

Drowning in Duplicates? Here's a Life Raft.



Duplicate data is a growing problem. Organizations large and small increasingly rely on electronic communication solutions—like email, word processing tools, and shared network drives—to manage and distribute information. In itself, this is a positive shift because of the dramatic increases these tools bring to worker productivity. On the other hand, the challenge of these tools is the ease with which they create duplicate information.

This concept is easy to understand when you think about your own document collection. For example, think of a time when you created a spreadsheet or text-based document and then emailed the file to a colleague or client. In doing so, you created at least three copies of the same document: one on your hard drive, one on your organization's email server, and one on the recipient's email server. In addition, your organization most likely backed the file up to a tape drive or storage server as part of its disaster recovery strategy (a fourth copy). And, if you saved the document to a network drive to share with your co-workers? You can probably see where this is going.

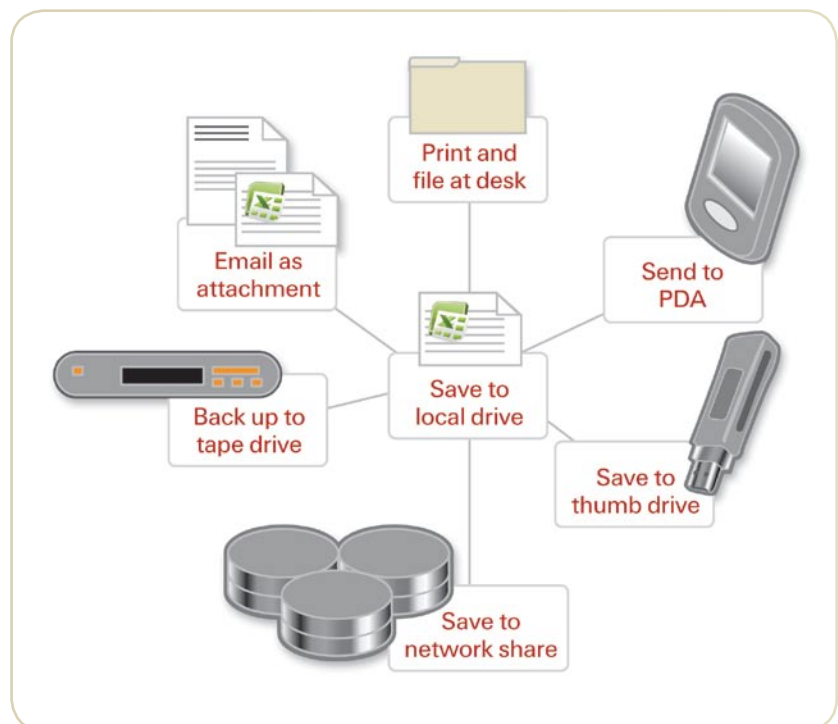
The problem with duplicate data is that it's expensive. And it's not just the cost associated with archiving data in an enterprise storage system—there's also the time and expense associated with manual review.

Many collections these days exceed a terabyte (1,024GB) of data, and the average collection consists of 30-50% duplicate data. Assuming a reviewer can mark 50 documents per hour at a cost of \$100 per hour, just identifying the duplicates on a 1 terabyte (1TB) collection would take human reviewers around 43,000 hours and would cost approximately \$4.3 million.¹ That doesn't even consider the review and processing costs for the rest of the collection!

About Automated Deduplication

Fortunately, there is a solution to the time and expense associated with manually identifying duplicates: automated deduplication. Automated deduplication technologies offer a proven way to quickly and accurately cull duplicates—and even near duplicates—from your document collection.

Exact Duplicates: Identifying exact duplicates (i.e., documents with a 100% match between content, metadata, and format) is a relatively straightforward process that is usually accomplished through software that uses MD5 hash-based technology to compare document content.



A hash value is like a document's fingerprint. It's created by an algorithm that analyzes the document's content and characteristics to assign a unique value to the document. Even a very subtle difference between two documents—for example, the addition of a dash or a pe-

Drowning in Duplicates? Here's a Life Raft.

riod—is enough of a variation to trigger the allocation of dissimilar MD5 hash values. So, an organization can safely consider those documents with matching values to be exact duplicates of each other.

The process of assigning MD5 hash values isn't perfect; there is a miniscule probability of the algorithm assigning the same hash value to two dissimilar documents. However, the likelihood of such failure is highly improbable.

Because deduplication software is so reliable at accurately identifying and culling exact duplicates, many organizations feel comfortable segregating or disposing of these documents without further manual review.

Near Duplicates: While identifying exact duplicates is pretty simple, identifying near duplicates is not so clear cut.

Near duplicates are documents with anything less than a 100% match between content, metadata, and format. While the differences between some near duplicate documents are immaterial, the subtle differences between other near duplicate documents are significant. For example, the 0.05% difference between two contracts might be the overall contract amount. Similarly, the slight difference between two emails might be the one sentence that gives rise to attorney-client privilege.

When it comes to identifying near duplicates, the best approach is usually to rely on automated deduplication software to identify similar documents, and then manually review those documents to make a final determination on near duplicate status. A good rule of thumb is to classify non-email documents that have a 70-80% content match as near duplicates that require human review. For email, a 30% content match is generally sufficient if the subject lines are identical.

Judicial Acceptance of Automated Deduplication

In the past, many attorneys were afraid to use automated deduplication technologies, believing they were unproven and therefore vulnerable to legal challenges. Today, that perception has widely changed.

Legal precedent shows that courts are frequently ordering deduplication. For example, in the Priceline.com Inc. securities litigation, when Plaintiffs moved to compel production of electronic data, the court ordered the Defendant to convert files into PDF or TIFF format and eliminate all duplicates. Likewise, in *Wiginton v. CB Richard Ellis, Inc.*, 229 F.R.D. 568 (N.D. Ill. 2004), the court ordered the producing party to deduplicate emails prior to production.

For large collections, automated deduplication is usually necessary to meet production deadlines and to reduce the financial burden of litigation. Fortunately, courts are ready to accept automated deduplication processes as long as they are highly transparent and based on reliable and proven technology.

The keys to successfully including automated deduplication processes in your litigation are:

- Gain acceptance from the court and opposing counsel before using automated deduplication techniques on a specific matter.
- Rely on experts to design and thoroughly document your automated deduplication process.
- Use hash-based deduplication software with a proven track record of accurate results.

If you follow these guidelines, you can feel comfortable knowing that you've effectively used automated deduplication to save time and money while still maintaining the defensibility of your litigation strategy.

¹ Assuming a 40% duplicate rate, the collection would include 410GB of duplicates or approximately 2.16 million duplicate documents.