On Smarter Document Retrieval

By Manuel Bulwa, Isausa, inc.

May 2020

This paper describes proven methodologies to **reduce costs and risks** involved in capturing and finding documents in a large digital collection, while **improving the overall search** experience. These methodologies rely on a combination of search techniques, barcode recognition, minimal indexing and visual browsing.

A digital collection usually consists of **scanned** (paper, microfilm, microfiche, etc.) and/or **ingested** documents (emails, digital files, websites, messages, faxes, etc.).

According to the American Society for Indexing a good indexing criterion must be accurate, clear, concise, consistent, sensical, comprehensive, accessible, reflexive, unbiased, natural and readable. I would add that the indexing structure produced must also be *governable* i.e., it should consist of a **minimally small** handful of index fields that require minimum maintenance, comply with records management policies, and should **not** include fields that:

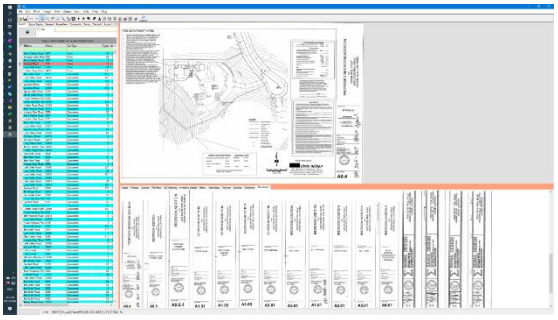
a) can be derived from true index fields captured for the same or other collection via lookups or screen scraping (RPA).
b) are needed only to narrow down numerous pages of a retrieved document.
c) may be better handled using *Brute Force Visual Browsing.*
d) can reliably assist finding documents using *Full Text Search*.
e) belong to the *Data Mining* realm i.e., beyond just search and find.
f) are volatile.

Virtually all modern digital search methods use data items captured either manually or through automated processes such as OCR, barcode recognition, digital extraction, etc. These fields are subject to database queries to locate and display one or more single or multipage **image** documents (unless none is found). This is usually represented as a tree-like hierarchical search process where the top-level node is a document class or type, and a small number of sub-nodes represent deeper levels of search details. Navigating this tree narrows down the search to a small hit list, which must be navigated through visualizing images and/or thumbnails to finally locate the actual sought target(s). The navigation of the tree can be *Brute Force* i.e., patiently looking at each node and branch one by one, *Random Access* i.e., using computer algorithms to instantly locate indexed records, or a compromise between the two known as *Binary Search* based on cutting in approximate halves the first phases of a search of an ordered list. This is how humans quickly find a word out of tens of thousands of words in a dictionary book.

“Good” index data is expensive to capture and maintain, but so is failing to timely find records where needed. By using the methodologies here described, records can be cost-effectively located with less dependency on digital data fields and subject matter expert indexers. Examples (graphics may have been blurred to protect sensitive information):

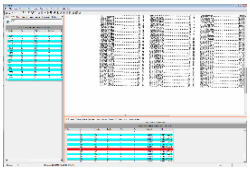
1. Engineering or Architectural drawings: These usually come in *Sets* of one or more related *Sheets*. A property set typically includes one or more sheets depicting construction tasks for various sections of a property. Once a hit list is quickly produced using random access search, the few sheets pertinent to the search can be brute force

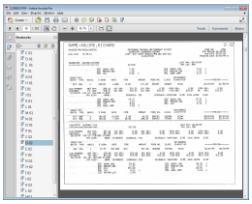
located by visually navigating through a small number of images of the title blocks. This makes possible the finding of records:



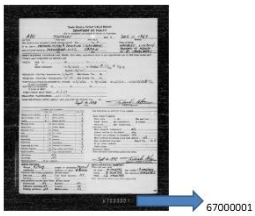
a. without the cost of keying in certain title block data fields such as sheet number, revision number, section, task, etc.
b. without the risk of keying or recognition errors and the cost of subject matter expertise to decide which ones to capture.
c. adding the convenience of improvising selection criteria on-the-fly by seeing stamps, signatures, handwritten annotations, logos, etc.

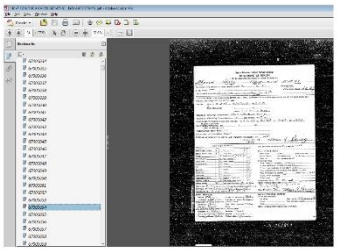
2. <u>COM Microfiche</u>: COM microfiche is a computer-generated report where each frame contains one or more report pages or fractions of a page. A sheet may contain hundreds of frames and has a title block indicating the range of indexed values contained in the sheet. Each frame is addressed by row, column coordinates. The bottom right last frame usually contains a list of coordinates for key data values in that sheet. We create a hyperlinked list of sheets using title block data. This list can be visually browsed, binary searched, random access searched and/or full text searched to retrieve multipage PDF files (one per sheet) with bookmarks named after each coordinate. Looking up the index frame we locate the coordinates of the first frame containing the index sought. We then visually binary-search the bookmarks and click on the first relevant one (random access). This brings up the first frame containing the starting report page(s). Finally, we brute force visually search consecutive frames looking for our target record(s). All this can be done in seconds or minutes even for large collections. This can be further improved by creating “hot spots” in the index frame images, so clicking on it will immediately display the corresponding frame.



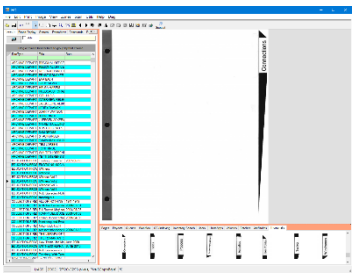


3. <u>Roll Microfilm</u>: A roll of microfilm may contain thousands of frames. A computer file known as “CAR” (Computer Assisted Retrieval) exists or can be created that acts as a index list similar to the COM fiche index frame described above, except that the coordinates here are the roll number and the location of a frame (accession number) on that roll. A search on the catalog will display the first frame pointed at by the catalog along with bookmarks named after accession number. Finally, a frame-by-frame brute force visual search of every frame until the next bookmark finishes the job. We can produce one PDF file per each roll, one PDF file for all rolls, or run software that does not use PDF at all.





4. Paper Documents: Paper documents frequently include sheets delimited by tabbed binder dividers, stapled or clipped pages, binder cover titles, etc. Each of these segments needs to be identified so they can become part of search criteria, so we use barcode separator sheets to signal them. Once a hitlist of multipage documents has been located using random access search index fields, any further searching and selection can be conducted using brute search visual browsing of each divider thumbnail image. This makes possible the finding of records without the cost of keying in data fields such as tab content, staple heading, clip heading, etc.



5. Video Inventory: On ocassions, we need to produce a manifest or inventory of boxes or cabinets and their folders, regardless of whether they would ever be scanned or not. Using smart glasses to operate hands free, we capture a video of every folder in a box in a few minutes. Our software subsequently turns the video into one chosen still frame per folder tab or cover. We then use visual binary search to determine in which box or cabinet a sought document is located, or that the document does not exist.

Of course, modern technology offers additional advantages using machine learning, AI, neural networks, natural language processing, computer vision, RPA, etc. can add great value through automated classification and data extraction.

In summary, a smart combination of random search, brute force search, binary search, full text search and visual browsing can significantly reduce costs and risks before and after capture, while enhancing the overall search experience.