



# DIGITAL DISCOVERY & E-EVIDENCE



VOL. 10, NO. 10

**REPORT**

JUNE 24, 2010

Reproduced with permission from Digital Discovery & e-Evidence, 10 DDEE 10, 06/24/2010. Copyright © 2010 by The Bureau of National Affairs, Inc. (800-372-1033) <http://www.bna.com>

## **BNA INSIGHTS**

IE Discovery's Chris May urges attorneys to consider moving from a custodian-based approach to e-discovery to a content-based model, despite the seemingly daunting obstacles.

### **The Leading Edge of Document Review: The Content-Based Review Process**



By CHRIS MAY

**W**hen it comes to litigation, most attorneys would rather focus their time and energy on crafting the best legal arguments and approaches that will give their clients the optimal outcome for each matter. However, attorneys often find themselves getting bogged down in the expense and effort of e-discovery instead. This is particularly true of the document review phase, which accounts for between 60 and 80 percent of

discovery costs. The need to review documents manually imposes considerable overhead, and it challenges the capacity of the legal system to perform e-discovery effectively.

The traditional custodian-based approach to document review—collecting and reviewing discretely for each person believed to have documents related to the case—has become increasingly cumbersome. The massive amounts of data that must be reviewed, along with the tight time frames of litigation, make this custodian-based approach increasingly problematic, not to mention risky if a key document was not in the possession of the custodians collected from.

*Chris May is CEO of IE Discovery, Inc.*

With the right processes and cutting-edge technology, though, litigators can take a different approach—one that focuses on content, where documents can be organized not by who created or handled them, but based upon the information they actually contain.

Changing from a custodian-based approach to a content-based approach is not a simple matter. It requires a paradigm shift for in-house and outside attorneys, IT staff, clients and vendors. Processes and technology must be appropriate for the matter at hand, thoroughly vetted and documented completely. They must also be cost-effective. Before launching a content-based approach to e-discovery, it's important to understand the pros and the cons and how such an approach can affect litigation management.

## ADAPTING PROCESSES, THEN ADOPTING TECHNOLOGY

In many ways, technology has been more a part of the problem than the solution when it comes to e-discovery. With the explosion in data amounts and types, including e-mails, IMs, Web pages, voice mails and texts, parties involved in litigations and investigations get buried under potentially responsive information.

There is still no technological magic black box where litigants can throw all their data, then have it automatically sort the non-responsive information from the responsive information and produce it in exactly the format that counsel wishes. So you must consider your discovery management approach, as well as your technology. In order to take advantage of the new tools that exist, attorneys must learn to adapt their processes before they can adopt new technologies.

The custodian-based approach is so deeply ingrained that as soon as a lawsuit appears on the horizon, most litigation teams immediately begin to think in terms of key players—who may have potentially responsive information, how can the most effective litigation holds be issued and how can that information be collected. Surveys are sent out, org charts are studied and interviews are conducted.

The content-based approach focuses instead on data, not people, in order to find out where information islands exist within the organization. Content analysis tools generally work in several different ways, as described below.

### Tools That Turn a Mountain Of Documents Into a Large Hill

Some technologies help limit the amount of data that must be reviewed, saving time and money.

**Clustering technologies** group similar types of documents together.

Consider the example of the phrase “birthday cake.” Generally speaking, e-mails that include phrases such as “birthday cake,” “celebrate” and “lunch room” will almost certainly have everything to do with recognizing an employee’s birthday and nothing to do with the matter under litigation. When technology can automatically cluster these types of documents together in a group, those files can be set aside and won’t need to be reviewed or produced.

However, there is a fine line between culling and setting aside. You must have complete confidence in the technology you are using, and you must be able to defend the technology in court if you need to. If your opponent decides to go on a fishing expedition and focuses on the “birthday cake” e-mails, you must be able to adequately explain to the judge why those particular e-mails were set aside and not subject to an extensive review. This is where a fully documented process can prove to be invaluable.

**De-duplication technology**, including near-dupe tools, represents an additional type of content-based tools that can narrow the number of documents requiring review. This is another area, though, where you must understand how documents are identified as identical or nearly identical, so you can defend your processes and technologies in court.

### Tools That Streamline The Review Process

These types of tools will save costs, and they will also offer more thorough results from the review.

One approach is **semantic clustering of batches**. This can be particularly valuable if you are using a subject matter expert for the review. Documents can be grouped in such a way that the experts can review all of the potentially responsive documents that relate to a particular subject matter, rather than having the experts review documents that don’t relate to their expertise or having reviewers without specialized knowledge mark those documents. Not only can you get a more defensible review, but you can increase efficiencies and the volume of documents that can be reviewed per hour.

Those who use Google’s Gmail™ are already familiar with the concept of **e-mail threading**, where e-mails are grouped by conversations, not by date. However, Microsoft’s Outlook® can do the same thing, with the right settings. Reviewing entire conversations, rather than picking through e-mails by date and custodian, can shave a great deal of time off a discovery.

**Automated document summarization** is another valuable tool that can help improve review processes.

### Tools that Focus on What Is Important

Several tools can help you more quickly find what you are looking for, including **thesaurical concept searching**, or technology that can automatically find documents that contain words synonymous with search terms your team has identified.

With **statistical clustering**, you can save time by organizing information prior to the review process. Again, this approach can allow for routing of documents to an appropriate subject matter expert rather than just a single reviewer for each custodian.

Some of the old familiar tools still have a place in content searching, such as **Boolean or keyword searching**. That’s the type of searching most attorneys are familiar with, and it’s still one of the best ways to search—as long as you know what you are looking for.

## Tools that Provide New Insights

Some of the new content-analysis tools offer different ways to look at data.

Some provide new ways to look at the **relationships between documents and people**. Certain technologies can distinguish how often documents and custodians have had contact with each other. If you can identify how many conversations people have had with each other or how often different people have altered a file, you can get a sense of whether those could be potentially relevant.

**Timelines and chronologies** represent another type of content searching, one that been used successfully for many years. In fact, the failure to take advantage of timelines and chronologies has caused many litigants a great deal of trouble.

Consider the *Zubulake v. UBS Warburg* case, where the defendant was sanctioned for gaps in data. If UBS Warburg attorneys had made use of a timeline, they may have noticed the significant gaps in e-mails from the plaintiff, which could have raised some questions about the thoroughness of the discovery.

## OBSTACLES TO IMPLEMENTATION

There are several reasons attorneys shy away from content-based technologies. They require a specific technical knowledge in order to validate the results and their practical application is poorly understood and usually ill-defined. And litigators are often uncomfortable with the paradigm shift that is required to move away from the custodian-based approach. Those attorneys tend to migrate back to their comfort zones. Rather than try to learn something new, it's much easier to revert to the tried and true, even if the tried and true is expensive, cumbersome and not necessarily effective.

This lack of understanding of the capabilities of the technology erodes confidence in the tools' results. Litigators have to be able to defend the results of their discovery methods. In fact, they stake their reputations and licenses on the thoroughness and defensibility of their methods. Gaining confidence in new technology is a chicken-and-egg problem. With an important case, there is too much at stake to experiment with new technologies, but with small cases, it's not worth the trouble.

However, nothing is perfect, not even a review that is done manually by highly qualified attorneys. So the goal should not be perfection. Instead, it should be defensibility. It is the reality we all have to live with.

It's also important to remember that you have to use the right tool for the right job. Consider concept or semantic search tools, which are very good for searching large, homogenous document collections when you aren't sure what those documents may contain. These tools work better for finding additional examples of concepts that are well-represented in the collection. On the other hand, under-represented concepts may not be as searchable. If you suspect that a so-called smoking gun exists in that collection, concept or semantic search tools will probably not provide the best approach for that particular collection.

## CHANGING YOUR APPROACH

There are several ways to adapt your processes in order to optimize the adoption of new tools. As an example, consider the electronic discovery reference, or EDRM<sup>1</sup>, model. In the EDRM approach, the different discovery steps focus on information management, identification, preservation, collection, processing, review, analysis, production and presentation. Several ways to adapt those processes and utilize new technologies for better results are described below.

**The Information Management Phase.** In the litigation readiness phase, it's important to have in place information management appliances, maintain maps of data stores and have a strategy for password repositories and management.

As part of the early case assessment, you can leverage tools to extract richer information. This is an area where chronologies and timelines can be particularly useful, to see if you are missing significant chunks of data from specific time periods.

You may also want to adapt your processes to perform investigations without the assistance of a custodian. After all, in many instances, especially in investigations, the custodian is an adverse party.

This is also the part of the discovery where you will be preparing for the meet and confer with your adversaries, and improved processes and the right tools can offer you a strategic advantage. If you enter the meet and confer with the right systems and technology in your pocket, you can often negotiate better terms from the outset of the litigation. And this is an excellent place to gain more confidence in your tools it is unlikely you will be called on to defend the technology you use at this point in the discovery.

**The Identification and Preservation Phase.** Here, you can adapt your processes to work with IT personnel, rather than custodians. With a content-based approach, you may not even need to involve the custodians, freeing them to focus on their jobs and eliminating much of the work and worry over whether custodians are properly managing litigation holds and turning over relevant data.

You can also focus on data stores, rather than people and saving relationships and timeline maps for further use in the analysis phase.

**The Collection Phase.** Here, your processes should maintain references to removed entries when culling de-dupe and near-dupe documents.

As the technology in this area improves, it's important to make sure your processes have evolved as well. If you are using more than one vendor for a single matter—a common occurrence when data stores are enormous and timelines are tight—consider how they are working together. With multiple vendors, you need to be sure their culling techniques and technologies are cooperating and not working against each other. If your vendors are not talking to each and sharing an understanding of each other's processes, you can quickly get into trouble. For example, it's important to be able to de-dupe across custodians, but you need to ensure that

<sup>1</sup> See, EDRM.net.

your vendors are keeping records of who was involved in each conversation and in what context.

**The Culling and De-Duplication Phase.** In this phase, you need to use the categorization and concept tools to mark—but not remove—particular data.

Your team, including your vendors, needs to understand the nature of a rolling collection, particularly in this phase. Many software vendors don't realize how the terms, issues and amount of data coming in changes constantly throughout the life of the discovery. Many of the technical tools assume that all information is available at all times which of course is not the case in a rolling collection or production process.

**The Processing/Loading Phase.** In this part of the discovery process, indexing time for rolling loads may be dramatically increased. You should also make sure your processes and technologies can handle bad OCR and image-only documents.

**The Review Phase.** In this phase, smart planning can really pay off. The vast majority of costs in e-discovery are incurred in this phase, where you are getting the data before the reviewers' eyes. The best way to save time and money is to get the most relevant documents before the reviewers.

This is also the phase where batching documents by issues and concepts, rather than custodians, allows you to better leverage your experts.

In one situation we were involved with, we had 28 different passes with documents. It was a technical case that revolved around infringement claims based on software code, so we needed technical experts to conduct the review. By batching on issue and concept, we were able to get the most relevant documents in front of those experts.

However, it's important to remember that concept tools need a minimum number of documents in order to be effective. Performing a batching on a small collection early in the lifecycle of the case often gives undesirable, and even strange, results. This may result in a loss of confidence in the product, when in reality the technology itself is highly trustworthy—it was just used in the wrong situation.

In order to boost confidence, you can also use clustering as a final quality control check. This aspect also bleeds over into the analysis phase.

**The Analysis/Research Phase.** In this part of the litigation, there are several content-based tools that can be particularly effective, including e-mail threading and content samples rather than keywords. But those keyword searches have their place and can supplement

content-search features. You can also use keyword searches for needle-in-the-haystack searching.

In fact, no matter how good other methods become, keyword searches will probably always have a place in e-discovery.

**The Production Phase.** In this phase, the paradigm shift away from custodians and towards content becomes particularly challenging. You must have confidence that your tools and processes have returned the most relevant documents. Judges may say that litigants need to utilize these types of tools in order to conduct efficient discovery, but the judges aren't the ones staking their reputations on the accuracy of the discovery.

This represents a major hurdle for litigators and the legal software industry, but planning ahead can help everyone cross into the unknown. This is an area where defensibility and black box concerns become particularly nagging, even if you considered those issues in the meet and confer.

At this phase, you can also perform statistical sampling to support the methods you have used, when necessary. While you should do this throughout entire process, it is critical to do so now.

## GETTING STARTED

For many attorneys, starting small with content-based approaches and technologies usually offers the best approach. E-mail threading represents an excellent place to begin. It's easy to understand and explain, so it is often a good jumping-off point.

Of course, attorneys aren't the only ones who need to be convinced. The client's upper management must buy into the new technology as well. In order to prove the ROI on these types of technology, you should start measuring costs in meaningful, and often different, ways. Consider what you really spend in costs and time for review and collection. Calculating the document-per-hour rate makes sense for many organizations. This type of data is easy to collect, and it offers a good framework for determining ROI. By limiting the number of documents that need to be reviewed, you can save considerable time and money, while boosting the quality of the review.

You can also gain valuable insight from your peers—find out what they are using, what has worked for them and what hasn't. Associations can offer good networking opportunities. But eventually you need to take a breath, jump the hurdle and just try something new. Start small, and once you have certainty and mastery of one area, branch out to the next. Technology has led to many of the current challenges of e-discovery; it's time to use technology to solve some of those problems.