

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

Whitney Street, Esq.<sup>1</sup>

---

The potential cost savings and increase in accuracy afforded by technology assisted review (“TAR”) have, in recent years, been widely discussed and well documented.<sup>2</sup> As courts increasingly approve (and, indeed, *sua sponte* order) the use of TAR,<sup>3</sup> it is imperative that litigators have a working understanding of the technology available and a thorough understanding of how that technology is being deployed in a given case. Getting the TAR process right from its inception and conducting quality control throughout the life of the project is critical, as any error in culling or coding will be magnified and extended to the entire production.

As one would expect, this has led to a spate of decisions in recent years as to the level of transparency that a producing party should provide regarding its TAR process. A particularly hot issue is whether a producing party should be required to disclose the documents that it used to train the algorithm. The following paper, which addresses this and related issues, proceeds in three parts. Part I explains the role of training sets across differing TAR methodologies. Part II provides a brief overview of the key cases that have discussed the disclosure of training sets. Part III examines the arguments made for and against the disclosure of training sets.<sup>4</sup>

**Part I. The Role of Training Sets in TAR**

Three of the most common TAR methodologies used in litigation are Simple Passive Learning, Simple Active Learning and Continuous Active Learning. Due to differences between these

---

<sup>1</sup> 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 any questions regarding this article.

<sup>2</sup> 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>.

<sup>3</sup> *EOHRB, Inc. v. HOA Holdings LLC*, No. 7409-VCL (Del. Ch. Oct. 19, 2012), transcript available at [http://ralphlosey.files.wordpress.com/2012/10/predictive\\_coding\\_order\\_sua\\_sponte\\_delaware.pdf](http://ralphlosey.files.wordpress.com/2012/10/predictive_coding_order_sua_sponte_delaware.pdf)

<sup>4</sup> The term “predictive coding” is sometime used instead of “technology assisted review.” According to a glossary of TAR terms prepared by Grossman and Cormack, “predictive coding” is “[a]n industry-specific term generally used to describe a Technology-Assisted Review process . . . .” 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*”). In other words “predictive coding is a species of the genus TAR.” See Paul Burns and Mindy Morton, *Technology-Assisted Review: The Judicial Pioneers*, The Sedona Conference (2014). Hence, the broader term TAR is used throughout except where the court uses the term “predictive coding” in its opinion. Similarly, the term “seed set” may be used to refer to a training set or more narrowly to documents selected through Judgment Sampling. See *TAR Glossary*. Hence, the broader term “training set” is used throughout unless the court uses the term “seed set” in its opinion.

methodologies (and in how any one method may be deployed) the particular transparency issues that may arise vary. A brief overview of the three methodologies and the corresponding role of training sets are described below.

### Simple Passive Learning

Simple Passive Learning (“SPL”) entails a training phase followed by a separate review phase. During the training phase, a set of documents is selected and coded for responsiveness and then put into the algorithm. Based on what it “learns” from the training set, the algorithm identifies the documents that are likely to be responsive, either by creating a subset of likely relevant documents or by assigning each document a relevance score.<sup>5</sup> A sample of the coded documents is then assessed to evaluate the accuracy of the algorithm’s coding decisions. Corrections are made as needed and the process repeated until the system stabilizes, *i.e.*, reaches a point where the algorithm’s coding of the documents is sufficiently consistent with how the (human) reviewer or reviewers would have coded the document. Once the system reaches stability, the documents deemed likely relevant by the algorithm are manually reviewed and coded for production. This is referred to as the review phase.

### Simple Active Learning

As with SPL, Simple Active Learning (“SAL”) begins with the creation of a training set that is used to teach the algorithm. The subsequent training documents are selected by the algorithm based on uncertainty sampling, *i.e.* the algorithm selects for further review/coding the documents about which it is least certain. 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 coding them.”<sup>6</sup> The algorithm is then run one more time to identify the likely relevant documents for manual review.

### Continuous Active Learning

In contrast to SPL, Continuous Active Learning (“CAL”) in a sense conflates the training and review phase. The process often starts with running key word searches.<sup>7</sup> Based on these terms, the algorithm ranks the documents from most to least likely relevant. An individual then reviews some of the top ranked documents and codes them for relevance. The system recalibrates its rankings based on this feedback and the individual then repeats the process of reviewing and coding the highly ranked documents that have not yet been coded. The process is repeated until

---

<sup>5</sup> 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 (July 2014).

<sup>6</sup> Maura R. Grossman and Gordon V. Cormack, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, Proceedings of the 37<sup>th</sup> International ACM SIGIR Conference on Research and Development in information retrieval (SIGIR’14) (July 2014).

<sup>7</sup> Although the CAL methodology may entail the use of key words to create a seed set, some tools allow the reviewer to simply begin coding by, for example, custodian 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.

“the number of top-ranked documents containing responsive information drops precipitously.”<sup>8</sup>

### The Creation of Training Sets and A Note About Key Words

The training sets used in TAR are created either through random generation, judgmental selection or some combination thereof. With random generation, the algorithm randomly selects a set of documents from the available universe. Those documents are then manually coded to create the training set. 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 entail some combination of random generation and judgmental selection.

Hence, key words may play a central role, a supporting role or no role at all within the TAR methodology.<sup>9</sup> Where key words *are* employed, parties frequently debate whether the requesting party should be provided with information regarding the search terms. This may include not only a list of the search terms, but also “hit reports” that provide insight into whether terms are overbroad or lead to anomalous results suggesting that terms may be too narrow or missing. Although the focus of this paper is on opinions regarding the discoverability of training sets, to the extent the parties or the courts also addressed issues related to key word searches, that analysis is included in the corresponding case summary.

### The Ramifications of the Selection of the Training Set

TAR offers the opportunity for greater accuracy at significantly lower costs to the producing party as compared to manual review. *See generally* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, Rich. J.L.& Tech., Spring 2011. Although the cost savings will vary depending upon the particular circumstances of the case, Grossman and Cormack have found that “The technology-assisted reviews require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review.” *Id.* at 43. These statistics underscore the importance of ensuring that the training set is reasonably accurate at the outset, as those decisions will be applied across the entire production. Thus, receiving parties have a compelling interest in achieving an acceptable level of transparency around the selection, composition and coding of training sets. A number of courts have weighed in on this issue within the last two years. The key cases are briefly summarized in Part II. Part III analyzes the arguments for and against disclosure of training sets.

---

<sup>8</sup> *Id.* at 290.

<sup>9</sup> A separate issue, which is beyond the scope of this paper, is whether it is appropriate to apply search terms prior to the application of TAR. Some academics argue that doing so improves the richness of the document set whereas others make a compelling argument that applying search terms before TAR results in an unacceptable number of relevant documents not being located and produced. For more on this issue, *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 (November 2013) and the response thereto, 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 (July 2014).

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

With one notable distinction, 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: the appropriate stage for establishing certain parameters, such as the relevance score cut-off; 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 algorithm, as well as the documents that were subsequently coded in the iterative training rounds.<sup>10</sup> The court noted that the defendants' transparency with respect to the training 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."<sup>11</sup> 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 "The issue regarding relevance standards might be significant if MSL's proposal was not totally transparent. Here, however, plaintiff will see how MSL has coded every email used in the seed set (both relevant and not relevant), and the Court is available to quickly resolve any disputes."<sup>12</sup> 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, "In 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 MSL."<sup>13</sup> 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*, 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 defendant provided the plaintiffs (and the court)

---

<sup>10</sup> *Id.* at 187.

<sup>11</sup> *Id.* at 192.

<sup>12</sup> *Id.* at 189.

<sup>13</sup> *Id.*

with a high degree of transparency with respect to its 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.<sup>14</sup> A log describing the withheld documents was also 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.*, 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” to work cooperatively to train the algorithm:

The Parties agree that as part of the Assessment and Training phases, all of the non-privileged and privilege-redacted documents reviewed by both parties’ experts will be categorized as, not relevant or skip (to be used for documents with technical problems). The privileged-withheld documents will be categorized by Defendants’ experts as relevant, not relevant, or skip, subject to the Parties’ right to have any privileged-withheld documents categorized as a “skip.”

\* \* \*

The Parties’ experts will then work collaboratively to determine the relevance of the non-privileged and privilege-redacted documents.<sup>15</sup>

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, 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.”<sup>16</sup>

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.”<sup>17</sup>

---

<sup>14</sup> Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding, *Global Aerospace Inc. v. Landow Aviation*, No. CL61040, 2012 EL 1419842 (Va. Cir. Ct. April 9, 2012).

<sup>15</sup> *Id.* at \*24, \*29. The Parties also agreed to meet and confer regarding the relevance score cutoff. *Id.* at \*31.

<sup>16</sup> *Id.* at \*23.

<sup>17</sup> *Id.* at \*1.

Defendants argued that the motion was premature<sup>18</sup> 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, "Based 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)."<sup>19</sup>

### Switching Horses Midstream<sup>20</sup>

Not uncommonly, a party may agree to one form of review only to later determine that the associated time and expense of the selected method warrant switching to TAR. One recent case, *Progressive Casualty Insurance Co. v. Delaney*, 2014 U.S. Dist. LEXIS 69166 (D. Nev. May 20, 2014) (J. Leen), reveals the potential risks a party faces if it chooses to unilaterally change a court-approved ESI protocol. Pursuant to the agreed-to protocol, the plaintiff, Progressive Casualty Insurance Company ("Progressive"), collected 1.8 million documents. After applying the agreed-to search terms, the number of documents subject to manual review was reduced to 565,000. After 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 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) with 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:

---

<sup>18</sup> 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.

<sup>19</sup> *Id.* at \*10.

<sup>20</sup> Notably, both cases discussed in this subsection, *Progressive Casualty Insurance Co. v. Delaney*, 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.

Progressive’s proposal would relieve it of the burden of manual review of ESI according to the ESI protocol it stipulated to and allow it to adopt a predictive coding alternative to a small subset of the universe of ESI collected. Its proposal would also give its counsel exclusive responsibility for training the predictive coding software, and it fails to comply with all of the best practices recommended by the Equivio software program. Progressive proposes a “do-over” of its own invention that lacks transparency and cooperation regarding the search methodologies applied.

\* \* \* \*

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.*<sup>21</sup>

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 code 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.”<sup>22</sup>

#### 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.***, 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.”<sup>23</sup> 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.”<sup>24</sup> However, the court 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

---

<sup>21</sup> *Id.* at \*28-29 (emphasis added).

<sup>22</sup> *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.

<sup>23</sup> *Id.* at \*3.

<sup>24</sup> *Id.* at \*5.

can lead a court to question why the uncooperative party is hiding something and such questions can affect the exercise of discretion.”<sup>25</sup>

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 court’s reasoning behind that decision.<sup>26</sup>

### **Part III. Training Sets: Analysis of the Arguments For and Against their Disclosure**

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 a requesting party is *not* entitled to disclosure of the documents used to train the algorithm. The following section examines arguments for and against the disclosure of training sets.

#### **A. 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.”<sup>27</sup> 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.”<sup>28</sup> 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.<sup>29</sup>

---

<sup>25</sup> *Id.* at \*5-6.

<sup>26</sup> 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 declined 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.

<sup>27</sup> 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.

<sup>28</sup> 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.

<sup>29</sup> 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 argued 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.*, 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.”)<sup>30</sup> 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.<sup>31</sup>

However, there is arguably a distinction between asking the producing party to reveal the documents that are *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.<sup>32</sup>

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 responsive to Plaintiffs’ document requests and does not delve into the thought processes of

---

<sup>30</sup> See also *FormFactor, Inc. v. Micro-Probe, Inc.*, 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).

<sup>31</sup> 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.

<sup>32</sup> See e.g., *FormFactor, Inc. v. Micro-Probe, Inc.*, 2012 U.S. Dist. LEXIS 62233, \*19-21 (N.D. Cal. May 3, 2012) (“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).

Defendants’ counsel.”<sup>33</sup> 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.

#### B. Disclosure of Non-Responsive Documents.

As discussed in the 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 nonresponsive as well as documents deemed responsive.<sup>34</sup> 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.”<sup>35</sup> 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.<sup>36</sup>

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 or not the defendant conspired to fix prices or whether a party 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-responsive documents particularly sensitive.<sup>37</sup> In that event, the parties may benefit from

---

<sup>33</sup> See also *In re Porsche Cars N. Am., Inc. Plastic Coolant Tubes Products Liability Litig.*, 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.)

<sup>34</sup> 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).

<sup>35</sup> *Biomet*, 2013 U.S. Dist. LEXIS 172570 at \*5.

<sup>36</sup> *Id.* at \*4.

<sup>37</sup> 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

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 order 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.

C. 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). 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.”<sup>38</sup> 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.

---

Learning and Other Advance Search and Review Methods), available at <http://www.umiacs.umd.edu/~oard/desi5/additional/Butterfield.pdf>

<sup>38</sup> *Da Silva Moore*, 287 F.R.D. at 192 and n. 4.