Predictive Coding: A Primer

by Amy Jane Longo, Esq.
and Usama Kahf, Esq.

O’Melveny & Myers LLP
Los Angeles, California

A commentary article reprinted from the March 2013 issue of Mealey’s Litigation Report: Discovery
Predictive Coding: A Primer

By
Amy Jane Longo and
Usama Kahf

[Editor’s Note: Amy Longo is a partner at O’Melveny & Myers LLP and a member of the Firm’s Financial Services and Electronic Discovery and Document Retention Practices. Usama Kahf is an associate and a member of the Firm’s Labor and Employment Practice. Both are resident in O’Melveny’s Los Angeles office. The opinions expressed in this article do not necessarily reflect the views of O’Melveny or its clients, and should not be relied upon as legal advice. Copyright © 2013 by Amy Jane Longo and Usama Kahf. Responses to this commentary are welcome.]

Having gained judicial approval — or acknowledgement — in fewer than a handful of cases, the method for collecting and reviewing electronic documents for discovery known as “predictive coding” appears to be “trending” — to borrow a term from social media culture. One way to understand the emergence of predictive coding — which is referred to alternately as computer-assisted review, technology-assisted review or intelligent review — is as an answer to the many critiques levied over the years by courts, litigants, and legal scholars about the adequacy of the “search terms” method of culling through repositories of electronically stored information (ESI) for relevant discoverable evidence.1 With the exponential increase in ESI discovery over the last decade, courts regularly address challenges to the sufficiency of producing parties’ efforts to search for and collect relevant ESI.2 Discovery disputes abound over the search terms utilized, number of custodians, types of repositories searched, and method of collection.3

To satisfy a party’s obligations in responding to discovery requests, the method the party uses to search, collect, and review ESI must be defensible as reasonable, if challenged.4 Importantly, while cost of discovery is certainly an important consideration for all parties, the legal adequacy of any search and review method must turn on its precision and accuracy in culling relevant documents and appropriately determining their responsiveness to discovery requests, not whether the method of choice provides costs savings relative to other methods.

Predictive coding involves “training” a computer (by way of an electronic discovery software or review tool) to recognize and identify the documents in a review set that are relevant and/or responsive to discovery requests. The software “learns” how to code from the case team’s attorneys, as it tracks their review decisions (i.e., which tags are checked) and uses mathematical algorithms to predict based on the contents and characteristics of documents which tags the attorneys would have checked.

Predictive coding first received judicial endorsement as a reliable discovery review method in a well-publicized article on October 1, 2011, by Magistrate Judge Andrew Peck of the Southern District of New York.5 Noting that at that point no court had approved of predictive coding, Judge Peck wrote, “Until there is a judicial opinion approving (or even critiquing) the use of predictive coding, counsel will just have to rely on this article as a sign of judicial approval. In my opinion, computer-assisted coding should be used in those cases where it will help ‘secure the just, speedy, and inexpensive’ determination of cases in our ediscovery world.”6

Only a few months later, the defendants in a gender discrimination case before Judge Peck, Da Silva Moore v. Publicis Groupe & MSL Group,7 sought his approval to use predictive coding without objection from plaintiffs to this method other than to the

[MEALEY’S LITIGATION REPORT: Discovery Vol. 10, #6 March 2013]
wording of the proposed stipulation and certain aspects of Judge Peck’s ruling. His ruling, which was later approved by the district court,8 became the first judicial opinion in which a court approved the use of predictive coding in searching for and reviewing ESI.

As Judge Peck explained, both in his article and in Da Silva Moore, the traditional method of using keywords to search through repositories of ESI for relevant documents has many flaws. Specifically, requesting parties often simply guess “which keywords might produce evidence to support its case without having much, if any, knowledge of the responding party’s ‘cards’ (i.e., the terminology used by the responding party’s custodians) ... [and] the responding party’s counsel often does not know what is in its own client’s ‘cards.’”9 Poorly designed search terms generally do not benefit any party (whether the requestor or the recipient), as the terms tend to be both over-inclusive (returning many documents that are not responsive) and under-inclusive (failing to identify responsive documents).

As a potential alternative, predictive coding enables the computer, with human input and coaching, to determine relevance based on sophisticated algorithms and its observations of the human reviewer’s review decisions. One important aspect of predictive coding is that in order to properly “train” the computer on what makes documents relevant (i.e., in order for predictive coding to be successfully used), a more senior attorney familiar with both the law and facts of the case — rather than junior attorneys or contract reviewers — must review and code a “seed set” of documents. This may require an initial investment of time by the more senior attorney, but the effectiveness of predictive coding highly depends on the reliability of the “seed set.” As the attorney codes the seed set, the computer identifies characteristics of the coded documents (including, for example, various metadata fields) and begins to associate those characteristics with how the attorney coded the documents. The attorney continues to review the “seed set” until the computer learns enough from the attorney about which documents are relevant. This happens when the computer starts to predict the attorney’s coding with increasing rates of precision and accuracy. Soon, the computer’s predictions begin to correlate with the attorney’s coding. At this point, review of the “seed set” is complete, and the attorney can be confident that the computer can now take over the initial coding of documents and be just as accurate and reliable.10

In Da Silva Moore, Judge Peck concluded that use of predictive coding software as agreed to by the parties in their ESI discovery protocol is more adequate to satisfy discovery obligations than keyword searching.11 Judge Peck acknowledged that predictive coding is not perfect and cannot be expected to be, but neither is using search terms.12 He concluded with the following message to practitioners:

What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review. Counsel no longer have to worry about being the “first” or “guinea pig” for judicial acceptance of computer-assisted review. As with keywords or any other technological solution to ediscovery, counsel must design an appropriate process, including use of available technology, with appropriate quality control testing, to review and produce relevant ESI while adhering to Rule 1 and Rule 26(b)(2)(C) proportionality. Computer-assisted review now can be considered judicially-approved for use in appropriate cases.13

Since Judge Peck’s endorsement of predictive coding in Da Silva Moore, several state and federal courts have discussed, and some have approved, use of this method in ESI discovery. For example, in National Day Laborer Organizing Network v. U.S. Immigration and Customs Enforcement Agency,14 a Freedom of Information Act case, Judge Shira Scheindlin of the Southern District of New York pointed to predictive coding as an example of what she called “emerging best practices” for proper search, collection, and review of ESI. Judge Scheindlin noted that “parties can (and frequently should) rely on latent semantic indexing, statistical probability models, and machine learning tools to find responsive documents. Through iterative learning, these methods (known as ‘computer-assisted’ or ‘predictive’ coding) allow humans to teach computers what documents are and are not responsive to a particular ... discovery request and they can significantly increase the effectiveness and efficiency of searches.”15
As of the end of 2012, besides this positive mention in *National Day Laborer Organizing Network*, predictive coding has been approved by one additional federal court and two state courts, and another federal court has held an evidentiary hearing on the method’s adequacy.

First, in a multidistrict case involving allegations that the diabetes medication Actos increases users’ risk of developing bladder cancer, *In re Actos (Pioglitazone) Products Liability Litigation*, U.S. Magistrate Judge Hanna Doherty of the Western District of Louisiana issued a case management order regarding discovery of ESI with comprehensive instructions to guide the parties’ use of technology-assisted review. The order, which was stipulated to by the parties, provided specific directions on how the parties should consider and treat data sources, custodians, costs, and format of production, among other discovery questions. Notably, the order included a “Search Methodology Proof of Concept” governing the parties’ use of technology-assisted review tools to search for, collect, and review ESI. The order stated that the parties “agree to meet and confer regarding the use of advanced analytics” as a “document identification mechanism for the review and production of . . . data,” as well as to select four key custodians whose email will be used to create an initial seed set, after which three experts would “train” the software on coding documents based on relevance. As a check on the reliability of this method, the order directed both parties to collaborate to train the software and to mutually decide upon the appropriate threshold of relevance, with the right to seek input from the court in case of a dispute. The order also provided that after sufficient training of the software, the documents coded by the software must be randomly sampled for quality control, and the defendants would retain the right to manually review documents prior to production for relevance, confidentiality, and privilege.

Second, in *Global Aerospace v. Landow Aviation*, the Virginia Circuit Court similarly approved a predictive coding protocol after defendants moved for a protective order due to the volume of data sought by plaintiffs. Relying on *Da Silva Moore*, defendants successfully argued that predictive coding was not only significantly less expensive than manual review and keyword searches, but also significantly more reliable in identifying relevant documents.

Third, a judge in the Delaware Court of Chancery surprised both plaintiffs and defendants in *EORHB v. HOA Holdings LLC*, when he issued an order from the bench that required both sides to use predictive coding with the same vendor for ESI discovery (or otherwise to “show cause why this is not a case where predictive coding is the way to go”). This appears to be the first time a judge has required both parties in a case to use predictive coding when neither requested it.

Finally, in *Kleen Products LLC v. Packaging Corp. of America*, the plaintiffs challenged the defendants’ proposed Boolean search methodology as likely to find less than 25% of responsive documents and asserted that content-based advance analytics search (i.e., predictive coding) would identify more than 70% of responsive documents at no greater cost. Plaintiffs argued that predictive coding would “not focus on matching words but instead on identifying relevant concepts out of the documents,” and would “provide a richer, substantially more accurate return than Boolean searches.” Plaintiffs criticized Defendants’ Boolean keyword search as per se “subject to the inadequacies and flaws inherent when keywords are used to identify responsive documents.” Defendants defended their search method on grounds that their quality control processes will ensure a “degree of accuracy” in line with industry standards. They also contended that predictive coding would involve additional costs and burdens not contemplated by the Federal Rules, local rules or case law. The *Kleen Products* court did not directly rule on the propriety of predictive coding because after an evidentiary hearing and five months of meeting and conferring, plaintiffs agreed to withdraw their demand that defendants use predictive coding.

Predictive coding technology continues to evolve rapidly, as concerns over its mechanics and reliability are addressed. One critique of this method is based on the notion that there is no substitute for an attorney’s judgment, as computer software may potentially replace humans in document reviews. A balanced approach that incorporates attorney judgment and discretion along the way may alleviate some of these
concerns. Until such a method is tried and tested, predictive coding will remain on the cutting edge.

Endnotes

1. See, e.g., Chura v. Delmar Gardens of Lenexa, Inc., 2012 U.S. Dist. LEXIS 36893, *33–35 (D. Kan. Mar. 20, 2012) (ordering an evidentiary hearing to explore the sufficiency of the defendant’s search for responsive ESI where it failed to produce emails and other electronic documents and allegedly failed to do more than run a search for terms in an email program on one computer); Custom Hardware Eng’g & Consulting v. Dowell, 2012 U.S. Dist. LEXIS 146, *7–8 (E.D. Mo. Jan. 3, 2012) (discussing the problems that arise from keyword search methodologies, particularly that the word choices may be arbitrary and unable to reach the relevant information).

2. E.g., Williams Mullen v. U.S. Army Criminal Investigation Command, 2012 U.S. Dist. LEXIS 93977, *12 (E.D. Va. July 6, 2012) (outlining the standard for whether a party’s search for responsive documents was adequate, including whether it was “reasonably calculated to uncover all relevant documents”); William A. Gross Const. Assoc., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 136 (S.D.N.Y. 2009) (“[W]here counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI’s custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of ‘false positives.’”).


4. See Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 262 (D. Md. 2008) (“Selection of the appropriate search and information retrieval technique requires careful advance planning by persons qualified to design effective search methodology. The implementation of the methodology selected should be tested for quality assurance; and the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented. In this regard, compliance with the Sedona Conference Best Practices for use of search and information retrieval will go a long way towards convincing the court that the method chosen was reasonable and reliable.”). See also William W. Belt, Dennis R. Kiker & Daryl E. Shetterly, Technology-Assisted Document Review: Is It Defensible?, XVIII RICH. J. L. & TECH. 10 (2012), available at http://jolt.richmond.edu/v18i3/article10.pdf.


6. Id.


10. Id.


12. Id. at *34.
13. *Id.* at 40.


15. *Id.* at *12.


20. *Id.* at *5.

21. *Id.*

22. *Id.* at *4.

23. *Id.*

24. *Id.*

25. *Id.* at *5. ■