

Your Data Is Like Ukrainian Nesting Dolls

You know those Ukrainian nesting dolls? Inside the first doll is a smaller doll, and inside that a tinier doll, and inside that an even smaller doll, and so on. Nesting dolls are a great analogy for enterprise data. A search engine like dtSearch needs to drill down through each doll level.

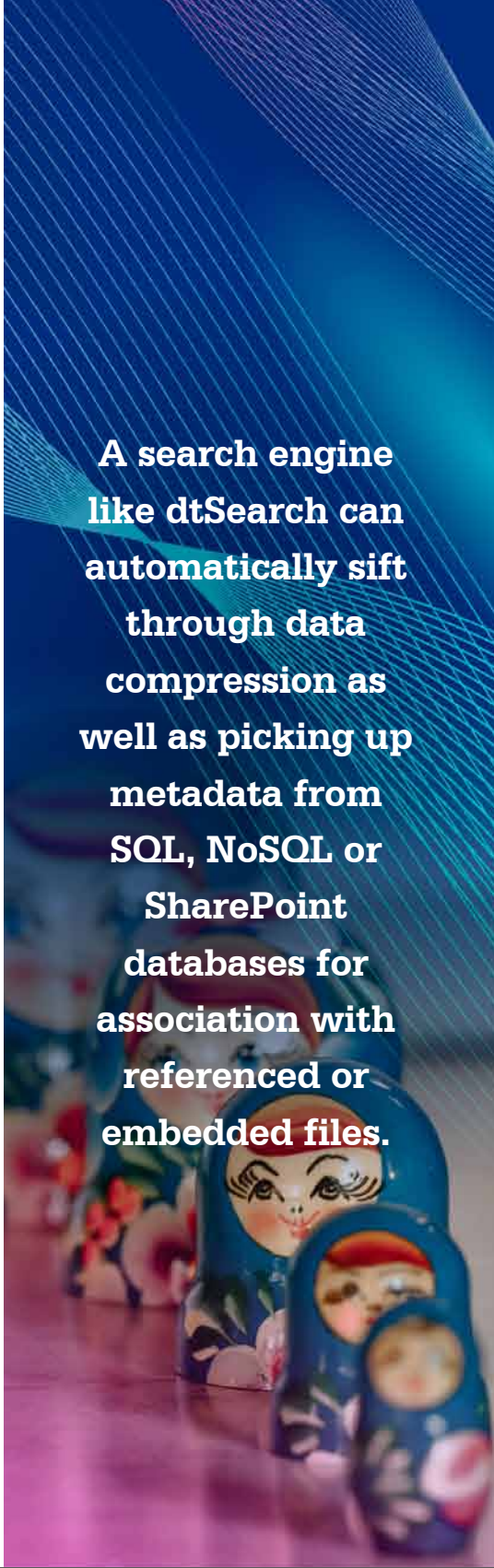
Nesting doll #1. Data archives would represent the outermost doll level. Email archives plus ZIP or RAR files hold individual files in a compressed format. Metadata archives like SQL, NoSQL or SharePoint store individual files as referenced or embedded BLOB data. A search engine like dtSearch can automatically sift through data compression as well as picking up metadata from SQL, NoSQL or SharePoint databases for association with referenced or embedded files.

Nesting doll #2. Individual files and email would represent the next nesting level down. Making the full-text content of these available for instant search is at the core of what a full-text search engine does. A search engine instantly searches terabytes after first indexing the data.

Indexing is easy. Just point to the folders and the like you want to index, and the search engine does everything else. Indexing enables over 25 different types of instant full-text and metadata search options across terabytes. Instant multiuser concurrent search can run over a network, from an "on premises" web server, or from a cloud server such as on Azure or AWS.

The first step to indexing the data is to recognize the file format of each item. PDF, Microsoft Word, Excel, Access, PowerPoint, OneNote, Outlook, Exchange, HTML, XML, etc. each have their own particular file format that the search engine needs to take into account when parsing these files. dtSearch looks *inside* each file in its binary format to figure out the correct

Article contributed
by [dtSearch®](#)



A search engine like dtSearch can automatically sift through data compression as well as picking up metadata from SQL, NoSQL or SharePoint databases for association with referenced or embedded files.

file format to apply. That way, the search engine can correctly parse files even if the file comes with a mismatched file extension, like a PDF that someone saved with a Microsoft Word extension.

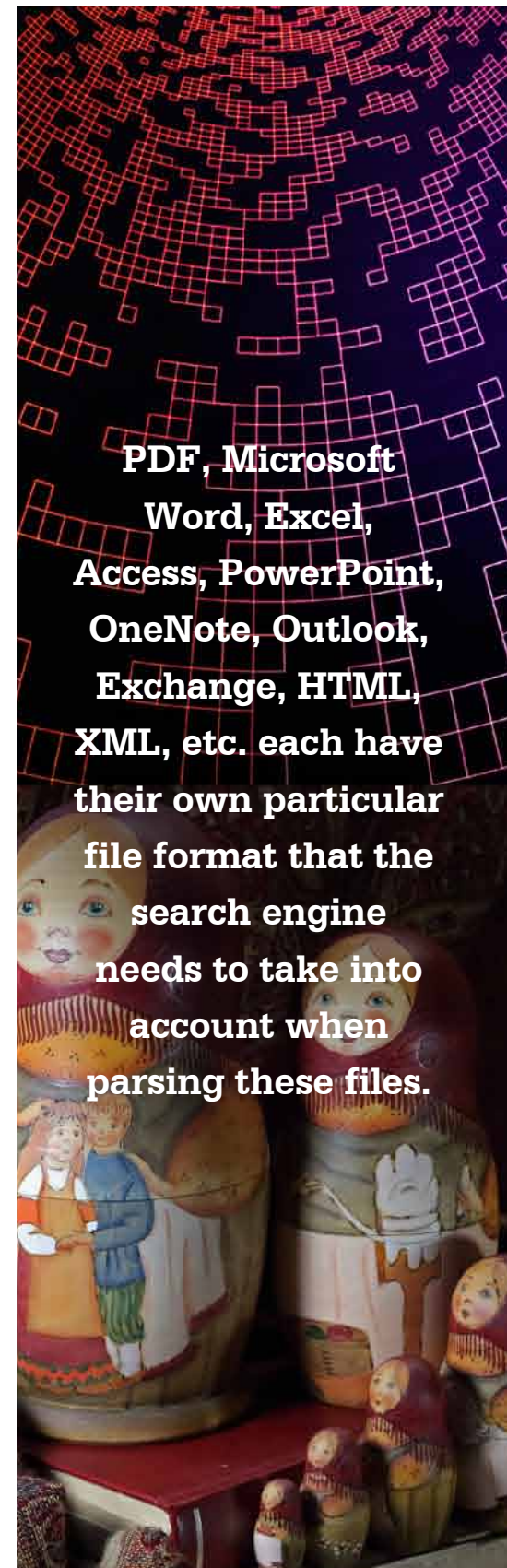
Nesting doll #3. Individual files can also embed hidden data. Text that blends in with a background color like black on black writing or white on white writing or red on red writing, for example, is by design extremely difficult to spot when viewing a file in its native application such as looking at a Word document in Microsoft Word. The binary format of the file, however, makes such text as readily apparent as any other text to a search engine. Viewing a file in its native application can also obscure certain metadata to the point that it can take a lot of clicking around to even know that it is there. But a search engine can easily locate all such metadata in a file's binary format.

Nesting doll #4. Beyond the individual file level, you can also have file attachments, like an email with a ZIP or RAR attachment containing individual files. And then you can have a file embedded inside another file, like an Access database embedding a Word document. The search engine has to go through all of these nested files.

Nesting doll #5. A particularly sneaky file type, “image only” PDFs might make up their own nesting doll level. Most PDFs that you create from an underlying file like printing to PDF from a Word document store the full text of the file plus all images. Likewise, the norm for creating a PDF from OCR'ed content is to use an OCR application like Adobe Acrobat to create a “searchable image” PDF. A “searchable image” PDF preserves a full picture of the original along with the OCR'ed text. dtSearch can then display this full picture through an embedded Adobe Reader window, with highlighted hits from the OCR'ed text superimposed on the image.

But sometimes, when you have a lot of files, an “image only” PDF will slip through the cracks. Because it is

Article contributed
by [dtSearch®](#)



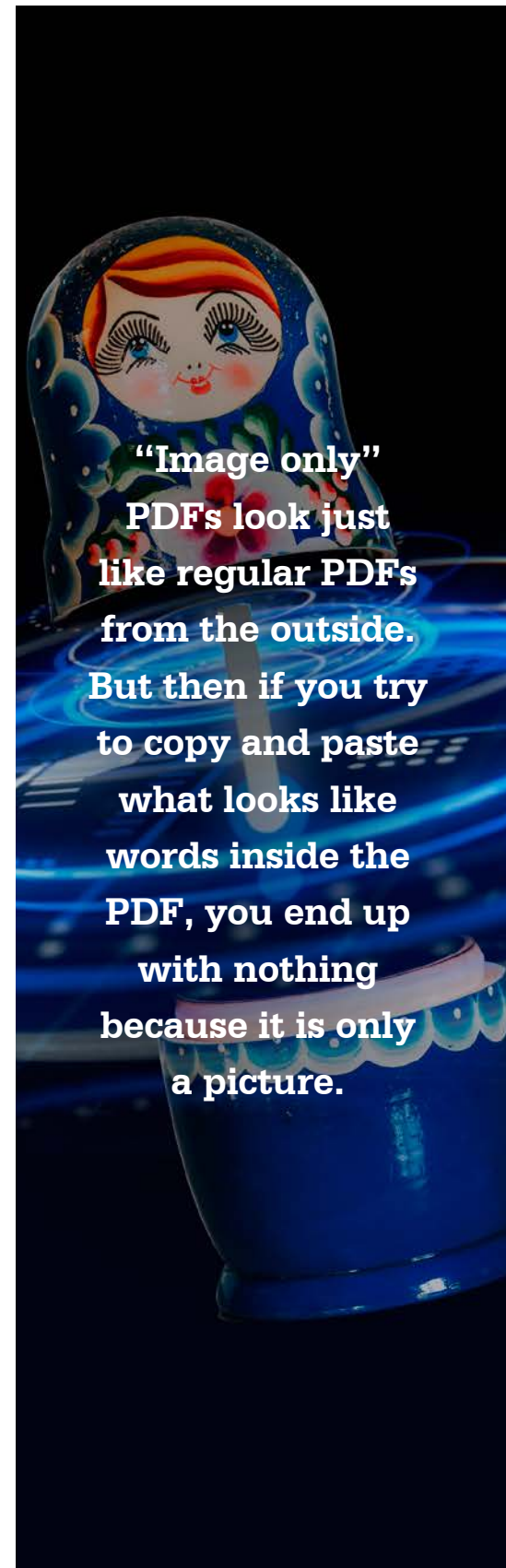
PDF, Microsoft Word, Excel, Access, PowerPoint, OneNote, Outlook, Exchange, HTML, XML, etc. each have their own particular file format that the search engine needs to take into account when parsing these files.

“image only,” there is no text in the file itself for a search engine like dtSearch to work with. “Image only” PDFs look just like regular PDFs from the outside. But then if you try to copy and paste what looks like words inside the PDF, you end up with nothing because it is only a picture.

Apart from going PDF-by-PDF trying to copy text, there is no general way to single out these files. However, dtSearch can flag these “image only” PDFs upon indexing. Once you know which PDFs are “image only,” you can run them through Adobe Acrobat OCR, for example, to turn these into “searchable image” PDFs.

Nesting doll #6. The last nesting doll level might consist of any unexpected content beyond standard text searches that a search engine might find. For example, in addition to searching for words or phrases in various natural language, Boolean and proximity-type search query configurations, a search engine can also look for specific numbers or numeric ranges. A search engine can further look for date and date ranges—even encompassing different date formats like February 2, 2019 to 10/11/22. dtSearch can even identify any valid credit card numbers that may appear in the indexed data. That way, you know that in one particular file, there is a credit card number that you may need to purge.

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 different search features, 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](#)



**“Image only”
PDFs look just
like regular PDFs
from the outside.
But then if you try
to copy and paste
what looks like
words inside the
PDF, you end up
with nothing
because it is only
a picture.**