New York Law Journal
Technology Today
WWW.NYLJ.COM
An ALM Publication
VOLUME 251—NO. 02
TUESDAY, FEBRUARY 4, 2014

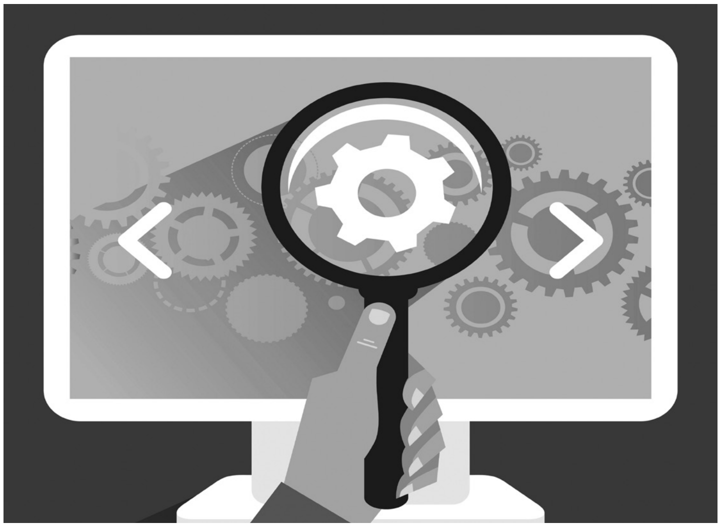
FEDERAL E-DISCOVERY

‘Seed Set’ Documents Should Not Be Discoverable

By
H. Christopher Boehning

And
Daniel J. Toal

Over the past two years, the issue of predictive coding—using computer-generated algorithms to aid in the determination of discoverable material—has attracted a great deal of attention. Yet this new frontier of legal technology has also raised significant issues, including how the work-product doctrine and its protections will translate as parties and courts confront the use of technology in the discovery process.[1] In this article, we briefly survey the recent decisions concerning the use of predictive coding, and explore the conflict over the extent to which the core processes of predictive coding are discoverable, including information about the “seed set” used to derive the relevant computer algorithm.



TAR: Predictive Coding

Predictive coding is a specific aspect of Technology-Assisted Review (TAR), which is an effort to make the e-discovery process more efficient through the effective use of technology. (For example, one commonly utilized TAR function is the use of keyword searches to reduce the volume of material to be reviewed.) Specifically, predictive coding employs computer algorithms to sort through the documents acquired through the e-discovery process, and to select the relevant material for production. This is not a purely automated process, however; the computer algorithms “learn” what documents are likely responsive through interactions with a human reviewer. The human reviewer codes a “seed set” of documents, noting which documents are responsive or non-responsive to specific issues. To be effective, the process will include identifying the “key” documents in a matter (to the extent known), as well as documents that are entirely irrelevant to the matter but likely to exist in the overall review population. The computer algorithm observes the properties of the relevant documents in this “seed set,” and, based on these observations, is subsequently able to classify the other documents in the set of potentially relevant material, without recourse to a human reviewer. This process does not stop with one round of analysis. Effective use of computer-assisted review requires an interactive process. After the seed set is developed, the computer algorithm provides tentative classifications, which are then confirmed or rejected by a human reviewer. This iterative feedback process, which requires human reviewers to again classify documents codified by technology, is a means to train the system. The documents used in this process are sometimes considered part of the underlying seed set, or are otherwise referred to as “training sets.” Some

H. CHRISTOPHER BOEHNING *and* DANIEL J. TOAL *are litigation partners at Paul, Weiss, Rifkind, Wharton & Garrison.* JACOB H. HUPART, *an associate at the firm, assisted in the preparation of this article.*

In some jurisdictions, this reprint may be considered attorney advertising. Past representations are no guarantee of future outcomes.

practitioners believe that in an appropriate case a properly devised computer algorithm can be faster, cheaper, and more accurate than other methods of document review.[2]

Of particular relevance here are the set of documents referred to as the "seed set," or "training sets," which are the array of documents referenced by the computer algorithm in "teaching" itself how to properly review the entirety of the potentially relevant material. The composition of the seed set and subsequent training sets are at the heart of the predictive coding process, as the individual document classifications employed at these stages guide the entire operation of the computer algorithm. The question of access to the seed set (and training sets) has been addressed infrequently by the courts, with divergent results. A detailed examination of those particular judicial decisions, and their potential relevance for the future use of predictive coding by parties engaged in substantial e-discovery, is provided below.

Predictive Coding in the Courts

Before analyzing recent decisions addressing the issue of the discoverability of the seed set, it is important to first provide a general overview of the treatment of predictive coding by the courts. There have been only a handful of judicial decisions addressing predictive coding. Most of these decisions have accepted the practice of predictive coding in general terms, viewing it as a reasonable method to review voluminous amounts of e-discovery.

- In *Moore v. Publicis Groupe*, a magistrate judge approved the use of predictive coding, stating that "[w]hat 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."[3]

- In *Global Aerospace v. Landow Aviation*, a Virginia state court ordered, over the objections of the plaintiffs, that defendants "shall be allowed to proceed with the use of predictive coding for the purposes of the processing and production of electronically stored information," but noted that the receiving party would still have the opportunity to "rais[e] with the court an issue as to the completeness or the contents of the production or the ongoing use of predictive coding."[4]

The "seed set," or "training sets," are an array of documents referenced by the computer algorithm in "teaching" itself how to properly review the entirety of the potentially relevant material.

- In *National Day Laborer Organizing Network v. U.S. Immigration & Customs Enforcement Agency*, Judge Shira Scheindlin advocated the use of predictive coding, stating: "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 FOIA or discovery request and they can significantly increase the effectiveness and efficiency of searches."[5]

- In *In re Actos (Pioglitazone) Products Liability Litig.*, the court approved a case management order agreed to by the parties, which included an agreement to engage in the predictive coding process, featuring the involvement of discovery experts from both parties in devising and implementing the seed set of documents.[6]

- In *EORHB v. HOA Holdings*, the court, after deciding a motion for partial summary judgment, ordered that "the parties shall … conduct document review with the assistance of predictive coding."[7] However, Vice Chancellor Laster reversed this compelled use of predictive coding the following year, stating that "Plaintiffs may conduct document review using traditional methods," while permitting defendants to "employ … computer assisted review tools to conduct document review."[8]

Overall, these decisions have evinced a general acceptance of the use of predictive coding as an available tool in managing e-discovery. This tentative approval by the courts, however, has provoked a number of related issues, including the extent of court intervention into disagreements between the parties concerning the limits of discovery involving the predictive coding process itself.

Recent Conflict: Discoverability

Certain decisions have suggested that the seed set used in predictive coding is discoverable to an adversary. In the first significant decision on predictive coding, *Moore v. Publicis Groupe*, Magistrate Judge Andrew J. Peck outlined an iterative process in which the parties would confer *multiple* times upon the documents in the seed set (and training sets) and the appropriate coding used.[9] Other decisions have followed the course outlined in *Moore*. In a pair of decisions from the Western District of New York, *Gordon v. Kaleida Health*,[10] and *Hinteberger v. Catholic Health System*,[11] the court appeared to contemplate that the parties would confer as to the documents used in the seed set for a document production based on predictive coding. Similarly, the document discovery plan endorsed by the court in *In re Actos (Pioglitazone) Products Liability Litig.* featured the involvement of both parties in devising and implementing the seed set for predictive coding.[12]

However, this aggressive approach to the discoverability of the seed set in predictive coding has been challenged by a recent decision, *In re Biomet M2a*

Magnum Hip Implant Prods. Liability Litig.[13] In this decision, Judge Robert L. Miller Jr. imposed limits on the discovery sought by plaintiff. In its brief, the defendant had argued that it "should not be required to produce non-relevant or privileged documents that were included in the seed set," arguing that such a production would be "outside the scope of discovery" and would "encroach on [Defendant's] work-product protections."[14] Miller agreed, denying the demand of the plaintiff for "the whole seed set [Defendant] used for the algorithm's initial training," ruling that "[t]hat request reaches well beyond the scope of any permissible discovery by seeking irrelevant or privileged documents used to tell the algorithm what to find. That [Plaintiff] has no right to discover irrelevant or privileged documents seems self-evident."[15] Miller based this decision upon the principles of Fed. Rule Civ. P. 26(b)(1), determining that the seed set of documents used in predictive coding is simply outside the scope of permissible discovery.[16]

Compelling forcible disclosure of seed set material is an unreasonable and unwarranted intrusion into the internal discovery processes conducted by a party, and arguably violates the core principles of the attorney work-product doctrine.

Need for Judicially-Imposed Limits

The conservative approach embraced by the court in *Biomet*, which limited the discoverability of the seed set, may render predictive coding more palatable to counsel and other participants in the e-discovery process. Yet this result will only occur if *Biomet*'s restrictive interpretation is followed and reaffirmed by subsequent decisions and other courts. As discussed above, the various prior decisions revolving around the seed set emphasized the access of opposing counsel to the documents and rationale underlying the seed set. Such an attitude by the courts only encouraged intrusions by adversaries into the details and process of an opponent's conduct of document review. As a result, counsel were rightly reluctant to adopt a document review system that would enable an adversary to entwine itself more readily into the internal decisions of counsel in the document review process.

The act of determining which documents to include within a seed set, or the classification and assigning of attributes to documents in a randomly generated seed set, both reflect significant attorney thought and effort, and should not be discoverable to an opponent.[17] In fact, it has long been settled that an attorney's selection and compilation of documents—segregating documents by relevance, privileged status, or overall importance to the case—is at the core of what is protected by the work product doctrine.[18] In effect, permitting discovery of the seed set of documents would enable an adversary to review documents deemed irrelevant, as well as documents deemed as "key," both of which are otherwise prohibited in the discovery process. The restrictive approach in *Biomet*, whether informed by the work-product doctrine or based solely upon the application of Fed. Rule Civ. P. 26(b)(1), preserves counsel's ability to conduct discovery away from the prying eyes of the adversary.

The exchange of seed set information can be done on a voluntary basis, according to an agreed-upon protocol between counsel, and respecting both the permissible scope of discovery and assertions of the work-product privilege. However, court intervention in compelling forcible disclosure of seed set material is an unreasonable and unwarranted intrusion into the internal discovery processes conducted by a party, and arguably violates the core principles of the attorney work-product doctrine. Furthermore, if courts follow *Moore* rather than *Biomet* by compelling such disclosures, judges will likely transform predictive coding into a less attractive option for counsel. Many attorneys, rightly or wrongly, still struggle with whether to exchange "search terms" used in the most commonly employed electronic discovery process—one in which algorithms and seed sets are not utilized. Predictive coding potentially requires an even greater intrusion into legal strategy and an attorney's internal assessment of a case, as it could involve court compelled disclosure of the documents deemed to be the most relevant by an attorney (those used in the "seed set"), and it is therefore unsurprising as to why counsel have been reluctant to adopt this new technology.

[1] See, e.g., Ralph C. Losey, "Predictive Coding and the Proportionality Doctrine: A Marriage Made in Big Data," 26 Regent U. L. Rev. 7, 28-30 (2014).

[2] See, e.g., Maura R. Grossman & Gordon V. Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," XVII Rich. J. L. & Tech. 11 (2011).

[3] 287 F.R.D. 182, 193 (S.D.N.Y. 2012).

[4] No. CL 61040, 2012 WL 1431215 (Vir. Cir. Ct. April 23, 2012).

[5] 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012).

[6] No. 6:11-md-2299, 2012 WL 7861249, at *3-*8 (W.D. La. July 27, 2012).

[7] 2012 WL 4896670 (Del. Ch. Oct. 15, 2012).

[8] 2013 WL 1960621 (Del. Ch. May 6, 2013).

[9] 287 F.R.D. 182, 186-88, 200-03 (S.D.N.Y. Feb. 24, 2012).

[10] No. 08-CV-378S(F), 2013 WL 2250579 (W.D.N.Y. May 21, 2013).

[11] No. 08-CV-380S(F), 2013 WL 2250603 (W.D.N.Y. May 21, 2013).

[12] No. 6:11-md-2299, 2012 WL 7861249, at *3-*8 (W.D. La. July 27, 2012).

[13] No. 3:12-MD-2391, 2013 WL 6405156 (N.D. Ind. Aug. 21, 2013).

[14] Defendants' Response to Plaintiffs' Demand for Defendants' Predictive Coding Seed Set, *In re Biomet M2a Magnum Hip Implant Prods. Liability Litig.*, No. 3:12-MD-2391 (N.D. Ind. Aug. 5, 2013), D.E. #722, at *3.

[15] Id. at *1.

[16] Id. at *2 ("Rule 26(b)(1) doesn't make such information disclosable."). It should be noted that the producing party here had voluntarily disclosed the identity of the relevant non-privileged documents used in the original seed set.

[17] See, e.g., Karl Schieneman & Thomas C. Gricks III, "The Implications of Rule 26(G) on the Use of Technology-Assisted Review," 7 Fed. Cts. L. Rev. 239, 262 (2013) ("To the extent that development of the seed set reflects attorney work product, the certification obligations of Rule 26(g) clearly do not require disclosure.").

[18] John Soumilas, "Compilations: Truth, Privacy, and the Work Product Doctrine," 73 Temp. L. Rev. 227 (2000).