Searching for Buried Treasure

September 25th, 2008

Searching and identifying relevant content is a common process for both electronic discovery and computer forensic investigations. But some people don't realize the challenges associated with indexing hundreds, or even thousands, of different file types and data structures. Mapping the data landscape may not immediately indicate where the textual "treasure" is located. Twenty years ago, full text searching was pretty simple. We usually had transcripts, and eventually optical character recognition (OCR) that was pretty straightforward to use (except for the less-than-perfect OCR results).

Today electronic based discovery requires our full text search engines to be able to extract the desired text from a wide variety of different file types, email formats, or the contents of the unallocated space on a hard drive. A common process mistake is assuming that all files are searchable. You hope to locate the relevant data by simply indexing the contents of a hard drive, DVD, or CD and performing a quick search on relevant keywords. Although sometimes it is this simple, there are several common exceptions that will prevent a complete search:

1. Encrypted and password protected files
2. Embedded files, sometimes at multiple levels
3. Archives (ZIP, RAR. and other compressed formats) and mail stores
4. Deleted files (some fully recoverable, and some with only minimal artifacts)

Both computer forensics and electronic discovery applications rely on full text search engines to locate relevant evidence. However, the common exceptions need to be handled to ensure that the content is available to the full text search software. I've been a fan of dtSearchfor many years because it handles large file collections of up to several terabytes, has extensive file type support, and great customer service.dtSearch is also integrated into several popular litigation support and computer forensic applications.