THE BUSINESS LAWYER

Business Law Section • American Bar Association

Electronically Stored Information in Litigation

Volume 69, November 2013

*By Timothy J. Chorvat and Laura E. Pelanek**

I. Introduction

Recent developments in the use of electronically stored information ("ESI") in litigation center on two emerging themes: one technological—the increasing prominence of predictive coding; and one legal—the increasing stress on proportionality as a way to limit the burden and expense of discovery. The overriding question is no longer whether electronic discovery is necessary, but how to conduct it most effectively in terms of scope and cost. Courts and litigants are focusing on tools to harmonize the requirements imposed by case law and procedural rules with the equally binding laws of economics.[1]

II. Teaching Computers to be Smarter Than We Are: Predictive Coding

A. Emerging Case Law Supports Predictive Coding but Does Not Mandate It

The current buzz phrase in electronic discovery is "predictive coding." A 2012 case, *Da Silva Moore v. Publicis Groupe,*[2] first brought wide attention to predictive coding, inspiring articles like *Why Hire a Lawyer? Computers Are Cheaper.*[3] *Da Silva Moore* was a gender discrimination case against an advertising conglomerate in which the parties agreed to a protocol that called for the use of predictive coding, but plaintiffs later changed their minds.[4] Magistrate Judge Peck encouraged the use of predictive coding, which uses sample e-mail data to train software to identify relevant documents, explaining that "every person who uses email uses predictive coding, even if they do not realize it. The 'spam filter' is an example of predictive coding."[5]

After the parties submitted the protocol and the court approved it,[6] plaintiffs challenged the protocol in several respects. Plaintiffs asserted that predictive coding "provides unlawful 'cover'" for defendant's counsel to escape their duty to certify that production is complete and correct under Rule 26(g) of the Federal Rules of Civil Procedure.[7] Judge Peck rejected that contention, noting that in "large-data cases like this, involving over three million emails, no lawyer using any search method could honestly certify that its production is 'complete'— but more importantly, Rule 26(g)(1) does not require that."[8] Plaintiffs also objected to the ESI protocol on the basis that it lacked a standard to determine whether the method is reliable.[9] Judge Peck described that objection as premature.[10]

Judge Peck observed that his opinion appeared to be the first to endorse the use of predictive coding, but he stressed that he was neither mandating predictive coding nor suggesting that predictive coding should be used in all cases.[11] Rather, "[w]hat the Bar should take away from this Opinion is that computerassisted 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."[12]

Following *Da Silva Moore*, a Virginia state court ordered the use of predictive coding in *Global Aerospace, Inc. v. Landow Aviation, L.P.*[13] There, defendants sought approval to use predictive coding for the purposes of production.[14] The court granted that motion, but noted that the order was "without prejudice to a receiving party raising with the court an issue as to the completeness or the contents of the production or the ongoing use of predictive coding."[15]

In *National Day Laborer Organizing Network v. U.S. Immigration & Customs Enforcement Agency,*[16] Judge Scheindlin resolved disputes about the production of metadata and also addressed predictive coding. Plaintiffs brought that action seeking records from five federal agencies under the Freedom of Information Act.[17] The parties disputed the adequacy of the agencies' searches for responsive information.[18] Judge Scheindlin explained that "[i]t is impossible to evaluate the adequacy of an electronic search for records without knowing what search terms have been used."[19] The court stressed that "[s]eemingly minor decisions—whether intentional or not—will have major consequences."[20] The court directed the parties to design new, targeted searches to address the deficiencies she noted and to provide full documentation of those searches, including custodians, sources, keywords, and Boolean search terms.[21] The court then recommended the use of predictive coding: "Through iterative learning, these methods (known as 'computerassisted' 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."[22] The court suggested that, "if [the parties]wish to and are able to, then theymay agree on predictive coding techniques and other more innovative ways to search."[23]

Vice Chancellor Laster of the Delaware Court of Chancery also issued a ruling supporting the use of predictive coding.[24] After deciding a motion for partial summary judgment in *EORHB Inc. v. HOA Holdings LLC*, he stated that "[t]his seems . . . to be an ideal non-expedited case in which the parties would benefit from using predictive coding."[25] Should the parties disagree with that assessment, Vice Chancellor Laster directed them to "show cause why this is not a case where predictive coding is the way to go."[26]

Another sign of the acceptance of predictive coding came in *Gabriel Technologies Corp. v. Qualcomm Inc.,*[27] where prevailing defendants sought attorney's fees with regard to plaintiffs' patent and trade secret claims.[28] The court awarded more than $12 million in fees,[29] including more than $2.8 million "for fees associated with a document review algorithm generated by [an] outside vendor."[30] In the motion seeking those fees, defendants explained that they had "collected almost 12,000,000 records—mostly in the form of . . . ESI . . . . Rather than manually reviewing the huge volume of resultant records, Defendants paid [a vendor] to employ its proprietary technology to sort these records into responsive and non-responsive documents."[31] The court found the "decision to undertake a more efficient and less time-consuming method of document review to be reasonable under the circumstances."[32]

On the other hand, Judge Miller applied considerations of proportionality to reject a request to direct a party solely to use predictive coding in *In re Biomet M2a Magnum Hip Implant Products Liability Litigation.*[33] There, defendant collected 19.5 million documents and used keywords and de-duplication to reduce the universe of potentially responsive materials to 2.5 million documents.[34] Biomet then used predictive coding to identify documents to be produced.[35] Plaintiffs contended that the use of keywords tainted the process and resulted in a reduced responsiveness rate.[36] Plaintiffs wanted defendant to apply predictive coding to the full universe of documents, but defendant objected on the basis of cost, noting that it spent more than $1 million on the initial production, and asserting that starting over would cost millions more.[37] The court noted that the issue was not "whether predictive coding is a better way of doing things than keyword searching prior to predictive coding," but "whether Biomet's procedure satisfies its discovery obligations."[38] The court found that Biomet's procedure fully complied with Rules 26(b) and 34(b)(2), as well as the Seventh Circuit Pilot Program's Principles Relating to the Discovery of Electronically Stored Information.[39] The court also rejected plaintiffs' request under the proportionality standard of Rule 26(b)(2)(C), based on its seven-figure cost.[40]

Magistrate Judge Nolan stressed the importance of cooperation while addressing a motion to compel the use of predictive coding in *Kleen Products LLC v. Packaging Corp. of America.*[41] There, plaintiffs complained about the defendants' use of a Boolean search method to identify responsive materials, arguing that the search was "subject to the inadequacies and flaws inherent when keywords are used to identify responsive documents."[42] Instead, plaintiffs requested the use of content-based advanced analytics to conduct natural language searches and subject matter searches.[43] Through two full days of evidentiary hearings related to ESI disputes, eleven status hearings, and three Rule 16 conferences, the parties resolved numerous issues, including search methodology.[44] Judge Nolan observed that Sedona Principle 6 states that "[r]esponding parties are best situated to evaluate the procedures, methodologies, and techniques appropriate for preserving and producing their own electronically stored information."[45] Ultimately, plaintiffs withdrew their request for predictive coding as to the document requests at issue, and they agreed to confer about appropriate search methodology for newly collected documents.[46]

B. The Predictive Coding Process

Courts seem favorably disposed toward predictive coding, but what is it? As a machine-learning technology, predictive coding requires litigants to take steps to "teach" the computer what documents are responsive.[47] The first step is to define the universe of materials to be reviewed, based on date range, file type, custodian, domain, or other parameters, excluding ranges known to be non-responsive.[48] Privileged documents also can be removed at the outset to avoid inadvertent production.[49] The next step is to estimate the yield of responsive documents from the collection by selecting and manually reviewing a statistically valid random sample from the universe.50 If the yield of responsive documents is low, then the size of the control set may need to be adjusted.[51] The size of the control set is important because a poorly sized control set can affect the software's performance, and the results may not be reliable.[52]

With a control set in place, it is time to teach the computer. A training set of documents is selected, reviewed, and coded to train the software.[53] Generally, the documents in the training set are not chosen randomly; rather, responsive documents, together with some non-responsive documents, are selected for the training set using keywords or concept-based searches.[54] Based on the training set, the software creates a model that assigns a prediction score to each document based on degree of responsiveness.[55] The model is then tested, using the control set.[56] The software's predictions about responsiveness are compared to the coding decisions made by humans on the same set of documents.[57] If the desired performance metrics are not met, additional training sets are selected, trained, and tested.[58] The iterative process continues until the software's performance meets the desired metrics, at which point the software can be applied to the universe of documents for production.

The computer learns by applying a mathematical algorithm to a data set.[59] Two common types of algorithms are latent semantic indexing and naÏve Bayes algorithms.[60] Latent semantic indexing essentially uses decision trees to look for a keyword or phrase that distinguishes responsive and non-responsive documents, based on a pre-coded set.[61] NaÏve Bayes algorithms categorize documents based on word use, and in practice tend to work much like a keyword search.62 In *Da Silva Moore*, the software used both Support Vector Machines, a latent semantic algorithm, and Probabilistic Latent Semantic Analysis, a naı¨ve Bayes algorithm.[63]

III. Proportionality and Developments in THE FEDERAL SYSTEM

A. PROPORTIONALITY AND THE FEDERAL RULES: GETTING LAWYERS TO PLAY NICE WITH OTHERS

Rule 26(b)(2)(C)(iii) provides that "the court must limit the frequency or extent of discovery . . . if it determines that . . . the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues."[64] That is the proportionality standard.[65]

As discussed above, in *In re Biomet M2a Magnum Hip Implant Products Liability Litigation*, the court invoked Rule 26(b) in determining whether to order defendant Biomet to re-run its production, using predictive coding on the entire universe of documents.[66] The court noted that "it would cost Biomet a million, or millions, of dollars to test [plaintiffs'] theory that predictive coding would produce a significantly greater number of relevant documents."[67] The court concluded that, "[e]ven in light of the needs of the hundreds of plaintiffs in this case, the very large amount in controversy, the parties' resources, the importance of the issues at stake, and the importance of this discovery in resolving the issues," the likely benefits would not equal or outweigh the costs.[68]

Similarly, in ADT *Security Services, Inc. v. Pinnacle Security, LLC,*[69] Chief Judge Holderman invoked the proportionality principle to affirm a ruling on a motion to compel. Plaintiff sought to compel defendant to re-do an ESI search, asserting that documents seemed to be missing based on the disparity in the volume of documents produced by the parties.[70] Magistrate Judge Kim granted limited relief, ordering Pinnacle to re-do its search with respect to seven employees' computers for which there was evidence that they contained correspondence missing from the initial discovery production.[71] Chief Judge Holderman affirmed, noting that plaintiff 's broadly worded inquiries "violate[d] the principle that '[t]o further the application of the proportionality standard in discovery, requests for production of ESI and related responses should be reasonably targeted, clear, and as specific as practicable.'"[72]

Conversely, in *Chen-Oster v. Goldman, Sachs & Co.*,[73] the court applied Rule 26(b)(2)(C) to direct extensive further discovery. There, plaintiffs sought database information dating back as far as twelve years.[74] Defendant objected based on phasing, accessibility, and proportionality arguments, contending that it would take hundreds of hours to extract and check the requested data.[75] The court noted that either sampling or a mass production of all data contained in the databases would resolve the issue, although neither party had endorsed those options.[76] Concluding that the information sought was central to the case, the amount in controversy was substantial, that defendant's resources were ample, and the litigation was important, the court granted the motion to compel.[77]

B. A Principled Approach to Proportionality: The Seventh Circuit Electronic Discovery Pilot Program's Proposed Case Management Orders

In an effort to provide concrete guidance in implementing the proportionality principle, the Seventh Circuit Electronic Discovery Pilot Program has proposed two case management orders to address issues related to the collection and production of ESI.

The first proposal, a Discovery Plan for Electronically Stored Information (the "Discovery Plan"), is a "framework that may be used by parties in cases with either limited or extensive discovery" of ESI.[78] The Discovery Plan follows from Rule 26 and Principle 2.01 of the Seventh Circuit Electronic Discovery Pilot Program, and it addresses scope, searching, production format, and third-party ESI.[79] It makes clear that the "parties are aware of the importance the Court places on cooperation and commit to cooperate in good faith" and that nothing in the Discovery Plan "shall supersede the provisions of any subsequent Stipulated Protective Order."[80] The Discovery Plan limits the scope of document collection to an agreed time range and allows for additional limitations, for example based on geographic or organizational factors.[81] The

Discovery Plan also distinguishes between cases with limited and extensive ESI.[82] For cases with extensive ESI, the Discovery Plan: (1) addresses the types of e-mail and unstructured data (like word documents or spreadsheets), custodians, and shared data to be produced, and provides for a search protocol that sets out whether technologyassisted review, like predictive coding, will be used;[83] (2) deals with the format for production materials, including metadata, load files, and de-duplication, as well as whether documents without standard pagination, like spreadsheets, will be produced in native, single page TIFs, or hard copy;[84] and (3) calls on each party to identify a knowledgeable e-discovery liaison.[85]

The second proposed order addresses privilege and work-product issues.[86] The proposed order requires a party that withholds ESI based on privilege to provide a spreadsheet listing the ESI withheld, with as much metadata as is reasonably available and a description of the categories of ESI being withheld.[87] The proposed order then provides a process for challenging privilege designations.[88] Finally, the proposed order includes a non-waiver and clawback protocol, specifying that production, whether inadvertent or intentional, does not waive any privilege.[89] The proposed order allows a producing party to assert the privilege "at any time," although affirmative use of a produced document waives any privilege as to that document.[90]

IV. Conclusion

With electronic discovery now a fact of everyday life, courts are moving to address the resulting cost and burden by more aggressively using the proportionality principle, even as technological advances like predictive coding increase the range of cost-saving measures available to attorneys. Both of those developments seem likely to continue in the future.

Endnotes:

[*] Mr. Chorvat is a partner and Ms. Pelanek is litigation counsel in the Chicago office of Jenner & Block LLP.

[1] For summaries of prior developments, see Timothy J. Chorvat & Laura E. Pelanek, *Electronically Stored Information in Litigation,* 68 BUS. LAW. 245 (2012) (surveying developments in 2011–2012, focusing on discovery of social media and cloud data); Timothy J. Chovat & Laura E. Pelanek, *Electronically Stored Information in Litigation*, 67 BUS. LAW. 285 (2011) (surveying developments in 2010–2011, focusing on state courts); Timothy J. Chorvat & Laura E. Pelanek, *Electronically Stored Information in Litigation*, 66 BUS. LAW. 183 (2010) (surveying developments in 2009–2010, focusing on federal courts).

[2] 287 F.R.D. 182 (S.D.N.Y. 2012).

[3] Joe Palazzolo, *Why Hire a Lawyer? Computers Are Cheaper*, WALL ST. J. (June 18, 2012), http://online.wsj.com/article/SB10001424052702303379204577472633591769336.html.

[4] *Da Silva Moore*, 287 F.R.D. at 183, 186–88.

[5] *Id.* at 184 n.2.

[6] *Id.* at 187 & n.6 (noting an objection by plaintiffs).

[7] *Id.* at 188 (citation omitted).

[8] *Id.*

[9] *Id.* at 189.

[10] *Id.*

[11] *Id.* at 193.

[12] *Id.*

[13] No. CL 61040, 2012 Va. Cir. LEXIS 50, at *2 (Apr. 23, 2012).

[14] *Id.* at *1–2.

[15] *Id.* at *2.

[16] 877 F. Supp. 2d 87 (S.D.N.Y. 2012).

[17] *Id.* at 93–94.

[18] *Id.* at 94.

[19] *Id.* at 106.

[20] *Id.* at 107.

[21] *Id.* at 111.

[22] *Id.* at 109.

[23] *Id.* at 111.

[24] Transcript of Oral Argument, EORHB Inc. v. HOA Holdings LLC, No. 7409-VCL (Del. Ch. Oct. 19, 2012) (order partially granting summary judgment for defendants and denying motion to dismiss counterclaims).

[25] *Id.* at 66.

[26] *Id.*

[27] No. 08cv1992 AJB (MDD), 2013 U.S. Dist. LEXIS 14105 (S.D. Cal. Feb. 1, 2013).

[28] *Id.* at *6–7.

[29] *Id.* at *46.

[30] *Id.* at *32.

[31] *Id.* at *34–35 (citation and quotations omitted).

[32] *Id.* at *35; *see also In re Actos* (Pioglitazone) Prods. Liab. Litig., No. 6:11-md-2299, 2012 U.S. Dist. LEXIS 187519, at *21, *25–26 (W.D. La. July 27, 2012) (approving protocol calling for

THE BUSINESS LAWYER
Business Law Section • American Bar Association

Volume 69, November 2013

predictive coding and specifying procedures).

[33] No. 3:12-MD-2391 (N.D. Ind. Apr. 18, 2013) (order regarding discovery of ESI).

[34] Id. at 2.

[35] Id. at 2–3.

[36] Id. at 3.

[37] Id. at 3–4.

[38] Id. at 4.

[39] Id.

[40] Id. at 5–6.

[41] No. 10 C 5711, 2012 U.S. Dist. LEXIS 139632 (N.D. Ill. Sept. 28, 2012).

[42] Id. at *14 (citation and quotations omitted).

[43] Id. at *14–15.

[44] Id. at *16–17.

[45] Id. at *18 (citation and quotations omitted).

[46] Id. at *19–20, *62–64.

[47] See, e.g., MATTHEW D. NELSON, PREDICTIVE CODING FOR DUMMIES 7–8 (2012).

[48] Id. at 15.

[49] Id.

[50] Id. at 16.

[51] Id.

[52] Id.

[53] Id. at 17.

[54] Id.

[55] Id. at 18.

[56] Id.

[57] Id.

[58] Id.

[59] Tim Leehealey, The Machine Learning/Predictive Coding Silver Bullet, EDISCOVERY INSIGHT (Sept. 24, 2012), http://ediscoveryinsight.com/2012/09/the-machine-learningpredictive-coding-silver-bullet.

[60] Id.

[61] Id.

[62] Id.

[63] James Hanft, Technology: Shedding Light on the Predictive Coding Black Box, INSIDE COUNS. (Mar. 16, 2012), http://www.insidecounsel.com/2012/03/16/technology-shedding-light-on-the-predictivecoding.

[64] FED. R. CIV. P. 26(b)(2)(C)(iii).

[65] The Standing Committee on Rules of Practice and Procedure of the United States Courts is considering amendments to the federal discovery rules, including Rules 26 and 37. Henry Kelston, Are We on the Cusp of Major Changes to E-Discovery Rules?, LAW TECH. NEWS (Apr. 17, 2013), http://goo.gl/CYpHf3. The revisions would modify Rule 26(b)(1) to restrict the scope of discovery so that it is "proportional to the needs of the case," based on the factors set out in Rule 26(b)(2)(C). Id.; see also CCL Addresses Federal Civil Rules Advisory Committee, CTR. FOR CONST. LITIG. (Apr. 17, 2013), http://www.cclfirm.com/blog/8333/ (charting proposed changes to the rules under consideration).

[66] In re Biomet M2a Magnum Hip Implant Prods. Liab. Litig., No. 3:12-MD-2391, slip op. at 3–4 (N.D. Ind. Apr. 18, 2013) (order regarding discovery of ESI).

[67] Id. at 6.

[68] Id.

[69] No. 10 C 7467, 2012 U.S. Dist. LEXIS 98948 (N.D. Ill. July 11, 2012).

[70] Id. at *5.

[71] Id. at *6.

[72] Id. at *7 (quoting SEVENTH CIR. ELEC. DISCOVERY PILOT PROGRAM COMM., PHASE ONE: OCTOBER 1,2009–MAY 1, 2010: STATEMENT OF PURPOSE AND PREPARATION OF PRINCIPLES 11 (Oct. 1, 2009) (setting forth Principle 1.03 of Principles Relating to the Discovery of Electronically Stored Information to be implemented and evaluated during the Phase One period from October 1, 2009 through May 1, 2010)).

[73] 285 F.R.D. 294 (S.D.N.Y. 2012).

[74] Id. at 297.

[75] Id. at 303–04.

[76] Id. at 304–05.

[77] Id. at 305–06, 308.

[78] SEVENTH CIR. ELEC. DISCOVERY PILOT PROGRAM COMM., INTERIM REPORT ON PHASE THREE: MAY 2012–MAY 2013 app. A, at 54 (2013), available at http://www.discoverypilot.com/sites/default/files/phase_three_interim_report.pdf.

[79] Id. at 55–61.

[80] Id. at 55.

[81] Id. at 55–56.

[82] Id. at 56.

[83] Id. at 56–57.

[84] Id. at 58–60.

[85] Id. at 61.

[86] Id. app. B, at 68.

[87] Id.

[88] Id. at 69.

[89] Id.

[90] Id.