

TechTalk:



Using Unique "Personalities" to Cull e-Discovery Data

By Kevin Carr, President, InterLegis

Editor's note: Each month, Kevin Carr will share his insights about discovery technologies in an easy-to-read straightforward manner. Check each edition of ALSP Update to stay abreast of his view on technology topics and trends. In this article, he discusses the unique attributes that discovery data contains and how to leverage that information to find relevant connections.

Have you ever been told you look like someone? Sometimes you can see it; sometimes you can't. For example, I've been told a few times that I look like Derek Jeter of the Yankees. Hmmm. I think there may be a couple of similar features, but overall I don't agree. However, I'll gladly trade him for his job and paycheck! I've also been told I look like Eddie Vedder of the band Pearl Jam. Don't really see that either ... maybe if my hair was long and I was screaming into a microphone ... but I'll take that too since I'm a huge fan (and will trade for his job as well).

Our Unique Personalities

For starters, there's the physical aspect of who we are. Sure, you may look like someone famous or someone somebody knows, but when it comes down to it, nobody looks EXACTLY like you (unless, of course, you are an identical twin). Each of us are unique in various aspects of our appearance such as complexion, eye color, shape of our nose, height, build, ethnicity, and many other physical features.

Then there are even more non-physical things that make us distinct from others, including: where we're from, our childhood, religion, hobbies, interests, education, skills, talents, career path, friends, family, taste in music and so on. In essence, there are so many attributes that make us who we are, it is impossible to find an exact match anywhere out there in the world.

Despite our uniqueness, we also share many random things in common with others. Picture yourself in a public place like a restaurant, movie theater, or a major sporting event. It's likely that just about everyone around you is a perfect stranger. However, if you started talking to these people you'd quickly find many things in common. How

As Seen in:



many share similar interests? How many are from your hometown or region of the world? How many are in similar industries? Went to the same college? Some may even live right down the street from you. And certainly, you'd be surprised how many mutual people you and these random strangers know or have connection to – the proverbial six degrees of separation in action.

It is the wide range of attributes we all possess that makes up our unique personality. And it's the certain aspects of our personalities – these common threads – we share with others that enable us to connect, cultivate friendships and build relationships. Think about your circle of friends, business associates or acquaintances. For each person you think of, there are certain attributes that connect you to each other.

In short, the potential to connect with others exists everywhere we go and with everyone we meet. And therefore, we take for granted just how many common threads we actually do share with most people around us – even with those that seem unlikely on the surface – if only we'd take a moment to discover those connections.

Now, this isn't an article about love, unity and world peace. There's no "Kum Ba Yah" moment here, and I promise not to channel John Lennon and start singing "Give Peace a Chance."

The "Personalities" of Electronic Data

However, I've spent time talking about this because, as silly a segue as it may seem – Electronically Stored Information (ESI) that gets collected for discovery actually has unique "personalities" as well. How so? Well, if you take an average document, it will possess a combination of aspects that make it unique to every other in the collection (unless, of course, it's been collected twice).

Just like people, each document's unique set of attributes gives it its own "personality". Taking it a step further – just like how we have connections to random people in a public place, there are common threads within these document personalities that share connections to other files in the database. And with proper analysis, these personalities and their connections can help legal teams streamline the e-Discovery process.

.....
... it's the certain aspects of our personalities – these common threads – we share with others that enable us to connect, cultivate friendships and build relationships.

.....
Just like people, each document's unique set of attributes gives it its own "personality".
.....

To illustrate, let's look at an e-mail for example. Its unique attributes would include: file type, sent date, custodian, author, recipients, attachment information, content, file location, header information and more. But despite its uniqueness, these attributes, along with others, can also represent relevant connections – or relationships – to other documents in the database that share some of these same elements. These possible common threads can include: concepts contained within, its position within an entire email thread, changes in recipients, relevant dates, as well as similarities (and differences) with other documents or related e-mail threads ... just to name a few.

Clearly ESI is significantly different from paper documents because of all these "moving parts" contained within the data – from content, to attributes, to metadata. And with the right technologies, we should be able to leverage all the data points inherent to ESI in order to discover all the stories contained in these collections. It's a process I call "**Relationship Mining.**"

The problem with existing tools commonly used for culling, processing and reviewing ESI today is that most do not fully use these unique personalities to their advantage. Instead, most legal teams take a limited and linear view of this data in order to find responsive documents. In other words, only basic criteria are typically utilized to flesh out important documents in most cases. But given the wide range of information built into every electronic file, there's so much more intelligence available that gets left out of the mix.

For example, let's look at the steps involved in your average ESI processing and review project. When electronic collections are initially delivered, they are typically in the form of fairly unstructured data residing on a hard drive. The next step in the process is to load the data in an e-Discovery processing tool in order to normalize the set. Once that step is completed, the data is usually filtered – or culled – using limited criteria such as keywords, file types, dates, custodian information, and other high-level attributes. And although this standard process **DOES** reduce the set somewhat in most situations, it still promotes the idea of casting a fairly wide net to ensure nothing important falls through the cracks. As such, many irrelevant documents make it to the next step: review. And therefore, tremendous time and money investments will be made in an already expensive process. More irrelevant data means more to process, more to load, more reviewers to assign, more data to host, and more time required to get through the review. And the costs rack up every step of the way.

Using Document Personalities To Cull Electronic Data

However, with the right relationship mining technologies, electronic collections could be dramatically culled to the smallest, most relevant set by analyzing all the common threads that exist within these document personalities. For example, a typical personality-based culling process could include the following steps:

1. Select relevant custodians:

By sub-dividing the collection by various attributes, you can easily start with that which is most obviously relevant. In this case, having the ability to choose only each relevant custodian's sub-set of documents is a great way to "trim the fat," so to speak. Let's assume that out of 9 custodians that produced documents, there are only three that we know were primarily involved in the issues of the case: Bob Smith, Susan Johnson and Stephen Davis.

2. Drill down into relevant concepts:

Depending on the issues pertinent to the case, it may be a good idea to analyze documents by subject matter using concept analysis tools. Doing so allows you to choose what data is potentially relevant based on what documents say. So, if the issues at play relate to a contract dispute, concept categories such as: "contract negotiation," "contract edits," "agreement status," "revised pricing levels," and "project XYZ engagement" could be chosen to further narrow your focus. As you see, there are technologies out there that can read and understand what documents are all about and group them together for you. These concepts are simply another attribute (out of many) that can be used to determine responsiveness.

3. See a listing of all file types:

This would allow you to quickly see the various formats of the filtered data at this point. Such information can illustrate many things, including: What type of files represent the lion's share of communication? Are there any unusual file types that need to be dealt with (CAD drawings, proprietary formats, etc)? Are there certain file types that you would expect but aren't seeing (a possible sign of incomplete data harvesting)? For our example, let's say that since we are dealing with contract negotiations that we will choose to focus only on e-mails, Word documents, and PDFs — the standard formats found in such situations.

.....
... with the right relationship mining technologies, electronic collections could be dramatically culled to the smallest, most relevant set by analyzing all the common threads that exist within these document personalities.

4. DataMap selected file types along a timeline:

This is where visual data mapping technologies are useful (a future TechTalk topic). In short, having the ability to see data in an illustrative format allows you to quickly identify trends in activity. So let's say you mapped the filtered collection along a timeline. Doing so allows you to quickly see what date ranges represent various spikes in activity, which usually fleshes out relevant communications. For this example, let's say based on what we see we decide to further cull the set by choosing the following date range: May 2007 – August 2007.

5. DataMap all communication threads between relevant entities:

Here's where the full benefit of both relationship mining and data mapping can be realized. Mapping communication threads simply charts out all conversations between two or more people. However, if presented properly, the technology could very well flesh out communications involving additional players than those originally selected. The obvious benefit here is that by analyzing these common threads, you've ensured that important activity didn't fall through the cracks based on initial assumptions. From here, you can either go back and edit some of your filtering criteria, or go ahead and select the conversations you want to further analyze. Let's say that the technology has culled the set to a total of 75 e-mail threads, yet we've decided to narrow our focus on 10 that are likely most relevant based on everything above.

6. Find other documents that share relevant connections:

So here's where we are at this point in our culling process:

- We started by selecting three custodians out of 9
- Within that sub-set, we've selected five concepts out of hundreds
- Within that, we've identified three file types that typically deal with contract negotiations – by following only the conceptual connections that were relevant
- Within that, we've focused only on a narrow range of dates that share certain selected attributes in common: concepts, file types, and specific spikes in activity
- And then uncovered all communications (and then some) within those common threads

.....
*... we've essentially created
 a unique "personality"
 of responsiveness*

... all made possible by the ability to analyze the unique personalities of individual documents, then following selected common attributes that are shared with others within the collection. And now that we have our short list of relevant documents, we've essentially created a unique "personality" of responsiveness. At this point, relationship mining can be taken a step further by asking the technology to "show me other documents from the entire collection that share these personalities."

Summary

As you can see, by having access to the right technologies that indexes all attributes in a collection, we can quickly uncover special relationships. However, the key to making this work is that you need to have hands-on access to such culling tools. It is important to be able to control the process so that you can immediately react to the results you see to help you drill down or widen your focus, change your criteria, follow tangents and make decisions on-the-fly.

And you'll notice many of the common steps of e-Discovery culling and processing are missing from the above example. Specifically, no keyword filtering has been made behind-the-scenes. This is not to say that keyword filtering does not have its place in the process, however, many "false positives" can come from this standard method of culling. Additionally, keyword filtering requires you to cast a fairly wide net. And this means more irrelevant information makes its way into the mix. But as you can see, the above scenario easily goes above and beyond the standard success rate of keyword filtering.

.....
*... given the nature of ESI, it
 makes little sense to not use
 ALL the moving parts contained
 within these collections
 to your advantage ...*

In essence, given the nature of ESI, it makes little sense to *not* use ALL the moving parts contained within these collections to your advantage ... unless of course, you like spending excess time and money on e-Discovery! When dealing with electronic data, it's all there, so why not use it?

Hopefully, I've helped you see ESI collections and the e-Discovery process in a slightly new way. Just remember that each document, like people, is unique in its own way. And within those unique personalities lie important information that connects all relevant facts together. We're just on the forefront of understanding how to effectively use all these unique document attributes to our fullest advantage.

However, now that you know all documents have unique personalities, please refrain from trying to strike up conversations with them. If you find yourself doing that, well ... that means this business is getting to you and some serious time off is needed!

'Til next time ...

KC



www.InterLegis.com

Kevin Carr, president of InterLegis Inc., has a wide range of Internet-based technology expertise. As the architect of the InterLegis system, he has developed cutting-edge discovery technologies and best practices relating to data mapping, conceptual analysis, electronic data culling/processing, similarity matching, streamlined document review, automatic categorizations, visual analysis, native review, document digitization, optical character recognition, compression, database indexing, advanced searching and document security.
