

The Implications of Rule 26(g) on
the Use of Technology-Assisted Review

Karl Schieneman¹ and Thomas C. Gricks III²

FOREWORD

The past few years have introduced a new “player” into the lexicon of ESI discovery—“Technology Assisted Review,” or “TAR,” aka “predictive coding.” Touted by some as the antidote for disproportionately expensive discovery of digital information, and condemned by others as an expensive, jargon-infested ruse to avoid legitimate production obligations, it is hard for lawyers, clients and judges to sort through the positive and negative hype, and reach a principled decision whether TAR is the long awaited solution, or just another exotic fad.

Enter Karl Schieneman and Tom Gricks, two lawyers possessed of a scientific and mathematical background and a talent for describing technical and statistical information in a way that generalists readily can understand, with a primer that is essential to understanding the proper framework to use in evaluating whether and how to employ TAR, or evaluate the appropriateness of its use by an adversary. This article accomplishes what has been sorely needed—a straightforward explanation that captures the governing procedural and substantive law, as well as sets out the essential technical and statistical concepts that will permit even computer shy lawyers and judges to understand enough to make the right choices.

Central to the paper is the primacy of Fed. R. Civ. P. 26(g)’s discovery certification obligations, grounded in the doctrine of proportionality required by Rule 26(b)(2)(C), and the importance of cooperation with opposing counsel, and transparency in the process used to collect and analyze the data set using TAR, to select the information to be produced in response to a Rule 34 request. The article fully explains the key phases of evaluating the proper way to use TAR: (1) *collection* (which embraces the concepts of recall, precision, prevalence/richness, and the statistical

1. B.A. University of Massachusetts, J.D. University of Pittsburgh, M.B.A. Carnegie Mellon University. President of Review Less, LLC.

2. B.S. Chemical Engineering, Carnegie-Mellon University, J.D. Duquesne University School of Law. Partner, Litigation Services, Schnader, Harrison, Segal & Lewis.

sampling concepts of confidence level and confidence interval); (2) *disclosure* to opposing counsel and the court the manner in which the documents have been collected, reviewed and produced; (3) the *training* of the TAR tool, through either judgmental or random selection of the training/seed set; (4) *stabilization* (the process of determining what levels of recall and precision are required to comply with the certification requirements of Rule 26(g), and entails considering the relative effectiveness of using TAR as opposed to the available alternatives—principally manual review or keyword searches) and , finally, (5) *validation* through statistical sampling that the production using TAR meets the mandates of Rule 26(g) that counsel undertake a reasonable inquiry before certifying the responsiveness of the production to opposing counsel.

This article is destined to be read and re-read by conscientious lawyers and judges as they strive to understand the promise, and avoid the pitfalls, of using TAR in ESI discovery. For that we can all be grateful.

Paul W. Grimm³

ABSTRACT

Technology assisted review has become common in cases where the collection of electronically stored information yields large quantities of potentially discoverable information. Rule 26(g) of the Federal Rules of Civil Procedure imposes unique obligations in this situation for counsel beginning with the reality that not all potentially relevant documents will be produced.

The authors explain the nature of technology assisted review and then apply the commands of Rule 26(g) to each phase of this review including collecting the data so that the search is truly reasonable and proportionate, disclosures to opposing counsel of methodology, creating the “seed set” that will guide the software’s collection of the relevant data and ascertaining whether the final production will meet the recognized standards of precision and recall and similar test of validation.

The authors insist that counsel must know how each step of this complicated process will impact the ultimate production so counsel will be able to meet the Rule 26(g) standards without fear of being sanctioned.

3. A.B. University of California, Davis, J.D. University of New Mexico School of Law, United States District Court Judge for the District of Maryland.

TABLE OF CONTENTS

| | | |
|------|---|-----|
| I. | THE LEGAL FRAMEWORK..... | 243 |
| II. | APPLICATION TO TECHNOLOGY-ASSISTED REVIEW | 248 |
| | A. Collection..... | 249 |
| | B. Disclosure | 254 |
| | C. Training..... | 259 |
| | D. Stabilization | 263 |
| | E. Validation..... | 269 |
| III. | CONCLUSION | 273 |

Slightly more than five years ago, Professor Richard L. Marcus⁴ anticipated the prominence of Federal Rule of Civil Procedure 26(g) in managing and controlling the conduct of attorneys in the burgeoning area of electronic discovery.⁵ In his Keynote Address given at *The University of Baltimore Law Review* Symposium on March 13, 2008, Professor Marcus observed:⁶

For me, having spent much of the last decade focused on e-discovery, it is interesting to consider how differently we might look at e-discovery in another decade. The rate of change is likely to abate somewhat, but given how different things are now from how they were a decade ago it seems dubious to expect that things will remain the same. So I'm not going to try to make predictions. Rather, I have some observations about how things may evolve and some questions about whether the fears of the past become the reality of the future.

* * *

(4) *There may be a new or enlarged role for Rule 26(g).* Rule 26(g) was added in 1983, at the same time that Rule 11 was substantially revised to strengthen its provisions. At the time, it was expected (perhaps hoped) that Rule 26(g) would be just as important as Rule 11. Needless to say, that did not happen. Amended Rule 11 mushroomed into the most prominent rule of its day, eventually being narrowed in 1993 to contain its effects. Rule 26(g) slipped from view, and had minimal effect.

It is possible that e-Discovery will breathe new life into Rule 26(g). The

4. Horace O. Coil ('57) Chair in Litigation Distinguished Professor of Law, University of California, Hastings College of Law, San Francisco.

5. Rule 26(g) embodies counsel's obligation to certify discovery responses, including responses to requests for production of documents and ESI, upon a "reasonable inquiry." FED. R. CIV. P. 26(g)(1).

6. Richard L. Marcus, *E-Discovery Beyond the Federal Rules*, 37 U. BALT. L. REV. 321, 344, 346 (2007-2008) (internal citations omitted).

extensive responsibilities of counsel in regard to consultations about e-discovery arrangements call for counsel to make representations to the other side, and sometimes to the court, about what can be done and when it can be done. Recently, some courts have reacted to unfounded (perhaps not entirely honest) statements as violating Rule 26(g). Maybe it will become the “new Rule 11.”

We are now at the midpoint of the decade, and it is obvious that Professor Marcus was correct in his “observations” of the role of Rule 26(g) in electronic discovery. The importance of the Rule 26(g) certification in the production of electronically stored information (“ESI”) is growing, as Wright, Miller & Professor Marcus himself reflected in the third edition of *Federal Practice and Procedure*.⁷

Although the addition of Rule 26(g) certification in 1983 did not produce a reaction like the simultaneous amendment of Rule 11, the expansion of discovery of electronically stored information has increased the importance of this certification. Parties may increasingly assert that their opponents have neglected to disclose or produce all of the electronically stored information they were supposed to provide, and that sanctions should therefore be imposed under Rule 26(g). Because the handling of this sort of discovery is a mandatory topic during the Rule 26(f) conference on the discovery plan, the opportunity to avoid miscommunication on this topic has increased, and judicial patience with later excuses about “misunderstandings” may accordingly decline.

Requests for these sanctions are escalating. By January 1, 2010, litigants had filed motions for sanctions in over 400 cases, resulting in 230 sanction awards on varying bases, including Rule 26(g).⁸ Moreover, sanctions for failing to comply with Rule 26(g) certification requirements are mandatory,⁹ and can be severe for both counsel and the client. For example, in *Qualcomm v. Broadcom* the trial judge initially issued an \$8.5 million sanction against Qualcomm for an inadequate search of electronically stored information, and further sanctioned six attorneys under Rule 26(g) for their role in the e-discovery process, requiring them to forward the sanction Order to the California Bar.¹⁰

In light of this trend, this paper addresses the implications of the Rule 26(g) certification requirements (and related Federal Rules) on technology-

7. 8A CHARLES WRIGHT, ARTHUR R. MILLER & RICHARD L. MARCUS, *FEDERAL PRACTICE AND PROCEDURE* § 2052 (3d ed. 2010) (internal citations omitted).

8. Dan H. Willoughby, Jr., Rose Hunter Jones & Gregory R. Antine, *Sanctions for E-Discovery Violation: By the Numbers*, 60 DUKE L.J. 789, 790, 799 (2010).

9. FED. R. CIV. P. 26(g)(3).

10. See *Qualcomm Inc. v. Broadcom Corp.*, 2008 U.S. Dist. LEXIS 911 (S.D. Cal. Jan. 7, 2008). The sanction against the attorneys was later lifted. *Qualcomm, Inc. v. Broadcom Corp.*, 2010 U.S. Dist. LEXIS 33889 (S.D. Cal. Apr. 2, 2010).

assisted review, the latest innovation in electronic document search methodology to come before the courts.¹¹ Technology-assisted review highlights, and renders eminently calculable (at least from a statistical perspective), one reality that lies at the heart of Rule 26(g) but has heretofore been largely ignored – the fact that some number of relevant documents *knowingly will not be produced*.¹² Recognizing the increasing role of technology-assisted review in the production of ESI, this paper considers the Rule 26(g) obligations as applied to several of the factors that can realistically be controlled by counsel during this process, in their efforts to properly manage this certainty and yet ensure “the just, speedy, and inexpensive” administration of litigation as mandated by the Federal Rules of Civil Procedure.¹³

I. THE LEGAL FRAMEWORK

The scope of the obligation to search for, and produce, ESI is circumscribed by Federal Rule of Civil Procedure 26(g), which provides as follows:¹⁴

(g) Signing Disclosures and Discovery Requests, Responses and Objections.

(1) Signature Required; Effect of Signature. Every disclosure under Rule 26(a)(1) or (a)(3) and every discovery request, response or objection must be signed by at least one attorney of record in the attorney’s own name – or by the party personally, if unrepresented – and must state the signer’s address, e-mail address, and telephone number. By signing, an attorney or party certifies that to the best of the person’s knowledge, information, and belief formed after a reasonable inquiry:

(A) with respect to a disclosure, it is complete and correct as of the time it is made; and

(B) with respect to a discovery request, response or objection, it is:

(i) consistent with these rules and warranted by existing law or by a

11. See, e.g., *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 183 (S.D.N.Y. Feb. 24, 2012) (“This judicial opinion now recognizes that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases.”).

12. See *id.* at 188 (“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’ . . .”).

13. FED. R. CIV. P. 1. See also *Da Silva Moore*, 287 F.R.D. at 191 (“Courts and litigants should be cognizant of the aim of Rule 1, to ‘secure the just, speedy and inexpensive determination’ of lawsuits”).

14. FED. R. CIV. P. 26(g)(1).

nonfrivolous argument for extending, modifying, or reversing existing law, or for establishing new law;

(ii) not interposed for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation; and

(iii) neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action.

Although it has occasionally been suggested that production must be “complete and correct” under Subsection (A) of Rule 26(g)(1), the more widely accepted view is that the search for, and production of, ESI in the context of discovery is controlled by Subsection (B).¹⁵ Thus, when ESI is produced using a technology-assisted review process, Rule 26(g) requires counsel to certify, upon reasonable inquiry, that the production is “consistent with [the discovery] rules. . . , not interposed for any improper purpose. . . , and neither unreasonable nor unduly burdensome or expensive.

15. *Compare* Kay Beer Distrib., Inc. v. Energy Brands, Inc., 2009 U.S. Dist. LEXIS 130595, at *14 (E.D. Wis. June 10, 2009) (“By signing a response to a discovery request, an attorney is certifying that to the best of his or her ‘knowledge, information, and belief formed after a reasonable inquiry’ the response is ‘complete and correct.’”) and *Doyle v. Gonzales*, 2011 U.S. Dist. LEXIS 20158, at *8 (E.D. Wash. Feb. 10, 2011) (“9. Additional phases of ESI discovery shall proceed on the same basis, until the City certifies that its ESI production is complete and accurate.”) with *Da Silva Moore*:

Plaintiffs’ objections to my February 8, 2012 rulings assert that my acceptance of MSL’s predictive coding approach “provides unlawful ‘cover’ for MSL’s counsel, who has a duty under FRCP 26(g) to ‘certify’ that their client’s document production is ‘complete’ and ‘correct’ as of the time it was made. 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. Plaintiffs simply misread Rule 26(g)(1). The certification required by Rule 26(g)(1) applies “with respect to a disclosure.” That is a term of art, referring to the mandatory initial disclosures required by Rule 26(a)(1). Since the Rule 26(a)(1) disclosure is information (witnesses, exhibits) that “the disclosing party may use to support its claims or defenses,” and failure to provide such information leads to virtually automatic preclusion, see Fed. R. Civ. P. 37(c)(1), it is appropriate for the Rule 26(g)(1)(A) certification to require disclosures be “complete and correct.”

Rule 26(g)(1)(B) is the provision that applies to discovery responses. It does not call for certification that the discovery response is “complete,” but rather incorporates the Rule 26(b)(2)(C) proportionality principle. Thus, Rule 26(g)(1)(A) has absolutely nothing to do with MSL’s obligations to respond to plaintiffs’ discovery requests. Plaintiffs’ argument is based on a misunderstanding of Rule 26(g)(1).

287 F.R.D. at 188 (internal citations omitted) and *Mancia v. Mayflower Textile Servs. Co.*, 253 F.R.D. 354, 357 (D. Md. 2008) (quoting Rule 26(g)(1) in its entirety).

...”¹⁶ The failure to adequately discharge these responsibilities subjects attorneys and parties to mandatory sanctions.¹⁷

The true breadth of counsel’s responsibilities under the Rule is explained in greater detail in the Advisory Committee Notes:¹⁸

Rule 26(g) imposes an affirmative duty to engage in pretrial discovery in a responsible manner that is consistent with the spirit and purposes of Rules 26 through 37. In addition, Rule 26(g) is designed to curb discovery abuse by explicitly encouraging the imposition of sanctions. The subdivision provides a deterrent to both excessive discovery and evasion by imposing a certification requirement that obliges each attorney to stop and think about the legitimacy of a discovery request, a response thereto, or an objection. The term “response” includes answers to interrogatories and to requests to admit as well as responses to production requests.

If primary responsibility for conducting discovery is to continue to rest with the litigants, they must be obliged to act responsibly and avoid abuse. With this in mind, Rule 26(g), which parallels the amendments to Rule 11, requires an attorney or unrepresented party to sign each discovery request, response, or objection. Motions relating to discovery are governed by Rule 11. However, since a discovery request, response, or objection usually deals with more specific subject matter than motions or papers, the elements that must be certified in connection with the former are spelled out more completely. The signature is a certification of the elements set forth in Rule 26(g).

Although the certification duty requires the lawyer to pause and consider the reasonableness of his request, response, or objection, it is not meant to discourage or restrict necessary and legitimate discovery. The rule simply requires that the attorney make a reasonable inquiry into the factual basis of his response, request, or objection.

The duty to make a “reasonable inquiry” is satisfied if the investigation undertaken by the attorney and the conclusions drawn therefrom are reasonable under the circumstances. It is an objective standard similar to

16. FED. R. CIV. P. 26(g)(1)(B). *See also* In re Porsche Cars N. Am., Inc. Plastic Coolant Tubes Prods. Liab. Litig., 2012 U.S. Dist. LEXIS 136954, at *15 (S.D. Ohio Sept. 25, 2012); Brock v. County of Napa, 2012 U.S. Dist. LEXIS 98579, at *11-12 (N.D. Cal. July 16, 2012).

17. *See* FED. R. CIV. P. 26(g)(3), which provides:

(3) Sanctions for Improper Certification. If a certification violates this rule without substantial justification, the court, on motion or on its own, *must* impose an appropriate sanction on the signer, the party on whose behalf the signer was acting, or both. The sanction may include an order to pay the reasonable expenses, including attorney’s fees, caused by the violation.

(emphasis added). *See also* Brock, 2012 U.S. Dist. LEXIS 98579, at *12.

18. FED. R. CIV. P. 26(g) Advisory Committee’s Notes, 1983 Amendments (internal citations omitted).

the one imposed by Rule 11. In making the inquiry, the attorney may rely on assertions by the client and on communications with other counsel in the case as long as that reliance is appropriate under the circumstances. Ultimately, what is reasonable is a matter for the court to decide on the totality of the circumstances.

Over the years since the inception of the Federal Rules relating to electronic discovery, the courts have developed several precepts to govern such discovery, all of which will likely extend to the use of technology-assisted review. First and foremost is the notion that the producing party must conduct a reasonable, good faith, and diligent search for responsive ESI, recognizing that the objective is not perfection.¹⁹ That obligation, however, is tempered by the proportionality concept in Rule 26(b)(2)(C)(iii) and the ultimate goal of “secur[ing] the just, speedy and inexpensive determination of every action and proceeding,” which is expressed in Federal Rule of Civil Procedure 1.²⁰ In other words, before signing a 26(g) certification, counsel is required to undertake a reasonable inquiry into the

19. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 191 (S.D.N.Y. 2012) (“While this court recognizes that computer-assisted review is not perfect, the Federal Rules of Civil Procedure do not require perfection.”); *Go v. Rockefeller Univ.*, 280 F.R.D. 165, 175 (S.D.N.Y. 2012) (“Under Federal Rules of Civil Procedure 26, a party must ‘conduct a diligent search’ for requested electronic documents.”) (quoting *Treppel v. Biovail Corp.*, 233 F.R.D. 363, 374 (S.D.N.Y. 2006)); *In re Porsche*, 2012 U.S. Dist. LEXIS 136954, at *29 (“[T]he Rules require only that Defendants and their clients engage in searches that are reasonable Defendants need not . . . affirm that their document searches are complete without qualification and that no additional responsive documents exist.”); *Zander v. Craig Hosp.*, 2011 U.S. Dist. LEXIS 29136, at *19 (D. Colo. Mar. 4, 2011) (“It should not require reiteration that litigants have an obligation, when discovery is sought from them, to make reasonable efforts to locate responsive documents”) (quoting *Fendi Adele S.R.L. v. Filene’s Basement, Inc.*, 2009 U.S. Dist. LEXIS 32615 (S.D.N.Y. Mar. 24, 2009)); *Moore v. Napolitano*, 723 F. Supp. 2d 167, 173 (D.D.C. 2010) (“Under Rule 26, a party is obligated to make a reasonable search for and produce documents responsive to the opposing party’s document requests.”); *Nycomed US Inc. v. Glenmark Generics LTD.*, 2010 U.S. Dist. LEXIS 82014, at *9 (E.D.N.Y. Aug. 11, 2010) (“This obligation to conduct a diligent search requires good faith on the part of the responding party and its attorneys. . . .”); *Kay Beer*, 2009 U.S. Dist. LEXIS 130595, at *14; *Treppel v. Biovail Corp.*, 233 F.R.D. 363, 374 (S.D.N.Y. 2006) (“Rather, [a party] must conduct a diligent search, which involves developing a reasonably comprehensive search strategy.”). *Accord*, MODEL RULES OF PROF’L CONDUCT R. 3.4(d) (“A lawyer shall not . . . in pretrial procedure, make a frivolous discovery request or fail to make reasonably diligent effort to comply with a legally proper discovery request by an opposing party.”)

20. *Da Silva Moore*, 287 F.R.D. at 188-89; *Cartel Asset Mgmt. v. Ocwen Fin. Corp.*, 2010 U.S. Dist. LEXIS 17857, at *28-29, *58 (D. Colo. Feb. 8, 2010) (noting that “all discovery is subject to the balancing test in Rule 26(b)(2)(C)(iii)” and that Rule 26(b)(2)(B) (relating to electronic discovery), must be construed and administered in accordance with Rule 1.); *Oracle USA, Inc. v. SAP AG*, 264 F.R.D. 541, 543-44 (N.D. Cal. 2009) (“[T]he Court has repeatedly emphasized that the scope of this case required cooperation in prioritizing discovery and in being mindful of the proportionality requirement of Federal Rule of Civil Procedure 26.”); *Mancia*, 253 F.R.D. at 359 (“The requirement of discovery being proportional to what is at issue is clearly stated in Rule 26(g)(1)(B)(iii)”).

relevant facts supporting adherence to these obligations, and his investigation and conclusions must be reasonable under the totality of the circumstances.²¹ Finally, the need to address the production of ESI has magnified the need for cooperation during the discovery process,²² resulting in an even greater demand for disclosure and transparency.²³

Ultimately, whether counsel has adequately discharged applicable Rule 26(g) obligations will typically be “a fact-intensive inquiry that requires evaluation of the procedures the producing party adopted during discovery. . . .”²⁴ Nevertheless, the manner in which the courts have dealt

21. *In re Delta/Airtran Baggage Fee Antitrust Litig.*, 846 F. Supp. 2d 1335, 1350 (N.D. Ga. 2012) (“This broad duty is satisfied when an attorney makes ‘a reasonable inquiry into the factual basis of his response, request or objection.’ Specifically, the attorney’s investigation and conclusions drawn therefrom must be reasonable under the circumstances.”); *Interpreter Servs., Inc. v. BTB Techs., Inc.*, 2011 U.S. Dist. LEXIS 149569 (D.S.D. Dec. 29, 2011) (“The pertinent inquiry pursuant to Rule 26(g) is whether the attorney made a reasonable inquiry before signing the discovery document.”); *Zander*, 2011 U.S. Dist. LEXIS 29136, at *19-20 (“It is equally the litigant’s responsibility (and that of its counsel) to determine beforehand the accuracy of any representations that production is complete. . . .”) (quoting *Fendi*, 2009 U.S. Dist. LEXIS 32615); *See also Chapman & Cole v. Itel Container Intl. B.V.*, 865 F.2d 676, 685-86 (5th Cir. 1989) (addressing traditional discovery).

22. As stated by the Sedona Conference:

Lawyers have twin duties of loyalty: While they are retained to be zealous advocates for their clients, they bear a professional obligation to conduct discovery in a diligent and candid manner. Their combined duty is to strive in the best interests of their clients to achieve the best results at a reasonable cost, with integrity and candor as officers of the court. Cooperation does not conflict with the advancement of their clients’ interests—it enhances it. Only when lawyers confuse advocacy with adversarial conduct are these twin duties in conflict.

The Sedona Conference, *The Sedona Conference Cooperation Proclamation*, 10 SEDONA CONF. J. 331, 331 (2009) (“*Sedona*” or “*The Sedona Cooperation Proclamation*”). *See also Da Silva Moore*, 287 F.R.D. at 192 (citing *Sedona*); *Anderson v. Reliance Std. Life Ins. Co.*, 2011 U.S. Dist. LEXIS 117058, at *2-3 (D. Md. Oct. 11, 2011); *Richardson v. Sexual Assault/Spouse Abuse Research Ctr., Inc.*, 270 F.R.D. 223, 227 (D. Md. 2010); *Mezu v. Morgan State Univ.*, 269 F.R.D. 565, 571 (D. Md. 2010); *Madison v. Harford County*, 268 F.R.D. 563, 565 (D. Md. 2010); *Cartel Asset Mgmt.*, 2010 U.S. Dist. LEXIS 17857, at *40-41 (citing *Sedona*); *Oracle*, 264 F.R.D. at 543-44 (citing *Sedona*); *William A. Gross Constr. Assocs. v. Am. Mfrs. Mutual Ins. Co.*, 256 F.R.D. 134, 136 (S.D.N.Y. 2009) (citing *Sedona*); *Mancia*, 253 F.R.D. at 359-60 (citing *Sedona*).

23. *In re Porsche*, 2012 U.S. Dist. LEXIS 136954, at *22-23 (“Transparency in the discovery process is necessary to ensure that all relevant information is made available to the litigants such that both parties have mutual knowledge of all relevant facts Full disclosure of Defendants’ efforts in collecting responsive documents will illuminate this issue so that the parties can resolve it.”); *Da Silva Moore*, 287 F.R.D. at 192 (“An important aspect of cooperation is transparency in the discovery process.”); *Cartel Asset Mgmt.*, 2010 U.S. Dist. LEXIS 17857, at *40 (“This Court has endorsed [*Sedona*] and its call for ‘cooperative, collaborative and transparent discovery.’”); *Mancia*, 253 F.R.D. at 363 (advocating the *Sedona* “drive to promote . . . the development of practical tools to facilitate cooperative, collaborative, transparent discovery.”).

24. *S2 Automation LLC v. Micron Tech., Inc.*, 2012 U.S. Dist. LEXIS 120097, at *99-100 (D.N.M. Aug. 9, 2012) (stating further that “it can become necessary to evaluate whether an attorney complied with his Rule 26(g) obligations and to evaluate the strategy an attorney used to provide responsive discovery, with relevant circumstances, including: (i) the number and complexity of the issues; (ii) the location, nature, number and availability of potentially relevant

with counsel's Rule 26(g) obligations to date provides substantial guidance for the increasing use of technology-assisted review.

II. APPLICATION TO TECHNOLOGY-ASSISTED REVIEW

Given the nature of discovery, adherence to the principles of Rule 26(g) is critical at every step of the process when technology-assisted review is used to cull relevant documents from an electronically-stored database. Practically speaking, any evaluation of the efficacy of a production is largely dependent upon the ability of the receiving party to identify deficiencies in that production.²⁵ Finding gaps in a document production, especially where the source is an electronic database, however, is not always a simple and straightforward task.²⁶ “At bottom, the discovery process *relies upon* the good faith and professional obligations of counsel to reasonably and diligently search for and produce responsive documents.”²⁷ As set forth below, that requires counsel to “stop and think” at every stage of the process of technology-assisted review, from the outset of collection to the completion of validation.

witnesses or documents; (iii) the extent of past working relationships between the attorney and the client, particularly in related or similar litigation; and (iv) the time available to conduct an investigation.”) (internal citations and quotations omitted).

25. *Accord S2 Automation*, 2012 U.S. Dist. LEXIS 120097, at 38 (“The Court related that, if S2 Automation certifies it has produced all responsive documents, there is little more the Court can do now to require them to produce documents, unless further discovery reveals that the search was inadequate.”); *Moore v. Napolitano*, 723 F. Supp. 2d 167, 172-73 (D.D.C. 2010) (In addressing the burden of proof associated with an inadequate production, the Court reflected that the burden is not placed upon the producing party until the receiving party discovers some flaw in the production.) *But see Apple, Inc. v. Samsung Elecs. Co.*, 2013 U.S. Dist. LEXIS 67085, at *38 (N.D. Cal. May 9, 2013) (requiring identification of search terms and custodians by subpoena respondent without first establishing deficiencies in production, and for the specific purpose of “evaluating the adequacy of [respondent’s] search . . .”).

26. *See, e.g., Wingnut Films, Ltd. V. Katja Motion Pictures, Corp.*, 2007 U.S. Dist. LEXIS 72953, at *11 (C.D. Cal. Sept. 18, 2007) (“Wingnut *fortuitously* discovered the existence of specific responsive documents that had not been produced.... All the while, New Line’s counsel persisted in belittling Wingnut’s concerns as ‘paranoia’ and ‘harassment’.”) (emphasis added). *See also Nat’l Day Laborer Org. Network v. United States Immigration & Customs Enforcement Agency*, in which the Court made the following observation:

It is impossible to evaluate the adequacy of an electronic search for records without knowing what search terms have been used. In earlier times, custodians and searchers were responsible for familiarizing themselves with the scope of a request and then examining documents individually in order to determine if they were responsive. Things have changed. Now custodians can search their entire email archives, which likely constitute the vast majority of their written communications, with a few keystrokes. The computer does the searching. But as a result, the precise instructions that custodians give their computers are crucial.

877 F. Supp. 2d 87, 106-07 (S.D.N.Y. 2012).

27. *Smith v. Life Investors Ins. Co. of Am.*, 2009 U.S. Dist. LEXIS 58261, *14-15 (W.D. Pa. July 9, 2009) (citing Rule 26(g) and the Advisory Committee Notes) (emphasis added).

A. Collection

The collection of ESI for culling with technology-assisted review requires an application of Rule 26(g) not previously associated with traditional review techniques. When ESI is culled using either linear review techniques or keyword search methodologies, the issue of whether a reasonable search was conducted typically turns on whether the collection was sufficiently broad and comprehensive.²⁸ In a linear review scenario, lawyers review every document in the collection, regardless of the size or scope of the collection. When keyword search techniques are used to cull particular documents from the collection, the ability to locate documents is not dependent upon the number of documents being searched, or the relevance of the documents containing the particular keywords.²⁹ Conversely, since technology-assisted review depends primarily upon statistics to validate effectiveness,³⁰ Rule 26(g) requires counsel to exercise greater care in the collection of ESI, in order to optimize the fraction of relevant documents processed into the tool.

To understand the potential impact of the initial collection of ESI upon technology-assisted review, it is first necessary to understand the various measures that are used to characterize and evaluate ESI collections. An ESI collection obviously contains both relevant and irrelevant documents. The fraction of relevant documents in an ESI collection is commonly referred to as “prevalence” or “richness.”³¹ Richness tends to be between five percent (5%) and ten percent (10%) of the total collection, but may be greater or even an order of magnitude less.³² Two common measures of the effectiveness of a technology-assisted review in locating these relevant

28. See *Nat'l Day Laborer Org.*, 877 F. Supp. 2d at 102 (S.D.N.Y. 2012) (“[T]he failure to search the files of seven former employees was not *de minimis* and made the unit’s search inadequate.”); *In Re: Delta/Airtran*, 846 F. Supp. 2d at 1351 (failing to collect back-up tapes and failure to process all hard drives); *Napolitano*, 723 F. Supp. 2d at 173 (limiting search to three specific Secret Service divisions); *Nycomed*, 2010 U.S. Dist. LEXIS 82014, at *26-27 (failing to search relevant laptop computers and database); *Wingnut Films*, 2007 U.S. Dist. LEXIS 72953, at *36 (failing to collect electronic documents from pertinent custodians).

29. Consequently, keyword searches have the potential to be over-inclusive, a propensity that is only exaggerated as the size of the ESI collection increases and the fraction of relevant documents decreases. See *Hodczak v. Latrobe Specialty Steel Co.*, 761 F. Supp. 2d 261, 279 (W.D. Pa. 2010) (A search of 206 mailboxes, consisting of 400,000 documents, revealed that only 2.7% of the documents contained the selected keywords, and only 0.2% of the total collection were in any way relevant.).

30. Herbert Roitblat, *Measurement in eDiscovery: A Technical White Paper*, ORCATEC (2013), <http://orcatec.com/sites/default/files/11%20Roitblat%20%20Measurement%20in%20eDiscovery.pdf>.

31. Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 2013 FED. CTS. L. REV. 7 (January 2013) (hereinafter *The Grossman-Cormack Glossary*).

32. Roitblat, *supra* note 29, at 5.

documents are “recall” and “precision.”³³ Recall represents the fraction of relevant documents within the ESI collection that are retrieved and identified as relevant by the technology-assisted review process; precision represents the fraction of the documents retrieved by the technology-assisted review process that are in fact relevant.³⁴

Given the typical size of ESI collections, ranging from tens of thousands to millions of documents, these measures are most often estimated by some means of statistical sampling.³⁵ “Two of the key ideas in sampling statistics are *confidence level* and *confidence interval*.”³⁶ Confidence level refers to the repeatability of the sample, and confidence interval is the range around the estimate that should contain the true value.³⁷ For example, if prevalence was estimated at 10% (the “point estimate”) by using a sample sufficient to obtain a confidence level of 95% and a confidence interval of $\pm 2\%$, then in 95 out of 100 samples, true prevalence would be between 8% and 12%.³⁸

When richness is low, the confidence interval can skew or mask the evaluation of the effectiveness of technology-assisted review. For example, in order to estimate the prevalence of a 1 million document ESI collection at a confidence level of 95% and a confidence interval of $\pm 2\%$, a random sample of 2395 documents would be required.³⁹ If the 2395-document sample exhibits a prevalence of only 1% (a low prevalence), the confidence interval would actually be $\pm 0.4\%$, which means that the true prevalence value lies somewhere within the range of 0.6% to 1.4%.⁴⁰ While the

33. *Id.* at 2 (“The two most commonly used measures of correctness and completeness are Precision and Recall.”).

34. *Id.*; *The Grossman-Cormack Glossary*.

35. *See* Roitblat:

eDiscovery protocols often call for the use of sampling because it would be prohibitively expensive to review all of the documents in a collection to determine which are truly responsive and which are not. With the ever-increasing size of document collections, it is simply not possible in many cases for the producing party to read every document in a reasonable amount of time or at a reasonable cost. Sampling allows us to assess a subset of documents, and on the basis of that subset make inferences about the collection as a whole.

supra note 29, at 6.

36. *Id.*

37. *Id.*

38. *Id.*

39. <http://www.surveysystem.com/sscalc.htm>. This is the maximum number of documents that would be required, because the estimation process assumes an equal distribution of relevant and irrelevant documents, *i.e.*, a richness of fifty percent (50%). For a richness that is either much greater or much less than 50%, the number of documents required to attain a confidence interval of $\pm 2\%$ becomes less or, correspondingly, the confidence interval becomes smaller if a 2395 document sample is used.

40. *Id.* *See also* *The Grossman-Cormack Glossary*. Cast in absolute terms, this means that the true number of relevant documents in the collection lies somewhere between 6,000 and 14,000

confidence interval appears to be tighter, the significance of this analysis lies in the fact that the actual number of relevant documents in the collection could exceed the statistically derived point estimate by *forty percent*. By comparison, if the prevalence was ten percent (10%), the confidence interval for a similar statistical sample would be greater in an absolute sense at $\pm 1.2\%$, but would only represent *twelve percent* of the point estimate.⁴¹ Since the calculation of recall depends on the total number of relevant documents in the collection, and the margin of error is greater when richness is low, there is a much wider range in potential recall values (relative to the point estimate), when richness is low.

Rule 26(g) requires counsel to recognize, understand, and endeavor to avoid high levels of uncertainty, and minimize the margin of error, in their use of technology-assisted review. Therefore, counsel “must conduct a diligent search, which involves developing a reasonably comprehensive search strategy. Such a strategy might, for example, include identifying key employees and reviewing any of their files that are likely to be relevant to the claims in the litigation.”⁴² Counsel is further obligated to “analyze[] the results of the search to assess its reliability, appropriateness for the task, and quality of its implementation.”⁴³ In analyzing the results, counsel is required to conduct a reasonable inquiry into the factual basis supporting Rule 26(g) compliance,⁴⁴ and “follow[] up on any perceived inadequacies with the production. . . .”⁴⁵ Collectively, these obligations require counsel to understand the impact of richness on the ability of technology-assisted review to cull relevant documents from the database, and to develop a strategy for avoiding as much uncertainty in the result as is reasonably possible.⁴⁶

There are several steps that can be taken to optimize and manage the

documents.

41. With a prevalence of 10%, the confidence interval for a 2395 document sample would be 1.2%, or only twelve percent (12%) of the point estimate. *See*, <http://www.surveysystem.com/sscalc.htm#two>.

42. *Treppel v. Biovail*, 233 F.R.D. 363, 374 (S.D.N.Y. 2006) (recognizing that “[d]efined search strategies are even more appropriate in cases involving electronic data, where the number of documents may be exponentially greater”). *See also S2 Automation*, 2012 U.S. Dist. LEXIS 120097, at *99-100 (identifying the circumstances relevant to the Court’s evaluation of counsel’s search strategy).

43. *Victor Stanley*, 250 F.R.D. at 259-60 (addressing proposed keyword search methodology). *See also William A. Gross*, 256 F.R.D. at 136 (requiring “quality control test[ing] to ensure accuracy in retrieval . . .”).

44. *In re Delta/Airtran*, 846 F. Supp. 2d at 1350.

45. *In re Porsche*, 2012 U.S. Dist. LEXIS 136954, at *28.

46. This obligation is particularly important in the context of Rule 26(g) certification, inasmuch as the producing party will be exclusively aware of the richness of the ESI collection at the outset.

richness of the database, consistent with the overall goals of Rule 26(g). First, counsel can interview custodians to identify personal ESI organizational conventions, thereby focusing the collection on relevant electronic documents, including emails as well as loose files, rather than collecting too broadly from locations that are not likely to contain relevant materials. It is not unusual for custodians to organize electronic files in folders, much the way paper documents were typically maintained, facilitating the collection of relevant ESI, and thereby enhancing the richness of the collection.

Second, counsel can, and indeed should, eliminate unnecessary file types before processing the data into the tool. For example, in many cases ESI collections will contain internet browsing files in .html format that have nothing to do with the litigation. Eliminating these unnecessary file types from the collection will therefore likely enhance richness.

Third, rather than controlling richness directly, counsel can effectively reduce the margin of error by increasing sample size. For example, for a 1 million document ESI collection with a prevalence of ten percent (10%), a random sample of roughly 3,450 documents would be needed to ensure that the confidence interval does not exceed ten percent of the point estimate (*i.e.*, a confidence interval of $\pm 1\%$). To achieve the same relative margin of error for an equivalent collection with a prevalence of only one percent (1%) would require a 34,000 document random sample. However, it would take nearly two months to review the random sample just to evaluate the richness of the ESI collection.⁴⁷ In addition, as sample sizes grow, the utility of independent review becomes questionable.⁴⁸ Therefore, whether this level of effort will ultimately be required to manage uncertainty will depend on the concept of proportionality in Rule 26(b)(2)(C)(iii), discussed *infra*.

47. See Nicholas M. Pace & Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND INSTITUTE FOR CIVIL JUSTICE, (Apr. 2012), http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf (“The most-expansive claims regarding review speed is about 100 documents per hour . . .”).

48. See Roitblat:

There is an added problem in trying to assess a large sample of documents. The sample may simply be too large for a single authoritative reviewer to assess. When we include multiple reviewers in finding our sample of responsive documents, we run into the fact that reviewers differ in their criteria for calling a document responsive. Professional eDiscovery reviewers agree with one another on relevance calls only about 50% of the time (Roitblat, Kershaw, & Oot, 2010).

So in addition to the variability inherent in sampling from a large population, we also have the variability due to differing judgments of responsiveness. Some of the documents that we intend to be truly responsive will actually be false positives. This is a shaky standard against which to compare the performance of our process.

Supra note 29, at 7.

Fourth, if done carefully and correctly, counsel can pre-cull the ESI collection with keywords to improve its richness.⁴⁹ However, keyword culling can be both over and under-inclusive, which may (1) further impair richness, and (2) exclude truly relevant documents, respectively.⁵⁰ Moreover, Rule 26(g) imposes additional obligations that must be evaluated and observed when keywords are used to locate relevant documents in the discovery process before running the collection through a technology-

49. There is certainly support for that exact procedure. See Order Regarding Discovery of ESI, In Re: Biomet M2a Magnum Hip Implant Prod. Liab. Litig., Cause No. 3:12-MD-2391 (N.D. In. Apr. 8, 2013), available at <http://www.inmd.uscourts.gov/mdl/ESI%20order.pdf> (hereinafter *Biomet I*).

However, while the *Biomet* decision authorizes pre-culling with keywords, the decision also exemplifies the difficulties associated with sparse collections. Biomet estimated the 19.5 million document collection to have a statistical richness of 1.93% (or roughly 375,000 relevant documents). In order to enhance richness, Biomet filtered the collection with keywords before running predictive coding, resulting in a culled set of 2.5 million documents, after de-duplication. The culled set, however, was estimated to have over 410,000 relevant documents – more than in the original collection.

Beyond this statistical anomaly, the *Biomet* decision should also serve as a case study in the dynamics underlying counsel's obligation to conduct a proper collection. The Plaintiff in *Biomet* sought to compel Biomet to run predictive coding on the entire collection of 19.5 million documents. Biomet refused, arguing that it would be exorbitantly expensive, especially given the time and expense associated with Biomet's ongoing ESI production procedures. The Court agreed with Biomet, on the basis of the proportionality considerations of Federal Rule of Civil Procedure 26(b)(2)(C). While the Court ultimately determined that Biomet's production "comple[d] fully with the requirements of Federal Rules of Civil Procedure 26(b) and 34(b)(2)," the Court did not even consider the implications of Rule 26(g). Nor did the Court consider the related admonition that a self-inflicted burden (*i.e.*, a very poor collection) cannot support a proportionality argument. See, e.g., *Escamilla v. SMS Holdings Corp.*, 2011 U.S. Dist. LEXIS 122165, at *16-17 (D. Minn. Oct. 21, 2011) (The Court favorably quoted *Quinby v. WestLB AG*, 245 F.R.D. 94, 104-05 (S.D.N.Y. 2006) for the proposition that, "if a party creates its own burden or expense by converting into an inaccessible format data that it should have reasonably foreseen would be discoverable material at a time when it should have anticipated litigation, then it should not be entitled to shift the costs of restoring and searching the data." The Court further noted the Sedona Conference determination that "[i]n assessing whether a particular discovery request or requirement is unduly burdensome or expensive, a court should consider the extent to which the claimed burden expense grow[s] out of the responding party's action or inaction." The Sedona Conference, *The Sedona Conference Commentary on Proportionality in Electronic Discovery*, 11 SEDONA CONF. J. 289, 298 (2010). See also *Pippins v. KPMG LLP*, 2011 U.S. Dist. LEXIS 116427, at *25-26 (S.D.N.Y. Oct. 7, 2011) (observing that the burden underlying KPMG's proportionality argument was "self-inflicted to a large extent.").

50. See, e.g., *Hodczak*, 761 F. Supp. 2d at 44 (finding less than 8% of the documents having the requested keyword were relevant); *Victor Stanley*, 250 F.R.D. at 257 ("common sense suggests that even a properly designed and executed keyword search may prove to be over-inclusive or under-inclusive . . ."); Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11, 18 (2011), <http://jolt.richmond.edu/v17i3/article11.pdf> (citing the Blair & Maron study in which keyword searching turned up only twenty percent (20%) of the relevant documents and failed to identify the remainder) (hereinafter *The JOLT Study*).

assisted review.⁵¹ Each of these considerations should be fully evaluated before using keywords to cull an ESI collection in advance of implementing technology assisted review.

Ultimately, if the collection cannot inherently be improved, counsel can implement an effective workflow to optimize the results of technology-assisted review for collections with low richness. For example, rather than implementing a technology-assisted review of the entire collection, counsel can initially focus the review on the custodians or ESI sources most likely to have relevant documents, and then expand the search to the balance of the collection.

Thus, counsel can, and indeed must, consider the manner in which ESI is collected for technology-assisted review, to optimize the richness of the collection and, in turn, enhance the certainty of the results generated by the review. Without a considered, reasonable and articulable basis, counsel cannot blithely cast a wide net without a view toward the impact of collection on the review. Technology-assisted review is no substitute for a well-crafted collection.

B. Disclosure

Nothing in Rule 26(g) obligates counsel to disclose the manner in which documents are collected, reviewed and produced in response to a discovery request.⁵² However, Federal Rule of Civil Procedure 26(f) *does*

51. See *Kay Beer*, 2009 U.S. Dist. LEXIS 130595, at *15 (discussing the need to include variants of search terms to ensure coverage); *William A. Gross Constr. Assocs. v. Am. Mfrs. Mutual Ins. Co.*, 256 F.R.D. 134, 136 (S.D.N.Y. 2009) (finding that counsel is required to “carefully craft the appropriate keywords, with input from the ESI’s custodians as to the words and abbreviations they use” and further that “the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of ‘false positives.’”). *Accord, Victor Stanley*, 250 F.R.D. at 257 (recognizing a number of considerations in evaluating a keyword search for privilege review, including the specific keywords, their rationale, the qualifications of the persons designing the search, the use of Boolean operators, and the analysis of the results). Indeed, in considering the nature of keyword search, Magistrate Judge John Facciola has observed:

Whether search terms or “keywords” will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics. Given this complexity, for lawyers to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread. This topic is clearly beyond the ken of a layman and requires that any such conclusion be based on evidence that, for example, meets the criteria of Rule 702 of the Federal Rules of Evidence.

United States v. O’Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008).

52. See *FED. R. CIV. P. 26(g); Benson v. St. Joseph Reg’l Health Ctr.*, 2006 U.S. Dist. LEXIS 28795, at *14 (S.D. Tex. May 1, 2006) (“Finally, it is unnecessary for Defendants to explain the details of their method of searching when they have certified and represented to the Court that they have complied fully with Plaintiffs’ requests and made reasonable efforts to find

require counsel to meet and confer, and prepare a discovery plan, which “must state. . . any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced. . . .”⁵³ This requirement has been read in conjunction with *The Sedona Cooperation Proclamation* to obligate counsel to cooperatively derive search methodology, at least in the context of keyword searching.⁵⁴ Indeed, as discussed above, *The Sedona Cooperation Proclamation* has been adopted and cited by a number of Courts in favor of cooperative and transparent electronic discovery.⁵⁵ Moreover, in the only opinion to address technology-assisted review protocols in any detail, Magistrate Judge Andrew J. Peck strongly recommended disclosure.⁵⁶

and disclose all responsive documents and emails.”). *Accord* In re Cathode Ray Tube Antitrust Litig., 2011 U.S. Dist. LEXIS 120218, at *34-35 (N.D. Cal. Mar. 21, 2011) (“Plaintiffs ask Hitachi to disclose the search terminology which Hitachi used to search for and examine relevant documents While desirable, such joint development is not a legal requirement in this case.”).

53. FED. R. CIV. P. 26(f)(3)(C).

54. In requiring counsel to confer and reach agreement on “search terms . . . , custodians . . . , date ranges . . . , and any other essential details about the search methodology they intend to implement for the production of electronically-stored information” the United States District Court for the Eastern District of Pennsylvania observed:

The Federal Rules of Civil Procedure, case law, and the Sedona Principles all further emphasize that electronic discovery should be a party-driven process. Indeed, Rule 26(f) requires that the parties meet and confer to develop a discovery plan. That discovery plan must discuss “any issues about disclosure or discovery of [ESI], including the form or forms in which it should be produced.” In fact, the commentary to the rule specifically notes that whether metadata “should be produced may be among the topics discussed in the Rule 26(f) conference.”

Indeed, the Sedona Conference has advised that:

Cooperation... requires... that counsel adequately prepare prior to conferring with opposing counsel to identify custodians and likely sources of relevant ESI, and the steps and costs required to access that information. It requires disclosure and dialogue on the parameters of preservation. It also requires forgoing the short term tactical advantages afforded one party to information asymmetry so that, rather than evading their production obligations, parties communicate candidly enough to identify the appropriate boundaries of discovery. Last, it requires that opposing parties evaluate discovery demands relative to the amount in controversy. In short, it forbids making overbroad discovery requests for purely oppressive, tactical reasons, discovery objections for evasive rather than legitimate reasons, and “document dumps” for obstructionist reasons. In place of gamesmanship, cooperation substitutes transparency and communication about the nature and reasons for discovery requests and objections and the means of resolving disputes about them.

Romero v. Allstate Ins. Co., 271 F.R.D. 96, 109 (E.D. Pa. 2010) (internal citations omitted).

55. *Cartel Asset Mgmt. v. Ocwen Fin. Corp.*, 2010 U.S. Dist. LEXIS 17857, at *40 (D. Colo. Feb. 8, 2010); *William A. Gross*, 256 F.R.D. at 136 (citing *The Sedona Cooperation Proclamation* for the proposition that “the best solution in the entire area of electronic discovery is cooperation among counsel”); *Mancia v. Mayflower Textile Servs. Co.*, 253 F.R.D. 354, 363 (D. Md. 2008).

56. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 184 (S.D.N.Y. 2012) (citing Andrew Peck, *Search Forward*, L. TECH. NEWS, Oct. 2011, at 25, 29).

Of course, the best approach to the use of computer-assisted coding is to follow the Sedona Cooperation Proclamation model. Advise opposing counsel that you plan to use computer assisted coding and seek agreement; if you cannot, consider whether to abandon predictive coding for that case or go to the court for advance approval.

In any event, there is no doubt that counsel will be required to disclose the use of a technology-assisted review process should the production of electronic documents appear to be deficient,⁵⁷ and the Courts disfavor unnecessary gamesmanship in the conduct of discovery, especially electronic discovery.⁵⁸

In addition to the possible disclosure during the meet-and-confer process, it may also become necessary to disclose in order to effectively comply with Federal Rule 34. “A proper response under Rule 34 of the Federal Rules of Civil Procedure to a request for production has two

57. See, e.g., *In re Porsche*, 2012 U.S. Dist. LEXIS 136954. The Plaintiffs in *In re Porsche* sought disclosure from the Defendants of “the parameters of their electronic searches and [] detail regarding the steps they took to locate and produce responsive documents” *Id.* at *22. In granting Plaintiffs’ request, the Court noted the following:

Transparency in the discovery process is necessary to ensure that all relevant information is made available to the litigants such that both parties have mutual knowledge of all relevant facts. Transparency is particularly important in this jurisdictional discovery, where the parties dispute almost every aspect of every request and response. There are several outstanding issues that the parties cannot resolve until they reach a level of understanding about the way in which Porsche AG maintains documents and the way in which Porsche AG’s counsel located and produced responsive documents.

* * *

Other factors support full disclosure of Defendants’ search methodology. For example, the parties dispute the definitions of several terms – such as “officer” and “directives” – and whether they have the same meaning in English and in German. It would frustrate the purpose of the discovery rules if Defendants were permitted to come up with their own well-crafted definitions of these terms and then refuse to disclose what those definitions are. Defendants assert that they have not hidden anything from Plaintiffs based on their objections to vague and ambiguous terms; full disclosure of the way in which Defendants applied these terms to their searches will either verify or refute this assertion.

* * *

...Disclosing the way in which Defendants searched for responsive documents will either confirm Defendants’ assertion that they have met their obligations under the Federal Rules or confirm Plaintiffs’ speculation that Defendants must do more – either way, it will shed light on the disconnect between the parties so that they can identify and resolve their dispute.

Id. at *22-26.

58. See *In re Facebook PPC Adver. Litig.*, 2011 U.S. Dist. LEXIS 39830, at *6-8 (N.D. Cal. Apr. 6, 2011) (citing The Sedona Conference to require negotiation of search terms); *Romero*, 271 F.R.D. at 109 (E.D. Pa. 2010) (citing The Sedona Conference to require negotiation of search terms); *Trusz v. UBS Realty Investors LLC*, 2010 U.S. Dist. LEXIS 92603, *15-18 (D. Conn. 2010) (citing The Sedona Conference to require negotiation of search terms).

components. First, the responding party must state in writing, for each category requested, that ‘inspection. . . will be permitted as requested’ or, if the responding party objects, the basis of that objection. . . . The second component of a proper response to a request for production requires the responding party – within a specified time – to actually produce the responsive documents for inspection or copying.”⁵⁹ “If the responding party cannot conduct a ‘careful and thorough’ search for all responsive documents within the thirty-day deadline imposed by this Rule, it ‘ha[s] an obligation to seek appropriate extensions. . . .’ Unilaterally deciding to conduct a cursory initial search to be followed by ‘rolling’ productions from subsequent, more thorough, searches is not an acceptable option.”⁶⁰ Since the implementation of technology-assisted review typically extends beyond the timeframe for an initial response to a request for production of documents, the need to obtain an extension and effect a rolling production may well require counsel to disclose the use of technology-assisted review so as not to run afoul of Federal Rule of Civil Procedure 34, which requires production of responsive documents within 30 days of service.⁶¹

If, however, disclosure fails to generate an agreement on the protocols that will be employed, there is support for either petitioning the Court to authorize the use of technology-assisted review, or implementing the process unilaterally. As Magistrate Judge Peck suggested in *Da Silva Moore*, rather than abandon the use of technology-assisted review, counsel

59. *Novelty, Inc. v. Mt. View Mktg., Inc.*, 265 F.R.D. 370, 375 (S.D. In. 2009) (internal citations omitted). *See also* *Kinetic Concepts, Inc. v. Convatec Inc.*:

[A]ttorneys comply with half of what Rule 34(b) requires – they file a written response, but they do not agree to the requested time, place and manner of production.. or propose a reasonable alternative method. Rather, they give the vague assurance that the requested documents will be produced [at an unspecified time]. This practice assures continued wrangling and negotiating before the documents ultimately are produced... [and] frequently derails the discovery process, because parties often wait to schedule depositions until after documents production has occurred.

268 F.R.D. 226, 246 (M.D.N.C. 2010) (internal citations omitted).

60. *Id.* at 376 (internal citations omitted).

61. Counsel should also be cognizant of the fact that the Rule 26(g) certification does not expire when the written response is made, but rather governs obligations throughout discovery even to the point of trial. For example, in expressly ordering a rolling production, Magistrate Judge John Facciola noted “I cannot emphasize too strongly that I will hold counsel to the obligations imposed upon them by Rule 26(g) of the Federal Rules of Civil Procedure” *Chevron Corp. v. The Weinberg Group*, 286 F.R.D. 95, 101 (D.D.C. 2012). *Accord* FED. R. CIV. P. 26(e)(1)(A) (“A party who has . . . responded to . . . [a] request for production . . . must supplement or correct its . . . response . . . in a timely manner if the party learns that in some material respect the . . . response is incomplete or incorrect”); *Bd. of Dirs. v. Hoover Treated Wood Prods., Inc.*, 136 F.R.D. 100, 107 (E.D. Va. 1991) (recognizing implicitly that the obligation to amend runs up to the time of trial, and that the failure to so amend would violate Rule 26(g)).

may always approach the Court for advance approval.⁶² That precise approach was taken by the Defendants in the Circuit Court of Loudon County, Virginia in the matter of *Global Aerospace, Inc. v. Landow Aviation, LP*.⁶³ In *Global Aerospace*, Defendants requested permission to implement a comprehensive technology-assisted review process that included concrete objectives and substantial transparency.⁶⁴ The Court approved the use of predictive coding, providing the receiving party with the opportunity to challenge the production if there were identifiable deficiencies.⁶⁵

Alternatively, a reasonable review process may be implemented unilaterally.⁶⁶ In *Treppel v. Biovail Corp.*, plaintiff served a request for production of documents upon defendant, and defendant in turn sought plaintiff's participation in establishing the scope and keywords for a search of electronically stored information.⁶⁷ "The plaintiff declined, apparently believing that 'the use of search terms has no application to the standard discovery process of locating and producing *accessible* hard copy and electronic documents.'"⁶⁸ While finding the plaintiff's position to be flawed, the Court determined that the defendant should have implemented its own proposed search:⁶⁹

Yet the plaintiff's recalcitrance does not excuse [defendant,] Biovail's failure to produce any responsive documents whatsoever. Biovail suggested a strategy by which it would search the computer files of Mr. Melnyk, Mr. Cancellara, and Kenneth Howling, its director of investor relations, using search terms: (i) Treppel, (ii) Jerry, (iii) Bank of America, (iv) Banc of America, (v) BAS, and (vi) BofA. Absent agreement from [plaintiff,] Mr. Treppel about a search strategy, Biovail should have proceeded unilaterally, producing all responsive documents located by its search. . . .

62. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 184 (S.D.N.Y. 2012).

63. *Global Aerospace Inc. v. Landow Aviation LP*, Consol. Case No. CL 00061040-00, (Va. Cir. Ct. Loudoun Cty., Apr. 23, 2012) (J. Chamblin.).

64. See Defendants' Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding to the Circuit Court for Loudoun County (2012), available at <http://www.clearwellsystems.com/e-discovery-blog/wp-content/uploads/2012/04/Memo-in-Support-of-Predictive-Coding.pdf> (hereinafter *Global Aerospace Memorandum in Support*).

65. See Order Approving the Use of Predictive Coding for Discovery, <http://pdfserver.amlaw.com/legaltechnology/2012%2004%2023%20-%20DJC%20-%20Order%20Approving%20Predictive%20Coding%20.pdf> (*Global Aerospace Order*).

66. See *Treppel v. Biovail*, 233 F.R.D. 363, 363 (S.D.N.Y. 2006). *But see*, *Novelty, Inc v. Mountain View Mktg. Inc.*, 265 F.R.D. 370, 376 (S.D. In. 2009) ("Unilaterally deciding to conduct a cursory initial search to be followed by 'rolling' productions from subsequent, more thorough, searches is not an acceptable option.").

67. *Treppel*, 233 F.R.D. at 374.

68. *Id.*

69. *Id.* (internal citations omitted).

As in *Global Aerospace*, the Court in *Treppel* recited the right of the receiving party to challenge the search methodology in the event of deficiencies in the production.⁷⁰

This ruling is not an endorsement of the methodology that Biovail has suggested, either in relation to the choice of files to be searched or the terms to be applied. It is, instead, an interim step that is subject to revision once Biovail has responded to the interrogatories relating to its electronic data and the plaintiff has articulated any specific concerns about the scope of the search.

Ultimately, counsel should disclose the use of technology-assisted review to his opponent. In fact, circumstances may make it necessary for counsel to do so in the context of planning discovery under Rule 26(f) or fully adhering to Rule 34. If, however, agreement cannot be reached on the use of technology-assisted review, such a protocol may nevertheless be implemented unilaterally so long as it is reasonable, subject to the right of the receiving party to seek redress for any deficiencies in production.

C. *Training*

There are two primary issues that may arise under Rule 26(g) in the context of training a technology-assisted review tool. The first relates to the manner in which the training set or seed set of documents is generated, *i.e.*, through judgmental selection or random selection. The second is whether that training set or seed set should or must be provided to the receiving party in the disclosure of the technology-assisted review protocol.

The training set” is “[a] sample of documents coded by one or more subject matter experts as relevant or non-relevant, from which a machine learning algorithm [used in the technology-assisted review tool] then infers how to distinguish between relevant and non-relevant documents beyond those in the training set.”⁷¹ Depending on perspective, the training set may or may not differ from the seed set, which is defined as:⁷²

The initial training set provided to the learning algorithm in an active learning process. The documents in the seed set may be selected based on random sampling or judgmental sampling. Some commentators use the term more restrictively to refer only to documents chosen using judgmental sampling. Other commentators use the term generally to mean any training set, including the final training set in iterative training, or the only training set in non-iterative training.

70. *Id.* at 374-75.

71. *The Grossman-Cormack Glossary*.

72. *Id.*

For the purpose of Rule 26(g), the distinctions between judgmental sampling and random sampling, and seed sets and training sets, lie in the fact that that a judgmental sample (and a seed set, to the extent reflective of judgmental sampling only), may not be representative of the entire population of electronic documents within a given collection.⁷³ A judgmental sample is:⁷⁴

A method in which a sample of the document population is drawn, based at least in part on subjective factors, so as to include the “most interesting” documents by some criterion; the sample resulting from such method. *Unlike a random sample, the statistical properties of a judgmental sample may not be extrapolated to the entire population.* However, an individual (such as a quality assurance auditor or an adversary) may use judgmental sampling to attempt to uncover defects. The failure to identify defects may be taken as evidence (albeit not statistical evidence, and certainly not proof) of the absence of defects.

Thus, by definition, a seed set comprised of a judgmental sample may be reflective of the “most interesting” documents but will not necessarily be reflective of the entire population.⁷⁵

Technology-assisted review, however, relies on the training set to establish the basis for differentiating relevant from non-relevant documents across the entire population.⁷⁶ This is true regardless of whether the technology-assisted review utilizes a machine learning approach or a rules-based approach.⁷⁷ For a machine learning approach, “supervised learning algorithms. . . are used to infer relevance or non-relevance of documents based on the coding of documents in a training set.”⁷⁸ Similarly, with a rules-based approach, criteria are developed by an expert to establish whether a document should be classified as relevant or non-relevant, and those rules are applied to the balance of the collection to “emulate the expert(s)’ decision-making process.”⁷⁹ Thus, a rules-based approach also relies on some manner of training set to develop the underlying model that will be propagated to the balance of the collection.

To the extent that the seed set reflects only the “most interesting” documents, and is less than representative of the entire population of

73. *Id.*

74. *Id.* (emphasis added).

75. *Id.*

76. By definition, technology-assisted review is “[a] process for prioritizing or coding a collection of documents using a computerized system that harnesses human judgments of one or more subject matter expert(s) on a smaller set of documents and then extrapolates those judgments to the remaining document collection.” *Id.*

77. *Id.*

78. *Id.*

79. *Id.*

relevant documents, the use of technology-assisted review to propagate the seed decisions throughout the collection may be under-inclusive and run afoul of Rule 26(g). For example, a support vector machine algorithm (one type of supervised learning algorithm), will find a mathematical hyperplane separating relevant from non-relevant documents in the training set.⁸⁰ The remainder of the collection will be classified as either relevant or non-relevant, based on which side of the hyperplane they fall.⁸¹ If the non-hyperplane is derived using a seed set that reflects something less than the full spectrum of relevant documents, the technology-assisted review process will not properly classify those documents which, while relevant, are not similar to the seed set documents.⁸² Knowingly, or even blindly, imposing this type of limitation on relevance violates counsel's obligation to conduct a reasonable effort to locate responsive documents under Rule 26(g).⁸³

This is not to say that the Federal Rules of Civil Procedure, or Rule 26(g) in particular, require perfection.⁸⁴ To the contrary, "the idea is not to make it perfect, it's not going to be perfect. The idea is to make it significantly better than the alternatives. . . ."⁸⁵ "The objective of review in ediscovery is to identify as many relevant documents as possible, while reviewing as few non-relevant documents as possible."⁸⁶ The creation of the seed set is simply another step in the process where counsel is required to "stop and think about the legitimacy of a discovery request, a response thereto, or an objection."⁸⁷

One of the ways to avoid disagreement over the methodology used to train the tool, and the resultant direction of production, is to share the training set with opposing counsel. Nothing in Rule 26(g) requires the exchange of the training set, and nothing in the Federal Rules of Civil Procedure requires the production of non-relevant documents.⁸⁸ However,

80. *Id.*

81. *Id.*

82. *Id.*

83. See *EEOC v. Dolgencorp, LLC*, 2011 U.S. Dist. LEXIS 35195, at *38-39, *51-52 (M.D.N.C. Mar. 31, 2011) (denouncing the EEOC's unilateral relevance determinations); *Kinetic Concepts, Inc. v. Convatec Inc.*, 268 F.R.D. 226, 248 (M.D.N.C. 2010) (same). *Accord* *Treppel v. Biovail*, 233 F.R.D. 363, 374 (S.D.N.Y. 2006) (holding counsel must develop a "reasonably comprehensive search strategy").

84. See, e.g., *Pension Comm. Of Univ. of Montreal Pension Plan v. Banc of Am. Sec.*, 685 F. Supp. 2d 456, 461 (S.D.N.Y. 2010).

85. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 187 (S.D.N.Y. 2012).

86. *Id.* at 189.

87. See FED. R. CIV. P.. 26(g) Advisory Committee's Notes, 1983 Amendments.

88. See FED. R. CIV. P.. 26(g); FED. R. CIV. P. 26(b)(1) ("Parties may obtain discovery regarding any nonprivileged matter that is *relevant* to any party's claim or defense") (emphasis added). See also Memorandum and Order Regarding Disclosure of Predictive Coding Documents, *In re Biomet M2a Magnum Hip Implant Prod. Liab. Litig.*, Cause No. 3:12-MD-2391 (N.D. In. Aug. 21, 2013), <http://www.innd.uscourts.gov/mdl/Disclosure%20of%20docs%20re>

the mere fact that the training set may contain irrelevant documents is not, in itself, a blanket prohibition against, or an impediment to, production.⁸⁹ Moreover, courts have increasingly adopted *The Sedona Cooperation Proclamation* and the call for transparency in electronic discovery in the context of interpreting Rule 26(g).⁹⁰ Consistent with this approach, Magistrate Judge Peck advocated full transparency and the exchange of the training documents in the conduct of technology-assisted review in *Da Silva Moore*.⁹¹

While not all experienced ESI counsel believe it necessary to be as transparent as MSL was willing to be, such 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. 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.

To the extent that development of the seed set reflects attorney work product,⁹² the certification obligations of Rule 26(g) clearly do not require disclosure. As provided in the Advisory Committee Notes:⁹³

Nor does [Rule 26(g)] require a party or an attorney to disclose privileged

%20predictive%20coding%20ord.pdf (hereinafter *Biomet II*)

89. *Accord* Chen-Oster v. Goldman, Sachs & Co., 285 F.R.D. 294, 305 (S.D.N.Y. 2012) (recognizing that there is no legal impediment to ordering production of a database containing both relevant and irrelevant information, as opposed to extracted portions containing solely relevant data) (citing High Point SARRL v. Sprint Nextel Corp., 2011 U.S. Dist. LEXIS 111158, at *39-40 (D. Kan. Sept. 28, 2011); Goshawk Dedicated Ltd. v. Am. Viatical Servs., LLC, 2007 U.S. Dist. LEXIS 82210, at *3-4 (N.D. Ga. Nov. 5, 2007)).

90. *See supra*, text accompanying note 22.

91. *Da Silva Moore*, 287 F.R.D. 182, 192 (internal notes omitted). *Cf. Biomet II* (“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.”).

92. Whether the selection or identification of relevant documents associated with the predictive coding effort comprise work product, however, is not a certainty. For example, in *SEC v. Collins & Aikman Corp.*, it was determined that the segregation of documents into folders corresponding to the allegations of the complaint did not constitute protected work product.:

It is first necessary to determine the level of protection afforded to the selection of documents by an attorney to support factual allegations in a complaint. Such documents are not “core” work product. Core work product constitutes legal documents drafted by an attorney – her mental impressions, conclusions, opinions, and legal theories. This highest level of protection applies to a compilation only if it is organized by legal theory or strategy. The SEC’s theory – that every document or word reviewed by an attorney is “core” attorney work product – leaves nothing to surround the core. The first step in responding to any document request is an attorney’s assessment of relevance with regard to potentially responsive documents. It would make no sense to then claim that an attorney’s determination of relevance shields the selection of responsive documents from production.

256 F.R.D. 403, 410 (S.D.N.Y. 2009).

93. FED. R. CIV. P.. 26(g) Advisory Committee’s Notes, 1983 Amendments.

communications or work product in order to show that a discovery request, response, or objection is substantially justified. The provisions of Rule 26(c), including appropriate orders after in camera inspection by the court, remain available to protect a party claiming privilege or work product protection.

However, this level of transparency can, in fact, be achieved without disclosing privileged or sensitive information that is not relevant to the case and would not otherwise need to be disclosed.⁹⁴

In sum, training the technology-assisted review tool is an important step in the process, and one that requires explicit consideration under Rule 26(g). In order to ensure that counsel is conducting a comprehensive, good faith search, the creation of the seed set should reasonably reflect the full breadth of relevance within the entire ESI collection. Disagreements over training methodology can reasonably be avoided if the producing party is transparent and exchanges the seed set with the receiving party – a process that can be achieved without unnecessarily disclosing sensitive information and without waiving any privileges.

D. Stabilization

Perhaps the most critical question attendant to the use of technology-assisted review for the production of documents is this: what levels of recall and precision are sufficient under Federal Rule of Civil Procedure 26(g)? Not surprisingly, given the incipience of technology-assisted review as a document review and production technique, neither the case law nor the Federal Rules provide a bright-line answer.

The starting point for the analysis of technology-assisted review is the consideration of the relative effectiveness of available alternatives. The preeminent study of the effectiveness of information retrieval techniques in the legal field was the JOLT Study, published in 2011.⁹⁵ The JOLT Study chronicled the results of the Legal Track Interactive Task of the 2009 Text Retrieval Conference (“TREC”), sponsored by the National Institute of Standards and Technology (“NIST”). The JOLT Study evaluated two predominant document review techniques employed in the legal field – linear manual review and keyword search techniques – in terms of both recall and precision, and compared those results against technology-assisted review.⁹⁶ The JOLT Study reported the results of an analysis of the effectiveness of keyword search techniques by David C. Blair and M. E.

94. See *Global Aerospace Memorandum in Support*, p. 11.

95. See *The JOLT Study*.

96. *Id.* The *JOLT Study* also evaluated the harmonic mean of recall and precision, known as the “F1” measure.

Maron in 1985 (the Blair-Maroon Study”).⁹⁷ According to the Blair-Maroon Study, unconstrained but unsophisticated keyword search techniques were able to achieve an average recall of only twenty percent (20%), with an average precision of roughly seventy-nine percent (79%).⁹⁸ The JOLT Study also detailed the TREC 2009 results for both the linear manual review and technology-assisted review efforts.⁹⁹ Under the conditions of the TREC 2009 program, linear manual review achieved an average recall of 59.3%, and an average precision of only 31.7%, although the range of results varied widely for both measures.¹⁰⁰ Conversely, the technology-assisted review techniques evaluated in TREC 2009 achieved an average recall of 76.7%, and an average precision of 84.5%.¹⁰¹ As a result, Grossman & Cormack concluded that “[t]echnology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort.”¹⁰² While the actual effectiveness of all review techniques can, and will, vary with any number of factors, in the context of Rule 26(g), the JOLT Study provides a baseline for comparison and a set of initial objectives against which to evaluate the effectiveness of a technology-assisted review.

In light of the recall figures generally associated with keyword search techniques (20%) and linear manual review (roughly 60%), the Court in *Global Aerospace* accepted the proposed standard of seventy-five percent (75%) recall for a technology-assisted review of roughly 1.3 million documents.¹⁰³ Since the proposed standard obviously exceeded the observed baseline effectiveness of the two traditional alternatives, and the Court expressly reserved the receiving party’s right to seek redress if the production was deficient, there was little debate over the efficacy of the proposed standard of 75% recall.¹⁰⁴

However, while these figures provide helpful benchmarks, the contours of ESI discovery and production are “more than a mathematical count.”¹⁰⁵ As Judge Peck explained in *Da Silva Moore*:¹⁰⁶

97. *Id.* at 18.

98. *Id.* at 18-19.

99. *Id.* at 35-44.

100. *Id.* at 37, Table 7.

101. *Id.*

102. *Id.* at 48.

103. See *Global Aerospace Memorandum in Support*, *Global Aerospace Order*.

104. *Id.*

105. *Accord* *Kleen Prods. LLC v. Packaging Corp. of Am.*, 2012 U.S. Dist. LEXIS 139632, at *46 (N.D. Ill. Sept. 28, 2012). In discussing the selection of custodians for ESI collection, the Court observed that “the selection of custodians is more than a mathematical count. The selection of custodians must be *designed* to respond fully to document requests and to produce responsive, non-duplicative documents during the relevant period.” (emphasis added). *Id.* This obligation to design a discovery protocol to respond fully to production requests applies

[W]here the line will be drawn as to review and production is going to depend on what the statistics show for the results, since proportionality requires consideration of results as well as costs. And if stopping at [producing any certain number of documents] is going to leave a tremendous number of highly responsive documents unproduced, [the] proposed cutoff doesn't work.

* * *

[I]t is unlikely that courts will be able to determine or approve a party's proposal as to when review and production can stop until the computer-assisted review software has been trained and the results are quality control verified. Only at that point can the parties and the Court see where there is a clear drop off from highly relevant to marginally relevant to not likely to be relevant documents. While cost is a factor under Rule 26(b)(2)(C), it cannot be considered in isolation from the results of the predictive coding process and the amount at issue in the litigation.

Consequently, the reasonable search and reasonable inquiry standards of Rule 26(g) will require counsel to evaluate both the effectiveness and the associated proportionality considerations of the review techniques before certifying the result of a technology-assisted review. And whether the results of a technology-assisted review, and a given recall, will constitute a "reasonable search" under Rule 26(g) is for the Court to decide, and is heavily dependent upon the specific facts and circumstances of each case.¹⁰⁷

Biomet is the only decision to address the relationship between proportionality and the effectiveness of a technology-assisted review, and then only implicitly.¹⁰⁸ The principal issue in *Biomet* was whether Biomet, after having first culled their 19.5 million document collection to 2.5 million documents using keyword search techniques followed by de-duplication, would be required to expand their technology-assisted review of the smaller set to encompass the entire collection.¹⁰⁹ In determining that the cost of ingesting and reviewing the entire collection outweighed the likely benefit, the Court focused on the number of relevant documents remaining in the balance of the original collection that had not been

equally to all aspects of document production pursuant to Rule 26(g). *See, e.g., Zander v. Craig Hosp.*, 2011 U.S. Dist. LEXIS 29136, at *20 (D. Colo. Mar. 4, 2011) ("It is equally the litigant's responsibility (and that of its counsel) to determine beforehand the accuracy of any representations that production is complete" (quoting *Fendi Adele S.R.L. v. Filene's Basement, Inc.*, 2009 U.S. Dist. LEXIS 32615 (S.D.N.Y. Mar. 24, 2009)).

106. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 185, 192 (S.D.N.Y. 2012) (internal quotations omitted).

107. *Moore v. Napolitano*, 723 F. Supp. 2d 167, 173 (D.D.C. 2010) (citing the Advisory Committee Notes to FED. R. CIV. P. 26); *Qualcomm, Inc. v. Broadcom Corp.*, 2010 U.S. Dist. LEXIS 33889, at *22 (S.D. Cal. Apr. 2, 2010) (same).

108. *See Biomet I.*

109. *Id.* at 2.

reviewed (an *elusion* test),¹¹⁰ and did not directly undertake a recall analysis.¹¹¹

In contrast, the Steering Committee's request that Biomet go back to Square One (more accurately Square Two, since Biomet first collected the 19.5 million documents) and institute predictive coding at that earlier stage sits uneasily with the proportionality standard in Rule 26(b)(2)(C). Doing so would entail a cost in the low seven-figures. The confidence tests Biomet ran as part of its process suggest a comparatively modest number of documents would be found.

* * *

It might well be that predictive coding, instead of a keyword search, at Stage Two of the process would unearth additional relevant documents. But it would cost Biomet a million, or millions, of dollars to test the Steering Committee's theory that predictive coding would produce a significantly greater number of relevant documents. Even 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, I can't find that the likely benefits of the discovery proposed by the Steering Committee equals or outweighs its additional burden on, and additional expense to, Biomet.

Indeed, the Court did not even address the fact that, *using Biomet's own figures*, Biomet was only producing, at most, some 60.8% of the relevant documents.¹¹²

Ultimately, what constitutes an acceptable level of recall will be based on the totality of the circumstances and, absent agreement, counsel should be prepared to make a proper showing under Federal Rule 26(b)(2)(C) to support any proposed target or actual result. "[T]he case law is replete that attorneys and parties have a duty to make their responses 'accurate and complete.'"¹¹³ As stated above, the objective is *not* perfection – it is to

110. "Elusion [is] the fraction of documents identified as Non-Relevant by a search or review effort, that are in fact Relevant." *The Grossman-Cormack Glossary* at 15.

111. *Biomet I* at 5.

112. See *Biomet I* at 2. Biomet's estimate of the prevalence of the original 19.5 million document collection was 1.92% ± 0.55% (or 374,400 relevant documents). Biomet's elusion estimate of the "unselected" documents (presumed to mean the 15.6 million documents remaining after the keyword cull) was 0.94% ± 0.39% (or 146,640 relevant documents). Thus, recall was [(374,400-146,640) ÷ 374,400], or 60.8%. This figure is likely an over-estimate of actual recall, since it does not account for technology-assisted review step, which would likely locate something less than one hundred percent of the relevant documents.

113. *Kleen Prods. LLC v. Packaging Corp. of Am.*, 2013 U.S. Dist. LEXIS 3016, at *18 (N.D. Ill. Jan. 9, 2013) (quoting *Johnson v. Cook Cnty. Bureau of Health Servs.*, 2010 U.S. Dist. LEXIS 22670, at *14 (N.D. Ill. Mar. 11, 2010)). See also *Chen-Oster v. Goldman, Sachs & Co.*, 285 F.R.D. 294, 306 (S.D.N.Y. 2012) ("[T]he standard for the production of ESI is not perfection.

achieve a significantly better result than available alternatives without nearly as much cost.¹¹⁴ Additionally, the party proposing the limitation must demonstrate the need to limit that discovery (*i.e.*, establishing a recall objective).¹¹⁵ Furthermore, since the results of the technology-assisted review are unknown at the outset, and typically evaluated by some measure of statistical sampling, it may be necessary to support the proportionality analysis through statistical means as well.¹¹⁶ Once that demonstration has

Rather, “[a] responding party must use reasonable measures to validate ESI collected from database systems to ensure completeness and accuracy of the data acquisition.”) (quoting The Sedona Conference, *The Sedona Conference Database Principles: Addressing the Preservation and Production of Databases and Database Information in Civil Litigation*, March 2011 Public Comment Version at 32); *Custom Hardware Eng’g & Consulting, Inc. v. Dowell*, 2012 U.S. Dist. LEXIS 146, at *10-11 (E.D. Mo. Jan. 3, 2012) (“Finally, because Plaintiff properly seeks information that is within the scope of discovery, it is Defendants’ burden to produce *all* responsive information.”) (citing FED. R. CIV. P. 34(b)(2)(A)) (emphasis added).

114. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 185, 189-190 (S.D.N.Y. 2012).

115. *In re Coventry Healthcare, Inc. v. This Document Relates*, 290 F.R.D. 471, 475 (D. Md. 2013) (“The party seeking to lessen the burden of responding to electronic discovery ‘bears the burden of particularly demonstrating that burden and of providing suggested alternatives that reasonably accommodate the requesting party’s legitimate discovery needs.’”) (quoting *Hopson v. Mayor & City Council of Baltimore*, 232 F.R.D. 228, 245 (D. Md. 2005)); *Kleen Products*, 2012 U.S. Dist. LEXIS 139632, at *48 (“[A] party must articulate and provide evidence of its burden.”); *Stambler v. Amazon.com, Inc.*, 2011 U.S. Dist. LEXIS 157125, at *36-37 (E.D. Tx. May 23, 2011) (“The burden of justifying non-production or reduced production should properly fall on [the producing party].”); *Ross v. Abercrombie & Fitch Co.*, 2010 U.S. Dist. LEXIS 47620, at *10-11 (S.D. Ohio May 14, 2010) (“The party claiming that discovery is burdensome does have an obligation to make that claim with specificity.”). *Accord Adair v. EQT Production Co.*, 2012 U.S. Dist. LEXIS 75132, at *14-16 (W.D. Va. May 31, 2012) (producing party obligated to justify limitations of keyword search); *Signature Dev’t, LLC v. Mid-Continent Cas. Co.*, 2012 U.S. Dist. LEXIS 132886, at *48-51 (D.S.D. Sept. 18, 2012) (same).

116. *Accord Larson v. AT&T Mobility LLC*, in which the Court observed:

Guidelines in the electronic discovery realm that contemplate statistical sampling to assist in the cost-benefit analysis required under Federal Rule of Civil Procedure 26(b)(2)(C)(iii) may also help determine what is a “reasonable effort” in the class action context under Rule 23(c)(2). In assessing whether to limit discovery, a court may be required to consider whether “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.” FED. R. CIV. P. Fed. R. Civ. P. 26(b)(2)(C)(iii). One of the Sedona Conference Principles of Proportionality, a set of guidelines that offer a framework for the best electronic discovery practices, provides that “[e]xtrinsic information and sampling may assist in the analysis of whether requested discovery is sufficiently important to warrant the potential burden or expense of its production.” The Sedona Conference® WG1, *The Sedona Conference® Commentary on Proportionality in Electronic Discovery* 291 (“Sedona Commentary”) (2010), available at <http://www.thesedonaconference.org/content/miscFiles/Proportionality2010.pdf>. The commentary to that principle provides as follows:

When asked to limit discovery on the basis of burden or expense, courts must make an assessment of the importance of the information sought. Discovery should be limited if the burden or expense of producing the requested information is disproportionate to its importance to the litigation. Performing such an

been made, the Court has “abundant resources to tailor discovery requests to avoid unfair burden or expense and yet assure fair disclosure of important information[,]” which includes setting an acceptable level of recall and precision.¹¹⁷

Unlike the affirmative obligations associated with recall, however, the Rule 26(g) certification requirement governs precision in a prohibitory manner.¹¹⁸

By signing, an attorney or party certifies that to the best of the person’s knowledge, information, and belief formed after a reasonable inquiry. . . with respect to a discovery. . . response. . . it is. . . not interposed for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation. . . .

Under Rule 26(g), counsel must evaluate the precision of the technology-assisted review process, and ensure that the process is not being conducted in such a way as to unnecessarily include irrelevant documents.¹¹⁹ As a practical matter, the precision of technology-assisted review processes is much greater than the precision of other review alternatives, making it unlikely that irrelevant documents are being produced for some improper purpose in violation of Rule 26(g).¹²⁰ Moreover, the documents identified for production through the technology-assisted review process are generally reviewed before production, further minimizing the production of irrelevant documents in violation of Rule 26(g).¹²¹

assessment can be challenging, given that it may be impossible to review the content of the requested information until it is produced.

In some cases, it may be clear that the information requested is important – perhaps even outcome-determinative. In other cases, courts order sampling of the requested information, consider extrinsic evidence, or both, to determine whether the requested information is sufficiently important to warrant potentially burdensome or expensive discovery.

Sedona Commentary 299 (internal footnote omitted); see Advisory Committee Notes to FED. R. CIV. P.. 26(b)(2) (“[T]he parties may need some focused discovery, which may include sampling of the sources, to learn more about what burdens and costs are involved in accessing the information, what the information consists of, and how valuable it is for the litigation in light of information that can be obtained by exhausting other opportunities for discovery.”).

687 F.3d 109, 130, n. 33 (3d Cir. 2012).

117. See *In re Coventry Healthcare*, 2013 U.S. Dist. LEXIS 39050, at *11-12.

118. FED. R. CIV. P.. 26(g)(1)(B)(ii).

119. *Id.* See also FED. R. CIV. P.. 26(g) Advisory Committee’s Notes, 1983 Amendments.

120. *The JOLT Study* at 18-19.

121. See, e.g., *Biomet I* at 2 (“After one round of ‘find more like this’ interaction between the attorneys and the software, the contract attorneys (together with other software recommended by Biomet’s ediscovery vendor) reviewed documents for relevancy, confidentiality, and privilege.”); *Global Aerospace Memorandum in Support*.

Even though the levels of recall and precision that are typically obtained through the use of technology-assisted review will most often equal or exceed those that can be achieved through either linear manual review or keyword search techniques, there simply is no standard for either measure that will universally satisfy Rule 26(g) certification obligations. Rather, because some fraction of relevant materials will necessarily be left behind, counsel will be required to evaluate the effectiveness of the review through the lens of the proportionality considerations in Rule 26(b)(2)(C). Only when the burden, generally expressed in terms of the cost of retrieving additional documents (or, in the case of precision, excluding additional documents), exceeds the benefits of those documents, will recall and precision levels be sufficient under the “reasonable search” requirements of Rule 26(g).

E. Validation

Finally, once the technology-assisted review process has been completed and the review set of relevant documents has been generated, Rule 26(g) requires some measure of validation to ensure the effectiveness of the process and the adequacy of the result. To do this consistent with Rule 26(g) certification and the obligation of “reasonable inquiry,” counsel must be thoroughly familiar with the various statistical alternatives, and must be prepared to justify the selected methodology and associated statistical parameters.

As discussed above, the Rule 26(g) certification can only be made upon a reasonable inquiry.¹²² The duty of reasonable inquiry is only satisfied if counsel undertakes an investigation of the facts supporting the certification, and the investigation and the conclusions drawn from that investigation are reasonable.¹²³ The goal of technology-assisted review is to achieve an acceptable level of recall and precision, in light of proportionality considerations, based on a statistical quality control analysis.¹²⁴ Thus, to satisfy Rule 26(g), it will be necessary for counsel to investigate and understand the statistical analysis of recall and precision, and to reach a reasonable conclusion that the technology-assisted review process was effective. This is not unlike the attorney’s obligation to ensure that keyword search techniques have effectively located responsive

122. FED. R. CIV. P.. 26(g)(1).

123. *Grider v. Keystone Healthplan Central, Inc.*, 580 F.3d 119, 140, n.23 (3d Cir. 2009) (citing 6 *Moore’s Federal Practice*, § 26.154[2][a]); *Chapman & Cole v. Itel Container Intl. B.V.*, 865 F.2d 676, 685-86 (5th Cir. 1989).

124. *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 185, 190, 191, (S.D.N.Y. 2012). *See also* Roitblat, *supra* note 29.

documents.¹²⁵

To ensure that a statistical validation protocol meets counsel's duty of reasonable inquiry, there are two principal factors that need to be considered. The first factor for consideration is the measure that will be evaluated, *e.g.*, recall, precision, elusion, etc.¹²⁶ Because it would be time-consuming and costly to review all of the documents in an ESI collection, which would obviously eliminate the primary benefit of technology-assisted review, these measures are typically estimated through the use of sampling.¹²⁷ The second factor for consideration is the extent to which the sample can be considered representative of the entire collection, which is reflected in the confidence level and confidence interval associated with the sample.¹²⁸

Regardless of the specific measure being evaluated, utilizing a sample having a confidence level of either 95% or 99%, and a nominal confidence interval of between $\pm 2\%$ and $\pm 5\%$ will satisfy the reasonable inquiry requirements of Rule 26(g).¹²⁹ Since the confidence level is only indicative of the repeatability of the sample, and is not indicative of the performance of the technology-assisted review itself, there is little benefit to be gained by using a 99% confidence level as opposed to a 95% confidence interval.¹³⁰ Similarly, "[b]ecause the size of the confidence interval tells us

125. *See, e.g.*, William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 135-36 (S.D.N.Y. 2009) (requiring quality control testing of keyword search results to assure accuracy in retrieval, and citing Magistrate Judge Facciola's warning from *United States v. O'Keefe*, 537 F. Supp. 2d 14, 24 (D.D.C. 2008), that the issue of whether a keyword search yields the anticipated results is a "complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics.").

126. Roitblat, *supra* note 29, at 1-4. In his paper, Dr. Roitblat identifies and discusses some sixteen statistical measures that can be used to varying degrees of effectiveness to evaluate the efficacy of a technology-assisted review.

127. *Id.* at 5.

128. *Id.* at 8.

129. *See In re Actos (Pioglitazone) Prods. Liab. Litig.*, 2012 U.S. Dist. LEXIS 187519, at *26-27 (W.D. La. July 27, 2012) ("The application's estimates of richness use a confidence level of 95%... with an error margin of plus or minus 4.3%."); *Da Silva Moore*, 287 F.R.D. at 186 ("The parties agreed to use a 95% confidence level [$\pm 2\%$] to create a random sample of the entire email collection") *See also Biomet I* at 2 (accepting statistical samples having a confidence level of 99% and a maximum confidence interval of $\pm 2\%$).

It is important to realize that the nominal confidence interval associated with a sample of any given size is a worst-case error margin as if the observed value were 50%, and the actual confidence interval will likely be less. *Actos*, 2012 U.S. Dist. LEXIS 187519 at p. 26-27. Thus, whereas the 500 document sample proposed in *Actos* would have a confidence level of 95% and a confidence interval of $\pm 4.3\%$ if the observed richness was 50%, the actual confidence interval if the richness was actually 10% would be $\pm 2.6\%$.

Id.

130. Roitblat, *supra* note 29, at 8. *See also The Grossman-Cormack Glossary* at 12 (defining Confidence Level).

so little about the quality of [the technology-assisted review], there may not be a lot of value derived from selecting narrower confidence intervals.”¹³¹ Increasing the confidence level and tightening the confidence interval multiplies the level of effort that must be expended to review the sample for validation purposes.¹³² Accordingly, establishing the confidence level and confidence interval for validation sampling invokes the proportionality considerations discussed above.

Although there are a wide variety of statistical measures of the efficacy of technology-assisted review, recall and precision are the most appropriate for the reasonable inquiry requirements underlying Rule 26(g) certification. Recall establishes the percentage of relevant documents in an ESI collection that are being produced to the requesting party as a result of the technology-assisted review.¹³³ Thus, establishing recall (in light of proportionality, as discussed above), is critical to demonstrating whether a reasonable search was conducted under Rule 26(g).¹³⁴ Precision, on the other hand, establishes the fraction of relevant documents in the production,¹³⁵ consequently establishing that the production is not being interposed for any improper purpose under Rule 26(g).¹³⁶

Other statistical measures are not as useful for assessing whether the Rule’s reasonable inquiry requirements have been met. For example, although the F_1 value incorporates *both* recall and precision, inasmuch as it is the harmonic mean,¹³⁷ F_1 does not directly reflect the level of compliance with Rule 26(g) and it also complicates the proportionality analysis. Similarly, accuracy is not an adequate indicator of the efficacy of technology-assisted review in the context of Rule 26(g). Accuracy simply reflects the percentage of documents that were properly coded as either relevant or not relevant.¹³⁸ Because the fraction of documents that are not

131. Roitblat, *supra* note 29, at 10.

132. For example, for a 1 million document set, a sample of 384 documents would be necessary to characterize the collection to a confidence level of 95% and a confidence interval of $\pm 5\%$. <http://www.surveysystem.com/sscalc.htm>. Increasing the confidence level to 99% and tightening the confidence interval to $\pm 2\%$ would increase the sample more than ten times, to 4143. *Id.*

133. *The Grossman-Cormack Glossary* at 27 (defining Recall); Roitblat, *supra* note 29, at 2.

134. *See supra*, text accompanying notes 18-19.

135. *The Grossman-Cormack Glossary* at 25 (defining Precision); Roitblat, *supra* note 29, at 2.

136. FED. R. CIV. P.. 26(g)(1)(B)(ii).

137. “ F_1 [is] the harmonic mean of recall and precision, often used in Information Retrieval studies as a measure of the effectiveness of a search or review effort, which accounts for the tradeoff between recall and precision. *The Grossman-Cormack Glossary* at 16. The harmonic mean is “the reciprocal of the average of the reciprocals of two or more quantities. If the quantities are named a and b, their harmonic mean is $2/(1/a+1/b)$.” *Id.* at 18.

138. *Id.* at 8 (defining Accuracy).

relevant in most cases significantly outweighs the fraction of relevant documents, accuracy can be very misleading and not at all indicative of the reasonableness of the search for relevant documents alone.¹³⁹ Finally, elusion can similarly be misleading, because it quantifies the number of relevant documents that have been missed by the technology-assisted review (*i.e.*, documents that are not being produced), and reveals nothing about the fraction of relevant documents that have been located and are in fact being produced.¹⁴⁰ While elusion can be used in conjunction with richness or prevalence to determine recall, it is an unnecessarily indirect calculation, given that recall can be calculated directly.¹⁴¹

In estimating recall, the Rule 26 proportionality considerations should guide the manner of calculation. Recall represents the fraction of the entire population of relevant documents that is correctly being identified as relevant by the technology-assisted review process.¹⁴² Therefore, the pristine method of calculating recall is to take a sample of *that population*, *i.e.*, those documents, and only those documents, that are believed by counsel to be relevant – and determine what fraction of that sample has been correctly identified as relevant through the technology-assisted review process.¹⁴³ The size of that sample is dictated by the confidence level and confidence interval being utilized in the validation, and is relatively manageable even at the most stringent confidence criteria.¹⁴⁴ For example, for a 20 million document collection it will only be necessary to collect a 4,159 document random sample to characterize the relevant document population with a confidence level of 99% and a maximum confidence interval of $\pm 2\%$.¹⁴⁵ However, the ability to *find* 4,159 relevant documents to serve as a sample of the population depends entirely on the richness of the collection. Thus, for example, counsel would be required to review roughly 42,000 documents to find 4,159 relevant documents among a collection with ten percent (10%) richness, and nearly 420,000 documents in a

139. *Id.*

140. *Id.* at 15 (defining Elusion).

141. Roitblat, *supra* note 29, at 4.

142. *The Grossman-Cormack Glossary* at 27 (defining Recall); Roitblat, *supra* note 29, at 2.

143. See *In re Actos (Pioglitazone) Prods. Liab. Litig.*, 2012 U.S. Dist. LEXIS 187519, at *28 (W.D. La. July 27, 2012). In *Actos*, the validation protocol required the parties to review documents during an initial Assessment Phase to create a Control Set which would be used to evaluate the results of the technology-assisted review. *Id.* The Assessment Phase was to continue until the Control Set contained at least 385 *relevant* documents – a sample of sufficient size to establish recall within the target confidence level of 95% \pm 5%. *Id.*

144. See <http://www.surveysystem.com/sscalc.htm>. See generally Roitblat, *supra* note 29.

145. Creative Research Systems, The Survey System Sample Size Calculator, <http://www.surveysystem.com/sscalc.htm>.

collection with one percent (1%) richness.¹⁴⁶ Since there are alternative means of calculating recall that do not require such a significant effort, whether this level of exactitude is required to satisfy the reasonable inquiry requirements of Rule 26(g) must be evaluated against the proportionality considerations in Rule 26(b)(2)(C)(iii).

For example, rather than looking solely at the relevant document population, counsel may choose to calculate recall indirectly by sampling (1) the production, and (2) either the original ESI collection, or the documents identified by the technology-assisted review as not relevant (*i.e.*, the null set). Sampling the production will establish the number of relevant documents produced. Sampling the original ESI collection will establish the total number of relevant documents available to be produced; and sampling the null set will establish the number of relevant documents that are not being produced. Recall can then be calculated as the fraction of the total number of relevant documents that are actually being produced in the production set.¹⁴⁷ This is a more cost effective validation protocol when richness is low, because the size of the samples that must be reviewed is dictated by the confidence criteria, and it is not necessary to review any documents other than those comprising the samples.

Ultimately, Rule 26(g) requires counsel to conduct a “reasonable inquiry” to confirm the effectiveness of a technology-assisted review in performing a “reasonable search” for documents requested in discovery. To satisfy the obligation of reasonable inquiry, counsel must undertake a statistical validation of the results of the technology-assisted review, generally focusing on the resultant recall and precision. The precise methodology, and the parameters surrounding that methodology, will be governed by the proportionality considerations of Rule 26(b)(2)(C)(iii).

III. CONCLUSION

Federal Rule of Civil Procedure 26(g) provides the governing principles for the implementation of technology-assisted review. At its core, Rule 26(g) requires two things: a reasonable search and a reasonable inquiry to ensure the efficacy of that search. Since every step of the technology-assisted review process impacts either the nature or efficacy of the search, or the level of inquiry, Rule 26(g) applies throughout the

146. See *The Grossman-Cormack Glossary* at 28, 26 (defining Richness and Prevalence).

147. For example, in *Global Aerospace*, recall was calculated by sampling the production and the set of documents presumptively identified by technology-assisted review as not relevant. See *Global Aerospace Memorandum in Support*. Recall equated to the number of relevant documents in the set being produced divided by the total number of relevant documents in both the set being produced and the set withheld as not relevant.

process, from collection through validation. As with all discovery, what is reasonable in the application of Rule 26(g) to technology-assisted review is governed primarily by the proportionality considerations of Rule 26(b)(2)(C)(iii). Consequently, before signing the Rule 26(g) certification, counsel must be certain that he has considered the impact of every step of the technology-assisted review process on the ability to locate relevant documents, as well as the cost and benefit of alternatives that might improve that ability. Failing to do so may not only impair the efficacy of the technology-assisted review, but may also result in the court's imposition of mandatory sanctions.