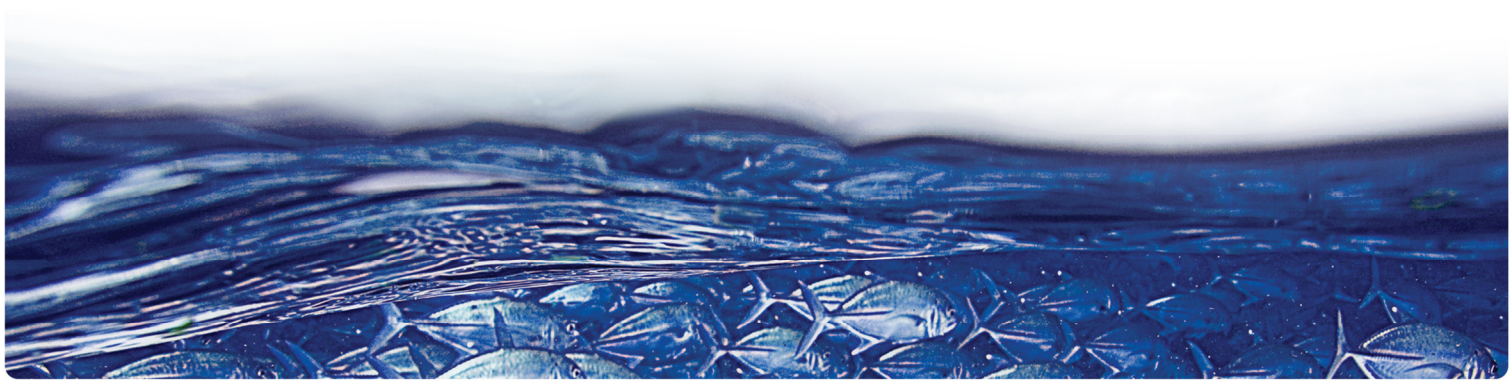
The best eDiscovery workflow in the world …ever[i]

Herbert L. Roitblat, Ph.D.

CTO, Chief Scientist, OrcaTec LLC

For more information:

info@orcatec.com
3200 Cobb Galleria Pkwy, Suite 200
Atlanta, GA 30339
Telephone: 888-335-2200 Ext 2



The best eDiscovery workflow in the world …ever

Of course, there are lots of workflows that could be used in eDiscovery and under the right circumstances they might be as good or better than the one I am about to propose. Still this workflow captures things that I have been thinking about for more than a decade—how to get the fastest, most accurate, most reliable, and least expensive eDiscovery possible.

Locate ESI

The first task in eDiscovery is to locate the data. What kind of devices is they stored on? Servers? Email systems? Laptops? Who has the information you're looking for? Is it better to collect from specific individuals or to collect it from email systems or other archives? Is it better for individuals to collect their own data or to have it collected independently? If the data are stored in an archive, it may be easiest and most cost effective to pull them directly from the archive.

Collect ESI

Although preservation holds may be necessary before the data are extracted from their locations, once extracted, these data are preserved by OrcaTec collection system. Regular data retention destruction schedules can be resumed without losing the required data.

Following collection, the data are indexed and analyzed. Duplicates and near duplicates are identified and clustered. Semantic clusters are computed, which organize the documents into groups of semantic similarity. A given document can be in more than one group, so documents that are about more than one thing are not in danger of being misclassified. For example, an email might say something like, "I'm bringing pizza to the party on Saturday, and by the way, the money that we stole is now in our Swiss bank account." You don't have to worry that this document will be assigned solely to the pizza party cluster where no one will look at it again.

Exploratory Data Analysis

At this point the data are ready for legal analysis. The OrcaTec Document Decisioning Suite contains a number of tools designed to help you understand what the collection is about and to identify the documents that are essential for formulating and advancing your case.

Concept search can be used to identify responsive documents, even if you don't know exactly the right keywords to search for. Concept search brings to the top of the list those documents that are most about the topic that you are searching for. It also expands the list of retrieved documents to include those that are about the same topic, but may not contain the specific word(s) that you searched for. The snippets for each search result help you to understand how the words were used in context and show the conceptual relationship between the search words and their semantic context.

Traditional Boolean search is also available and can be combined with concept search and any other criteria. You can search by specific fields, tags, and combinations.

Email threads provide further context for understanding what a message is about. They place each email in the conversational context in which it was sent.

Near duplicates pull together documents that are nearly identical so you can focus on their differences or treat them as a group. The difference visualizer lets you see at a glance exactly how two documents differ from one another.

Suggesters are available to help you to identify alternative forms of the key terms that you are interested in and alternative spellings. For example, you can enter plan into one tool and have it tell you all of the words in the collection that start with those letters (e.g., plan, plans, planning, planet, plant). You can then select the ones that you want to search for. The alternative spelling suggester allows you to input a word and it returns the words in the collection that are spelled similarly. People do not always spell things in the proper form, but if you don't know what spelling they did use, you could otherwise miss important information.

The interesting phrases tool provides words and phrases that are characteristic of a document. What makes this document stand out from the rest of the collection? These phrases can be used as additional search terms or as a quick summary of the document.

Timeline visualizations let you focus in a specific periods of time. You can identify specific documents that were sent soon after significant events. You get an overall impression of the volume of documents for each time period and can widen or narrow your focus as appropriate.

Social network analysis and the OrcaTec Sonar display let you identify who communicated with whom about specific topics, and when. Email senders and recipients are identified. For example, you could identify who sends the most emails about a specific topic or who receives the most emails about the topic. These visualizations can help you to identify potential candidates for depositions or identify new custodians whose data should be collected.

In the course of this exploratory analysis, you can tag documents. You have complete control over the tags that available and used. Express tags mean that the most frequently used tags can be applied to documents without having to open a tag tree. Individual documents, email threads, near duplicate clusters, and entire search results can be tagged with one or a few clicks.

Intelligent Culling

You can include and exclude documents from further analysis by custodian, sender domain, recipient domain or using any search criteria. For example, you may be able to exclude any emails from or to Travelocity.com or Amazon.com. Two special tags are always available for culling—"in" indicates that this document is worthy of consideration and "out" means that this document can be excluded from consideration. Documents that are tagged in may need further review to mark them as either responsive or nonresponsive. Documents that are marked out are not removed from the system, they could be marked in at another time.

You can use the semantic clusters to quickly identify large groups of documents that may be either essential or irrelevant to the matter. OrcaClustering gathers together documents that are about the same topic, without requiring any search. If a particular cluster is found to be irrelevant to the matter, the whole group of documents can be dismissed (marked out) with one click. Because a given document can be in more than one cluster, however, you don't have to be concerned that marking a document out will prevent it from being seen if it contains other

information that could be important. No matter how many times a document appears in a cluster that is marked out, if it ever appears in a cluster marked in, it will be marked in for further analysis. And, these decisions are reversible.

Extensive rules are built in to the system and can be augmented by project managers to control the order of precedence of tags. For example, the special "in" tag always takes precedence over the "out" tag. Tags that have been applied specifically to individual documents always over-ride those that have been applied through bulk tagging. Tag groups can be set up to make them mutually exclusive, for example, so that a document cannot be marked both "responsive" and "nonresponsive." These rules help to make tagging more reliable and systematic, while maintaining the user's level of control.

Categorization / predictive coding

Once you are familiar with the documents in the case, predictive coding can be a very effective method to apply that knowledge to categorize all of the remaining documents in the collection. An expert in the subject matter of the case, for example, the attorney who must sign the Rule 26(g) declaration evaluates a series of documents for responsiveness. After a few documents have been categorized by the expert, the computer comes to make responsiveness predictions about subsequent unseen documents (hence, the name predictive coding) and the expert can either confirm the computer's prediction or over-ride it. Over time, the computer comes to identify the evidence in each document that distinguishes the responsive from the nonresponsive documents—in the opinion of the expert.

Some cases merit a more lenient or generous tendency to categorize documents as responsive. Other cases suggest a more narrow definition. In any case, the computer comes to approximate the judgments of the expert reviewer. At the same time, the predictions help to improve the expert's own consistency. When a document comes up that the computer predicts is responsive, it says, in essence, the last time you saw a document like this one, you judged it responsive. This prediction helps to ensure the expert's consistency and alertness.

Because the documents presented to the expert are randomly selected, they are a representative of the collection as a whole. There is, therefore, good evidence on each successive sample of documents that the rules being learned will apply to the whole set.

The computer does not make up the rules for deciding between responsive and nonresponsive documents. Instead, it captures the opinion of the expert and the evidence in the documents to implement that expert's judgments.

The time it takes to "train" the predictive coding to identify documents depends on the richness of the collection. The more responsive documents there are in the collection, the faster is the training. The time needed for training does not, however, depend on the size of the collection. The small amount of time spent training the system can be applied to collections of any size.

The documents identified as responsive by the computer and the expert constitute the output of first-pass review. In a few hours to a few days, a single expert can effectively identify the responsive documents in a collection of a million or two million documents or more. Rather than using common culling methods, such as keyword selection, which is known to be only about 20% effective at identifying responsive documents, or having multiple teams of readers examine these documents at 70% effectiveness, this technique can achieve accuracies that can be much higher at much lower cost.

Quality sampling

The final stage in this work flow is to sample the documents that were not selected during the initial exploratory data analysis or during predictive coding. This sampling is intended to assure that we did not leave an unreasonable number of responsive documents in the set that was not selected. A random sample of documents is drawn from those that were not selected for further review. If none of these documents is found to be responsive, then we can say with a certain level of confidence that we did not leave an unreasonable number of responsive documents behind. The size of the sample that is needed depends on the confidence level and on our specification of what the maximum number of reasonable documents might be. The higher the level of confidence we want, or the lower the maximum number of reasonable documents, the larger the sample size has to be.

Conclusion

What makes this approach so valuable is that a single person or a very small group of people can explore the ESI in the matter, understand the content, identify the major constituents (in terms of content and custodians), and conduct a demonstrably effective first pass review in days, rather than weeks or month. This process is highly defensible in that it is transparent and in that it produces results of measured quality. It dramatically reduces the cost of first pass review while simultaneously raising the quality.

---

[i] My apologies to Virgin / Blue Note Records for cribbing the title to their 1999 Album, *The Best Blue Note Album in World …Ever!*