LAW IN THE AGE OF EXABYTES: SOME FURTHER THOUGHTS ON 'INFORMATION INFLATION' AND CURRENT ISSUES IN E-DISCOVERY SEARCH

By Jason R. Baron*

Cite as: Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, XVII RICH. J.L. & TECH. 9 (2011), http://jolt.richmond.edu/v17i3/article9.pdf.

I. INTRODUCTION

[1] In 2007, in the pages of this Journal, George L. Paul and I posed a question to the legal profession at large, to wit: can the legal system adapt to the new reality of an era of rapid inflation in the amount of electronically stored information (ESI) at issue in civil litigation?[1] After surveying the history of technological innovation that led to an explosion

* Director of Litigation, National Archives and Records Administration, Washington, D.C.; Co-Chair, The Sedona Conference® Working Group on Electronic Document Retention and Production; Adjunct Professor, University of Maryland. B.A. *magna cum laude,* Wesleyan University (1977), J.D., Boston University School of Law (1980). This article represents a reworking and expansion of a presentation at the 2010 Conference on Civil Litigation held at Duke Law School in May 2010. *See* "Controlling for 'Discovery Excess' in the Age of ESI: Some Thoughts on The Legal Profession Embracing Search Analytics, Process Quality, and Structured Cooperation" (unpublished paper), *available at* http://civilconference.uscourts.gov/LotusQuickr/dcc/Main.nsf/h_Discussion/898C52B07ADD17DD8525773B004941AC/?OpenDocument. The views expressed herein are my own, and do not necessarily represent any institution, public or private, with which I am associated.

[1] *See* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, ¶ 2 (2007), http://law.richmond.edu/jolt/v13i3/article10.pdf. I remain grateful to George Paul, who was responsible for bringing to my attention the potential application to the legal space of the cosmological metaphor. *See generally* ALAN H. GUTH, THE INFLATIONARY UNIVERSE: THE QUEST FOR A NEW THEORY OF COSMIC ORIGINS (1997).

of new data, we proceeded to discuss various legal strategies for success in our current inflationary epoch.[2] These strategies included: consideration of new and emerging ways in which to think about search and information retrieval in light of the limitations of traditional keyword searching the legal profession had come to rely upon;[3] greater use of sampling and iteration so as to ensure greater quality;[4] the use of multiple meet-and-confers to produce a "virtuous feedback cycle" of cooperation amongst counsel;[5] predicting congressional enactment of Federal Rule of Evidence 502, enabling parties to leverage resources by providing large amounts of data in open discovery;[6] and finally, making tentative predictions about the future of artificial intelligence as applied to information law problems.[7]

[2] In connection with a symposium held at the University of Richmond School of Law in March 2011, the Editors of this Journal invited me to provide some further thoughts on the arc of what has happened in the past four years, with respect to at least some of the ideas first presented in the 2007 article concerning advances in search and information retrieval law. What follows is at best a brief and informal "interim progress report," sketched out with a greater interest in being provocative than comprehensive.

II. THE EXPANDING ESI UNIVERSE

[3] With respect to the central underlying assumption of our 2007 article, *Information Inflation: Can the Legal System Adapt?*, the intervening four years should put to rest any doubts concerning the accelerating expansion of the overall universe of data in which the world

[2] Paul & Baron, *supra* note 1, ¶¶ 7-56.

[3] *Id.* ¶¶ 36-40.

[4] *Id.* ¶¶ 47-49.

[5] *Id.* ¶¶ 50-56.

[6] *Id.* ¶ 62.

[7] Paul & Baron, *supra* note 1, ¶ 65.

is awash.[8] The article's proposals for re-engineering the discovery process are playing out against the backdrop of profound, transformational change.[9] This change is occurring both in the legal profession as well as in the greater world at large, due to advances in computer science, the wealth of cyber-networks in which society is immersed, and the inflationary growth of ESI.[10] The world as we know it is simply awash in exabytes[11] of data, and the pace of its accumulation is only accelerating.[12]

[4] In the world of investigations and litigation, information inflation has manifested itself with a new "watermark" in terms of volume. For example, in the Report, in multiple volumes, of the Examiner in the Lehman Brothers Holdings Chapter 11 case in Bankruptcy Court in NY, dated March 11, 2010, the examiner was tasked with culling down a universe of 350 billion pages–three petabytes of data–to review for the

[8] *See generally* John F. Gantz et al., *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011*, EMC (Mar. 2008), http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.

[9] *See* Richard L. Marcus, *E-Discovery & Beyond: Toward Brave New World or 1984?*, 25 REV. LITIG. 633, 635-66 (2006).

[10] *See id.*

[11] 1 exabyte = 1,000,000,000,000,000,000, or, 10^{18} bytes = 1 billion gigabytes = 1 million terabytes = 1 thousand petabytes. *See Exabyte*, TECHTERMS, http://www.techterms.com/definition/Exabyte (last visited Mar. 22, 2011).

[12] *See* Martin Hilbert & Priscila López, *The World's Technological Capacity to Store, Communicate, and Compute Information*, SCIENCEXPRESS, 1 (Feb. 10, 2011), http://www.sciencexpress.org/10February2011/Page1/10.11126/science.1200970; *see also* Jason R. Baron & Ralph C. Losey, *e-Discovery: Did You Know?*, YOUTUBE (Feb. 11, 2010), http://www.youtube.com/watch?v=bWbJWcsPp1M. The Hilbert study found that as of 2007, mankind had the capability to store 290 exabytes (2.9×10^{20} bytes), and that the rate of increase in globally stored information is 23% a year. Hilbert & López, *supra.* For some perspective, this combined capacity "is approaching order of magnitude of the roughly 10^{23} bits stored in the DNA of a human adult, but it is still miniscule compared to the 10^{90} bits stored in the observable universe." *Id.* at 5 (internal citations omitted). The authors concluded the study by noting that "the world's technological information processing capacities are quickly growing at clearly exponential rates." *Id.*

purpose of writing his report.[13] The Examiner narrowed the collection by selecting custodians and using dozens of separate Boolean searches to collect in excess of five million documents–representing more than forty million pages–for review, which were then subject to two further levels of manual review by seventy-plus attorneys.[14] It is not clear whether any of the most up to date, sophisticated search techniques were employed.[15]

[5] As Judge Scheindlin recently opined in *Pension Committee v. Banc of America*, all lawyers now live "[i]n an era where vast amounts of electronic information is available for review," and as such "discovery in certain cases has become increasingly complex and expensive."[16] While the Judge's observations are obviously sound, advocacy by others that e-discovery is an all-encompassing problem needing correction through major rules reform is, in my view, ill-advised.[17] The challenge facing the legal profession today is *not* a matter of dealing with discovery abuse or excessiveness *per se*, at least not to the extent that e-discovery is considered the culprit to blame.[18] Rather, the greater challenge is how

[13] *See* 7 Report of Anton R. Valukas, Examiner app. 5 at 1, *In re* Lehman Brothers Holdings Inc., No. 08-13555 (JMP) (Bankr. S.D.N.Y., Mar. 11 2010), ECF No. 7531, *available at* http://lehmanreport.jenner.com/VOLUME%207%20-%20APPENDICES%202-7.pdf.

[14] *Id.* at 30-31.

[15] *See id.* at 30-32.

[16] Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC, 685 F. Supp. 2d 456, 461 (S.D.N.Y. 2010).

[17] *See* INST. FOR THE ADVANCEMENT OF THE AM. LEGAL SYS., FINAL REPORT ON THE JOINT PROJECT OF THE AMERICAN COLLEGE OF TRIAL LAWYERS TASK FORCE ON DISCOVERY AND THE INSTITUTE FOR THE ADVANCEMENT OF THE AMERICAN LEGAL SYSTEM 2 (2009), *available at* http://www.actl.com/AM/Template.cfm?Section=Home&template=/CM/ContentDisplay.cfm&ContentID=4053 (recommending significant changes to the Federal Rules of Civil Procedure).

[18] Litigation has seen its fair share of abuse or excessiveness with respect to the discovery of ESI, but these instances stem from a party's action or inaction and find suitable remedy in the courts. *See, e.g.*, *In re* Seroquel Prods. Liab. Litig., 244 F.R.D. 650, 665 (M.D. Fla. 2007) (sanctioning a defendant for its "failure to [timely] produce 'usable' or 'reasonably accessible' documents").

best to reasonably (not perfectly) manage the exponentially growing amount of ESI caught in, and subject to, modern-day discovery practice.[19] The answer lies principally in culture change (i.e., fostering cooperation strategies), combined with savvier exploitation of a range of sophisticated software and analytical techniques.

[6] Indeed, as The Sedona Conference recognized in its 2009 piece *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process*, "[t]he legal profession is at a crossroads: the choice is between continuing to conduct discovery as it has 'always been practiced' in a paper world – before the advent of computers [and] the Internet . . . or, alternatively, embracing new ways of thinking in today's digital world."[20] To meet the challenge of the exploding volume and complexity of potential electronic evidence, lawyers must embrace new technologies, adopt high quality standards for their work that ensure accuracy and completeness, and begin to think about new ways of approaching structured cooperation within the bounds of the adversary system.[21] Importantly, all of the above can be accomplished within the framework of the existing Federal Rules of Civil Procedure.[22]

[7] For the purposes of this Article, the reader should view the new methods, tools and techniques as falling into three groupings, which are not mutually exclusive. The first grouping pertains to methods of

[19] *See generally* Patrick Oot et al., *Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. U. L. REV. 533, 534-35, 545 (2010).

[20] The Sedona Conference, *The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process,* 10 SEDONA CONF. J. 299, 302 (2009) [hereinafter *Sedona Quality Commentary*].

[21] *See id.* at 303-04.

[22] For thoughtful critiques of the call for rules reform, see Milberg LLP & Hausfeld LLP, *E-Discovery Today: The Fault Lies Not In Our Rules* . . ., 2011 FED. CTS. L. REV. 4 (Feb. 2011), http://www.fclr.org/fclr/articles/html/2010/hausfeld.pdf; Paul W. Grimm, The State of Discovery Practice in Civil Cases: Must the Rules Be Changed to Reduce Costs and Burdens, or Can Significant Improvements Be Achieved Within the Existing Rules? (unpublished article), *available at* http://civilconference.uscourts.gov (follow "Papers Submitted by Conference Panelists"; then follow second "Papers" hyperlink) (last visited Feb. 24, 2011).

advanced search techniques, including those now associated with various forms of concept searching, the use of clustering algorithms, and what has come to be termed "predictive coding."[23] The second grouping involves techniques of greater interaction with opposing counsel using an iterative, tiered or phased approach to discovery search issues that advances the idea of proportionality, helps solve possible ethical issues, and otherwise achieves the overall aims of Federal Rules of Civil Procedure 1.[24] The third grouping concerns new ways of approaching the process of discovery through better project management, sampling of data, and quality control.[25]

[8] With only limited acknowledgement in the form of reported cases, such tools and methods have been shown to be cost effective in managing the burdens of discovery associated with huge volumes of ESI found on client or client-controlled computer servers and networks, including those found in "cloud computing" environments.[26] Use of these methods would help mitigate the high costs of e-discovery in the disproportionately small group of cases that, due to their e-discovery aspect and inherent

[23] *See, e.g.*, *E-Discovery Institute Survey on Predictive Coding*, EDISCOVERY INST., 2 (Oct. 1, 2010), http://www.ediscoveryinstitute.org/pubs/PredictiveCodingSurvey.pdf [hereinafter *Survey on Predictive Coding*] (explaining that predictive coding is "a combination of technologies and processes in which decisions pertaining to the responsiveness of records gathered or preserved for potential production purposes . . . are made by having reviewers examine a subset of the collection and having the decisions on those documents propagated to the rest of the collection without reviewers examining each record.").

[24] *See* FED. R. CIV. P. 1 (noting that the Federal Rules of Civil Procedure are intended "to secure the just, speedy, and inexpensive determination of every action and proceeding").

[25] *See Sedona Quality Commentary*, *supra* note 20, at 302-03.

[26] *See* The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 195 (2007) [hereinafter *Sedona Search Commentary*]; *see also* William Jeremy Robison, *Free at What Cost?: Cloud Computing Privacy Under the Stored Communications Act*, 98 GEO. L.J. 1195, 1199 (2010) (defining "cloud computing" as "the ability to run applications and store data on a service provider's computers over the Internet, rather than on a person's desktop computer").

complexity, tend to take up an inordinate amount of the courts' time and resources.[27]

III. THE UNFOLDING LAW OF SEARCH AND INFORMATION RETRIEVAL

[9] As discussed in *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, beginning in the 1980s (at the dawn of the age of large computerized databases), through the 1990s and the introduction of e-mail, networks, and the Internet, and arguably right up until December 2006, lawyers placed themselves in the equivalent of the Dark Ages with respect to the use of search terms.[28] This hindrance unfolded for two distinct reasons. First, through experience with Westlaw and Lexis, attorneys had limited knowledge of how to utilize Boolean commands or anything more advanced than the use of simple keywords.[29] Second, rarely, if ever, did counsel consider having an honest exchange to determine how one or both parties in a case would go about the very basic job of searching for relevant electronic evidence, as practitioners considered such a discussion tantamount to discussing attorney work-product.[30] The ramifications of such thinking were evident in *United States v. Philip Morris*, in which government attorneys searched a database of thirty-two million Clinton-era White House e-mail records.[31] This review was based, for the most part, on unilateral selection of keywords without prior disclosure to counsel for the combined tobacco

[27] *See*, *e.g.*, Helmert v. Butterball, LLC, No. 4:08CV00342 JLH, 2010 WL 2179180, at *4-6 (E.D. Ark. May 27, 2010). For further discussion of the time the court devoted to the e-discovery aspects of this case, see *infra* notes 52-61 and accompanying text.

[28] *See Sedona Search Commentary*, *supra* note 26, at 197-203.

[29] *See id.* at 199-202.

[30] *See* The Sedona Conference, *The Sedona Conference Cooperation Proclamation*, 10 SEDONA CONF. J. 331, 332 (Supp. 2009).

[31] *See* Allen Weinstein, Archivist of the United States, Ask the White House (Jan. 7, 2006) http://archives.gov/about/archivist/ask-the-white-house.html.

defendants.[32] The result was one percent of the universe (approximately 200,000 "hits" based on keyword terms used) to manually search through.[33]

[10] With the expected size of the cumulative archived White House e-mail universe on track to swell to one billion by 2017,[34] manual searching, *even after automated methods have been used based on keywords,* will become an increasingly resource-intensive endeavor.[35] This is due to the burden of wading through the false positive noise of irrelevant documents, as well as the arguably greater problem of false negatives (i.e., the search method missing relevant evidence).[36] Extrapolating this out, the legal profession likely will see cases experience a "1% crisis," meaning that even one percent of a large quantity of data (a quantity measuring in terabytes, petabytes and beyond) that remains after automated keyword searching, is *still* too large a haystack for manual searches (even at contract attorney rates).[37]

[11] In line with the 2007 article's prediction that information retrieval would be an increasing area of prominence, the past four years have seen a growing cottage industry of case law,[38] commentaries,[39] and research,[40]

---

[32] *See* Jason R. Baron, *Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery*, 6 SEDONA CONF. J. 237, 239 (2005).

[33] *See id*.

[34] Paul & Baron, *supra* note 1, ¶ 19.

[35] *See id*. ¶ 20.

[36] *See id.* ¶ 56 n.134.

[37] *See id*. ¶ 20.

[38] *See, e.g.*, Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 256-61 (D. Md. 2008) (discussing the *Sedona Search Commentary's* practice pointers and stating that "while it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying exclusively on such searches for privilege review"); *see also* William A. Gross Constr. Ass'n v. Am. Mfrs. Mut. Ins. Co., 256

acknowledging the limitations of keyword searching and discussing

F.R.D. 134, 134 (S.D.N.Y. 2009) ("This Opinion should serve as a wake-up call to the Bar . . . about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or 'keywords' to be used to produce emails or other electronically stored information."); Equity Analytics, LLC v. Lundin, 248 F.R.D. 331, 333 (D.D.C. 2008) (citing *O'Keefe*, 537 F. Supp. 2d at 22-23) (questioning the effectiveness of keyword terms used in searches of ESI); United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008) (finding that in light of the interplay between computer technology, statistics and linguistics, properly constructing a search protocol calls for some measure of special expertise); *In re* Seroquel Prods. Liab. Litig., 244 F.R.D. 650, 660 n.6, 662 (M.D. Fla. 2007) (criticizing the defendant's failure to cooperate on search terms, or to "assure reasonable completeness and quality control" in the search for relevant material).

[39] *See Sedona Search Commentary*, *supra* note 26, at app. 217; *Sedona Quality Commentary, supra* note 20, at 302, 313-18; Lori Heilman, Comment, *Federal Courts' Reactions to Inadequate Keyword Searches: Moving Toward a Predictable and Consistent Standard for Attorneys Employing Keyword Searches*, 78 U. CIN. L. REV. 1103, 1105-06, 1127 (2010); *EDRM Search Guide*, ELECTRONIC DISCOVERY REFERENCE MODEL, http://edrm.net/resources/guides/edrm-search-guide (last visited Feb. 22, 2011); *see also* RALPH C. LOSEY, INTRODUCTION TO E-DISCOVERY: NEW CASES, IDEAS, AND TECHNIQUES 245-46, 261-62 (2009); Jason R. Baron & Edward C. Wolfe, *A Nutshell on Negotiating E-Discovery Search Protocols*, 11 SEDONA CONF. J. 229, 232 (2010); Maura R. Grossman & Terry Sweeney, *What Lawyers Need to Know About Search Tools*, NAT'L L. J., Aug. 24, 2010, *available at* http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202470952987&slreturn=1&hbxlogin=1; Gregory L. Fordham, *Using Keyword Search Terms in E-Discovery and How They Relate to Issues of Responsiveness, Privilege, Evidence Standards and Rube Goldberg*, 15 RICH. J.L. & TECH. 8, ¶¶ 5-8, 11, 56 (2009), http://jolt.richmond.edu/v15i3/article8.pdf; H. Christopher Boehning & Daniel J. Toal, *In Search of Better E-Discovery Methods*, N.Y. L.J., Apr. 23, 2008, *available at* http://www.law.com/jsp/cc/PubArticleCC.jsp?id=900005509469; Mia Mazza et al., *In Pursuit of FRCP 1: Creative Approaches to Cutting and Shifting the Costs of Discovery of Electronically Stored Information,* 13 RICH. J.L. & TECH. 11, ¶¶ 46-51 (2007), http://jolt.richmond.edu/v13i3/article11.pdf.

[40] *See, e.g.*, Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review,* 17 RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf; *see also* Douglas W. Oard et al., *Evaluation of Information Retrieval for E-discovery*, 18 ARTIFICIAL INTELLIGENCE & L. 347, 353-54, 365 (2010), http://www.springerlink.com/content/m700w2k26n264u01/; Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70 (2010), *available at* http://onlinelibrary.wiley.com/doi/10.1002/asi.21233/full.

alternative forms of search. This sub-genre of ESI law has made some headway in identifying the power and future promise of automated methods used in e-discovery to reduce cost and improve results,[41] but the pace of the reported opinions remains well behind technological progress as practiced by the most agile firms and individuals.[42] Indeed, many lawyers are practicing "arbitraging" by exploiting new technologies in their everyday e-discovery practice, as compared to yesterday's methods used by their opponents.[43] Doing so results in achieving a competitive advantage in litigation while at the same time acting consistently with the aim of Federal Rule of Civil Procedure 1.[44]

[12] The one percent issue is addressed in the first practice pointer of The Sedona Conference's *Search Commentary,* in which the following overarching observation is made:

> In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary.[45]

However, the Commentary went on to add the following caveat in Practice Point 5:

> The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may

[41] *See Sedona Search Commentary*, *supra* note 26, at 195.

[42] *See* Heilman, *supra* note 39, at 1127.

[43] *See infra* note 110.

[44] *See* FED. R. CIV. P. 1.

[45] *Sedona Search Commentary*, *supra* note 26, at 208.

> produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.[46]

[13] How have such observations played out in reported cases? Notwithstanding all the newfound attention surrounding the subject of search, there appears to be a substantial gap between what constitutes best practices, and what passes for the norm in many reported cases.[47] This discrepancy, along with difficult fact patterns and sloppy lawyering, predominate in published decisions. Counsel have run the gamut between flamboyantly proposing up to 1,000 keywords for searching,[48] to saying with a straight face they had met the reasonable expectations of a judge in a different case by searching against only *one* keyword.[49] In the case of *In re Fannie Mae Securities Litigation*, the Court of Appeals found that government counsel had stipulated to an open-ended list of keywords, which produced 660,000 hits to be recovered from disaster recovery backup tapes, at a cost of $6 million, which constituted nine percent of the agency's budget; however, the agency was still held in contempt.[50] These examples show a wide gap in the understanding of what present day e-discovery demands entail. In each case, counsel's apparent lack of savvy contributed heavily to either the excessiveness of the effort or the prolonging of discovery at undue expense and delay.[51]

[14] The cases also fall along a spectrum of activism, in which judges, with various degrees of exasperation regarding the inability of parties to reach agreement or live up to prior representations, intervene and "get

---

[46] *Id.* at 211.

[47] *Compare* William A. Gross Constr. Assocs. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 134 (S.D.N.Y. 2009), *with* Capitol Records, Inc. v. MP3tunes, LLC, 261 F.R.D. 44, 47-48 (S.D.N.Y. 2009).

[48] *See William A. Gross Constr. Assocs.*, 256 F.R.D. at 134.

[49] *See Capitol Records*, 261 F.R.D. at 48.

[50] *See In re* Fannie Mae Litig., 552 F.3d 814, 817, 824 (D.C. Cir. 2009).

[51] *See Fannie Mae,* 552 F.3d at 817-18; *Capitol Records*, 261 F.R.D. at 47; *William A. Gross Constr. Assocs.*, 256 F.R.D. at 135.

their hands dirty" in crafting a search protocol.[52] On one side of the spectrum is the garden-variety case of *Helmert v. Butterball*,[53] in which Judge Holmes for the United States District Court for the Eastern District of Arkansas took enormous time and effort in adjudicating both the matter of search terms and custodians in response to a motion to compel further discovery.[54] Plaintiffs filed an action under the Fair Labor Safety Act alleging that they and others "were not fully compensated for time spent donning, doffing, and sanitizing protective gear and equipment."[55] After limited production and multiple meet-and-confers, defendant Butterball objected to plaintiffs' seventy-odd keyword requests as overbroad and claimed that it did not have the capability to undertake proximity searching on its e-mail system, which allows searching for two terms within the same sentence.[56] The court parsed plaintiffs' various requests keyword by keyword (many of which consisted of first or last names of individuals with prior donning and doffing cases) and generally agreed that additional responsive documents would be produced beyond "'donning and doffing' or 'don* and doff*.'"[57] Although the court ruled that defendant Butterball did not have to engage in proximity searches, the result seems unfortunate.[58] Automated methods with the assistance of an e-discovery vendor might have reduced defendants' overall search burden.[59] Neither the court nor the parties seemed to be aware of the existence of alternative search methods that would have expedited the

[52] *Compare* Helmert v. Butterball, LLC, No. 4:08CV00342 JLH, 2010 WL 2179180, at *4-5 (E.D. Ark. May 27, 2010), *with* D'Onofrio v. SFX Sports Grp., Inc., 254 F.R.D. 129, 131 (D.D.C. 2008), *and* Flying J Inc. v. Pilot Travel Ctrs. LLC, No. 1:06-CV-00030 TC, 2009 WL 1834998, at *3 (D. Utah June 25, 2009).

[53] *See Helmert*, 2010 WL 2179180.

[54] *See generally id*.

[55] *Id.* at *1.

[56] *Id.* at *2.

[57] *Id.* at *4-5.

[58] *See id.* at *5.

[59] *See Sedona Quality Commentary*, *supra* note 20, at 319.

review.[60] *Helmert* is only a recent example of a growing litany of case law in which judges have felt similarly compelled to make "Solomon-like" decisions on the number and type of keywords because parties have been unable to agree on a search protocol.[61]

[15] In contrast, Judge Facciola was perhaps the first to caution judges against venturing too far into the weeds in adjudicating search term disputes.[62] *United States v. O'Keefe* involved a defendant indicted on the charge that, as a State Department employee living in Canada, he had received gifts and other benefits from a co-defendant in return for expediting visa requests for his co-defendant's company employees.[63] The district court previously had required that the government "conduct a thorough and complete search of both its hard copy and electronic files in a good faith effort to uncover all responsive information in its possession[,] custody or control."[64] This involved a search inclusive of electronic files and e-mail "prepared or received by any consular officers" at various named posts in Canada and Mexico "that reflect[ed] either policy or decisions in specific cases with respect to expediting visa applications."[65] Through a declaration, the government documented its search of custodian files, accomplished using the following search string: "early or expedite* or appointment or early & interview or expedite* & interview."[66] Although only documents clearly about unrelated matters

[60] *See Helmert*, 2010 WL 2179180, at *4.

[61] *See, e.g.*, D'Onofrio v. SFX Sports Grp., Inc., 254 F.R.D. 129, 131 (D.D.C. 2008) ("Since I have gotten so little help from counsel, I will create a protocol of my own using as best I can my understanding of the limited agreement that the parties reached . . . ."); Flying J Inc. v. Pilot Travel Ctrs. LLC, No. 1:06-CV-00030 TC, 2009 WL 1834998, at *3 (D. Utah. June 25, 2009) (providing a search protocol and requiring that the requesting party justify twenty-eight specific terms).

[62] *See* United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008).

[63] *Id.* at 15-16.

[64] *Id.* at 16 (quoting United States v. O'Keefe, No. 06-CR-0249, 2007 WL 1239204, at *3 (D.D.C. Apr. 27, 2007)) (internal quotation marks omitted).

[65] *Id.* (internal quotation marks omitted).

[66] *Id.* at 16-18 (internal quotation marks omitted).

were removed ("e.g., emails about staff members' early departures or dentist appointments"), defendants objected that the search terms were inadequate.[67] Recognizing limits in the ability of judges to know how to intervene, Judge Facciola cited both to *Information Inflation: Can the Legal System Adapt?* as well as the *Sedona Search Commentary*, stating:

> 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 and judges 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.[68]

He went on to invite the defendants to file a further motion to compel.[69]

[16] Consistent with Judge Facciola's general caveats, a second line of recent cases represent decisions in which judges expressed wariness in entering the fray, especially where large amounts of ESI are at issue and there is the possibility of cooperation. For example, in *Trusz v. UBS Realty Investors LLC*, a case involving an "alleged concealment of overvaluing real estate investments,"[70] the plaintiff claimed that

[67] *Id.* at 18, 23-24.

[68] *O'Keefe*, 537 F. Supp. 2d at 24 (citations omitted); *see* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260 n.10 (D. Md. 2008) (providing commentary and analysis on the significance of Judge Facciola's citation to Federal Rule Evidence 702 and the role experts may yet play in testifying on the vagaries of search protocols); *see also* Heilman, *supra* note 39, at 1122-26 (proposing a three-tier system of evaluating search protocol disputes, where third tier of complexity involves the use of experts).

[69] *See O'Keefe*, 537 F. Supp. 2d at 24.

[70] *See* Trusz v. UBS Realty Investors LLC, No 3:09 CV 268(JBA), 2010 WL 3583064, at *1 (D. Conn. Sept. 7, 2010).

"defendants engaged in a 'massive document dump' by producing 1.8 million documents" placed on a series of disks that, in the plaintiff's view, contained ninety-nine percent irrelevant material.[71] In response to a motion to compel, the court stated:

> Among the items about which the court expects counsel to "reach practical agreement" without the court having to micro-manage e-discovery are "search terms, date ranges, key players and the like." Moreover, "[t]he use of key words has been endorsed as a search method for reducing the need for human review of large volumes of ESI[,]" to be followed by "a cooperative and informed process [which includes] sampling and other quality assurance techniques."[72]

The Court went on to say that "the issues raised in this motion largely could have been eliminated had counsel *actually* conferred with each other about refining the search terms,"[73] and that they were henceforth to

> [H]old an *in person conference with one another,* at which they *shall* conduct themselves in a professional and constructive manner, in order to ascertain if there are more discrete search terms, or combinations of search terms, that can be applied by defendants, so that the volume of documents produced are substantially less than 1.8 million, with the expectation that as a result, a significantly higher percent of the documents captured by the searches will be relevant.[74]

[17] Similarly, in *Romero v. Allstate Insurance Company*,[75] a class

[71] *Id.* at *2-3.

[72] *Id.* at *5 (alterations in original) (citations omitted) (quoting Thomas Allman, *Conducting E-Discovery After the Amendments: The Second Wave*, 10 SEDONA CONF. J. 215, 217, 223 (2009)).

[73] *Id.* (emphasis in original).

[74] *Id.* (emphasis in original).

[75] *See* Romero v. Allstate Ins. Co., 271 F.R.D. 96 (E.D. Pa. 2010).

action case alleging age discrimination, the court faced a motion from the plaintiffs seeking an order not only compelling the defendant to confer regarding additional relevant custodians and search terms, but also seeking information on what searches the defendant had conducted in the past.[76] *Romero*, referencing *Trusz*, similarly compelled the parties to meet and agree on the search terms, custodians, and date ranges the defendants intended to use and "any other essential details about the search methodology they intend to implement . . . ."[77] *Romero*'s result suggests that courts are approaching a ceiling in tolerating intervention to resolve simple keyword disputes. Moreover, the court denied plaintiffs' request for details on past searches in light of the court's confidence in the parties' cooperation "on a forward-going basis to share information about what has already been completed and what needs to be done . . . ."[78]

[18] Finally, following in the footsteps of Judge Facciola, Magistrate Judge Thynge of the United States District Court for the District of Delaware, in a patent drug infringement action, held that the defendants' arguments for further discovery of ESI would "force the court into the mysteries of keyword search techniques, specifically the efficacy of various methods used to search electronically stored information . . . ."[79] She went on to say that "[n]either lawyers nor judges are generally qualified to opine that certain search terms or files are more or less likely to produce information than those keywords or data actually used or reviewed."[80] The court stated that it "[would] not enter the wilderness of keyword search usage and is not directing the appropriate search terms for plaintiffs to employ,"[81] and instead ruled on pending motions by fashioning what patent issues, rather than specific search terms, remained

[76] *See id.* at 98, 109.

[77] *Id.* at 109-10.

[78] *Id.* at 110.

[79] Eurand, Inc. v. Mylan Pharm., Inc., 266 F.R.D. 79, 84 (D. Del. 2010).

[80] *Id.*

[81] *Id.* at 85 n.31.

fair game for further exploration in discovery.[82]

[19] To the extent the judiciary takes a position adverse to one of hands on micro-management of how parties develop search protocols (in terms of ruling on specific search terms), a greater spotlight is placed on the expectation of non-adversarial cooperation among counsel.[83] The questions that naturally arise at such juncture include: How exactly is cooperation manifested? And, are there situations in which some level of cooperation proves necessary as a matter of professional ethics?

IV. COOPERATION, ITERATIVE DISCOVERY, AND A QUESTION OF ETHICS

[20] Over six decades ago, the Supreme Court in *Hickman v Taylor*[84] stated "[m]utual knowledge of all the relevant facts gathered by both parties is essential to proper litigation."[85] This long-held aspiration is made all the more challenging today, given the growing asymmetry in access and knowledge to data sets between requesting and producing parties in discovery practice.[86] It is also the case, arguably, that even *producing* parties now face increasing technological hurdles in obtaining reasonable intellectual control over their own ESI repositories, notwithstanding their profoundly greater access rights to the data subject to discovery requests.[87] In short, counsel in any substantial litigation must

[82] *See id.* at 85.

[83] *See* Balboa Threadworks, Inc. v. Stucky, No. 05-1157-JTM-DWB, 2006 WL 763668, at *5 (D. Kan. Mar. 24, 2006) (directing parties to meet and confer on the use of a search protocol, including keywords).

[84] 329 U.S. 495 (1947).

[85] *Id.* at 507.

[86] Jason R. Baron, Remarks at the Ninth Annual Georgia Symposium on Ethics and Professionalism: Ethics and Professionalism in the Digital Age (Nov. 7, 2008), *in* 60 MERCER L. REV. 863, 866 (2009).

[87] *See* Mazza et al., *supra* note 39, ¶ 3 ("The explosive growth of ESI has changed the very nature of discovery, with new electronic complexities making the preservation and production of evidence far more challenging. It is an accepted fact that 'the discovery of computer-based information [can] cost more, take more time and create more headaches than conventional paper based discovery.'") (footnotes omitted).

deal with searching through vast quantities of documents and ESI greater than anything imagined in the decades since *Hickman v. Taylor.*[88]

[21] In the spirit of *Hickman,* the 2007 *Information Inflation* article called for "invoking a new strategic cooperation paradigm,"[89] and mapped out a more complex, structured meet and confer process to discuss the parties' search demands, and accommodating recognized asymmetries in knowledge.[90] Employing the phrase "'Virtuous Cycle' Iterative Feedback Loop," the article suggested staging multiple meet and confers, where: (1) the parties first meet to discuss initial searches, including the possibility of discussing sophisticated keyword searching using proximity operators, stemming terms, wildcard, and truncation, as well as the use of alternative search methods; (2) the parties then use sampling techniques to conduct searches in the period between meet and confers; (3) the parties return to discuss the results and fine-tune requests, based on under- or over-inclusiveness; and (4) the process repeats under mutually agreed conditions.[91] The idea of holding multiple meet and confers has indeed been embraced in the interim.[92]

---

[88] *See* Kevin A. Griffiths, *The Expense of Uncertainty: How a Lack of Clear E-Discovery Standards Put Attorneys and Clients in Jeopardy*, 45 IDAHO L. REV. 441, 442 (2009) (noting that technological advancements in the practice of law have "led to an exponential increase in electronically stored information (ESI)").

[89] Paul & Baron, *supra* note 1, ¶ 26.

[90] *Id.* ¶¶ 26-33, 52.

[91] *Id.* ¶¶ 51-55.

[92] *See, e.g.*, ClearOne Commc'ns, Inc. v. Chiang, No. 2:07 CV 37 TC, 2008 WL 920336, at *2 (D. Utah Apr. 1, 2008) (suggesting parties can review the efficacy of keyword searches and other aspects of electronic discovery in multiple stages). As part of spin-off research from the TREC Legal Track, *see TREC 2010 Legal Track*, U. MD., http://trec-legal.umiacs.umd.edu/ (last visited Feb. 27, 2011), Charlie Zhao was able to demonstrate, at least tentatively, the utility of using iterative methods through one or two rounds of meet and confers and of fine-tuning search algorithms for the matter of producing a richer set of relevant documents. Feng C. Zhao et al., *Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback*, U. PITT. L., 6-7, 9 (June 8, 2009), http://www.law.pitt.edu/DESI3_Workshop/Papers/DESI_III.Zhao_Oard_Baron.pdf. However, Zhao also found that diminishing returns would be expected after two rounds. *See id.* at 7.

[22] In harmony with the *Information Inflation* article, *The Sedona Conference Cooperation Proclamation*, now signed by over 100 members of the federal and state judiciary,[93] contains as one of its major planks a call for parties to "jointly develop[] automated search and retrieval methodologies to cull relevant information."[94] However, that is as far as it goes.[95] Does *jointly* developing necessarily mean *simultaneous* exchange of initial keyword terms or search protocols, or is there reasonable freedom to engage in a staged or phased process, acknowledging the greater access the responding party has to its own data set (i.e., asymmetry in knowledge), and thus a presumption the responding party will take the lead? And to what extent does a responding party's greater knowledge regarding the contents of its own data universe give rise to a duty to disclose certain defects in a proposed search methodology?

[23] As the keyword search case law abundantly illustrates, Ralph Losey is correct in noting that "[m]ost lawyers [today] [d]o [s]earch as if it were a game of *Go Fish*."[96] Losey believes that the prevalent negotiated keyword search model is no better than a fishing expedition, where the "party requesting ESI guesses what key words might produce evidence to support [its] case," and "cannot see the responding party's cards."[97] Losey posits that this keyword search model exists because "[t]he way the game is now often played, the requesting party also keeps [its] secrets[,]" since it "do[es] not want to reveal exactly what it is that [the party is] looking

[93] *See generally* The Sedona Conference, *The Sedona Conference Cooperation Proclamation*, 10 SEDONA CONF. J. 331, 334-38 (Supp. 2009) (providing a list of judges endorsing the *Cooperation Proclamation* as of September 30, 2010).

[94] *Id.* at 332. Also of relevance, the *Cooperation Proclamation* suggests that cooperation might include "[e]xchanging information on relevant data sources, including those not being searched, or scheduling early disclosures on the topic of Electronically Stored Information." *Id.*

[95] *Compare id.*, *with* Paul & Baron, *supra* note 1, ¶¶ 49-55 (suggesting multiple meet and confers as well as specific technical search methods for parties to develop an agreed upon and effective discovery process).

[96] Ralph C. Losey, *Child's Game of 'Go Fish' Is a Poor Model for E-Discovery Search*, *in* ADVENTURES IN ELECTRONIC DISCOVERY (forthcoming 2011).

[97] *Id.*

for."[98] The requesting party hides its discovery intentions "by using broad, general requests" that turn out to be unhelpful.[99] Losey notes that this approach is both "unreliable and inefficient."[100] His alternative: a new game in which responding parties control the search because "their data, their IT systems, their data custodians, their employees, their agents, their attorneys, their language, their retention policies, [and] their retention practices" are at stake.[101]

[24] Furthermore, Losey acknowledges that

> This new game also requires cooperation and transparency by the responding party, moreover it requires [the party's] initiative and leadership. The responding party can no longer just sit back and watch poor guesses being made. [The responding party] must take the lead in getting the truth out. This is a burden, but the responding party is more than compensated for this burden by the protection this provides from over-broad, expensive, inefficient search. It also protects the responding party from having to show [its] whole deck of cards, [its] entire ESI collection. . . .
>
> The party responding to requests for production must be proactive. [It] must design the search. . . . [T]his only makes sense because it is [the responding party's] data[and the responding party has] unfettered access to it. [The responding party] know[s] the language. [It] know[s] the people involved. For these reasons, the responding party is always in the best position to search the data and, if

[98] *Id.*

[99] *Id.*

[100] *Id.*

[101] *Id.*

> asked, to fully explain how and why the search met the needs of the requesting party.[102]

[25] In creating a presumption in which the responding party "first moves," Losey acknowledges that "[t]he process may still sometimes be iterative," noting that "[a] careful study by the requesting party of the ESI received may lead to new goals, new issues, and new more focused requests."[103] He also adds, "[o]nce the cards responsive to the request are found, they *all* have to be disclosed, the bad as well as the good," with exception only for privileged documents.[104] Losey concludes that "[h]onesty and good faith are critical in all discovery processes."[105]

[26] While I am very much in agreement with Ralph Losey that joint negotiations, in the absence of knowledge, tend toward being meaningless, inefficient exercises, and that solutions to the asymmetry problem argue in favor of responding parties taking the lead in developing appropriate search protocols, there is still the possibility that requesting parties have a voice at the table from the start.[106] As my article with Ed Wolfe, *A Nutshell on Negotiating E-Discovery Search Protocols* argues,

> A requesting party may have a legitimate, good faith belief that they are sufficiently informed regarding the causes of action at issue and underlying facts so as to be able to propose well-formed search queries, including through the use of keywords. As the propounder of the eventual discovery requests, the requesting party is in the best position to know what it believes are the most salient aspects of the case that are in need of discovery in the first place. To the extent the producing party is willing to allow

[102] Losey, *supra* note 96.

[103] *Id.*

[104] *Id.* (emphasis in original).

[105] *Id.*

[106] *See* Baron & Wolfe, *supra* note 39, at 233.

> the requesting party to control all or some of the keyword search, without raising a threshold objection, doing so holds out the possibility of significantly reducing the level of conflict and subsequent motion practice in discovery.[107]

Further, early agreement may be strategically valuable in diminishing the other side's ability to object, and "[c]ooperation in the form of reaching agreement on search terms ultimately reduces the legal risk in having to undertake new and different searches through large collections of data."[108]

[27] The 2007 *Information Inflation* article noted that there are game theoretic aspects involved in exercising cooperative behavior, and that lawyers would benefit from greater knowledge of "winning strategies," as developed in a variety of disciplines utilizing game theory.[109] This idea is still worthy of further systematic study and development.[110]

---

[107] *Id.*

[108] *Id.*

[109] *See* Paul & Baron, *supra* note 1, ¶ 56 & n.134 (citing ROBERT AXELROD, THE EVOLUTION OF COOPERATION: AGENT-BASED MODELS OF COMPETITION AND COLLABORATION (1997); Robert Axelrod, *The Emergence of Cooperation Among Egoists*, 75 AM. POL. SCI. REV. 306 (1981)) (discussing Robert Axelrod's theories of cooperation).

[110] As the case law has played out, there appear to be a number of standard, opening-move variations, akin to standard chess openings, that range from (i) the requesting party refusing to supply keywords because it perceives doing so as bidding against itself, *see, e.g.*, William A. Gross Constr. Assocs. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 135 (S.D.N.Y. 2009); (ii) the requesting party priming the pump by demanding 1,000 or more keywords, perhaps as a negotiating tactic to get to a reasonable number, *see id.* at 134; (iii) the responding party unilaterally performing discovery searches that result in a data dump on the other side, *see, e.g.*, Romero v. Allstate Ins. Co., 271 F.R.D. 96, 109 (quoting Trusz v. UBS Realty Investors LLC., No. 3:09 CV 268(JBA), 2010 WL 3583064, at *4 (D. Conn. Sept. 7, 2010) (quoting The Sedona Conference, *The Case for Cooperation*, 10 SEDONA CONF. J. 339, 344-45 (2009))); and (iv) the responding party refusing to negotiate and ending up with the requirement to respond to massively overbroad requests, *see, e.g.*, Kipperman v. Onex Corp., 260 F.R.D. 682, 697 (N.D. Ga. 2009). For obvious reasons, successful jointly negotiated approaches are rarely seen in reported cases. *See Romero*, 271 F.R.D. at 101 ("Although the parties engaged in a 'meet and confer' in an effort to resolve this dispute, they were unable to reach any mutually agreeable solution. As such, the Court now rules on the issues . . . ."). I am of the view

[28] Also worthy of further reflection and analysis are the potential ethical pitfalls presented by the responding party's failure to share known information regarding the inefficacy of particular proposed search terms.[111] This is particularly critical when corporate culture and private language result in the use of code names or nick names that could never be known *a priori* to the requesting party, but also applies in a variety of other circumstances arising due to the fuzziness of language (misspellings of individuals' names, failure to recognize known synonyms, etc.).[112] This author's position on the ethical implications was adequately captured in The Sedona Conference article, *The Case for Cooperation*, which states, "because knowledge of the producing party's data is usually asymmetrical, it is possible that refusing to 'aid' opposing counsel in designing an appropriate search protocol that the party holding the data knows will produce responsive documents could be tantamount to concealing relevant evidence."[113] That article went on to note that this author,

> has argued that in circumstances where a party is certain that opposing counsel's proposed search protocol would not capture documents it knows would be responsive violates Rule 3.4 of the Model Rules of Professional Responsibility by failing to suggest or use additional search

---

that there is probably a limit on the available permissible permutations of at least the very opening one or two moves between would-be adversaries; thus, over time, with enough "games" being played in the form of reported cases, there should be a standard set of moves known to each side and pretty much adhered to as representing best practices in the area. The further caveat here is that, just as in any exponentially expanding universe of possibilities, such opening moves lead to rapid complexity (and case uniqueness), depending on the legal setting and the permutations that subsequently arise in any litigated matter. None of the above discussion is limited to negotiations over keywords *per se*; a typology of cases is likely to apply for future information retrieval methods employed as well.

[111] *See* Baron, *supra* note 86, at 866. The 2008 Mercer Law School Symposium raised the "ethics conundrum" for the first time, and readers of this article are invited to consider the extended argument with all of its caveats. *See id.* at 866-80.

[112] *See id.* at 875-76.

[113] The Sedona Conference, *The Case for Cooperation*, 10 SEDONA CONF. J. 339, 344 (2009).

> terms that would result in production; such conduct is tantamount to suppression.[114]

[29] While there are no reported cases discussing the matter of "keyword search ethics," it is only a matter of time before courts are faced with deciding difficult issues regarding the duty of responding parties and their counsel to make adequate disclosures, given better knowledge of their data sets, in line with the "honesty and good faith" Ralph Losey calls for in the new game of discovery.[115]

V. A TIPPING POINT: THE EMERGENCE OF SOPHISTICATED "PREDICTIVE" AND "PRIORITIZATION" SOFTWARE ACCELERATING SEARCH AND DOCUMENT REVIEW

[30] As one of its main tenets, the 2007 *Information Inflation* piece recognized the profound importance of language, in all of its subtlety and ambiguity, as posing a central challenge to achieving perfection in e-discovery searches.[116] George Paul and I turn challenged the legal profession to make better use of, and delve into, the parallel universes of information retrieval science and artificial intelligence – disciplines that, while containing foreign concepts and scary mathematics, represented a pathway to better, more efficient outcomes in future e-discovery engagements.[117] In that vein, a portion of the 2007 article was devoted to looking beyond keywords, given their inherent limitations, and predicting a time in the near future where lawyers would be more comfortable using

---

[114] *Id.* at 344 n.19 (citing Baron, *supra* note 86, at 877); *see* MODEL RULES OF PROF'L CONDUCT R. 3.4 (2007) ("A lawyer shall not: (a) unlawfully obstruct another party's access to evidence . . . (b) falsify evidence, counsel or assist a witness to testify falsely . . . (f) request a person other than a client to refrain from voluntarily giving relevant information to another party . . . ."); *cf.* MODEL RULES OF PROF'L CONDUCT R. 1.6 (2007) ("A lawyer shall not reveal information relating to the representation of a client unless the client gives informed consent . . . .").

[115] Losey, *supra* note 96 ("Honesty and good faith are critical in all discovery processes.").

[116] Paul & Baron, *supra* note 1, ¶ 38.

[117] *Id.* ¶¶ 65-66.

increasingly sophisticated software to perform search and document review functions as a supplement to, if not a replacement of, status quo search methods.[118] Perhaps that time is now arriving, for at least a few, and perhaps many, e-discovery practitioners.

[31] Indeed, in a 2010 article, entitled *Mandating Reasonableness in a Reasonable Inquiry*, Patrick Oot and his colleagues suggest "[t]he [e]nd of [k]eyword [s]earch [m]ethods."[119] By this, the authors meant that a rough consensus has developed supporting the idea that simple use of selected keywords, without lawyers considering the use of additional automated technologies, or using or applying a degree of sophistication to how a search protocol is to be constructed, refined, and tested, should be considered a thing of the past.[120] Case law demonstrates that reports of the demise of keyword search (and linear review) may be somewhat exaggerated, though there are indeed a host of new and promising ways to organize and search through large data sets.[121]

[32] Easily fitting under the umbrella term "[n]ew '[i]nformation [c]oncepts' [i]n the [p]ractice of [l]aw"[122] is a fusion of various techniques well known for decades by information retrieval scientists as forms of latent semantic indexing,[123] and going under such currently fashionable

[118] *See id.* ¶¶ 37-40, 64-68.

[119] Oot et al., *supra* note 19, at 553.

[120] *See id.* at 553-56.

[121] *See id.* at 557-58.

[122] Paul & Baron, *supra* note 1, ¶ 40.

[123] *See* Peter W. Foltz, *Using Latent Semantic Indexing for Information Filtering*, *in* PROCEEDINGS OF THE CONFERENCE ON OFFICE INFORMATION SYSTEMS 40 (Frederick H. Lochovsky & Robert B. Allen eds., 1990), *available at* http://www-psych.nmsu.edu/~pfoltz/cois/filtering-cois.html ("Latent Semantic Indexing . . . takes advantage of the implicit higher-order structure of the association of terms with articles to create a multi-dimensional semantic structure of the information. . . . Retrieving information in LSI overcomes some of the problems of keyword matching by retrieval based on the higher level semantic structure rather than just the surface level word choice."). Latent Semantic Indexing (LSI) is a form of Latent Semantic Analysis (LSA), which "is an approach to automatic indexing and information retrieval that attempts to overcome [the

names as "predictive coding," "clustering" technologies, "content analytics," and "auto-categorization," among many others.[124] The software involved is used both in early case assessment as well as traditional document review for production to opposing counsel.[125] Such methods present the possibility for *greatly* increasing present rates of document review because they provide the possibility to reduce the overall manual search burden on counsel, thereby dramatically reducing review costs.[126] Reduced to its essence, "predictive coding" and its equivalents (i) start with a set of data, derived or grouped in any number of variety of ways (e.g., through keyword or concept searching); (ii) use a human-in-the-loop iterative strategy of manually coding a seed or sample set of

---

drawbacks of search queries] by mapping documents as well as terms to a representation in the so-called *latent semantic space*." Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, BROWN U., http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf (last visited Mar. 1, 2011). Moreover, with respect to LSA,

> [t]he general claim is that similarities between documents or between documents and queries can be more reliably estimated in the reduced latent space representation than in the original representation. The rationale is that documents which share frequently co-occurring terms will have a similar representation in the latent space, even if they have no terms in common. LSA thus performs some sort of noise reduction and has the potential benefit to detect synonyms as well as words that refer to the same topic."

Id. These methods come in many varieties, including Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichet Allocation. *See id*.; *Sedona Search Commentary, supra* note 26 *Appendix* (describing advanced search methods); David M. Blei et al., *Latent Dirichlet Allocation*, 3 J. MACHINE LEARNING RES. 993, 993-97 (2003), *available at* http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf.

[124] *Survey on Predictive Coding*, *supra* note 23, at 8, 25-26 (In response to the survey question, "If you think there is a better generic term than 'predictive coding,' what would it be?," responses from vendors included: "Prognostic Document Profiling," "Predictive Ranking," "Relevance Assessment," "Suggested coding," "Predictive Categorization," "Automatic Categorization," "'Propagated Coding' or 'Replicated Coding,'" and "Automated Document Classification." *Id*. at 5.

[125] *See id*. at 26.

[126] *See id; see also* John Markoff, *Armies of Expensive Lawyers Replaced by Cheaper Software*, N.Y. TIMES, March 5, 2011, at A1.

documents for responsiveness and/or privilege; (iii) employ machine learning software to categorize similar documents in the larger set of data; (iv) analyze user annotations for purposes of quality control feedback and coding consistency.[127]

[33] Growing evidence suggests that sophisticated players' use of these types of selection methods at early stages of collection and culling have led to substantial bottom line savings in a wide variety of complex litigation.[128] For example, Bennett Borden has reported "extraordinary results" with various clustering and categorization technologies in terms of the speed in which they affect document review.[129] "In one review of about 5,000 documents, a team of five reviewers . . . . required 110 working hours at a rate of about [forty-five] documents per hour" to review the set.[130] While using clustering software, the same group of reviewers took fifty-five working hours to review 7,500 additional

[127] *See id.* at 6-10 (listing vendor descriptions of predictive coding overall process); *see also Sedona Quality* Commentary, *supra* note 20, at 319 (describing how using automated methods can reduce the initial burden of review). For an in depth discussion of the use of one type of predictive coding, see Caroline Privault et al., *A New Tangible User Interface for Machine Learning Document Review,* 18 ARTIFICIAL INTELLIGENCE & L. 459 (2010), http://www.springerlink.com/content/45185q2641807340/fulltext.pdf (describing CategoriX software).

[128] At LegalTech 2011, held in New York City, I participated on two panels specifically addressing the subject of "predictive coding," including presentations by a variety of law firm counsel pointing to examples of real-world litigation savings in costs and time expended on review using clustering and categorization technologies. *See, e.g.*, Jason R. Baron, Dir. of Litig., Nat'l Archives and Records Admin., et al., The Emerging "Smart Discovery" Paradigm for Cost Management in E-Discovery – Case Studies and Case Law (Jan. 31, 2010) (PowerPoint on file with author); *see also* Oot et al., *supra* note 19, at 551 ("The use of auto-categorization systems can potentially reduce document request response times from over four months to as little as thirty days for even the largest datasets.").

[129] Bennett B. Borden, *E-Discovery Alert: The Demise of Linear Review*, CLEARWELL SYSTEMS, 3 (Oct. 2010), *available at* http://www.clearwellsystems.com/e-discovery-blog/wp-content/uploads/2010/12/E-Discovery_10-05-2010_Linear-Review_1.pdf.

[130] *Id.* Borden noted that "[t]his is a bit slower than the usual [fifty] to [sixty] documents per hour that is used as an e-discovery industry average, but the documents were fairly technical and required more than the average level of scrutiny." *Id.*

documents from the same overall collection, "at a rate of about 136 documents per hour."[131] In a second experiment, which involved about 60,000 documents and two groups of reviewers with roughly equal experience, the reviewers using categorization software completed a review six times faster than reviewers using Boolean techniques and a straight linear review.[132] Finally, in a larger experiment, twenty-two reviewers completed a review of 1.5 million documents in 124 hours using the same software, averaging 318 documents per hour.[133] Quality control analysis showed in all cases rates of error equal to or less than that of traditional review (three percent in the case of 1.5 million documents).[134] While Borden is careful to note certain caveats for these experiments, they are in line with what are emerging as new best practices for document review in terms of the ability to efficiently process large data sets.[135]

[34] Subject to two exceptions, case law lacks discussion of parties' use of predictive software, or, for that matter, any form of alternative or more sophisticated search methods, for document review. Judge Facciola's opinion in *Disability Rights Council of Great Washington v. Metropolitan Transit Authority* was the first published case to suggest that parties should contemplate the use of an alternative to keyword search, in the context of searching restored backup tapes.[136] The court wished to "bring to the parties' attention recent scholarship that argues that concept searching . . . is more efficient and more likely to produce the most comprehensive results."[137]

---

[131] *Id*.

[132] *Id*.

[133] *Id*.

[134] *See id.* at 3-4.

[135] Borden, *supra* note 129, at 3-4.

[136] *See* Disability Rights Council of Greater Wash. v. Wash. Metro. Transit Auth., 242 F.R.D. 139, 148 (D.D.C. 2007).

[137] *Id.* at 148 (citing Paul & Baron, *supra* note 1, ¶ 41-43).

[35] The second case is the heralded opinion of Judge Grimm in what is now referred to as *Victor Stanley I*, in which the question before the court was whether defendants had waived their right to attorney–client privilege by mistakenly producing 165 privileged documents after employing a faulty keyword filter to discriminate between privileged and non-privileged content.[138] After finding that defendants had been "regrettably vague" in informing the court of how keywords were developed, how the search was conducted, and what quality controls were employed,[139] and after noting that "all keyword searches are not created equal," [140] Judge Grimm went on to find that waiver of privilege had occurred.[141] In an extended footnote, for the benefit of the parties, Judge Grimm, citing the *Sedona Search Commentary*, set out to describe his knowledge of alternatives to keyword searching, including fuzzy search models, Bayesian classifiers, clustering, and concept and categorization tools.[142]

[36] Clearly, *Victor Stanley I's* acknowledgement of an information retrieval world beyond keyword searching was ahead of its time.[143] But, given the practice of law by the most agile e-discovery practitioners today, it is only a matter of time before the case law catches up.

VI. QUALITY PROCESSES AND THE NEED FOR STANDARDS

[37] Case law and commentaries have also focused on quality controls, especially given what is now recognized to be "industrial size" e-

[138] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 253-54 (D. Md. 2008).

[139] *Id.* at 256.

[140] *See id.* at 256-57.

[141] *See id.* at 267-68.

[142] *See id.* at 259 n.9 (citing *Sedona Search Commentary*, *supra* note 26, at 217-23).

[143] *Compare* Interview by T. Jayaraman with Edward Witten, Professor, Inst. For Advanced Study, *in On the Right Track*, FRONTLINE, Feb. 3-16, 2001, *available at* http://www.hinduonnet.com/fline/fl1803/18030830.htm ("'[S]tring theory is a piece of 21st century physics that happened to fall into the 20th century'"), *with Victor Stanley*, 250 F.R.D. at 259 n.9.

discovery processes, which involve the increasing frequency of gigabytes and terabytes of data in litigated matters.[144] Judge Scheindlin opined in *Pension Committee of the University of Montreal Pension Plan v. Banc of America Securities, LLC* that a party's "failure to assess the accuracy and validity of selected search terms" amounted to negligence *per se*.[145] Moreover, as noted earlier, Judge Peck has issued his "wake-up call" to the Bar that quality control and sampling matter.[146]

[38] There is indeed a variety of available project management, sampling, and quality control techniques aimed at *reducing* the cost of discovery.[147] These methods and techniques invariably dovetail with the need for multiple counsel interactions in phased or tiered discovery and the multiple meet and confers described earlier, where samples of ESI are proffered after a reasonable search is made, so as to iteratively refine the search protocol in the interest of reaching a consensus on what constitutes the universe of relevant evidence.[148]

[39] In line with "Principle 2" of the *Sedona Quality Commentary,* parties should also be prepared to employ reasonable forms or measures of quality throughout the e-discovery process, including "sampling at different phases of the process; using independent testing to report whether results can be replicated and confirmed; adopting reconciliation measures for different phases of the e-discovery process; and employing inspection[s] to verify and report . . . discrepancies."[149] Prior to 2006, none of these steps were ever rigorously contemplated as routine. Beyond

[144] *See Sedona Quality Commentary*, *supra* note 20, at 305.

[145] Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC, 685 F. Supp. 2d. 456, 465 (S.D.N.Y. 2010).

[146] William A. Gross Constr. Assocs. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 134 (S.D.N.Y. 2009); *see supra* note 38 and accompanying text.

[147] *See Sedona Quality Commentary*, *supra* note 20, at 303, 310-12.

[148] *See id.* at 304.

[149] *Id.* at 303.

those points, the *Sedona Quality Commentary* emphasized project management by the "active leadership" of a senior attorney.[150]

[40] Quality control is not necessarily a sure-footed area for the judiciary. Compare *Mt. Hawley Insurance Co. v. Felman Production, Inc.*, a dispute over insurance proceeds in which the plaintiff produced over 346 gigabytes of data in response to the defendants' requests for production,[151] with *Victor Stanley I*, in which the court addressed a question of waiver of attorney-client privilege subsequent to inadvertent production (and request for clawback) of two key e-mails alleged to be evidence of fraud.[152] Despite the fact that the *Mt. Hawley* court made extensive findings regarding the measures of testing, sampling, and quality control the plaintiff employed in an effort to prevent disclosure of privileged materials, the inadvertent disclosure of 377 privileged documents out of 346 gigabytes apparently constituted sufficient evidence for the Court to find that the plaintiff and plaintiff's counsel exhibited a "lack of care."[153] One key, though problematic, finding the court made is that "the number of inadvertent disclosures is large, more than double the number discussed in *Victor Stanley* [*I*]."[154] While it appears that in *Victor Stanley I* the 165 inadvertently disclosed documents were out of a universe of approximately 9,000 PDF files, the 377 documents in *Mt. Hawley* were out of a universe of 5,536,000 files: a percentage error figure orders of magnitude better than that of *Victor Stanley I.*[155] Courts must understand the statistical properties of the ESI universe in which e-discovery is

[150] *Id.*

[151] *See* Mt. Hawley Ins. Co. v. Felman Prod., Inc., 271 F.R.D. 125, 126, 135 (S.D.W. Va. 2010).

[152] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 253 (D. Md. 2008).

[153] *Mt. Hawley*, 271 F.R.D. at 136.

[154] *Id.*

[155] *See Mt. Hawley*, 271 F.R.D. at 135-36; *Victor Stanley*, 250 F.R.D. at 253; Ralph Losey, *The Good, the Bad, and the Ugly: "*Mt. Hawley Ins. Co. v. Felman Production, Inc.", E-DISCOVERY TEAM (June 10, 2010, 7:11 AM), http://e-discoveryteam.com/2010/06/10/the-good-the-bad-and-the-ugly-"mt-hawley-ins-co-v-felman-production-inc-"/.

practiced, but the *Mt. Hawley* opinion does not suggest that the court evaluated waiver using an appropriate mathematical yardstick.[156] In cases involving as much as one billion objects – or a petabyte of data – courts should approach inadvertent disclosure issues in relative, rather than absolute terms, and consider whether to set reasonableness standards that require near-perfect accuracy.[157]

[41] The notion that there is a measure of quality to achieve in the e-discovery process through testing and sampling techniques suggests a further question: are there quality *standards* that the profession can coalesce so as to represent best practices in e-discovery search and information retrieval? "Essentially, the idea is that . . . we agree on how each performer of E-discovery services should design measures to gain insight into the quality of the results achieved by their particular process. The design of their process, and of their specific measures, is up to each performer."[158] Apart from the research performed as part of the TREC Legal Track,[159] this author has elsewhere suggested that the legal profession should look for inspiration to other standards and benchmarking entities as a platform for further development, such as the ISO 9000 family of international quality management system standards and the Capability Maturity Model Integration ("CMMI") used in evaluating software engineering.[160] Working on standards may in turn lead to the further possibility of industry and academia developing some form of certification entity entrusted by both bench and bar to resolve questions concerning the quality of particular software or services offered in the context of specific litigation. The DESI IV workshop taking place

[156] Losey, *supra* note 155.

[157] *See id.*

[158] Oard, et al., *supra* note 40 at 381.

[159] *See generally* TREC 2010 Legal Track, *supra* note 92.

[160] *See, e.g.*, Oard et al., *supra* note 40, at 382. Other examples include the Statement on Auditing Standards No. 70: Service Organizations (SAS 70), which guides effective audit reporting, and the Payment Card Industry Data Security Standard (PCI DSS), which enforces uniform compliance in processing credit card payments. *See id.*

at the ICAIL 2011 conference in Pittsburgh will focus on these aspects of e-discovery.[161]

VII. CONCLUDING THOUGHTS

[42] The Sedona Conference recognized that "just as *Moneyball* demonstrated the value of applying new statistical measures to assess baseball talent, even if running counter to 'tried and true practices' based on intuition and culture," the legal profession must employ best-in-class thinking from various disciplines (including project management, quality control, statistics and information retrieval) to optimize efficiency in modern-day discovery.[162] As litigation continues, "cost-conscious firms, corporations and institutions of all kinds intent on best practices, as well as over-burdened judges, will demand that parties undertake new ways of thinking about how to solve discovery problems."[163] Maximizing the use of automated technologies in search and document review to achieve a true quality outcome "is consistent with the highest ethical calling of the legal profession."[164] All legal practitioners should strive to be more agile, efficient and technically savvy to work within the existing rules structure so as to best pursue this noble end.

[161] *See ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, U. MD., http://www.umiacs.umd.edu/~oard/desi4/ (last updated Jan. 15, 2011). The workshop is scheduled for June 6, 2011. *Id*.

[162] *Sedona Quality Commentary*, *supra* note 20, at 325 (footnote omitted).

[163] *Id.*

[164] *Id.*