

⌘ Cull Your Original Collections Before Processing

by Kevin Carr of InterLegis

Determining document relevancy presented an inescapable burden to litigators long before the era of electronic data discovery (EDD). While originating in paper environments with manual processes, the burden of determining document relevancy has continued to grow even with the advent of new technologies and automated processes. More than ever, legal and IT professionals are reminded of this burden as they face audits, investigations and litigation.

Relevancy and the Challenges of ESI

In today's electronically stored information (ESI)-intensive workplace, the sheer volume of data — the garbage mixed with the gold — intimidates electronic discovery professionals. In fact, almost any suit or inquiry that encompasses ESI includes relevant documents held in many different locations overseen by many different custodians. Recent regulatory changes have increased the challenge of e-discovery as they have lengthened the required hold time for certain documents. This increased hold time translates directly into an increase in information that e-discovery teams must collect and cull. Also, this increased hold time may increase e-discovery complexity as it sometimes means collecting outmoded, legacy-formatted data.

Powerful computers that can quickly process vast stores of information coupled with enabling technology that converts paper documents into electronic formats have made it much easier for today's litigators to manage ESI. However, it is ultimately a human analyst's decision that determines what information is useful and what information is not. With mountainous data collections that far exceed those of the paper days and the growing use of communications technologies to include e-mail, instant messaging and text messaging, effectively determining what is relevant information can quickly become an incredibly time-consuming and expensive proposition for even the most talented of analysts.

In the process of determining relevant information, organizations are paying legal teams hundreds of thousands — in some cases millions — of dollars to find meaningful fact patterns within terabytes of electronic data. Legal teams work to process and load collected data into a review platform so attorneys can perform first-level review. Many times, these processing and loading actions are accomplished by legal and IT specialists who are not schooled in the latest e-discovery technologies. These new technologies can help legal teams deduplicate, analyze and cull discovery data sets well before entering the more expensive process of attorney review. Use of these new technologies is critical for organizations seeking to decrease the time and costs associated with e-discovery.

New technologies that enable “pre-culling” can indicate the most relevant documents based on keywords, date ranges, custodians, concepts and other characteristics. Additionally, pre-culling technologies can protect the integrity of company information and help prevent the inadvertent production of fragile trade secrets in unculted data sets provided to opposing counsel.

The Value of New Pre-Culling Technologies

Today's pre-culling tools come in several forms with primary features that include matching, hashing, clustering, plotting and visual analysis.

Similarity matching tools can compare documents in an ESI collection and find similar sentences, phrases, concepts and even page layouts. For example, they can locate documents that contain references to “\$1.4 million paid to XYZ Corporation on June 12, 2005,” stated in various ways (*e.g.*, “ABC Associates paid \$1.4 million to XYZ Corporation on June 12, 2005,” “XYZ Corporation received \$1.4 million in June 2005 from ABC Associates.”) While similarity matching is useful for culling purposes, it also can help in the identification of the original source of a particular piece of information.

In e-discovery, a “hash” can be considered as a digital fingerprint for a document. Any change in a document — as minor as the deletion or insertion of a comma — alters its hash value. One of the most common hash formulas in use today is the MD5 algorithm. An MD5 message hash (*i.e.*, digest) helps e-discovery professionals both verify the integrity of transferred files and check the digital signature of those files. By applying hash functions to MD5 digests, legal teams can quickly locate documents in different formats within a sizeable data collection. Additionally, through the use of pre-culling hashing tools, they can rapidly identify duplicate documents by comparing hash values.

The finding and grouping of documents in e-discovery has also been enhanced by new pre-culling tools that go beyond query methodology in concept and fuzzy searching. Not long ago, document sets were compiled with keyword searches and then narrowed by using fewer search terms. Now, with the advent of concept clustering (*i.e.*, foldering), advanced document analysis can help organize information more effectively by subject. This clustering capability greatly facilitates the review process by showing attorneys which subjects warrant the greatest attention.

Plotting is another valuable pre-culling tool. By plotting the attributes of a document, legal teams can study and ensure only potential responsive information is identified for inclusion in the EDD review set. Plotting also helps ensure that certain relevant documents, which may be meaningless

•• at first blush, are not inadvertently discarded. A few examples of document attributes include keywords, dates, document types, native file types, original document locations, metadata, recipients, coding fields, authors and custodians. Taken together, these and other attributes constitute the “personality” of a document. Just as humans have individual traits, so do documents. With the pre-culling technology of plotting, legal teams can now index these data traits and analyze them in terms of unique relationships with other documents.

Finally, new visual analysis technologies are also having a major impact on culling strategies. As legal teams derive a condensed, workable subset from the original mountain of collected data, they can now better evaluate data as they can actually “see” relationships among documents. Visual analytics can reveal the complex relationships among documents based on factors that include file format, author, dates, custodian and concept, just to name a few. Legal teams can try various data combinations to expose increases or decreases in activity and to highlight noteworthy relationships. In very short order — perhaps only a few minutes — legal teams can isolate a small group of e-mail messages pertaining to a given subject that were exchanged within a critical time period among specific individuals.

The Importance of Data Mapping

Data mapping software is perhaps the most powerful pre-culling tool. It provides the framework for visual analysis, showing users the different “points” across their continent of data. A good mapping program can extract and index metadata and text from native files, create clusters based on any combination of attributes (including metadata, content, concepts and communication threads) and enable users to search and analyze document collections prior to full EDD processing. Data mapping applications should be able to remove duplicates in advance and, correspondingly, can help attorneys and case administrators reduce irrelevant documents by as much as 80 percent. However, to achieve optimal results, users should apply data mapping technologies before processing collected data.

Legal teams should also consider features such as flexible graphing, robust reporting and Web-style interfaces as they evaluate data mapping solutions. Additionally, it is important for legal teams to consider recent data mapping improvements centered on pretagging. Pretagging enables legal teams to code documents categorically before exporting them. This capability provides another important way that users can cull down data on the front end of e-discovery before the project enters the costly review phase.

Another benefit of data mapping software is that it provides litigators direct control over the document collection. They can manipulate data themselves, in real time, without the need for vendor assistance or external processing. Litigators can also examine different “what-if” scenarios, prioritize document groups and immediately identify documents that are most logical and promising to their specific matter. With this direct control, they can analyze available information and determine the most appropriate data strategy before submitting the streamlined data collection for full processing.

Working Smarter, Not Harder

The objective of pre-culling is to save time and money while ensuring thorough, accurate e-discovery results. Simply stated, pre-culling provides a smart way to process and review an extreme volume of electronically stored documents.

Designed to identify unique document relationships, pre-culling technology allows for faster document review and more effective document coding. While legal and IT professionals can obtain individual pre-culling tools as separate software applications, ideally they may want to consider a unified pre-culling application. A unified pre-culling application should help users avoid multiple (and redundant) processing steps, help reduce time required for pre-culling and help trim overall project costs.

By taking control of mountainous data collections at the outset, legal teams can focus on the work that needs to be accomplished rather than wading through unnecessary data. Additionally, by taking control of data collections at the outset with pre-culling tools, legal teams can proactively address the ever-present burden of determining document relevancy with a renewed sense of purpose.

About our author :: :: ::

Kevin Carr, President of InterLegis, Inc., has a wide range of mission-critical Internet-based technology expertise. He began development of the InterLegis system in 1998 in order to provide powerful and cost-effective solutions for the litigation discovery field. Today, he's considered a thought leader in developing best practices and effective uses of cutting-edge technologies designed to streamline every phase of the discovery life cycle. He can be reached at kcarr@interlegis.com.

This article was first published in ILTA's May, 2008 issue of Peer to Peer and is reprinted here with permission. For more information about ILTA, visit their website at www.iltanet.org.