Skeletons Welcome for Halloween! Not So Welcome in Data, But Enterprise Search Can Help

Article contributed by dtSearch®

Most people enjoy a good scare for Halloween. Much less delightful are skeletons like intellectual property risks, employment liability concerns and misplaced credit card numbers in an organization's data. While enterprise search generally maximizes efficiency by letting multiple workers at once instantly search terabytes, it can also proactively locate skeletons before they pop out and scare someone to death.

Enterprise search can instantly search terabytes after first indexing the data. Carving a pumpkin for Halloween? Hard. Indexing terabytes with enterprise search? Easy. Just point to the folders and the like to index and the software will do the rest. With dtSearch®, for example, a single index can hold up to a terabyte, and you can build and simultaneously search as many terabyte indexes as you want.

For efficiency, enterprise search approaches files in their binary formats. (Retrieving files in their native applications would take way too long. We'd all be skeletons first.) Parsing specifications for PDF, Word, Access, Excel, PowerPoint, OneNote, emails and other binary formats can vary dramatically. Enterprise search needs to determine the exact right file format of each item before indexing. Luckily, the indexer can do that by itself, looking inside each binary format.

Indexing should be comprehensive, with no potential skeletons slipping through the cracks.

- A file with a misplaced extension like a Word document with a .PDF extension won't affect the indexer as the indexer looks to the binary format to determine file type, not to the file extension.
- Data like black writing against a black background that can hide in a native application view of a file is just plain text to the indexer.
- Metadata requiring exhaustive clicking around in a native application to spot is on the same level as any other metadata to the indexer.

Reprinted with permission of IT Business Net, itbusinessnet.com

- The indexer can also tackle multilevel nested files like an email with a ZIP or RAR attachment holding an Excel spreadsheet which itself contains an embedded Word document.
- Finally, the indexer can handle remote content like SharePoint attachments and MS Office 365 files so long as these present as part of the Windows folder system.

After indexing, one or more concurrent users can instantly search the same data. Concurrent search is less like multiple people trying to simultaneously bob in the same small barrel for apples and more like enumerable incorporeal ghosts seamlessly haunting the same mansion. While indexing is resource-intensive, concurrent searching is resource-light. Multithreaded search can proceed with no limits on the number of concurrent search threads, with Internet/Intranet search running statelessly for easy scalability. Instant concurrent searching can continue even while automatically updating an index to account for new, modified or deleted content.

Enterprise search can support over 25 different types of search requests. The most basic query types are natural language "all words" or "any words." An "all words" search request for *skeletons halloween witches candy* would look for files that contain all of those words. An "any words" search request would find files that match even one of those items, like just mentioning *skeletons*.

Boolean, proximity and phrase options bring search to the next level, looking for precise configurations of words or phrases: *("Halloween night" w/16 "candy bars")* and *(skeletons or witches) and not ghosts.* (The w/16 connector would require the two initial phrases within 16 words of each other.) Stemming finds different word endings on the same root word like *witch* and *witching* in a search for *witches*. The wildcard character ! would find one missing letter or number and * would find any number of missing letters or numbers: *halloween night**.

Concept searching finds similar concepts from a built-in thesaurus or your own custom synonym rings. Metadata-specific searching requires one or more portions of a search request to appear in certain metadata, like limiting a search to only files that contain *halloween night* in subject metadata. Fuzzy searching adjusts from 1 to 10 to accommodate typographical deviations, such as those that may occur in emails or OCR'ed PDFs. A fuzzy search for *skeletons* would also pick up *skelerons*.

Article contributed by dtSearch®

Reprinted with permission of IT Business Net, itbusinessnet.com

Beyond word and phrase-oriented searching, enterprise search supports number and numeric range queries as well as date and date ranging searching. A date range of *October 29, 2023* to *November 31, 2023* would further pick up date variants like *10/31/23* and *Oct 31 2023* in full-text or metadata. Enterprise search can even flag any credit card numbers in data. It does so by taking any number of digits that might represent a credit card and running them by an internal credit card verifier.

Because enterprise search supports Unicode, it can sift through not only English text, but also other European language text, right-to-left languages like Hebrew and Arabic, and double-byte Asian text like Chinese, Japanese and Korean. A file or email can switch from one language to another multiple times, and Unicode and enterprise search will follow that progression. Enterprise search can even find specific Halloween and other Unicode emojis 👻

By default, enterprise search will use "vector space" relevancy-ranking. If *halloween, witches* and *candy* are prevalent across indexed data, but *skeletons* are scarce, *skeletons* would get a higher relevancy ranking with denser *skeletons* getting an even greater ranking. Enterprise search also supports custom variable term weighting, like an additional positive weight of 7 for *halloween* and a negative weight of 3 for *candy*. Or it can give extra "meat on the bones" weight to *skeletons* but only appearing at the top or bottom of a file or in certain metadata.

After a search, enterprise search can show a full copy of retrieved files, emails and other retrieved items with highlighted hits for convenient navigation. For a different view of search results, instantly re-sort by some other criteria, like file date or file location. You should have no problem retrieving those skeletons.

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®

Reprinted with permission of IT Business Net, **itbusinessnet.com**