Digging for Digital Gold

Article contributed by dtSearch®

The discovery of gold in 1848 in California spawned the Gold Rush of 1849. But finding gold today has less to do with wading through streams and more to do with effective text retrieval across terabytes of data. Here's how to get started – no Western migration required.

The first step is to build an index across the data. An index is just an internal tool that allows the search engine to quickly sift through terabtyes, the digital equivalent of panning for gold. A single dtSearch index, for example, can hold up to a terabyte of text across one or more data repositories.

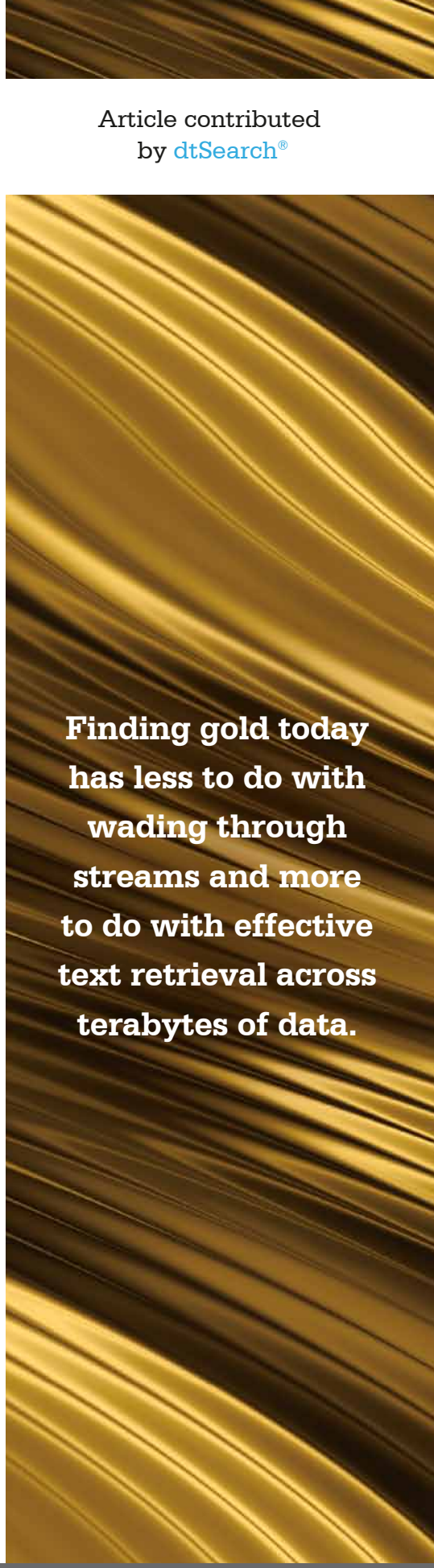
There are no limits on the number of search indexes that dtSearch can build and instantly search. In a concurrent-access environment like an Intranet or Internet site, search can run in a completely stateless manner, making it very easy to scale. Multiple concurrent search threads can operate with instant response time. Each user can then review search results with highlighted hits.

To start indexing, just point to the data you want to index and dtSearch does everything else. No need to even tell the software what types of data it is working with. The software will automatically recognize whether it is working with PDFs, emails, web-based formats, compressed data like ZIP or RAR, Microsoft Office files such as PowerPoint, Excel, Access, Word, OneNote, etc.

The final index contains a compilation of each unique word and number across the indexed data and its position in the data. For continually evolving data, dtSearch can update indexes automatically as often as you like using the Windows Task Scheduler. Updating an index simply adds new data, deletes old data and reindexes modified data. Index updates do not lock out searching, so instant search—even instant multithreaded concurrent searching—can continue unaffected.

Turning to search options, for the '49ers, the critical determination was binary: gold or not gold. Luckily, dtSearch has moved beyond a simple binary framework to provide over 25 different types of search options. You can search for words or phrases in any type of Boolean and/or/not configuration: *gold nuggets and (Sacramento Valley or San Francisco) and not fool's gold*. Or you can add a proximity element, such as taking the entire previous search request and adding a requirement that *gold nuggets* appear within 34 words of *silver bars*, or say 17 words before *silver bars*.

Concept searching finds thesaurus or user-defined synonyms. Add on fuzzy searching adjustable from 0 to 10 to accommodate potential typographical errors whether relating to old document scans or current email misstypings. That way, even if *San Francisco* is mistyped *San Franmisco*, you can still pick that up with a low level of fuzziness. By default, dtSearch will search the full text and metadata of all items, or you can limit one or more elements of a search to specific metadata.



Article contributed by dtSearch®

After a search, dtSearch will automatically rank search results for display with highlighted hits by so-called vector-spaced relevancy-ranking. With that type of default ranking, if you search for *silver or gold*, and *silver* appears millions of times in the indexed data but *gold* just a few times, *gold* mentions would rank more highly and items with the densest *gold* mentions would rank even higher. Or you can customize ranking "on the fly" at search time, like giving *silver* a negative weight of 7 and *gold* a positive weight of 8 regardless of the terms' prevalence in indexed data. Or you can choose to adjust the weightings if one or both terms appears in certain metadata or near the top of a file, for example.

Beyond word and phrase queries, dtSearch also supports number and numeric range searches, letting you search for any number between *1948* and *2022*. And dtSearch also supports date and date range searches, letting you search for any date between *July 4, 1949* and *April 17, 2023* even if the dates appear in different formats, like *July 11, 1949* and *5/18/21*. And dtSearch can even find credit cards in data, just to make sure that that a stray credit card number that paid for a prospecting pan isn't still sitting out there.

But now suppose someone has found gold and is trying to actually hide that fact in text. Notably, a search engine like dtSearch can also search through many types of hidden data. Some examples:

- Metadata may be "hidden" so it is really hard to spot looking at a file in its native application, requiring an extraordinary amount of clicking around to locate. But all metadata is right there in the binary format of each file, and hence "clear as day" to a search engine.
- Embedding documents inside other documents will also not work to obscure text. dtSearch can seamlessly parse multilevel container files, like an email with a ZIP attachment including a OneNote file with a fully embedded PowerPoint.
- Saving a file with a misleading file extension also won't work as a means of hiding text. You can have an Access database saved with an Excel spreadsheet extension and dtSearch will still recognize the correct file format and search it appropriately.
- Black text against a black background, white text against a white background or gold text against a gold background is just text for a search engine.
- And dtSearch can even identify "image only" PDFs. These are files that look externally like normal PDFs, but really consist of just a straight-up picture, with no actual digital text. dtSearch can flag those at indexing time so you know you need to run them through an OCR application like Adobe Acrobat to make them full-text searchable.

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 can download a fully-functional 30-day evaluation copy from dtSearch.com to find buried digital gold.