"Just De-Dup it!" – Practical Considerations When Applying De-Duplication Filtering in E-Discovery

By Thomas E. Bonk
Vice President of Professional Services
DTI

At first glance, de-duplication filtering is perhaps one of the easier concepts to understand for newly minted E-Discovery practitioners. Most of us are taught that individual items that comprise Electronically Stored Information (ESI) each have a unique "fingerprint" that may be used as the basis to automatically identify and filter exact duplicates that inevitably exist within a universe of collected ESI. Most modern ESI processing software platforms have the ability to inexpensively log these fingerprints, commonly referred to as hash codes, and programmatically suppress the advancement of duplicate items for document review and other downstream activities. After all, why would you want to spend money on expensive Attorneys reviewing a document more than once?

But, alas, those making decisions that appear to be the easiest to confidently make often do not take into consideration some of the nuanced implications that can impact the integrity of document review workflow and the critically important endgame activities such as Production and Presentation.

For instance, there are two common approaches to defining the scope of how de-duplication filtering is applied, either within a defined custodian or across all custodians, i.e., "per project". The former method results in at most one appearance of a document for each custodian, but could result in multiple occurrences of a duplicate document if the document was collected from multiple custodians. The latter method is more restrictive. Only one appearance of a unique file is ensured, regardless of how many custodians the document was collected.

This sounds like a straight forward decision, right? De-duplication across all custodians reduces the total number of documents to review and possibly produce so that must be correct. Not so if a custodial based production protocol is ordered which requires that at least one document be produced for each occurrence collected for a custodian. This production order is common in investigations where the requesting party is particularly interested in establishing exactly which custodians had a particular document. If a relevant document was collected from six different custodians, then it would be produced at least six times, one occurrence per custodian.

There are workarounds available that require maintaining a running ledger of all custodians and file path locations for each document collected. But it can be messy to administrate and cumbersome to update this information for existing documents already migrated to a review workspace. The rolling nature of processing and

productions could also result in incomplete information or costly rework to update this information.

Sometimes it is difficult to define exactly what defines a “project” to establish the basis for per project de-duplication. Litigation projects often morph into sub-projects, offshoots and other unanticipated configurations. Because the concept of the “project” is often dynamic, decisions made early on about de-duplication filtering across an entire project may result in complex and error-prone rework or a decision to simply start over the processing of ESI.

In summary, applying de-duplication filtering on a project basis may seem like the correct call, but de-duplicating within custodians may prove to be a safer decision especially if flexibility is desired for downstream changes and custodial-based production protocols.

Another commonly encountered problem is the appearance of documents within a universe of de-duplicated ESI that are seemingly exact duplicates. The root cause for this may be related to how the ESI was either maintained by its source software application or the existence of multiple application platforms.

For instance, email messages sent from Outlook to a recipient using a different email platform such as Lotus Notes or Cloud based commercial systems or even different devices may undergo nuanced revisions in how certain metadata fields are maintained and stored. These metadata fields form the basis for the hash code algorithms. An example could be the format of the sent date\time values in email is often inconsistent across different platforms resulting in the generation of different hash codes.

Further, email archiving methods that remove attachments from the corresponding message for storage efficiencies, a process commonly referred to as stubbing, may slightly alter the metadata and message text and can result in the generation of different hash codes.

Defining time zone offsets as a processing specification to display a perceived “accurate” sent date may also unexpectedly impact the generation of different hash codes for seemingly identical email messages. Some E-Discovery practitioners normalize all processing protocols to a common time zone such as GMT to obviate this problem, but some litigators may object to the practice of presenting displayed sent date\time on email messages to a normalized time zone unrelated to the location of where the ESI was maintained in the normal course of business.

For some processing platforms the order by which ESI is processed may impact whether an individual document is filtered as a duplicate. If an email with an attachment efile is processed first, the subsequent processing of that same item maintained as a loose efile may result in the loose efile version of the document being filtered as a duplicate. Many

litigators would not consider it proper to suppress a standalone version of a document from a sub-item of a compound document. It is important to understand whether the settings used during processing allow for this scenario.

Lastly, the exact methods used by various ESI processing applications to extract and log hash codes are not necessarily consistent across software platforms. The rude upshot, often learned under duress and late in the process, is that the lack of conformity across software applications does not allow ESI processing activities to be easily transferred to a different platform without incurring significant risk that the precision of ongoing de-duplication filtering will be severely compromised. That is, once you make your bed early in the process, you may be destined to sleep there for the long run.