Ghosts of Data Past

Article contributed
by dtSearch®

Every year, an organization accumulates more data. A search engine can instantly bring past data to light, but there are some *ghostly cautions* to keep in mind.

A search engine like dtSearch® can instantly search terabytes after first building a search index. A search index is not like a back-of-the-book index. Rather, it is just an internal guide for the search engine that stores each unique word and number across the data, as well as the location of each word and number in the data.

Building an index is easy. All you have to do is point to different folders and the like you want to cover in the index, and leave the rest to the search engine. dtSearch will automatically figure out the data types it is working with, whether web-ready formats; emails; compressed content like ZIP or RAR; Microsoft Word, Access, Excel, OneNote or PowerPoint files; PDFs; etc. After indexing, the search engine can instantly search terabytes using over 25 different search options. dtSearch can then display each retrieved item with highlighted hits.

Along with an option for individual search, dtSearch can also run in a shared concurrent-search environment, with each search thread processing independently. That way, multiple end-users at once can instantly search across terabytes on a network, through a local web server, or via the cloud as on Azure or AWS.

Ghostly Caution #1. The first caution involves ghostly file format extensions. Accidentally or sometimes even intentionally someone will, for example, save a PDF with a Microsoft Word extension or a OneNote file with a PowerPoint extension.

dtSearch can "see through" these ghostly file extensions because the dtSearch document filters, or the component that parses each file, works directly with the binary version of each document. The document filters use the binary contents to determine file type and which parsing standard to apply. The file extension is irrelevant to this equation.

Ghostly Caution #2. The next big caution is literally the ghost of documents past, or more specifically, tracked changes. Even after editing is long over, tracked changes can remain in a document. If you don't want a search engine to find these, from inside a document, delete all tracked changes. Or just print a final copy of a document to PDF and archive it that way.

Reprinted with permission of USA Daily Chronicles, usadailychronicles.com

Ghostly Caution #3. You can also effectively have ghostly PDFs where you see words on a page, but the PDF is only an image. Have you ever had a PDF where you tried to copy and paste text out of it, only to end up with nothing? If so, that was probably an image-only PDF. In that case, you need to apply a separate OCR process, such as running the document through Adobe Acrobat OCR, to turn the word images into computer-recognized text that a search engine can index and search.

The final PDF can retain a full image of the original PDF. But beneath the image will also be the computer-recognized text that a search engine like dtSearch can work with. After OCR, dtSearch can highlight the hits through Adobe Reader and have those highlights superimposed on top of the original image. That way, even if there is a doodle or a barely legible note on the page, the doodle or note would remain visible in the file along with the now searchable text.

Although tricky to spot in a collection of mixed documents, dtSearch can flag image-only PDFs during the indexing process. After dtSearch flags these, you can run the image-only PDFs through a product like Adobe Acrobat to make them full-text searchable.

Ghostly Cautions #4 and #5. Modern document formats can include obscure metadata that does not readily appear when you view a document. It is also possible to embed a document completely inside of another document. You could have, for example, a Microsoft Word file with a Microsoft Excel spreadsheet inside where portions of the spreadsheet may not be readily visible from within Microsoft Word. A search engine like dtSearch can seamlessly retrieve with highlighted hits all such below-the-surface content.

About dtSearch. dtSearch has enterprise and developer products that run "on premises" or on cloud platforms to instantly search terabytes of "Office" files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch can instantly search terabytes with over 25 precision search options, many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com to instantly find whatever lurks in the data.

Article contributed by dtSearch®