For Earth Day, Recycle Those Stacks of Paper and Instantly Find the Forest Through the Trees

Article contributed by dtSearch®

Yes, there is a connection between these two seemingly unrelated items. The missing steps are scanning, OCR, PDFs and enterprise search. This article will fill in the blanks, leaving you all set for Earth Day.

Before you toss those stacks of paper into the recycling bin, you'll want to scan them. Scanning takes a picture of the pages. While your picture may be worth 1,000 words, you can't do much with it other than look at it. To take an image of the word forest on the page and turn it into something you can do more with than just gaze at requires OCR, or optical character recognition.

An application like Adobe Acrobat will OCR the text from the scanned image, turning the text into something that you can copy, paste and otherwise work with. Now suppose you are scanning a copy of a memo. OCR will take the main typed text and digitally store that. But what if there are some notes that someone may have scribbled in the margins of one of the pages?

The gold standard—or should I say the green standard—for combining text and images is "searchable image" PDF. The format superimposes the OCR'ed text onto the original image of the page. Plus you can add on metadata. Let's say the memo you are digitizing came from a *Project Waterways* binder. Even if *Project Waterways* appears nowhere on the memo itself, you can add a metadata element containing that phrase and that too will be part of the "searchable image" PDF.

The last step is to install an enterprise search engine. An enterprise search engine isn't a span-the-internet search engine like Google. Rather, it is an application like dtSearch® that goes deep into your organization's own data to retrieve anything anywhere—in the full-text and metadata—that matches your query. For "searchable image" PDFs, a search will show an OCR "hit" overlaying the original image including items like margin notes.

Enterprise search can instantly span terabytes only after first indexing the data. The index isn't like a reference book index; rather it is just an internal guide holding each word and number in the data and its location for the sole purpose of enabling instant search. To get the search engine to

Reprinted with permission of Successful Business News, successfulbusinessnews.com

build its index, all you need to do is point to the folders and the like to index, and the search engine will do everything else.

The search engine will automatically recognize and index PDFs, both "searchable image" and otherwise. And it can also automatically recognize and support other formats like Microsoft Word, Excel, PowerPoint, OneNote and Access files; web-based formats; compressed formats like ZIP or RAR; and even emails plus nested attachments. For example, if you have an email with a compressed attachment that includes a PDF and an Excel spreadsheet with a Word document embedded inside, enterprise search can automatically index and search the whole thing.

It's not just instant individual queries that enterprise search enables but also concurrent network or web-based queries as well. Online, search can proceed in a stateless matter, making multithreaded text retrieval fully scalable without affecting search speed. Search features encompass over 25 different full-text and metadata word, phrase and number-oriented options, so everyone can find the *forest*, the *trees* as well as the *woodland creatures*.

A catalog of all of the search features, relevancy-ranking and sorting options is beyond the scope of this article. But I'll end with 3 search tips relating to "searchable image" PDFs specifically.

Search tip #1: before you make these old stacks of paper instantly searchable by the entire office, do a quick indexed search for credit card numbers to make sure that these do not appear in the newly digitized collection. The search engine can flag valid credit cards that appear in the OCR'ed PDFs – or anywhere else across indexed data.

Search tip #2: turn on fuzzy searching to a low level when you search through OCR'ed PDFs to sift through any minor OCR errors. For example, if the word *activate* is mis-OCR'ed as *actiwate*, a fuzziness level of 1 would pick that up in a search for *activate*. Fuzzy searching is also very helpful for searching emails, where mistypings can be common.

Search tip #3: Glancing at a collection of files in the folder system, it is impossible to distinguish "searchable image" PDFs from "image only" PDFs. The latter are not full-text searchable; just the filename and metadata are searchable. The search engine can flag "image only" PDFs when it builds its index. If you find "image only" PDFs, just run them through Adobe Acrobat to turn them into "searchable image" PDFs.

Article contributed by dtSearch®

Reprinted with permission of Successful Business News, successfulbusinessnews.com