



Effectively Managing Duplicate Electronic Documents In Discovery An Ounce Of Prevention ...

By Ramana Venkata

One of the most common and vexing challenges of e-discovery is that of duplicate documents.

And the problem is as old as it is widespread. Company archives have always contained duplicate records, and reviewers have long struggled to keep track of them during document review. In decades past, a reviewer may have encountered a paper document that had a duplicate somewhere else in the stack of boxes that comprised the document collection, but there was no easy way to know which box contained the duplicate. In large collections, where numerous reviewers would examine many boxes over the course of several weeks or months, it was virtually impossible to identify every duplicate document.

Dealing with electronic documents magnifies the problem. Electronic storage is so inexpensive and ubiquitous that more documents are saved in more locations than ever before. For instance, an e-mail message may be located in the outbox of the sender as well as in the inboxes of direct, *cc* and *bcc* recipients. Alternatively, many people can easily download a single document from a server onto their personal hard drives, laptops or thumb drives. Finally, additional copies are created when companies back up their system on a regular basis for document-retention or

disaster-recovery purposes. Even forgotten documents that are never accessed or viewed can be present on multiple backup tapes.

You surely know from the e-discovery practice grapevine, and may have firsthand experience that informs you, that reviewing duplicate documents can be expensive and inefficient. This is especially true for electronic-document review, where large documents, like spreadsheets and presentations, must be tediously checked to determine whether they are truly identical or contain minor variations. The task becomes onerous when multiple reviewers are involved. Not only is it difficult to coordinate duplicate recognition among different reviewers, but it's possible that different people will categorize the same document differently. Obviously, litigants have a strong interest in consistently categorizing documents.

Fortunately, it is far easier to automatically identify and organize duplicate electronic documents than it is to do the same for duplicate paper documents. Proper identification of duplicates can result in significant savings in review time and make for more accurate review.

THREE TYPES OF DUPLICATE DOCUMENTS

The underlying concept of duplicate documents is intuitively simple, but responding to the seemingly routine question "What is a duplicate document?" and defining exactly what duplicate documents are can be unexpectedly tricky.

Much of the subtlety regarding different types of duplicates emerges from the concept of document "metadata." Document metadata is information about a

document, such as its author, the date it was last modified or its logical or physical location. Metadata is fundamentally different from the contents of the document but provides a wealth of descriptive information about the document itself.

In contrast to electronic documents, few paper documents come with any sort of explicit metadata. While an attorney may know that a piece of paper came from a certain person's file cabinet, that knowledge is rarely written on the paper document itself. The metadata of an e-document identifying the source of that document, however, forms an intrinsic part of the electronic document.

The importance of metadata during review is common knowledge. Less well known is the role that document metadata plays in duplicate-detection and -sorting. The combination of the document's content and its metadata can be used to create an "electronic fingerprint" for every document. You may have heard the technical term *MD5 hash* in discussions about duplicate documents. This is one method for creating the "electronic fingerprint." Additional statistical and semantic algorithms can be used to enhance this "electronic fingerprint."

Electronic documents contain their own metadata as part of their internal structure. As a result, it is best to preserve and process electronic documents in their native format to ensure the access to their metadata that is required to analyze duplicate documents. If documents are printed out, or if they are converted to .PDF or .TIFF images, their metadata is typically lost. The first step in properly analyzing duplicate documents is to process documents in electronic form to preserve the metadata that is a part of every

Ramana Venkata is co-founder and CEO of Stratify Inc., an electronic-discovery company that has been retained by leading law firms and corporations in litigations and governmental investigations. You can reach him at ramana@stratify.com. The firm is on the Internet at www.stratify.com.

electronic document.

There are three types of duplicate documents that play an important role in discovery: exact duplicates, content duplicates and near duplicates. The specifics are outlined below.

Exact duplicates Documents are "exact duplicates" if they have identical content and metadata. Exact duplicates have an identical MD5 hash. For instance, if a single document (that has not changed) is saved on multiple backup tapes, then an exact duplicate of that document will be stored on every tape.

Before a set of documents is reviewed, attorneys often will request that exact duplicates be eliminated from the document universe. These documents are the same in every aspect, and usually there is little reason to spend valuable reviewer time on them. Electronic-discovery software can perform this operation automatically during initial document-processing.

Content duplicates. "Content duplicates" have identical content but different metadata. Documents with identical content but different metadata can arise in all sorts of situations. For instance, when copying a document from a laptop to the file server, the content remains identical, but the "location" of the document changes. Or consider an example common in our use of e-mail: An e-mail message sent from Ann to Bob will be found in Ann's outbox and in Bob's inbox. If Ann attached a document from the corporate server to her message, then the attachment will be located on the corporate server, in her outbox and in Bob's inbox. If Bob saves the attachment to his computer's hard drive, then a copy of the attachment will appear there, too. If, 6 months later, Bob opens the copy of the document on his personal hard drive and re-reads it without making any changes, then the metadata on his hard drive will reflect the fact that he accessed the document.

These simple examples show the potential value of document metadata, and the significance of content duplicates. Even if the documents have identical content, the fact that a document was found on Bob's hard drive might have significance, indicating that he read the e-mail from Ann. The ability to perform this analysis requires that all content duplicates be maintained in the electronic-discovery environment. Identifying content duplicates and enabling reviewers

to review and tag them together enables a more accurate understanding of the document while saving the time of examining a single document multiple times. There are matters, however, in which reviewers simply need to identify whether the document, regardless of its multiple locations, is responsive, or not. In such situations the content duplicates can be removed from the document universe.

Near duplicates. "Near duplicates" is the final category of duplicate documents. Near duplicates include documents that are highly similar to each other, such as document versions or messages in an e-mail thread. Similarity between documents can be determined on a statistical basis when the content of the document is approximately 99% similar, although this parameter is adjustable.

For instance, if a few cells in a financial spreadsheet are updated monthly, the different versions of that spreadsheet would be near-duplicate documents. A similar situation can occur when a contract, article or presentation moves through several revisions. Near duplicates are especially common in e-mail. If Bob replies to a message from Ann by appending "Right" to her message, then the two e-mails would be near-duplicate documents. If Ann sent Bob an e-mail with an attachment, and Bob made a few modifications before sending it on to Liz, then the two attachments would be recognized as near-duplicate documents.

During discovery, the identification of near-duplicate documents is extremely valuable. In the above example, imagine that Ann e-mails a document to Bob, who makes a few changes, and saves it to his computer's hard drive. A few months later, Bob then revises the document again before sending it to Liz. Three versions of the document exist: 1) attached to the e-mail from Ann to Bob; 2) on Bob's hard drive; and 3) attached to Bob's e-mail to Liz. If that document turned out to be important in the case, then a reviewer would want to examine all three copies at once to compare the differences in the three versions of the document.

The documents can be best analyzed when one person looks at them at the same time, because multiple people may not immediately recognize and appreciate the significance of the modifications.

Moreover, if the documents are stored in three different locations, they may be widely separated in the document set. They could even be produced at different times. Identifying and organizing near duplicates provides attorneys with a powerful tool to accelerate their review while increasing the consistency of document-tagging.

CONCLUSION

The benefits of managing these three different types of duplicate documents are strongest when they are used together. The ease with which electronic documents can be moved, stored, communicated and revised creates situations in which all instances occur. It is not unusual to encounter situations where:

- A specific version of a document is stored in multiple locations (creating content duplicates);
- Selected storage devices are backed up regularly to tape (creating exact duplicates);
- A document undergoes a series of revisions (creating near duplicates); and
- Some revisions are then included as attachments in e-mail messages (creating additional content and near duplicates).

The ubiquity of electronic storage and e-mail has made managing and reviewing duplicate documents in discovery a greater challenge than ever. An electronic-discovery system should be able to identify between these different types of duplicate documents and to deliver them in an organized fashion. The accurate identification of different types of duplicate documents can significantly decrease the burden they pose on reviewers. With this capability, reviewers can review duplicate documents at once, instead of over days or weeks, increasing the consistency and efficiency of their review.



This article is reprinted with permission from the September 2005 edition of the LAW JOURNAL NEWSLETTERS - THE BANKRUPTCY STRATEGIST. © 2005 ALM Properties, Inc. All rights reserved. Further duplication without permission is prohibited. #055/081-09-05-0006