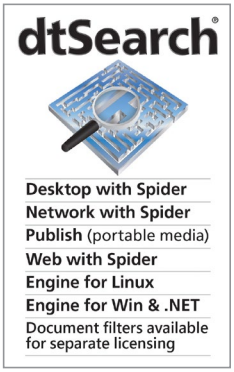
Getting to know dtSearch

Find out more about the company's broad product line in this Q&A



Q: The dtSearch product line includes enterprise and developer text retrieval and document filters. Can you explain these components?
A: Starting with text retrieval … Text retrieval can take the form of searching without an index or searching with an index. dtSearch products can perform both indexed and unindexed searches, displaying retrieved text with highlighted hits in context. However, the vast majority of customers use indexed search for the simple reason that, after indexing, searching across even terabytes of data typically takes less than a second.

Q: What type of data can a dtSearch index hold?
A: An index can contain multiple data sources – including files, emails and attachments, static and dynamic online data, and other databases. dtSearch products can create and simultaneously search any number of indexes, each holding a terabyte of data from any of these data sources. Federated searching can span all of these data sources at once, including integrated relevancy ranking across all retrieved data.

Q: What about online and other concurrent searching?
A: The product line offers efficient multithreaded searching, with no limit in the software on the number of concurrent search threads. For online search, the products can also run in a completely stateless manner, making it very easy to scale.

Q: And search options?
A: dtSearch products have over 25 full-text search options, including display of retrieved data (whether document data, email data, database data or other online data) with highlighted hits.

Q: What about metadata?
A: dtSearch developer products offer multiple advanced data classification and search display options integrating full-text and metadata, including advanced features like positive and negative variable term weighting, and faceted search.

Q: That covers text retrieval. But what then are the document filters?
A: The document filters allow dtSearch products to automatically identify and parse documents and other data.

Q: How does that work?
A: If you looked at an MS Word file in binary format—as a search engine needs to review it—the format would be so complex as to make it nearly impossible to pick out any words. Making this task even more complex, documents can have a nested structure. For example, an individual MS Word file could embed an entire MS Excel file, which itself could embed an MS Access file. The document filters take this binary data and parse all full-text content—including all recursively nested content—as well as all metadata, nested metadata, and even potentially hidden field content.

Q: So, prior to search, the document filters parse the entire file, including metadata and nested file objects?
A: Yes, then after the initial parsing and search, the document filters enable display of retrieved files with highlighted hits. And even apart from search, the filters can also perform data extraction and conversion.

Q: What MS Office file formats do the document filters support?
A: Support covers MS Word, PowerPoint, Excel, Access, OneNote and RTF. And support covers not only parsing all these files (whether "flat" or recursively nested) and display with highlighted hits, but also display of browser-ready embedded images along with the highlighted hits in these formats.

Q: Do the document filters support other file formats beyond MS Office?
A: Support covers PDF, HTML, XML/XSL, including display with integrated images and text in all these formats, along with highlighted hits. The document filters also provide hit-highlighted support for other databases like CSV and XBASE, as well as other "Office" formats like OpenOffice and Ichitaro. And the document filters also support compression formats like RAR, ZIP, GZIP/TAR, etc.

Q: Do the document filters cover emails?
A: Yes, including MS Exchange/Outlook (PST/MSG) and Thunderbird (MBOX/EML), along with display of highlighted hits and browser-ready images in these formats.

Q: What about email attachments?
A: Support covers the full-text of all attachments, including recursively embedded attachments. For example, the document filters would support an email with a ZIP attachment containing a PowerPoint that embedded an Access document and various images.

Q: What about online data?
A: The dtSearch Spider supports public and private or secure dynamic web data (MS SharePoint, ASP.NET, CMS, PHP, etc.) and static web data. Online data can consist of, or embed, document data such as HTML, PDF, XSL/XML, or even MS Office files, all of which the document filters would support.

Q: And other databases?
A: The dtSearch Engine APIs can also work with SQL-type databases. Databases can include BLOB file data, which the document filters would also support. On the development end, the dtSearch Engine's multiple options for working with integrated full-text and metadata are a key feature here.

Q: What programming languages does the dtSearch Engine support?
A: SDKs cover native 64-bit and 32-bit APIs in C++, Java and .NET through current versions. Additionally, the document filters are available for separate licensing for developers requiring data parsing, conversion and extraction functionality only, without searching.