Technology Assisted Review:
The Disclosure of Training Sets and Related Transparency Issues

Whitney Street, Esq.[1]

The potential cost savings and increase in accuracy afforded by technology assisted review ("TAR") have been widely discussed and well documented in recent years.[2] As a result, parties in complex litigation increasingly rely on TAR to assist with their efforts in responding to document requests. TAR does not eliminate the role of the human reviewer. Rather, the human reviewer trains the document review system to distinguish between responsive and non-responsive documents, thus reducing the number of documents that must be manually reviewed and coded by the human reviewer.

Although there are a variety of methods that may be used to train the system, one common approach is for a human reviewer to code a subset of documents for responsiveness. That subset of documents, commonly referred to as the "training set" or "seed set," is then put into the system.[3] Based on what it learns from the training set, the system identifies the likely-relevant documents, which are then reviewed and manually coded by a human reviewer. Any documents deemed not relevant by the system will not be included in the manual review process and will not be considered for production. Hence, the selection, composition and coding of the training set is of critical importance, as any errors or omissions will be magnified and extended to the entire production. The requesting party therefore has a compelling interest in achieving an acceptable level of transparency into this process. Although the courts that have weighed in on the use of TAR have generally favored transparency, there is currently much debate specifically with respect to the discoverability of training sets. The following paper, which examines this debate, proceeds in three parts. Part I explains the creation and role of training sets. Part II provides an overview of the relevant case law. Part III examines the arguments made for and against the disclosure of training sets.

---

[1] Whitney Street is a partner of the law firm of Block & Leviton LLP where she focuses her practice on antitrust and securities class action litigation. Ms. Street can be contacted at whitney@blockesq.com or (617) 398-5600 with questions regarding this article.

[2] *See* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, Vol. XVII, Rich. J. L. & Tech. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.

[3] The term "seed set" may be used to refer to a training set or more narrowly to documents selected through judgmental sampling (discussed below). Maura R. Grossman and Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 Fed. Courts L. Rev. 1 (2013) (the "TAR Glossary"). The broader term "training set" is therefore used throughout unless a court uses the term "seed set" in an opinion discussed herein. Similarly, the term "predictive coding" is sometimes used instead of "technology assisted review." According to the TAR Glossary, "predictive coding" is "[a]n industry-specific term generally used to describe a Technology-Assisted Review process . . . ." *Id.* In other words, "predictive coding is a species of the genus TAR." Paul Burns and Mindy Morton, *Technology-Assisted Review: The Judicial Pioneers*, The Sedona Conference (2014), *available at* http://www.americanbar.org/content/dam/aba/administrative/litigation/materials/2014_sac/2014_sac/technology_assisted_review_the_judicial_pioneers.authcheckdam.pdf. Hence, the broader term TAR is used throughout except where a court uses the term "predictive coding" in an opinion discussed herein.

Part I. The Role of Training Sets in TAR

The term "training set" is broadly used to describe the documents that are used to train the system to find the likely-relevant documents in the universe of documents. The creation of training sets and the role they play in differing TAR methodologies are discussed below.

Creation of Training Sets and a Note about Key Words

One approach for the creation of a training set is to randomly select the documents from the available universe and to code them for responsiveness. In contrast, judgmental selection relies upon the judgment of individuals to find a suitable set of documents to train the system, often through key word searches. Or, the protocol may call for some combination of random selection and judgmental selection.[4]

Hence, key words may play a central role, a supporting role or no role at all within the TAR methodology. Where key words are employed, parties frequently debate whether the producing party should disclose information regarding such search terms. This may include not only a list of the search terms, but also "hit reports," which provide insight into whether terms are overbroad, too narrow or missing. The debate regarding the discoverability of information related to key words relates to, but also pre-dates, the use of TAR. Although the focus of this paper is on the discoverability of training sets, to the extent courts also address key words, that information is noted in the case summaries, which comprise Part II of this paper.

Common TAR Methodologies

With a Simple Passive Learning ("SPL") algorithm, the training process starts with the creation of a training set, which is then fed into the system. The training process is typically iterative and stops once the system stabilizes, *i.e.,* reaches a point where the algorithm's coding of the documents is sufficiently consistent with how the (human) reviewer would have coded the document. The next step is a separate review phase, during which documents deemed likely relevant by the algorithm are manually reviewed and considered for production.

Simple Active Learning ("SAL") similarly begins with a training set, but relies upon uncertainty sampling to select the subsequent training documents. With uncertainty sampling, the algorithm selects the documents about which it is least certain for further review/coding. The newly coded documents are added to the training set and the process is repeated "until the benefit of adding more training documents to the training set would be outweighed by the cost of reviewing and

[4] Certain commentators believe that random selection results in a training set that is more representative of the data set. *See* Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review,* 2013 Fed. Cts. L. Rev. 7 (2013). In contrast, other studies have suggested that omitting judgmental selection decreases the quality and increases the cost of the review. *See* Maura R. Grossman & Gordon V. Cormack, *Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review,"* 2014 Fed. Cts. L. Rev 1 (2014) (hereinafter, "Grossman-Cormack Comments on Schieneman-Gricks Article").

coding them."[5] After reaching this point, the algorithm makes a final determination about the likely relevant documents; those documents are then manually reviewed and coded.

In contrast, Continuous Active Learning ("CAL"), sometimes referred to as TAR 2.0, is based on a ranking and continuous review system. More specifically, CAL often begins by running key word searches.[6] Based on these key words, the algorithm ranks the documents from most to least likely relevant. An individual then reviews the top ranked documents and codes them for relevance. As the human reviewer codes documents, the system recalibrates its rankings based on this feedback. The process continues until "the number of top-ranked documents containing responsive information drops precipitously."[7]

Given the variations between the available methodologies, the specific transparency issues that may arise will differ. This includes variations in the ways in which errors may be introduced into the process and in the feasibility of providing transparency to ameliorate those risks. For example, the disclosure of iterative training sets, which may reveal discrepancies between parties' assessments of new information not known at the outset of the review, would potentially be more straightforward when using SPL as compared to when a party employs an active learning algorithm. Due to the relatively recent embrace of TAR by parties and by courts, the treatment of these specific issues is still developing. However, a handful of opinions have addressed or discussed the disclosure of training sets more generally. A summary of those cases follows.

Part II. Case Law Regarding the Disclosure of Training Sets and Related Issues

As discussed below, the trend appears to be toward the disclosure of training sets to the requesting party. Additional issues of interest addressed by courts in recent decisions include: when certain parameters, such as the relevance score cut-off, should be decided upon; practical methods for dealing with sensitive or confidential business information; and whether a party may unilaterally change its ESI protocol.

Cases Entailing Relatively High Transparency

In the first case to address the acceptability of TAR, *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012) (M.J. Peck), the producing party agreed to provide a high degree of transparency into its TAR process. This included the disclosure of the initial set of documents used to train the system, as well as the documents that were subsequently coded in the iterative training rounds.[8] The court noted that the defendants' transparency with respect to the training

[5] Maura R. Grossman and Gordon V. Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, Proceedings of the 37th International ACM SIGIR Conference on Research and Development in information retrieval (SIGIR'14) (July 2014).

[6] Although the CAL methodology may entail the use of key words to create a training set, some CAL tools allow the reviewer to simply begin coding without employing key word searches. Kroll Ontrack is an example of a CAL tool that does not rely on key word searches to start the protocol.

[7] *See* Grossman-Cormack Comments on Schieneman-Gricks Article, *supra n.* 4, at 290.

[8] *Id.* at 187.

sets "allows the opposing counsel (and the Court) to be more comfortable with computer-assisted review, reducing fears about the so-called 'black box' of the technology."[9] However, the court did not go so far as to say that training sets must always be produced to the requesting party. *Id.* at 192 ("This Court highly recommends that counsel in future cases be willing to at least discuss, if not agree to, such transparency in the computer-assisted review process.").

The court also addressed two related issues: predetermination of relevance scores (*i.e.*, the cut-off point for relevance) and the error level to be applied at the quality control ("QC") phase. With respect to the first issue, the court found that the matter was premature. The court noted that "[t]he issue regarding relevance standards might be significant if [the defendant's] proposal was not totally transparent. Here, however, plaintiff will see how [the defendant] has coded every email used in the seed set (both relevant and not relevant), and the Court is available to quickly resolve any disputes."[10] Similarly, the court did not require the defendants to disclose how many errors (*i.e.,* relevant documents incorrectly coded as irrelevant) would be tolerated in the final quality control random sample. The court explained that, "[i]n order to determine proportionality, it is necessary to have more information than the parties (or the Court) now has, including how many relevant documents will be produced and at what cost to [the defendant]."[11] Factors noted by the court as relevant to the proportionality consideration included whether the case would become a collective or class action and whether any of the incorrectly coded documents in the final QC set would prove to be "hot" or "smoking gun" documents.

In *Global Aerospace Inc. v. Landow Aviation*, No. CL 61040, 2012 Va. Cir. LEXIS 50 (Va. Cir. Ct. April 23, 2012), the court went a step further, approving the defendants' proposed use of TAR over the plaintiffs' objection. As in *Da Silva Moore*, the defendants provided the plaintiffs (and the court) with a high degree of transparency with respect to their TAR protocol. This included the disclosure of the complete set of coded, non-privileged training documents (relevant and non-relevant), minus irrelevant documents containing confidential and/or sensitive information.[12] A log describing the withheld documents was also prepared and produced to the requesting party, allowing opposing counsel to evaluate and raise any objections. A similar process was followed with respect to the QC phase.

The producing party in the *In re Actos (Pioglitazone) Products Liability Litig*., MDL No. 6:11-md-2299, 2012 U.S. Dist. LEXIS 187519 (W.D. La. July 27, 2012) (J. Doherty) agreed to an even higher level of transparency, indeed including the opposing party in the training process. More specifically, each party proposed three attorneys to act as "experts," who then worked cooperatively to train the system. In order to address confidentiality issues, plaintiffs' experts signed a nondisclosure and confidentiality agreement providing that if "exposed to information that would be subject to withholding or redaction under the Protective Order in this matter,

[9] *Id.* at 192.

[10] *Id.* at 189.

[11] *Id.*

[12] Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding, *Global Aerospace Inc. v. Landow Aviation,* No. CL61040, 2012 WL 1419842 (Va. Cir. Ct. April 9, 2012).

Plaintiffs' experts agree not to disclose such information to co-counsel, client, any Party, or any third party without obtaining prior written consent of the other Party regarding the particular piece of information sought to be disclosed."[13]

Although the issue of the production of training sets was raised by the plaintiffs in *Hinterberger v. Catholic Health System*, No. 08-CV-380S, 2013 U.S. Dist. LEXIS 73141 (W.D.N.Y., May 21, 2013) and the companion case *Gordon v. Kaleida Health*, No. 08-CV-378S, 2013 U.S. Dist. LEXIS 73330 (W.D.N.Y. May 21, 2013), the issue was not expressly decided by Magistrate Judge Foschio in his opinions. There, the plaintiffs moved to compel the defendants "to meet and confer with respect to establishing an agreed protocol for implementing the use of predictive coding software; alternatively, Plaintiffs request the court to adopt and impose such protocol."[14] Defendants argued that the motion was premature[15] and that *Da Silva Moore* did not *require* that a producing party disclose the training set to opposing counsel. The court agreed with the defendants that under the circumstances, the plaintiffs' motion was premature. With respect to the training sets, the court simply noted that plaintiffs did not contest that *Da Silva Moore* does not *require* parties to meet and confer regarding the producing parties' selection of the seed set and further noted that, "[b]ased on Defendants' expressed awareness of Defendants' discovery obligations. . . the court also need not, as Plaintiffs' request, remind Defendants of relevant considerations regarding Defendants' use of predictive coding regarding ESI document production obligations under Fed. R. Civ. P. 34(a)."[16]

The Need For Transparency When Switching Horses Midstream[17]

Although not directly relevant to the issue of the discoverability of training sets, two recent opinions discuss the need for transparency in the context of a producing party's decision to switch to TAR after previously agreeing to a different method of document review. In *Progressive Casualty Insurance Co. v. Delaney*, No. 2:11-cv-00678, 2014 U.S. Dist. LEXIS 69166 (D. Nev. May 20, 2014) (J. Leen), the plaintiff, Progressive Casualty Insurance Company ("Progressive") initially proposed to apply key word searches to the universe of documents and to manually review and code the culled documents. Application of the agreed-to search terms reduced the number of documents subject to manual review from 1.8 million to 565,000. After

[13] *Id.* at *23.

[14] *Id.* at *1.

[15] It appears that, prior to the filing of the motion, the parties' discussions regarding the TAR process had stalled due to defendants' objection to plaintiffs' ESI vendor, who had previously performed services for defendants. *See Hinterberger*, 2013 U.S. Dist. LEXIS 73141 at *8-10.

[16] *Id.* at *10.

[17] Notably, both cases discussed in this subsection, *Progressive Casualty Insurance Co. v. Delaney*, No. 2:11-cv-00678, 2014 U.S. Dist. LEXIS 69166 (D. Nev. May 20, 2014) (J. Leen) and *Bridgestone Americas, Inc. v. International Business Machines Corp.*, No. 3:13-1196 (M.D. Tenn. July 22, 2014) (M.J. Brown), raise the issue of whether it is appropriate to apply TAR after reducing the universe of documents through key word searching. This is currently a hotly contested issue that will benefit from further examination by the courts.

manually reviewing approximately 125,000 documents, Progressive determined that the process was too time consuming and too costly. It therefore abandoned the manual review and employed predictive coding on the remaining 440,000 documents.

Defendants took issue with (i) the unilateral change to the court-approved ESI protocol, (ii) the decision to employ the predictive coding on a subset of the documents rather than on the 1.8 million and (iii) the lack of transparency regarding the new ESI protocol. The court agreed with defendants and ordered Progressive to produce all 565,000 documents that were generated by the search terms, without further review for responsiveness (although the court did permit the plaintiff to apply a filter for privilege). In so ordering, the court suggested that it would require a substantial amount of transparency before allowing a party to use TAR, including, potentially, the disclosure of the training set:

> The cases which have approved technology assisted review of ESI have required an unprecedented degree of transparency and cooperation among counsel in the review and production of ESI responsive to discovery requests. . . . *In the handful of cases that have approved technology assisted review of ESI, the courts have required the producing party to provide the requesting party with full disclosure about the technology used, the process, and the methodology, including the documents used to "train" the computer*.[18]

As in the *Progressive* case, the plaintiff in *Bridgestone Americas, Inc. v. International Business Machines Corp.,* No. 3:13-1196 (M.D. Tenn. July 22, 2014) (M.J. Brown) proposed switching to predictive coding after agreeing to search terms and manual review. Defendants objected to the change and in particular to the decision to employ predictive coding to the subset of documents generated through the use of search terms. The court permitted the plaintiff to proceed with employing predictive coding on the subset of documents, but stressed the importance of transparency, including the disclosure of the training set, in its decision: "[O]penness and transparency in what Plaintiff is doing will be of critical importance. Plaintiff has advised that they will provide the seed documents they are initially using to set up predictive coding. The Magistrate Judge expects full openness in this matter."[19]

Cases Not Requiring Disclosure of Training Sets

In contrast to the foregoing cases, the court in *In re: Biomet M2a Hip Implant Products Liability Litig.*, No. 3:12-MD-2391, 2013 U.S. Dist. LEXIS 172570 (N.D. Ind. Aug. 21, 2013) (J. Miller) held that the plaintiffs' request for the seed set "reaches well beyond the scope of any permissible discovery by seeking irrelevant or privileged documents used to tell the algorithm what not to find."[20] The court further stated its view that it did not possess the "authority to compel discovery of information not made discoverable by the Federal Rules."[21] However, the court

[18] *Id.* at *28-29 (emphasis added).

[19] *Bridgestone Americas, Inc. v. International Business Machines Corp.,* No. 3:13-1196 (M.D. Tenn. July 22, 2014) (M.J. Brown), ECF No. 89, p. 2.

[20] *Id.* at *3.

[21] *Id.* at *5.

recognized that the "[Plaintiffs'] Steering Committee is right that Biomet's cooperation falls below what the Sedona Conference endorses. An unexplained lack of cooperation in discovery can lead a court to question why the uncooperative party is hiding something and such questions can affect the exercise of discretion."[22]

The court similarly declined to require the producing party to provide the training sets at the outset of the document collection process in *In re Drywall Antitrust Litig.*, 13-md-2437 (E.D. Pa.), although the court's order did not detail the reasoning behind its decision.[23]

Part III. Analysis of the Arguments For and Against the Disclosure of Training Sets

Although the trend appears to be toward the disclosure of training sets, no court has decisively ruled that the producing party must produce this information and indeed one court has ruled that training sets are not discoverable under the federal rules. The following section examines arguments for and against the disclosure of training sets.

Attorney Work Product.

In *Biomet*, defendants argued that they should not be required to produce the training sets in part because "the process for identifying relevant documents is protected as work-product."[24] Plaintiffs countered that "Biomet should not be heard to argue that the [plaintiffs'] request encroaches on attorney work product. Identification of the documents used to train the predictive coding algorithm is no different from identification of the keywords or search terms used by Biomet to filter/cull the corpus of documents it collected before it applied predictive coding."[25] In denying plaintiffs' request for the training sets, the court discussed the attorney-client privilege (among other arguments), but did not expressly address the attorney work product doctrine.[26]

---

[22] *Id.* at *5-6.

[23] *See In re Drywall Antitrust Litig.,* No. 13-md-2437 (E.D. Pa. Nov. 26, 2013), ECF No. 88 ("Counsel presented argument concerning defendants' protocols for the production of electronic discovery. The Court appreciates the constructive dialogue among all parties, and directs all parties to maintain documents used to determine the collection and production of electronically stored information and other materials, but declines to require defendants to take any additional steps, pending production on January 15, 2014.") The author has been appointed co-lead counsel on behalf of the indirect purchaser plaintiffs in the *Drywall* matter.

[24] 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), ECF No. 722, p. 3.

[25] Plaintiffs' Motion for Relief from Defendants' Refusal to Disclose Relevant Documents Used in Predictive Coding, *In re Biomet M2a Magnum Hip Implant Prods. Liability Litig*., No. 3:12-MD-2391 (N.D. Ind. Aug. 5, 2013), ECF No. 723, p. 4.

[26] The defendants reasoned that plaintiffs were requesting privileged and irrelevant documents and information as to how "Biomet used certain documents before disclosing them," which the court ruled as beyond the scope of Rule 26(b)(1). *Biomet,* 2013 U.S. Dist. LEXIS 172570 at *3-*4.

As suggested by the requesting party in *Biomet,* it appears to be generally accepted that key words are not protected attorney work product. *See Apple, Inc. v. Samsung Electronics Co. Ltd*., No. 12-CV-0630, 2013 U.S. Dist. LEXIS 67085, *40 (N.D. Cal. May 9, 2013) ("During their meetings, Google maintained that its search terms and choice of custodians were privileged under the work-product immunity doctrine, an argument it has abandoned no doubt in part because case law suggests otherwise.") *See also William A. Gross Const. Assoc., Inc. v. American Manufacturers Mutual Insurance Co.*, 256 F.R.D. 134 (S.D.N.Y. 2009) ("This Opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or 'keywords' to be used to produce emails or other electronically stored information.").[27] Thus, the question is whether key words and training sets are substantially similar enough to warrant similar treatment.

In an article earlier this year, authors Richard H. Lowe, James G. Welch and Kimberly G. Lippman provided a cogent argument as to why training sets should be viewed differently than key words:

> In contrast [to producing key words], the producing party would argue, tagging documents for relevance to develop the seed set may involve greater complexity; actively culling through a seed set to determine which specific documents are relevant (and which are not) arguably demands more of an application of the attorney's mental impressions of the claims than coming up with search terms for documents not yet reviewed.[28]

However, there is arguably a distinction between asking the producing party to reveal the documents it believes to be *responsive* to the document requests as opposed to what documents it believes are *relevant* to the case. Moreover, courts have recognized that the set of facts underlying whether a document is responsive does not implicate the thought processes of opposing counsel.[29]

For example, in *Romero v. Allstate Ins. Co.*, 271 F.R.D. 96, 110 (E.D. Pa. 2010), the court found that it was reasonable to require the parties to meet and confer regarding "any other essential details about the search methodology they intend to implement for the production of electronically-stored information." The court reasoned that such information was not protected by attorney work product because "it goes to the underlying facts of what documents are

---

[27] *See also FormFactor, Inc. v. Micro-Probe, Inc.*, No. C-10-03095, 2012 U.S. Dist. LEXIS 62233, *7, n. 4 (N.D. Cal. May 3, 2012) (listing cases in which search terms were deemed not to be work product).

[28] Richard H. Lowe, James G. Welch and Kimberly G. Lippman, *Disclosure of Seed Sets: Required to Cooperate or Protected as Attorney Work Product?,* THE LEGAL INTELLIGENCER, Feb. 18, 2014.

[29] *See e.g., FormFactor, Inc.*, 2012 U.S. Dist. LEXIS 62233 at *19-21 ("To the extent Plaintiff argues that disclosure of search terms would reveal privileged information, the Court rejects that argument. ***Such information is not subject to any work product protection because it goes to the underlying facts of what documents are responsive to Defendants' document request, rather than the thought processes of Plaintiff's counsel***. . . . There is simply no way to determine whether Plaintiff did an adequate search without production of the search terms used.") (emphasis added).

responsive to Plaintiffs' document requests and does not delve into the thought processes of Defendants' counsel."[30] Although the case related to the use of search terms as opposed to TAR, the basic rationale (*i.e.,* facts regarding responsiveness are not work product) would seem to apply equally to the disclosure of training sets.

As noted, to date it does not appear that any court has squarely addressed whether training sets are protected attorney work product. Thus, this is an area that will benefit from further guidance from the courts.

Disclosure of Non-Responsive Documents.

As discussed in Part II, *supra,* it is not uncommon for a producing party to agree to disclose all non-privileged documents used to train the algorithm, including documents deemed non-responsive as well as documents deemed responsive.[31] However, the court in *Biomet* held that the requesting party had no right to the production of irrelevant documents and that the court lacked the "authority to compel discovery of information not made discoverable by the Federal Rules."[32] In so holding, the court focused on the language of Fed. R. Civ. P. 26(b)(1) and presumably on the word "relevant" as used therein.[33]

However, one could argue that the court focused on information that is *substantively* relevant to the exclusion of information that is *procedurally* relevant, the latter of which is also regularly produced in litigation. For example, it is routine for a party in complex litigation to produce its document retention policy. While the substance of that document is not directly relevant to the matters at issue in the litigation (*e.g.*, it does not bear on whether the defendant conspired to fix prices or whether it breached its contract), it is nonetheless relevant to the document collection and production process, and therefore within the scope of information to be produced pursuant to the Federal Rules of Civil Procedure.

In addition, courts routinely engage in a balancing of interests, as well as a balancing of costs and benefits, in litigation. The concern that errors in the training sets will be extended across the entire production resulting in a material number of responsive documents being left behind arguably outweighs the producing party's interest in withholding a small set of irrelevant documents. Authors William P. Butterfield, Conor R. Crowley & Jeannine Kenney correctly observed that there may be instances where the nature of the litigation renders the non-

[30] *See also In re Porsche Cars N. Am., Inc. Plastic Coolant Tubes Products Liability Litig.*, No. 2:11-md-2233, 2012 U.S. Dist. LEXIS 136954, *26-27 (S.D. Ohio Sept. 25, 2012) (Finding that "the facts underlying the way in which [the producing party] identified and produced responsive documents" were not protected attorney work product.)

[31] *See e.g., Da Silva Moore v. Publicis Groupe,* 287 F.R.D. 182 (S.D.N.Y. 2012), *Global Aerospace Inc. v. Landow Aviation,* 2012 Va. Cir. LEXIS 50 (Va. Cir. Ct. April 23, 2012) and *In re Actos (Pioglitazone) Products Liability Litig.*, 2012 U.S. Dist. LEXIS 187519 (W.D. La. July 27, 2012).

[32] *Biomet,* 2013 U.S. Dist. LEXIS 172570 at *5.

[33] *Id.* at *4.

responsive documents particularly sensitive.[34] In that event, the parties may benefit from adopting a protocol similar to that relied upon in *In re Actos* and *Global Aerospace* (discussed *supra*), pursuant to which sensitive and/or confidential information in the training set was withheld (or redacted) and logged.

Despite the foregoing arguments in favor of producing the (non-privileged) non-responsive documents used in the training set, no court has expressly ordered the production of this information over a producing party's objection. Hence, it remains to be seen whether courts will agree with the rationale of *Biomet* or adopt a different approach.

Transparency With Respect To TAR As Compared To Transparency With Respect To Manual Review.

As a number of commentators have noted, the level of transparency expected of a party employing technology assisted review is arguably greater than what would be asked if the producing party were to employ a linear review. *See e.g.,* Jeane A. Thomas & David D. Cross, *Predictive Coding: How Much Transparency and Cooperation Is Required When Using Technology Assisted Review In Litigation?*, CROWELL & MORING'S DATA LAW INSIGHTS (January 31, 2013).[35] As Thomas and Cross observed:

> The concern among some TAR advocates is that these practices exceed what is required under the Federal Rules of Civil Procedure and that, if these levels of transparency come to represent the minimum legal threshold of cooperation for using TAR, producing parties will be dissuaded from using TAR as a result of the added costs and litigation risks. *Id.*

On the other hand, as the court reasoned in *Da Silva Moore,* "transparency allows the opposing counsel (and the Court) to be more comfortable with computer-assisted review, reducing fears about the so-called 'black box' of the technology" and the risk of "garbage in, garbage out."[36] If the trend toward the disclosure of training sets is ratified by the courts, it remains to be seen whether this will affect parties' willingness to employ TAR as opposed to other methodologies.

---

[34] William P. Butterfield, Conor R. Crowley & Jeannine Kenney, *Reality Bites: Why TAR's Promises Have Yet to be Fulfilled*, (DESI V: Workshop on Standards for Using Predictive Coding, Machine Learning and Other Advance Search and Review Methods), *available at* http://www.umiacs.umd.edu/~oard/desi5/additional/Butterfield.pdf.

[35] *Available at* http://www.crowelldatalaw.com/2013/01/predictive-coding-how-much-transparency-and-cooperation-is-required-when-using-technology-assisted-review-in-litigation/

[36] *Da Silva Moore*, 287 F.R.D. at 192 and n. 4.