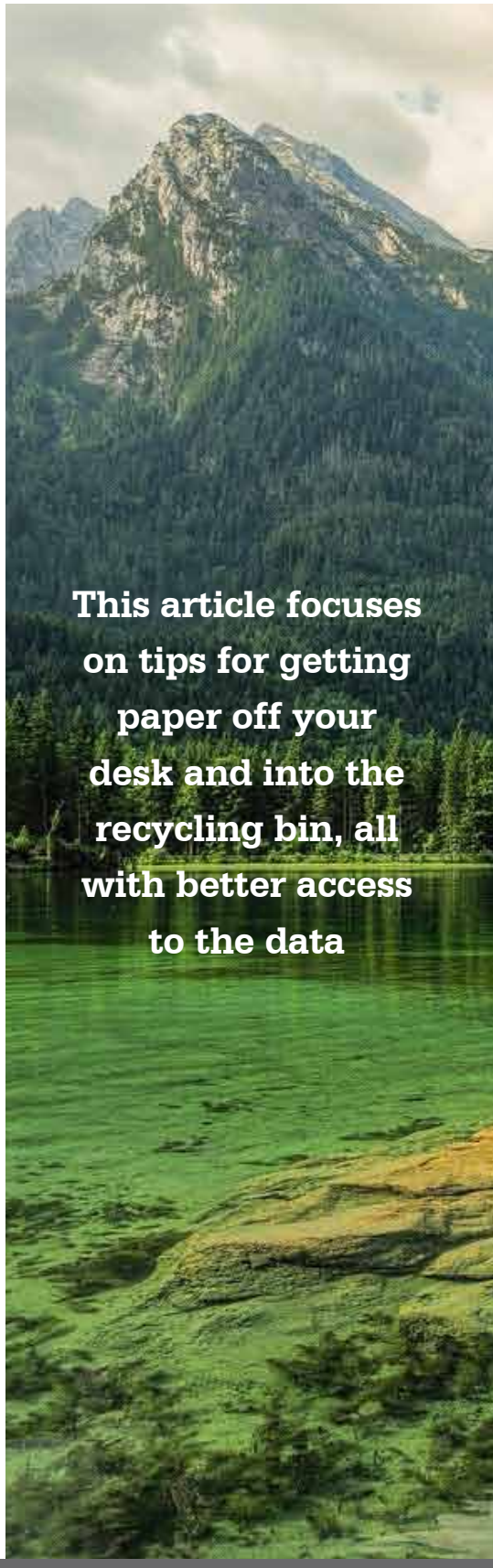
Going Green: Tips for Turning Paper into Instantly Searchable Electronic Text

Article contributed by dtSearch®



This article focuses on tips for getting paper off your desk and into the recycling bin, all with better access to the data. For search access to this data, this article discusses dtSearch.

As general background, dtSearch enterprise and developer products instantly search terabytes of "Office" files, PDFs, popular compression formats, emails with multilevel nested attachments, databases and Internet or Intranet data. dtSearch software can run "on premises" or in a cloud environment like Azure or AWS. Because dtSearch can instantly search terabytes, many dtSearch customers are large enterprises like Fortune 100 companies and federal, state and international government agencies.

However, in addition to offering enterprise-level search, dtSearch also lets **you** instantly search **your own** files and emails. With that in mind, below are tips for getting paper off your desk and into the recycling bin while ensuring optimal access to the data.

Let's start when you are feeding paper into the scanner. You can scan the page as a straight-up image file, like TIF or image-based PDF. While that will retain the full picture of the original page, you can have a tough time leveraging the contents. You can't cut and paste text from an "image only" file. And you can't electronically search the contents.

A better way is to scan into a format that combines a picture of the page with optical character recognition or OCR. Adobe, which provides OCR in Adobe Acrobat, calls this combination "searchable image" PDF. With that format, you can still see the full image of the original page including graphics. But the format also saves an OCR'ed copy of the text along with the page image. You can cut and paste the OCR'ed text. And you can electronically search the OCR'ed PDFs alongside your Microsoft Office documents, emails and other files, jumping right to highlighted hits inside the file views.

The next tip relates to how you use a search engine like dtSearch. The tip here is to make sure that you have your search engine build an index prior to searching, as opposed to performing "blunt force" unindexed searching. As it happens, dtSearch can perform both indexed and unindexed searching. But while unindexed searching is slow, indexed searching is typically instantaneous even across terabytes of data.

Reprinted with permission of The Daily Blaze, thedailyblaze.com | Original article link (Feb 2020)

All you have to do is point dtSearch at the relevant folders and other data you want dtSearch to index, and dtSearch will go off on its own and build the index. You don't even need to tell dtSearch what types of files you have. dtSearch will figure that out all by itself.

After indexing, you can instantly search all of the indexed data and see the full contents with highlighted hits. In an enterprise setting, multiple users can all search the same index concurrently and see retrieved data with highlighted hits. As a technical matter, each concurrent search can run on a separate thread so that it can still proceed instantaneously without affecting other search threads.

Another advantage to indexing is that the indexing process can flag "image only" PDFs that may be hidden among "searchable image" PDFs. There is no filename label that distinguishes the two formats. And sometimes, despite best efforts, "image only" PDFs can creep into an otherwise full-text searchable document collection. dtSearch will identify "image only" PDFs during indexing, so that you can go back and OCR them subsequently using Adobe Acrobat, for example. After OCR, dtSearch can update its index or indexes to add the new files without disrupting individual or concurrent multithreaded searching.

The final tip relates to search options. dtSearch products have over 25 different search features, and one of these is especially important for OCR'ed text. Particularly if the original page is not in mint condition, say because of a coffee stain, a word like caffeine can be mis-OCRed into a word like caffeime. dtSearch has an adjustable fuzzy search setting which you can add on to pretty many any of dtSearch's 25 plus other search options. With a low level of fuzziness, if you search for caffeine you'll also pick up the smeared word caffeime.

And fuzzy searching also works well for other files that may have typographical errors like emails. Almost everyone mistypes on occasion in emails. A low level of search fuzziness will typically sift through these mistypings, just like it will sift through OCR errors.

At dtSearch.com, you can download a fully-functional 30-day evaluation version to search through your own files and emails. And for more advanced tips on everything from credit card search to developer options like metadata-driven faceted search, go to dtSearch.com, and select Features Map.

Article contributed by dtSearch®