr>
<tr>
<td>Text layers or corresponding text files</td>
<td>X</td>
</tr>
<tr>
<td>Tiff or PDF versions</td>
<td>X</td>
</tr>
<tr>
<td>Redactions log for each redaction:</td>
<td></td>
</tr>
<tr>
<td>• Hash value</td>
<td>X</td>
</tr>
<tr>
<td>• Page coordinates</td>
<td>X</td>
</tr>
<tr>
<td>• Text of redacted content</td>
<td>X</td>
</tr>
<tr>
<td>• Purpose of redaction</td>
<td>X</td>
</tr>
<tr>
<td>Redacted TIFs or PDFs</td>
<td>X</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
</tr>
<tr>
<td>Authority lists for specific attributes</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 6. Deliverables in Different Visual Classification Use Cases
### Different BR Functions Performed in Different Visual Classification Use Cases

<table>
<thead>
<tr>
<th>BR Functions</th>
<th>Use Cases</th>
<th>Scoping</th>
<th>Granular Classification w/Attribution</th>
<th>Big Bucket Classifications w/o Attrib.</th>
<th>Redaction</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collector</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hash</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dedupe</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deNIST</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copy</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unzip</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hash</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log New</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dedupe</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deNIST</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyph Catalog</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Cluster</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classify</td>
<td>X X X X X X X X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#Classes</td>
<td>99+ 99+ 99+ 99+ 99+ &lt;10 &lt;10 &lt;10 *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribution</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redaction</td>
<td>X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edit Attr. Values</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edit All Text</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Export</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalize Email</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email MAG ID</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email Thread ID</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAG Relev. Text</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thematic Analysis</td>
<td>X X X X X X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Redaction can be accomplished without formally designating labels for document types

Table 7. Different BR Function in Different Visual Classification Use Cases
V. Selecting File or Document Classification Systems

The real challenge when selecting what file classification system to use is to separate marketing hype from what any particular system can actually deliver. For that reason, a “proof-of-concept” trial can be invaluable.

Here are thoughts on using proof-of-concept trials or pilots:

- **Scalability.** If you have petabytes of data to classify, it’s reasonable to ask that at least a terabyte of data be processed in the trial. Providers who can’t quickly turn around a petabyte are not believable when promising petabytes.

- **Capacity/Turnaround Time.** A petabyte is not an enormous amount of data, and providers should be able to provide a turnaround time of a day or two for the initial classifications.

- **SLAs.** The POC and any final agreement should involve written service level agreements that specify how accuracy will be determined for both the initial classification and extracted attributes and/or content enablement.

- **Compare Vs. Existing System.** If you have used an existing classification system, have the prospective providers auto-classify what you’ve already classified and compare the results. This can provide insight into the consistency of the existing classifications and the capabil-
Vendor Fly-Offs. If you’ve narrowed the selection to two or three vendors, have each classify the same TB of information and compare the results. A gap analysis of where the differences lie will tell you which system is being the most consistent and dependable.

Land and Expand. Sometime software license fees are just the tip of the iceberg because the customer must hire an ever-expanding army of consultants to perform the classifications, attribute extraction, and QC needed for initial implementation and ongoing maintenance. Require that the provider detail all resources needed to accomplish the POC.

Checklist. The earlier portions of this book describe various things to look for in an automated file classification system, and many are summarized on the following checklist.

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>Product or Offering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BR</td>
</tr>
<tr>
<td>1</td>
<td>Rationalize - How many &amp; where</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subject to 260-character path limit</td>
<td>No (2600)</td>
</tr>
<tr>
<td></td>
<td>Hash &amp; log all files</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Onboard deNISTing</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Onboard deduping</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Compress copied files</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Encrypt copied files</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Time to collect 1 PB</td>
<td>7-10 days</td>
</tr>
<tr>
<td></td>
<td>Local connect option</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Network connect option</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Resume after interruption w/o start over</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Requires installation of collection S/W</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Collect from Unix OS</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Collect from Apple OS</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Collect from Microsoft OS</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Classify - What’s in the files</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Require text in files</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Require full sentences for classification</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Require limited languages</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 8. Checklist for File Classification Technology
<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>Product or Offering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BR</td>
</tr>
<tr>
<td></td>
<td>Require creation &amp; tuning of taxonomy</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Require selecting exemplar</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Require writing scripts for classification</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Require consultants to group files</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Scalable to 100+M file collections</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Cloud option</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Behind firewall option</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Time (days) from collect to start classify</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Proof-of-concept trial size (TB)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Granularity - No. Classes</td>
<td>2-100+</td>
</tr>
</tbody>
</table>

**Attrition - What to extract**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Page zone specifiers</td>
<td>Y</td>
</tr>
<tr>
<td>Single-instance glyph/text edit</td>
<td>Y</td>
</tr>
<tr>
<td>Non-textual (e.g., logo/signature) extraction</td>
<td>Y</td>
</tr>
<tr>
<td>Font-size specifier</td>
<td>Y</td>
</tr>
<tr>
<td>Convert all text for content enablement</td>
<td>Y</td>
</tr>
<tr>
<td>CSV output option</td>
<td>Y</td>
</tr>
<tr>
<td>PDF w/text option</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Validation**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Full tracking of all files</td>
<td>Y</td>
</tr>
<tr>
<td>Analytics for all metrics</td>
<td>Y</td>
</tr>
<tr>
<td>Locate page location for all attributes</td>
<td>Y</td>
</tr>
<tr>
<td>Glyph-text value confirmation</td>
<td>Y</td>
</tr>
<tr>
<td>Minimum accuracy rate stated in SLA</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Additional**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Zoned redaction</td>
<td>Y</td>
</tr>
<tr>
<td>Text-string redaction</td>
<td>Y</td>
</tr>
<tr>
<td>Redaction log w/redacted content/purpose</td>
<td>Y</td>
</tr>
<tr>
<td>Graphic search</td>
<td>Y</td>
</tr>
</tbody>
</table>
VI. Getting Started: One Bite at a Time

The best way to get started using visual classification is with a discrete set of files that will provide experience using the new technology. The usual implementation progression is as follows:

- **Free Trial.** For qualified organizations, BR processes a TB of content and reviews the results with the organization. The review includes examining the larger groups of visually-similar files and demonstrating how classification and attribution can be applied.

- **Paid Pilot.** A discrete file collection is processed and BR and the client develop a classification tree or matrix along with lists of attributes, and classification and attribution is applied to the collection.

- **Production.** BR and the client enter an ongoing licensing or service agreement.

Regardless of which project is selected as the first, much of the time and effort invested in classifying and attributing documents on the first implementations can be re-used in subsequent operations. For example, if document-type classes are established for a paper-archive conversion project, those same document types can be used for file share remediation or ECM migration. The file types that both collections had in common will both be setup when the second project starts.

Over time, more and more of the document types in an organization will be classified and attributed, making it easier and easier to onboard additional projects and processes.
VII. Further Information

For further information about visual classification and how it can help you manage your unstructured content, contact John Martin via email at John.Martin@BeyondRecognition.net. More information is also available on-line at http://www.BeyondRecognition.net. The site includes free downloadable resources and a contact form.

We would also welcome any comments or suggestions on ways to improve this Guide to Managing Unstructured Content.