Easter Eggs in Enterprise Data

Article contributed by dtSearch®

With Easter coming up, it's a good time to explore Easter Eggs in enterprise data. These Easter Eggs are not colorful plastic shells holding candy. Rather, today's topic starts with the alternate Easter Egg meaning of a hidden message in a movie, album, video game and the like. Easter Eggs in enterprise data, however, are less happy finds and more "gotchas" lying in wait. Fortunately, enterprise search can help see through these "gotchas."

Can you give some examples of Easter Egg "gotchas" in enterprise data?

An Easter Egg can take the form of a whole largely hidden file. Someone can save a document with a misplaced file extension, like giving a OneNote file a PowerPoint extension or giving an Access database a PDF file extension, making such a file hard to access in the ordinary course of review. Or an Easter Egg can result from recursively nested files, such as an email with a ZIP or RAR extension with an Excel spreadsheet that itself embeds a Word document. Both examples could be from a deliberate attempt to obscure a file, an unintentional accident, or—particularly in the recursively nested example—normal file manipulation. Probably more common than whole-file Easter Eggs would be text snippets hidden intentionally or accidentally in an otherwise normal-looking file.
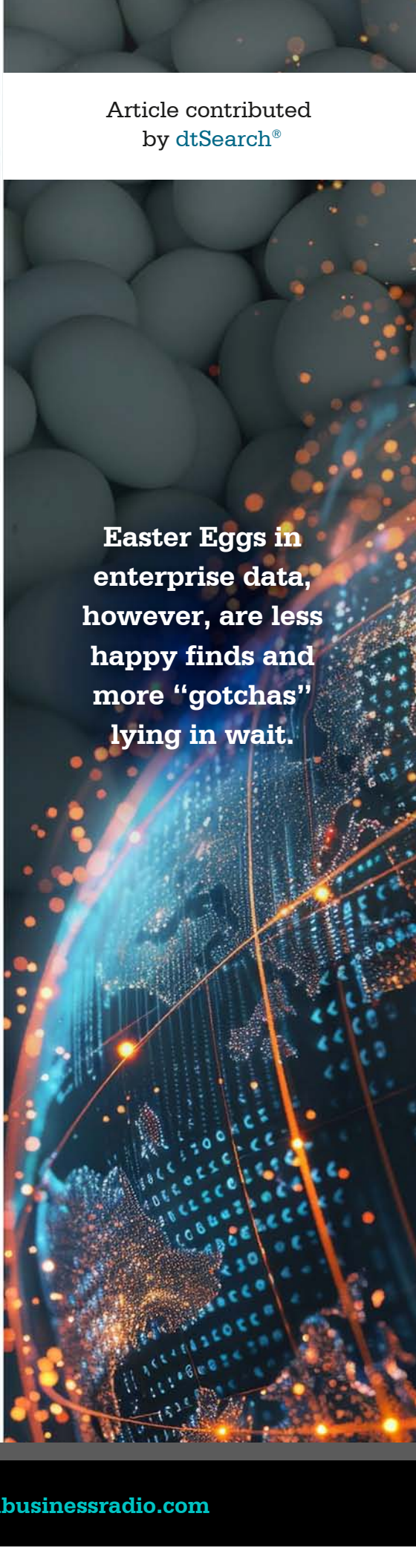
Like what?

A file can have certain metadata that an end-user might never find just clicking around the file in its associated application. Text can also blend in with its background color such as black writing against a black background or eggshell-color writing against an eggshell-color background, making such text nearly impossible to spot in its associated application. A file can also have track changes with the edits not fully accepted so the earlier text remains, even if not immediately apparent. A file can further have text that someone tried to redact but that persisted behind a black rectangle. Amazingly enough, this example has even made an appearance in some major court filings in recent years.

Are there other Easter Egg examples?

While text that blends in with a background color can be hard to spot in a file's associated application, the opposite problem can also occur, most notably in PDFs. A PDF can look like a normal PDF from the outside and read like a normal PDF in a PDF viewer, but really be an "image only" without text available to copy and paste and to search. "Image-only" PDFs can sneak into a collection of normal text-based PDFs by accident. Or "image-only" PDFs can be a deliberate attempt to obscure content, e.g. in an eDiscovery context.

Files can also have what I'll call micro Easter Eggs. Mistyping in an email or an OCR error scanning a file can result in minor misspellings, like *Easmer Eggs* instead of *Easter Eggs*. And then you can have items that shouldn't be in open archives all but are nonetheless sitting there in enterprise data as "gotchas" such as credit card numbers.

> **Easter Eggs in enterprise data, however, are less happy finds and more "gotchas" lying in wait.**

So what do you do about these?

The key to seeing through all these Easter Eggs is enterprise search. In an associated application, a file displays through the lens of that application. In contrast, enterprise search bypasses the associated application view and heads straight to the binary format of files. Enterprise search like dtSearch® can use the binary format to determine the file type, so a misplaced file extension will not affect the parsing of a file. The binary formats further let enterprise search seamlessly dig through recursively nested files like the email with a ZIP or RAR attachment with an Excel spreadsheet that itself embeds a Word document.

What about your other Easter Egg examples?

Text that blends in with a background color in a native application view is just ordinary text in binary format. And metadata, no matter how obscure in a file's native application, is readily apparent in the binary version. dtSearch can flag image-only PDFs as an alert that these require an OCR program like Adobe Acrobat Reader to turn them into fully-operational PDFs. For the micro Easter Eggs, dtSearch's fuzzy searching adjusts from 0 to 10 to sift through minor typographical and OCR errors. dtSearch can even identify any credit card numbers that may be lurking in enterprise data.

How does dtSearch work?

dtSearch lets multiple users simultaneously and instantly search terabytes only after indexing the data. Indexing is as easy as unwrapping a chocolate bunny. Simply point to the folders to index, and the software will take it from there. dtSearch can even seamlessly index remote files like Office365 or SharePoint attachments that present as part of the Windows folder system. A single index can hold up to a terabyte of text, and there are no limits on the number of terabyte indexes that the software can create and concurrently search.

What about reindexing data and ongoing concurrent searching?

dtSearch can automatically reindex data as often as desired using the Windows Task Scheduler. Updating an index does not affect continued instant concurrent searching, so there is no "down time" associated with keeping indexes current. The index structure is optimized for instantaneous concurrent searching even in high-traffic Intranet or Internet sites. dtSearch has over 25 different search features for precision indexed searching. After a query, dtSearch can display highlighted hits in a range of Easter Egg pastels for convenient browsing.

What about different languages?

dtSearch works with Unicode in files and emails. Unicode supports hundreds of international languages. A file or email can go from English; to a different European language; to double-byte Asian text like Chinese, Japanese or Korean; to a right-to-left language like Hebrew or Arabic; and then back to English; and dtSearch will track all of that.

Final thoughts?

Find Easter Eggs and more in your own enterprise data. dtSearch.com has evaluations ready for immediate download.

Article contributed by dtSearch®

Probably more common than whole-file Easter Eggs would be text snippets hidden intentionally or accidentally in an otherwise normal-looking file.