

May Is Zombie Awareness Month; Now Search That Enterprise Zombie Data

In honor of the fact that May is official Zombie Awareness Month, today's topic is what I'll call zombie data. I ordinarily describe how enterprise search can retrieve information you probably have some idea is out there. For Zombie Awareness Month, I'll address items buried in the deep underground of data that enterprise search can unearth. The principles of enterprise search are the same across all data so I'll start with the basics before getting into some zombiesque examples.

So how does enterprise search work?

Enterprise search can instantly search terabytes after initially indexing the data. To those unfamiliar with enterprise search, indexing can sound hard. And it is a lot of work, but only for enterprise search, not for humans, living or otherwise. All we humans need to do is point to the relevant folders, email archives and the like to index, and enterprise search will take it from there. No need to even tell the indexer if it is encountering PDFs, Word documents, Excel spreadsheets, Access databases, PowerPoints, OneNote files, emails, etc. The dtSearch® indexer, for example, can on its own figure out the file format of each item looking at the binary presentation of each file.

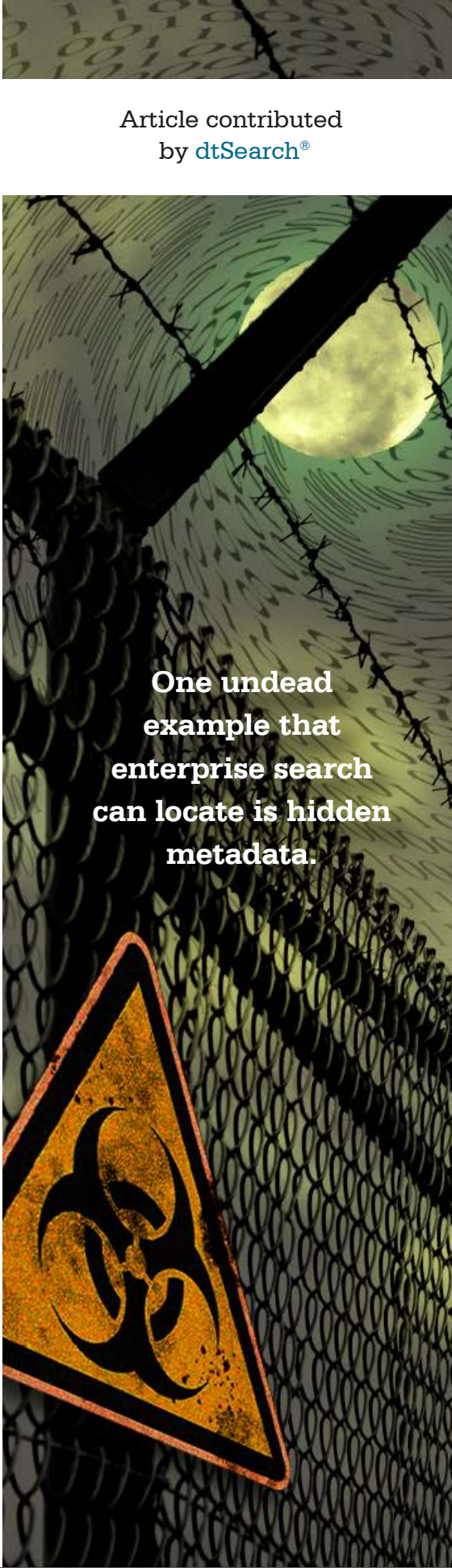
Does the data have to be local for indexing to work?

Files inside folders to index can be on the same machine as the indexer, on a network drive, or on a remote cloud server. As long as the indexer can see the contents of such folders as part of the Windows file system, dtSearch can index and search the contents, including SharePoint attachments, Office 365 files and the like.

How much data can a single index hold?

A single index can hold up to a terabyte, and there are no limits on the number of indexes that the software can build and instantly concurrently search. While indexing is resource intensive,

Article contributed
by [dtSearch®](#)



One undead example that enterprise search can locate is hidden metadata.

searching is resource-light with search threads operating independently. There are no limits on the number of instant search threads that dtSearch can simultaneously process. As data evolves, dtSearch can automatically update its indexes to reflect new, modified and deleted items. Updating an index does not affect continuing concurrent search. After indexing, the product has more than 25 different concurrent search options for precision retrieval. Unicode support extends searching to hundreds of international languages, including right-to-left languages like Arabic and Hebrew as well as double-byte text like Chinese, Japanese and Korean. dtSearch enables multiple relevancy ranking and other sorting options. End-users can browse retrieved files with highlighted hits.

What about zombie data?

Text retrieval can extend beyond information that you have some idea is out there to deeply buried content that you may never have known was present before enterprise search resurrects it zombielike from the dead. One undead example that enterprise search can locate is hidden metadata. Files and emails have tons of metadata, only a small portion of which a standard file or email view may readily expose. But all metadata is apparent in the binary format of a file that enterprise search looks to, making all metadata easily available to enterprise search.

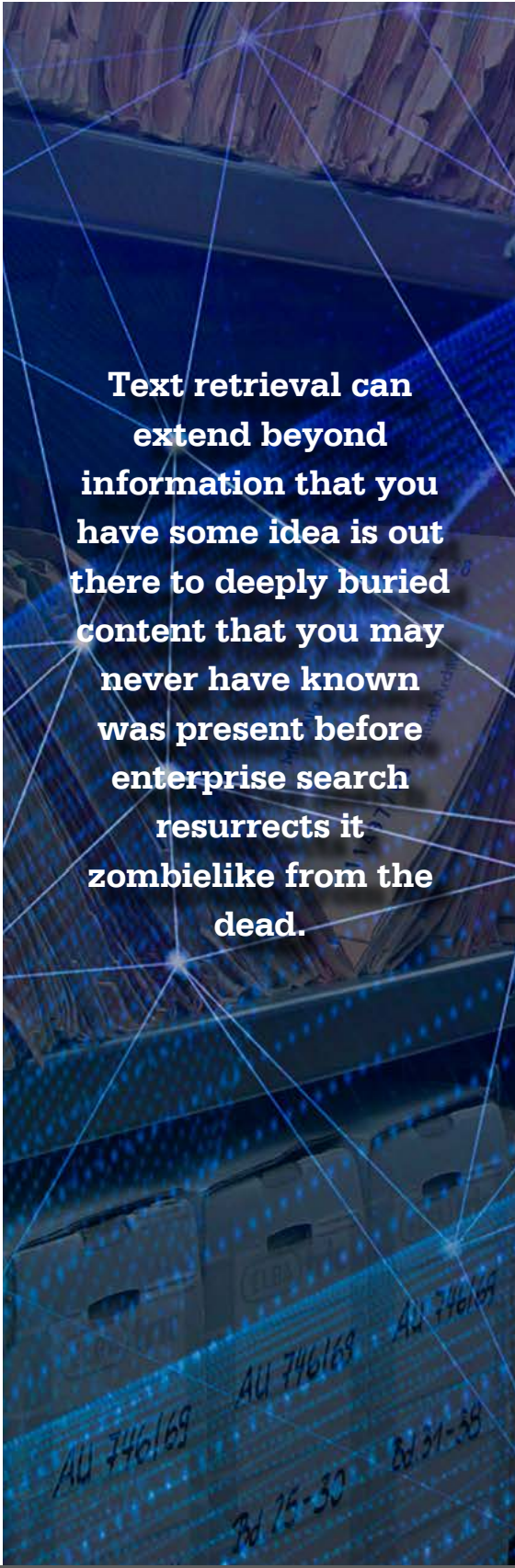
Do you have other examples of zombie data?

Another example of zombie data is text that blends in with its background color in a native application file view such as zombie black text against a zombie black background. While you might completely miss such text in a standard application file view, in the binary format of a file that enterprise search works with, all text is on the same footing, regardless of contrast with background color in a normal file display.

Are there more zombie data examples?

Recursively nested files can also lead to zombiesque data. If you are looking at a Word document in Microsoft Word, you may very well miss portions of an Excel spreadsheet that the Word document fully embeds. But because enterprise search relies on the binary format of files, indexed search can seamlessly cover all nested layers. In fact, dtSearch can automatically parse through multiple file levels such as an email with a ZIP or RAR attachment holding a Word document with an Excel spreadsheet inside of that.

Article contributed
by dtSearch®



**Text retrieval can
extend beyond
information that you
have some idea is out
there to deeply buried
content that you may
never have known
was present before
enterprise search
resurrects it
zombielike from the
dead.**

What about files that have mismatched file extension?

Files with mismatched extensions can stymie ordinary file review. But dtSearch uses the binary format to determine the correct format of each item. You can have a PowerPoint with a .PDF file extension and an Access database with a .ONE OneNote extension, and dtSearch will correctly sort through all of that.

Are there other zombielike examples people should be aware of?

OCR'ing documents can lead to typographical errors. And it is easy to mistype a word or two when dashing off an email. In dtSearch, fuzzy searching adjusts from 0 to 10 to sift through such typographical errors. With a low level of fuzziness, a search for *zombie* would pick up not only the word itself but also *zomdie* with the “b” mis-typed or mis-OCR'ed as “d.”

Anything else?

The final category of zombie data I'll mention are credit card numbers. It is easy to have these accidentally left undead in enterprise data. But dtSearch can run any digits that may represent a credit card number past an internal credit card validator included with the software. If the digits do indicate a credit card, dtSearch can flag that so you know to remove it from open data.

Final thoughts?

At dtSearch.com you'll find fully-functional 30-day evaluation downloads, enabling everybody—living or undead—to simultaneously and instantly search across terabytes of zombie and other data.

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

Article contributed
by dtSearch®



Recursively nested
files can also lead to
zombisque data.