The Sedona Conference Working Group Series

wgs℠

The Sedona Conference

Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery

A Project of The Sedona Conference
Working Group on Electronic Document Retention
& Production (WG1)

December 2013

The Sedona Conference®

THE SEDONA CONFERENCE BEST PRACTICES COMMENTARY ON THE USE OF SEARCH AND INFORMATION RETRIEVAL METHODS IN E-DISCOVERY

A Project of The Sedona Conference Working Group on Electronic Document Retention & Production (WG1)

Author:

The Sedona Conference

2013 Editors-in-Chief:

Jason R. Baron
Maura R. Grossman

2007 Editor-in-Chief:

Jason R. Baron

2007 Senior Editors:

Thomas Y. Allman
M. James Daley
George L. Paul

We thank all of our Working Group Series Sustaining and Annual Sponsors, whose support is essential to our ability to develop Working Group Series publications. For a listing of our sponsors, click on the "Sponsors" navigation bar on the homepage of our website.

REPRINT REQUESTS:
Requests for reprints or reprint information should be directed to The Sedona Conference info@sedonaconference.org or 602-258-4910.

Preface and Acknowledgments
2013 Edition

Welcome to the 2013 Edition of The Sedona Conference Best Practices *Commentary on the Use of Search and Information Retrieval Methods in E-Discovery.* Since the publication of the 2007 Public Comment Version of this Search Commentary, there have been significant developments in both case law and technology in the area of search and retrieval. Indeed, the 2007 Search Commentary itself has been prominently cited in a number of reported cases as an authoritative source on best practices in this area.

The 2013 Edition of the Commentary reflects changes in legal practice with a new section on computer- or technology- assisted review, as well as citations to more recent case law. Certain of the original eight Practice Points have been revised to reflect developments in law and practice, including recognition of the key principles of cooperation and proportionality advanced by The Sedona Conference. The Appendix on Information Retrieval Methods has also been modified to reflect changes in technology. The text of the 2007 Version of this Commentary otherwise remains largely intact, except for the deletion and/or updating of outdated information, and for minor stylistic and grammatical edits. The text was not edited with an eye towards being a fully-revised "Second Edition" of the original Commentary. Nevertheless, The Sedona Conference recognizes that the rapidly evolving nature of automated techniques calls for continuing close attention to further changes in professional practice in this area, especially with respect to defending the process used, and we will endeavor to meet that need through future publications.

I want to thank the entire Working Group for all their hard work and contributions, and especially the 2013 Edition Editorial Committee for leading this effort to update the existing Search Commentary. I wish to acknowledge the contributions of Jason R. Baron and Maura R. Grossman for taking the lead in revising and updating the prior version, as assisted by Bobbi Basille, Todd Elmer, Amir Milo, Priya Keshav, and James Sherer. Finally, but certainly not least, the Working Groups of The Sedona Conference could not accomplish their goals without the financial support of the sustaining and annual sponsors of the Working Group Series listed at *www.thesedonaconference.org/sponsors*.

Kenneth J. Withers
Deputy Executive Director
The Sedona Conference
December 2013

Table of Contents

Overview

Traditional Approaches To Searching For Relevant Evidence Are No Longer Practical Or Financially Feasible

Discovery of relevant information about a topic in dispute is at the core of the litigation process.[1] However, the advent of "e-discovery" is causing a rapid transformation in how that information is gathered. While discovery disputes are not new, the huge volume of available electronically stored information ("ESI") poses unique challenges. Some years ago, a party facing a review of information for production to the other side in a document-intensive case might have been concerned with hundreds of "banker's" boxes of documents.

Today, that same amount of data is easily found on a single computer hard drive.[2] Moreover, as the ability to create and store massive volumes of electronic information mushrooms, the cost to store that information inversely plummets. In 1990, a gigabyte of storage cost about $20,000; as of 2013, two-terabyte drives readily sell for less than $70, or 3.5¢ per gigabyte, with even lesser rates charged for hosting gigabytes in the "cloud." As a result, more individuals and organizations are generating, receiving, and storing more data, which in some cases means more information must be identified, collected, reviewed, and produced in litigation.

With billable rates for associates at many law firms averaging between $200 and $500 per hour,[3] the cost to review just one gigabyte of data can easily exceed $30,000.[4] These economic realities—i.e., the huge cost differential between the nominal cost to store a gigabyte of data and the $30,000 to review it—act as drivers to change traditional attitudes and approaches of lawyers, clients, courts, and litigation support providers forced to search for relevant evidence during discovery and investigations. Data volumes now numbering in the billions of ESI objects, review costs, and shrinking discovery timetables have created the need for a profound change in practice.

As discussed in this Commentary, just as technology has given rise to these new litigation challenges, technology can help to solve them. The emergence of new discovery strategies, best practices, and processes, as well as new search and retrieval technologies are transforming the way lawyers litigate. Collectively, they provide opportunities for huge volumes of information to be reviewed faster, more accurately, and more affordably than ever before. The good news is that search and retrieval systems are improving in effectiveness and expanding their capacities,

1 *Hickman v. Taylor,* 329 U.S. 495, 507 (1947) ("Mutual knowledge of all the relevant facts gathered by both parties is essential to proper litigation.").

2 Here's why: One gigabyte of electronic information can generate approximately 70,000 to 80,000 pages of text, or 35 to 40 banker's boxes of documents (at 2,000 pages per box). Thus, a 250 gigabyte storage device (e.g., a laptop or hard drive), theoretically, could hold as much as the equivalent of 8,750 to 10,000 banker's boxes of documents. In contrast, in 1990, a typical personal computer held just 200 megabytes of data—less than 1/1000 the capacity of a typical hard drive today. Even if only 10% of a computer's available capacity today contains user-created information (as distinguished from application programs, operating systems, utilities, etc.), attorneys still would need to consider and potentially review 1,750,000 to 2,000,000 pages per device.

3 See Alex Vorro, *Law firm billing rates steadily climbing despite down economy*, (April 17, 2012) (citing TyMetrix Legal Analytics' 2012 Real Rate Report™), http://www.insidecounsel.com/2012/04/17/law-firm-billing-rates-steadily-climbing-despite-d.

4 *See infra* note 17 and accompanying text.

buoyed by a tsunami of activity aimed at improving the "search" experience for users generally.[5] For example, advanced forms of machine learning—including supervised and unsupervised document and content classifiers—can automatically organize ESI in new ways not achieved by the more familiar methods of the past (which include the use of simple "keywords" as an automated aid to conducting manual searches). And not only can these new techniques increase accuracy and efficiency, through the proper use of statistical sampling and validation techniques, practitioners can measure the accuracy of the results of either traditional or alternative forms of search, retrieval, and review.

New challenges require new solutions. This Commentary aspires to present the bench and bar with an intelligible picture of the new challenges associated with the search and retrieval of ESI. The Commentary also presents alternative ways to address those challenges and to select the best solution for a given set of circumstances, taking into account the just, speedy, and inexpensive determination of every action (consistent with Federal Rule of Civil Procedure 1).

5 One indication of the amount of ongoing effort and investment to improve search and retrieval capabilities is evidenced by the research and development spending of internet and technology giants Google, Microsoft, Apple, and IBM. According to published reports, Google spent $3.76 billion, Microsoft spent $8.7 billion, Apple spent $1.78 billion, and IBM spent $6.03 billion on core research and development activities in 2010. See Booz & Company, *The 2011 Global Innovation 1000 – Why Culture Is Key* (October, 2011), http://www.booz.com/media/file/BoozCo-Global-Innovation-1000-2011-Webinar.pdf.

Executive Summary

Discovery has changed. For a growing number of cases, the process of identifying, reviewing, and producing information has evolved from the manual review of paper documents to an evaluation of vastly greater volumes of ESI.

A perfect review of the resulting volume of information is impossible. It is also not economical. But governing legal principles and best practices do not require perfection in making disclosures or in responding to discovery requests.[6] Instead, best practices focus on reasonable and proportional actions taken by practitioners as part of their duties, which must include an appreciation for the particular challenges of electronic information.

The Sedona Conference has helped establish the benchmarks governing the evolution and refinement of reasonable, good faith practices for searching intimidating amounts of data. Principle 6 of The Sedona Principles, Second Edition (2007) notes that "[r]esponding parties are best situated to evaluate the procedures, methodologies and technologies appropriate for preserving and producing their own electronically stored information," and Principle 11 amplifies the point by stating that "[a] responding party may satisfy its good faith obligation to preserve and produce relevant electronically stored information by using electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonably likely to contain relevant information."

This Commentary discusses the existing and evolving methods by which a party may choose to search unprecedented volumes of information. As the practice of using these "search and retrieval" technologies—the generic term we will utilize in this Commentary—advances, a new understanding of what is "reasonable" and "proportional" under any particular set of circumstances will advance as well. Therefore, the challenges addressed by this Commentary go beyond litigation and also encompass the full breadth of the search and retrieval of information from large volumes of data.

The Revolution in Discovery

Not long ago, all information was stored on physical records such as paper. There was typically a single "original" document, and the number of duplicate copies and their locations were generally limited. Administrative assistants, file clerks, records managers, and archivists developed expertise in managing that storage, generally pursuant to pre-existing file systems. In the case of litigation, it was reasonable and relatively easy (in all but the exceptional case), for a legal representative to gather, manually review, and prepare each individual item prior to its production.

The digital revolution did more than make documents truly portable—it also created a review-process paradigm shift in terms of what is truly feasible regarding document review in

6 *See, e.g.*, *Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC,* 685 F. Supp. 2d 456, 461, 465 (S.D.N.Y. 2010), *overruled in part on other grounds by Chin v. Port Authority of New York & New Jersey,* 685 F.3d 135 (2nd Cir. 2012).

litigation. This revolution has shifted nearly all information storage (as a percentage of existing information) to the digital realm and has caused an explosion in the amount of information that resides in any organization. And not only did the information's volume and format change, the very geography of where information "lives" moved from a file cabinet to a broad distribution amongst many different storage devices: from large mainframe computers to hand-held devices.

Each device may be capable of storing the equivalent of several warehouses of paper documents; each device may also have networking capabilities which allow it to integrate into complex systems. These systems are intricate, interdependent, and evolve spontaneously, behaving nearly like living ecosystems. To further complicate the picture, a legal professional who completely understands the workings of this new form of "information ecosystem" is rare indeed.

Finally, in addition to the search and retrieval challenge, a large percentage of the records searched in litigation are written in human language, not just numbers. Human language is an inherently elastic, ambiguous, "living" tool of enormous power. Its elasticity allows for jargon, private codes, and discrete vocabularies to exist in different subcultures in any organization, thereby making the identification of search terms much more challenging.

Essential Conclusions of this Commentary

This Sedona Conference "Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery" strives to set forth state-of-the-art knowledge defining the challenges associated with searching enormous databases for relevant information (the "Problems"), and presents methods and tools to retrieve that information with a minimum of wasted effort (the "Solutions").

By way of summary, we set forth our conclusions about the Problems and their Solutions, and summarize our Practical Advice articulated in the balance of this Commentary.

Problems

- The exponential growth in digital information is a critical challenge to the justice system.
- Parties are frequently unable to identify ESI that is likely to contain information relevant to the claims and defenses in the dispute.
- Electronically stored information consists of human language, which challenges computer search tools. These challenges are posed by the ambiguity inherent in human language; the imprecision resident in human use of logic; and the tendency of people within organizations or networks to speak in metaphor, to invent their own words, and to communicate in jargon, short forms, or code.
- The application of simple keyword search, while still a valuable tool, has well-documented deficiencies. There are also documented problems with manual document review.

Solutions

- Educating clients that good information governance reduces e-discovery costs by reducing the volume of ESI that is kept, and effective management of the ESI that is kept results in collecting a smaller data set with a higher concentration of relevant information.
- Counsel should work closely with their clients to identify and then narrow sources of ESI that are likely to contain information relevant to the dispute.
- The proper selection of information for production in discovery can benefit from the learning from a variety of other disciplines, including, but not limited to, Information Retrieval science, linguistics, and the implementation of effective project management processes.
- Alternative search tools may properly supplement simple keyword search and Boolean search techniques. These include using various forms of computer- or technology-assisted review, machine learning, relevance ranking, and text mining tools which employ mathematical probabilities, as well as other techniques incorporating supervised and unsupervised document and content classifiers.[7]
- Parties and their counsel should cooperate and seek ways to agree on measurements to evaluate the effectiveness of the search and retrieval process. The metrics currently used in information retrieval science, most notably "precision" and "recall," may serve as key points of reference.

Practical Advice

Practice Point 1. *In many settings involving large amounts of relevant electronically stored information ("ESI"), relying 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 under certain circumstances.*

Practice Point 2. *The successful use of any automated search method or technology will be enhanced by a well-thought-out process with substantial human input on the front end.*

Practice Point 3. *The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed. Parties and their counsel must match the use case with the tools and best practices appropriate to address it, and must incorporate proportionality considerations involving the overall costs and the stakes of the litigation.*

7 There is great variation in the description and application of these technologies, whether for technical, sales, or marketing differentiation or other business purposes. Some of the terms currently in use as of the date of this Commentary include: computer-assisted review; technology-assisted review; predictive coding; relevance ranking; text mining; tools that employ mathematical probabilities; as well as other techniques, including fuzzy logic to capture variations on words, and conceptual search, which makes use of taxonomies and ontologies assembled by linguistic means. For a glossary of terms relating to these technologies *see* Maura R. Grossman and Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review with Foreward by John M. Facciola, U.S Magistrate Judge*, 2013, Fed. Cts. L. Rev. 7 (January 2013), http://www.fclr.org/fclr/articles/html/2010/grossman.pdf.

Practice Point 4. *Parties and their counsel should perform due diligence when choosing a particular information retrieval product or vendor service.*

Practice Point 5. *Because of the characteristics of human language, no search and information retrieval tool can guarantee the identification of all responsive documents in large data collections. Moreover, different search methods may produce different results, subject to a measure of statistical variation inherent in the science of information retrieval.*

Practice Point 6. *Parties and their counsel should make a good faith attempt to cooperate when determining the use of particular search and information retrieval methods, tools, and protocols (including keywords, concepts, computer- or technology-assisted review and other types of search parameters and quality control measures).*

Practice Point 7. *Parties and their counsel should expect that their choice of search methodology (and any validation of it) will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and at trial).*

Practice Point 8. *Parties, counsel, and the courts should be alert to new and rapidly evolving search and information retrieval methods. Moreover, parties and their counsel should recognize that information retrieval is a distinct field of study that includes expertise in such areas as computer science, statistics, and linguistics, and that consultation with or utilization of experts in information retrieval may improve the quality of search results in complex cases involving large volumes of ESI.*

How The Legal Community Can Contribute to The Growth of Knowledge

A consensus is forming in the legal community that human review of documents in discovery is expensive, time consuming, and error-prone. There is also a growing awareness that, used correctly, linguistic and mathematically-based content analysis, embodied in new forms of search and retrieval technologies, tools, techniques, and processes in support of the review function, can effectively reduce litigation cost, time, and error rates.

Recommendations

1. *The legal community should continue to support collaborative research with the scientific and academic sectors aimed at establishing the efficacy of a range of automated search and information retrieval methods.*

2. *The legal community should encourage the establishment of objective benchmarking criteria, to assist lawyers in their evaluation of the competitive legal and regulatory search and retrieval services market.*

Members of The Sedona Conference community have and will continue to participate in collaborative workshops and other forums dedicated to information retrieval issues. The Sedona Conference intends to remain in the forefront of the efforts of the legal community aimed at seeking out centers of excellence in this area, including the possibility of fostering private-public partnerships focusing on continued research.

I. Introduction

The exponential growth in the volume and complexity of ESI found in modern organizations poses a substantial challenge to the justice system. Today, even routine discovery requests can require searches of, and retrieval from, the storage devices found on servers, networked workstations, desktops and laptops, home computers, removable media (such as CDs, DVDs, and USB flash drives), handheld devices (such as PDAs, smart phones, cell phones, and iPods), and the "cloud." Complicating things further, such information is now almost always flowing robustly throughout a "network," in which it has likely been replicated, distributed, modified, linked, attached, accessed, backed-up, overwritten, deleted, undeleted, fragmented, defragmented, morphed, and multiplied. Complying with preservation or discovery obligations in some ESI cases may require a process to identify relevant emails from among thousands, millions, or even tens-of-millions of individual messages, with attachments in various file formats.

The volume and complexity of ESI highlights several issues: First, whether automated search and information retrieval methods are reliable and accurate, and if so, how accurate. Second, whether the legal profession has the skills, knowledge, and processes required to use such automated search and retrieval methods intelligently in conjunction with huge data sets, in ways that are defensible under the rules governing discovery.

In The Sedona Principles, Second Edition (2007), The Sedona Conference endorsed several highly pragmatic and relevant consensus best practices relevant to this discussion.[8]

First, Principle 6 provides that parties responding to discovery are in the best position "to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information." Principle 11 expands this concept to include the use of "electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonably likely to contain relevant information."

Second, the Commentary to Principle 11 provides that the "selective use of keyword searches can be a reasonable approach when dealing with large amounts of electronic data," and states that it "is also possible to use technology to search for 'concepts,' which can be based on ontologies, taxonomies, or data clustering approaches, for example."[9] This exploits a unique feature of electronic information: the ability to conduct fast, iterative searches for the presence of patterns of words and concepts in large document populations. The Commentary to Principle 11 also states that "[c]ourts should encourage and promote the use of search and retrieval techniques in appropriate circumstances," and suggests that "[i]deally, the parties should agree on the search methods, including search terms or concepts, to be used as early as practicable. Such agreements should take account of the iterative nature of the discovery process and allow for refinement as the parties' understanding of the relevant issues develops."[10]

8 *The Sedona Principles, Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2d ed. 2007) (hereinafter *The Sedona Principles*), *available at* https://thesedonaconference.org/download-pub/81.

9 *Id.* at Comment 11.a.

10 *Id.*

Third, The Sedona Conference has recognized that "there are now hundreds of companies offering electronic discovery services."[11] This is also true of search and information retrieval products and services for use in legal contexts—which form a subset of a burgeoning sector of the economy devoted to improving users' "search" experience. However, there remains some confusion as to the strengths and weaknesses of such tools. Legal practitioners have a need for guidance as to the appropriate use of search and information retrieval technologies. Such guidance can help practitioners judge the relative costs and benefits of such tools in specific cases.

This Commentary is designed to help educate the justice system—attorneys, judges, and litigants alike—on "state of the art" search and retrieval tools, techniques, and methodologies, and how they can best be used as part of an overall process to more efficiently manage discovery. This discussion includes the critically important concept of an integrated process of search and retrieval; the ability to differentiate among different search methods; how to evaluate such differences; and what questions to ask before using any particular method or product in a specific legal setting.

For the past three decades, the legal community has had familiarity with simple keyword and natural language searches on Westlaw® and Lexis® in the context of legal research, and to a lesser extent the use of "Boolean" logic to combine keywords and "operators" (such as "AND," "OR," and "AND NOT," or "BUT NOT") that produce broader or narrower searches. Over time, lawyers have applied this knowledge to employ simple keyword, Boolean, and other search and retrieval tools to reduce the amount of information to be reviewed for production in discovery.[12] In the past few years, the relative efficacy of competing search and retrieval tools used to accomplish review for production has begun to be measured. However, the field is still wide open for the development of more advanced search and information retrieval best practices. These methods include merging keyword search with more sophisticated systems that use computer- and technology-assisted techniques, and incorporating mathematical algorithms and various forms of linguistic techniques, to help find, group, and present related content.

What follows is an in-depth analysis of the problems lawyers confront in managing massive amounts of data in discovery, including how search and retrieval techniques are used in everyday practice and the key element of "process." This Commentary also provides background on the field of information retrieval and at least partially describes the world of search tools, techniques,

---

11 *The Sedona Conference Best Practices for the Selection of Electronic Discovery Vendors: Navigating the Vendor Proposal Process* (2007), *available at* https://thesedonaconference.org/download-pub/80.

12 There may be a role for the use of some type of search and retrieval technology in discharging obligations to preserve ESI, as well as during the initial pre-review data culling or "collection" phases, in anticipation of complying with specific ESI and document requests. During the collection phase, for example, the goal is to maximize the amount of potentially relevant evidence in a subset of the greater universe of available ESI, without necessarily selecting only the more relevant information that might be the focus of the review phase preceding production. Accordingly, parties may well end up using (and agreeing to use) differing search methods in the initial collection and later review phases of litigation. While we acknowledge that use of advanced search tools during earlier phases of litigation (e.g., during early case assessment, at preservation, etc.) remains cutting edge and worthy of future discussion, the primary focus of this Commentary will be on search tools as they are used in the review process. *See generally* Thomas Y. Allman, Jason R. Baron, and Maura R. Grossman, *Preservation, Search Technology, and Rulemaking*, 30 The Computer and Internet Lawyer No. 2 (February 2013); Mia Mazza, Emmalena K. Quesada, and Ashley L. Sternberg, *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 (2007), at [53] & [60], *available at* http://law.richmond.edu/jolt/v13i3/article11.pdf (discussing the use of concept searching in regard to preservation); *The Sedona Principles*, *supra* n.8 ("Organizations should internally address search terms and other filtering criteria as soon as possible so that they can begin a dialogue on search methods as early as the initial discovery conference.").

and methodologies that are currently commercially available. It also includes "practice pointers" on the factors to consider in making an overall legal evaluation among different search methods, both on a conceptual and practical level. In a concluding section, the future of search and retrieval efforts is discussed. A more technical discussion of various search methodologies is included in an Appendix. Where appropriate, reference will be made to technical definitions found in the updated Sedona Glossary.

II. The Search and Information Retrieval Problem Confronting Lawyers

Discovery today is drowning in an exponential flood of potential sources of information. This increase in volume, especially since the mid-1990s, is principally due to the combined effects of the PC revolution, the widespread use of email and other new forms of communication, and the growth of mobile device and social networks. Indeed, the implication of this growth in volume is that it places at severe risk the justice system's ability to achieve the "just, speedy and inexpensive" resolution of disputes, as contemplated by Rule 1 of the Federal Rules of Civil Procedure.

The Rise of Crushing Volumes of Information in the Digital Realm

A history of the computer and information technology advances occurring since the mid-1970s is beyond the scope of this Commentary. Suffice it to say that over the last 40 years, there has been a fast-paced and widespread shift from physical information storage technologies to new, digital information storage technologies. This "digital realm" was created by an accretion of technological advances, each built on preceding advances, which together have resulted in as fundamental a shift in the way information is shared, such as that which occurred in 1450 when Johannes Guttenberg invented the printing press. Included among the advances contributing to the new "digital realm" are the invention of the microchip, the development and diffusion of the personal computer, the spread of various types of networks linking together both computers and other networks, the rise of email and its dominant use in the business world, the plunging cost of computing power and storage, and of course, the spread of the Internet and with it, the World Wide Web.[13]

By the mid-1990s, networked computers and their storage devices had created a true information-based-society, with a constant flow of messages in all forms exchanged on a 24/7 basis. For example, studies reflect that the typical corporate worker sends and receives about 105 emails per day.[14] The size and nature of the attachments to these emails is also growing, with increased integration of image, audio, and video files. More recently, there has been a similar explosion in the use of instant and text messaging throughout organizations, including increasingly, through the use of mobile devices. In many organizations, the average worker maintains several gigabytes of stored data.[15] At the same time, the costs of storage have plummeted from $20,000 per gigabyte in 1990 to less than 3.5¢ per gigabyte in 2013.[16] Existing technologies are only

13 *See* George L. Paul and Jason R. Baron, *Information Inflation: Can the Legal System Adapt?,* 13 Rich. J. L. & Tech. 10 (2007), at [1], n.2, *available at* http://law.richmond.edu/jolt/v13i3/article10.pdf ("Organizations have thousands if not tens of thousands of times as much information within their boundaries as they did 20 years ago."); Peter Lyman and Hal R. Varian, *How Much Information*, 2003, *available at* http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

14 *See* Sara Radicati, Ed., *Email Statistics Report, 2011-2015*, The Radicati Group, Inc., May 2011, *available at* http://www.radicati.com/wp/wp-content/uploads/2011/05/Email-Statistics-Report-2011-2015-Executive-Summary.pdf.

15 One gigabyte is equivalent in volume to between 70,000 and 80,000 pages of material. At 2,000 pages per box, one gigabyte is therefore equivalent to 35 to 40 boxes of documents. *See supra* n.2.

16 *See Memory Storage Density*, Wikipedia, http://en.wikipedia.org/wiki/Memory_storage_density#Effects_on_price (last visited November 25, 2013); Michelle Kessler, *Days of officially drowning in data almost upon us*, USA Today, Mar. 5, 2007, www.usatoday.com/tech/news/20070305data_N.htm. Cloud storage reduces the costs of storage even further. *See, e.g.,* https://developers.google.com/storage/docs/pricingandterms (last visited November 25, 2013); http://mashable.com/2012/08/21/amazon-glacier.

beginning to grapple with providing a viable automated means for applying records retention requirements, including the ability to implement legal holds, in the new digital world.

Organizations have continued to aggressively leverage technology to increase productivity. But leveraged technology sometimes comes with a lack of oversight or control. In many organizations, no one really controls how, where, how many times, and in how many forms information is stored. For example, copies of the same Word document can be found in email attachments, local hard drives, network drives, document management systems, websites, and on all manner of backup and removable media, such as USB flash drives, CDs, DVDs, and so on.

Discovery In the Recent Past: Manageable Amounts of Physically Stored Information

Historically, outside counsel played a key role in a comparatively simpler discovery process: Litigants, assisted by their counsel, identified and collected information that was relevant to pending or reasonably foreseeable litigation. Counsel manually reviewed the information and produced any information that was responsive and not otherwise protected from disclosure by the attorney-client privilege, the attorney work product, or by trade secret protections.

This worked fine in the days where most of the potentially relevant information was created in or was stored in printed, physical form, and in reasonable volumes, so that it required only "human eyes" to review and interpret it. However, with increasingly complex computer networks, and the exponential increase in the volume of information existing in the digital realm, the venerated process of "eyes only" review is no longer generally workable or economically feasible.

The cost of manual review of such volumes is prohibitive, often exceeding the damages at stake. Anecdotal reports indicate that the cost of reviewing information can easily exceed thousands of dollars per custodian, per event, for collection and attorney review. Litigants often cannot afford to review all available ESI in the time permitted for discovery.[17] Accordingly, the conventional document review process is poorly adapted to a growing percentage of today's litigation.[18] Lawyers of all stripes therefore have a vital interest in utilizing automated search and retrieval tools where appropriate. The plaintiff's bar has a particular interest in being able to efficiently extract key information received in mammoth document productions, and in automated tools that facilitate the process. The defense bar has an obvious interest in reducing attendant costs, increasing efficiency, and in better risk management of litigation (including reducing surprises). All lawyers, clients, and judges have an interest in reducing cost and barriers to entry to the justice system, and maximizing the quality of discovery, by means of using automated tools that produce a reliable, reproducible, and consistent result.

---

17 Compare mere pennies to store a gigabyte of data with $32,000 to review it in a traditional, linear fashion (i.e., assuming one gigabyte equals 80,000 pages and assuming that an associate billing $200 per hour can review 50 documents per hour at 10 pages in length, such a review would take 160 hours at $200/hr., or approximately $32,000). *See generally* Nicholas M. Pace & Laura Zakaras, Rand Study, Where the Money Goes: Understanding Litigant Expenditures for producing Electronic Discovery 17-27 (2012) [hereinafter Rand 2012 Study], http://www.rand.org/pubs/monographs/MG1208.html.

18 Not all cases are equally reliant on electronic discovery—from time to time, counsel may even forgo the production of electronically stored information and rely solely on hard-copy documents.

Ideally, then, judges and litigants should strive to increase their awareness of search and retrieval sciences generally, and of the sciences' appropriate application to discovery. Some technologies have been used for years to produce documents from large litigant document databases, but often without much critical analysis. The legal system may benefit from the rich body of research available through the Information Retrieval and library science disciplines. The discussion that follows is designed to provide a common framework and vocabulary for proper application of search and retrieval technologies in this new "age of information complexity" in the legal environment.

The Reigning Myth of "Perfect" Retrieval Using Traditional Means

It is not possible to discuss this issue without noting that there appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible—perhaps even perfect—and constitutes the gold standard by which all searches should be measured. Even assuming that the profession had the time and resources to continue to conduct manual or "linear" review of massive sets of electronic data (which it does not), the relative efficacy of that approach versus utilizing newly developed automated methods of review appears to be increasingly in question.[19] Moreover, past research demonstrates the gap between lawyers' expectations and the true efficacy of certain types of keyword searches. The Blair and Maron study (discussed below) shows that human beings are far less accurate and complete than they believe themselves to be when searching and retrieving information from a heterogeneous set of documents (i.e., in many data types and formats), using ad hoc, simple keywords as the sole means to identify potentially relevant documents. The importance of this point cannot be overstated, as it provides a critical frame of reference in evaluating how new and enhanced forms of automated search methods and tools may benefit litigation practices.

The Intelligent Use of Tools

Although the continued use of manual search and review methods may be infeasible or even indefensible in discovery involving significant amounts of ESI, merely adopting sophisticated automated search tools, alone, will not necessarily lead to successful results. Lawyers must recognize that the process by which a legal team uses such tools, including close involvement of lead counsel, is just as important as the automated tools themselves. This may require an iterative process which importantly incorporates feedback and learning and allows for measurement and validation of results. The time and effort spent up front on designing a sophisticated discovery process that targets the real needs of the litigation must be viewed as a condition precedent to deploying automated methods of search and retrieval.

19 *See infra* text accompanying note 47.

III. Lawyers' Current Use of Search and Retrieval Methodologies

Attorneys across all disciplines are generally familiar with search and retrieval methodologies based on their exposure over the past thirty-plus years to automated means of searching of caselaw and other databases provided by LexisNexis® and Westlaw®. More recently, lawyers are using Google® and other Web-based search engines to hunt down the increasing amounts of online information relevant to their practices. Additionally, law firms and corporate legal departments use search methods for administrative matters, such as locating data on personnel, supporting billing functions, managing conflicts of interest, and for contact management. Many products employing search methods of various kinds exist in the legal marketplace to assist lawyers in these functions.

Current Database Tools in the Practice of Law

Litigators use automated search and retrieval tools at many stages of the litigation process. PACER and other automated means are used to uncover data on opposing counsels' pleadings, motions, and pretrial filings in similar litigation, as well as showing how a judge has ruled on similar issues even if unreported in legal reporting services. Lawyers also use a variety of search methods involving online, CD-ROM, client-developed, and "cloud" databases to unearth facts on opposing parties, witnesses, and even potential jurors. At later stages of litigation, lawyers use various litigation support software applications to search through potential exhibits in connection with proceedings held in "electronic courtrooms." But until recently, litigators seldom used automated search and retrieval methods with their clients' or their opponents' growing collections of unstructured ESI.

"Deduplication" in the Processing of ESI

With the exponential increase in the volume of data subject to e-discovery, lawyers have begun to take steps towards employing automated search tools to manage the discovery process. One example of this is "deduplication" software used to find duplicate electronic files, since ESI often consists of a massively redundant universe. For example, the same email can be copied tens or even hundreds of times in different file locations on a network or on backup media. Deduplication software reduces the time attorneys must spend reviewing a large document set and helps to ensure consistent classification of documents for responsiveness or privilege.[20] Increasingly, "email threading" and "near deduplication" tools are used to assist in organizing and expediting overall document reviews, even if the technique is not used to reduce the actual number of unique documents subject to review.[21]

20 "Deduplication" tools tag identical documents as duplicates by means of a "binary hash function" (a mathematical way of comparing the text of two documents represented in the underlying digital 1's and O's actually stored on the computer to see if the documents are perfectly alike). Deduplication by binary hash has been widely used without much notice in court opinions to date. See *Wiginton v. CB Richard Ellis*, Inc., 229 F.R.D. 568, 571 (N.D. Ill. 2004) (referring to deduplication process); *Medtronic Sofamor Danek Inc. v. Michelson*, 229 F.R.D. 550, 561 (W.D. Tenn. 2003) (same).

21 "Email threading" refers to a particular message and a running list of all subsequent replies pertaining to that original email. David D. Lewis & Kimberly A. Knowles, *Threading electronic mail: A preliminary study* 33 Information Processing and Management 209–217 (1997) ("Near deduplication" involves files that "are not hash value duplicates but are materially similar."), *available at* http://pdf.aminer.org/000/936/211/threading_electronic_mail_a_preliminary_study.pdf.

The Use of "Keywords"

The most commonly used search methodology today still entails the use of "keyword searches" of full text and metadata as a means of filtering data for producing responsive documents in civil discovery. For the purpose of this Commentary, the use of the term "keyword searches" refers to set-based searching using simple words or word combinations, with or without Boolean and related operators (see below and Appendix for definitions). The ability to perform keyword searches against large quantities of evidence has become a widely accepted practice, as recognized by the courts. As one United States Magistrate Judge stated in 2004, "the glory of electronic information is not merely that it saves space but that it permits the computer to search for words or 'strings' of text in seconds."[22]

Courts have not only accepted, but in many cases have ordered, the use of keyword searches to define discovery parameters and resolve discovery disputes.[23] Early on, one court suggested that a party might satisfy its duty to preserve documents in anticipation of litigation by conducting a system-wide keyword search and preserving a copy of each "'hit."[24]

Because of the costs and burdens associated with the review of increasingly vast volumes of electronic data, it makes sense in appropriate cases for producing parties to negotiate with requesting parties in advance to define the scope of discoverable information. For example, parties could agree on conducting a search of only files maintained by key witnesses, in certain data sources, and/or for certain date ranges. They may negotiate and agree to a set of keywords relevant to the case. Both sides might see the advantage to using such protocols or filters to reduce the volume of extraneous information, such as spam, routine listserv notifications, and personal correspondence, typically found when searching through electronic data collections.[25]

---

22 *In re Lorazepam & Clorazepate,* 300 F. Supp. 2d 43, 46 (D.D.C. 2004); *see also In re CV Therapeutics, Inc.,* 2006 WL 2458720 (N.D. Cal. Aug. 22, 2006) (court endorses employment of search terms as reasonable means of narrowing production); *J.C. Associates v. Fidelity & Guaranty Ins. Co.,* 2006 WL 1445173 (D.D.C. 2006) (requiring search of files using four specified keywords); *FTC v. Ameridebt, Inc.,* 2006 WL 618563 (N.D. Cal. Mar. 13, 2006) ("email could likely be screened efficiently through the use of electronic search terms that the parties agree upon"); *Windy City Innovations, LLC v. America Online, Inc.,* 2006 WL 2224057 (N.D. Ill. July 31, 2006) ("[k]eyword searching permits a party to search a document for a specific word more efficiently"); *Reino de Espana v. Am. Bureau of Shipping*, 2006 WL 3208579 (S.D.N.Y. Nov. 3, 2006) (court approves of keyword search for names and email addresses as a "targeted and focused discovery search"); *U.S. ex rel. Tyson v. Amerigroup Ill.*, Inc., 2005 WL 3111972 (N.D. Ill. Oct. 21, 2005) (referencing agreement by parties to search terms); *Medtronic*, 229 F.R.D. at 561 (court orders defendant to conduct searches using the keyword search terms provided by plaintiff); *Alexander v. FBI,* 194 F.R.D. 316 (D.D.C. 2000) (court places limitations on the scope of plaintiffs' proposed keywords to be used to search White House email).

23 *See generally* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, 17 Rich. J. L. & Tech. 9 (2011) (compiling case law), *available at* http://jolt.richmond.edu/v17i3/article9.pdf.

24 *Zubulake v. UBS Warburg, LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004); *see also Zakre v. Norddeutsche Landesbank Girozentrale*, 2004 WL 764895 (S.D.N.Y. Apr. 9, 2004) (court denies plaintiff's request for additional indexing of records, holding that defendant's production of CD-ROMS in a text searchable form was sufficient, citing to Guideline 11 of *The Sedona Principles*, 2004 Edition). *cf. Cache La Poudre Feeds, LLC v. Land O'Lakes, Inc.,* 2007 WL 684001 (D. Colo. Mar. 2, 2007) (where court denied motion for sanctions based on an allegation that the opposing party failed to properly monitor compliance with its discovery obligations by not conducting keyword searches, court also stated that *The Sedona Principles*, 2004 Edition and *Zubulake* were not to the contrary).

25 *See also* R. Brownstone, *Collaborative Navigation of the Stormy e-Discovery Seas*, 10 Rich. J. L. & Tech. 53 (2004), *available at* http://law.richmond.edu/jolt/v10i5/article53.pdf (arguing that parties must agree to search terms and other selection criteria to narrow the scope to manageable data sets); *The Sedona Principles*, *supra* note 8, Comment 11.a ("For example, use of search terms could reveal that a very low percentage of files (such as emails and attachments) on a data tape contain terms that are responsive to 'key' terms. This may weigh heavily against a need to further search that source, or it may be a factor in a cost-shifting analysis. Such techniques may also reveal substantial redundancy between sources (i.e., duplicate data is found in both locations) such that it is reasonable for the organization to preserve and produce data from only one of the sources."). *See generally* Kenneth J. Withers, *Computer-Based Discovery in Federal Court Litigation*, 2000 Fed. Cts. L. Rev. 2 (2000), http://www.fclr.org/fclr/articles/html/2000/fedctslrev2.shtml (suggesting parties adopt collaborative strategies on search protocols).

In *Treppel v. Biovail Corp.*,[26] the defendant refused to produce documents because the plaintiff would not agree to keyword search terms. Citing to Principle 11 of the Sedona Principles for Electronic Document Production, the Court held that the defendant was justified in using keyword search terms to find responsive documents and should have proceeded unilaterally to use its list of terms when the plaintiff refused to endorse the list. The Court held that plaintiff's "recalcitrance" did not excuse defendant's failure to produce any records and ordered the company immediately to conduct the automated search, produce the results, and explain its search protocol. Another early case emphasized the need to confer after plaintiff was successful in obtaining a "mirror image" of data on all of defendant's computers.[27]

Issues With Keywords

There are nonetheless a number of notable limitations to the effectiveness of traditional or basic keyword search. Keyword searches work best when the legal inquiry is focused on finding particular documents and when the use of language is relatively predictable. For example, basic keyword searches work well to find all documents that mention a specific individual or date, regardless of context. However, while basic keyword search techniques have been widely accepted both by courts and parties as sufficient to define the scope of their obligation to perform a search for responsive documents, the experience of many litigators is that simple keyword search alone is inadequate in at least some discovery contexts. This is because simple keyword searches are both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages).[28]

Traditional keyword searches identify all documents containing a specified term regardless of context, often capturing many documents irrelevant to the user's query. For example, the term "strike" could be found in documents relating to a labor union tactic, a military action, options trading, or baseball, to name just a few (illustrating "polysemy," or ambiguity in the use of language). The problem of the relative percentage of "false positives" or noise in the data is potentially huge, amounting in some cases to enormous numbers of files which must be searched to find responsive documents.[29]

On the other hand, basic keyword searches have the potential to miss documents that contain a word that has the same meaning as the term used in the query, but is not specified. For example, a user making queries about labor actions might miss an email referring to a "boycott" if that particular word was not included as a keyword, and a lawyer investigating tax fraud via options trading might miss an email referring to "exercise price" if that term was not specifically searched (illustrating "synonymy" or variation in the use of language). And of course, if authors of records

26 233 F.R.D. 363 (S.D.N.Y. 2006).

27 *Balboa Threadworks v. Stucky,* 2006 WL 763668 (D. Kan. Mar. 24, 2006) (court ordered parties to meet and confer on the use of a search protocol, including keyword searching).

28 Some case law has held that keyword searches were either incomplete or over inclusive. *See Alexander, 194. F* F.R.D. at 316; *Quinby v. WestLB AG*, 2006 WL 2597900 (S.D.N.Y. Sept. 5, 2006) (court narrows party's demand for 170 proposed search terms in part due to the inclusion of commonly used words).

29 *See, e.g.*, G. Paul and J. Baron, *Information Inflation*, *supra* n.13, at [20] (discussing potential time and cost of searching through 1 billion emails); Craig Ball, *Crafting A More Effective Keyword Search*, Law Technology News (June 24, 2009), http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202431693400&slreturn=20130302113909.

are inventing words "on the fly," or using short-forms or code names (as they have done throughout history, and with increasing frequency in electronic communications), such problems are compounded.[30]

Keyword searches can also exclude common or inadvertently misspelled instances of the term (e.g., "Phillip" for "Philip," or "strik" for "strike") or variations on "stems" of words (e.g., "striking"). Even the best of optical character recognition (OCR) programs introduce a certain rate of random error into document texts, potentially transforming would-be keywords into gibberish. Finally, using keywords alone results in a return set of potentially responsive documents that are not weighted or ranked based upon their potential importance or relevance. In other words, each document is considered to have an equal probability of being responsive subject to further manual review.

More advanced keyword searches using "Boolean" operators and techniques borrowed from "fuzzy logic" may increase the number of relevant documents and decrease the number of irrelevant documents retrieved. These searches attempt to emulate the way humans use language to describe concepts. In essence, they simply translate ordinary words and phrases into a Boolean search argument. Thus, a natural language search for "all birds that live in Africa" is translated to something like ("bird* + liv* + Africa").

At the present time, it would appear that the majority of automated litigation support providers and software rely on some form of keyword search, although the legal landscape is changing (see discussion below). Such methods are limited by their dependence on matching a specific, sometimes arbitrary choice of language to describe the targeted topic of interest.[31]

However, these challenges can be at least partially overcome by employing a more methodical and informed approach to defining keywords. Such a process begins with a clear definition of relevance, outlining criteria to identify relevant documents for each issue and subtopic. The problem of false positives can be minimized by combining key terms within a certain proximity of one another or in a specified order. Singular keywords are often ambiguous, but disambiguating (for verbs) and specifying words (for nouns) when joined to the central keyword with Boolean operators can reduce over-inclusiveness. Additionally, gaps in keywords can be a big issue in early stages or when the issues at stake are relatively unknown. One approach to identify such gaps trains a software system on initial custodians (i.e., utilizes a computer-assisted review workflow) and uses the trained system to generate a set of potential keywords that can then be used to confirm or add to the original assumptions regarding keywords in subsequent stages of review.[32]

30 Philosophers use colorful imagery to describe the dynamism and complexity of human language. *See, e.g.*, Ludwig Wittgenstein, The Philosophical Investigations, Section 18 (G.E.M. Anscombe, trans., 3d ed. 1973) ("[T]o imagine a language is to imagine a form of life. … [L]anguage can be seen as an ancient city; a maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude of new boroughs with straight regular streets and uniform houses").

31 *See infra* Part IV; *see generally* S.I. Hayakawa, Language in Thought and Action (5th ed.1990) (stating that such methods are inherently limited by their specific choice of language to describe a specific object or reality).

32 *See generally The Sedona Conference Commentary on Achieving Quality in E-Discovery* (2013), https://thesedonaconference.org/download-pub/3556 (discussing how to construct a better quality search process).

Recent judicial opinions, including several citing to the 2007 Version of this Commentary, have examined many of the limitations of keyword search.[33] Notably, the Court in *William A. Gross Construction Associates* declared a

> [W]ake-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 e-mails or other electronically stored information ('ESI'). ... Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of 'false positives.'[34]

Use of Alternative Search Tools and Methods

Lawyers are beginning to feel more comfortable using alternative search tools to identify potentially relevant ESI. These more advanced text mining tools include, but are not limited to, "conceptual search methods," which rely on semantic relations between words, and/or use "thesauri" to capture documents that would be missed in keyword searching. Specific types of alternate search methods are set out in detail in the Appendix.

"Concept" search and similar information retrieval technologies attempt to locate information that relates to a desired concept without the presence of a particular word or phrase. The classic example is the concept search that will recognize that documents about Eskimos and igloos are related to Alaska, even if they do not specifically mention the word "Alaska." The first reported case referencing the possible use of "concept search" as an alternative to strict reliance on keyword search was decided in 2007.[35]

Other automated tools rely on "taxonomies" and "ontologies" to help find documents conceptually related to the topic being searched, based on commercially available data or on specifically compiled information. This information is provided by attorneys or developed for the business function or specific industry (e.g., the concept of "strike" in labor law vs. "strike" in options trading). These tools rely on the information that linguists collect from the lawyers and witnesses about the key factual issues in the case—the people, organization, and key concepts relating to the business as well as the idiosyncratic forms of communication that might be lurking in documents, files, and emails. For example, a linguist would want to know how union organizers or company officials might communicate plans, any special code words or lingo used in the industry, the relationships of collective bargaining units, the company's management structure, and other issues and concepts.

---

33 *See, e.g.*, *United States v. O'Keefe,* 537 F. Supp. 2d 14, 24 (D.D.C. 2008) (Facciola, M.J.); *Equity Analytics, LLC v. Lundin,* 248 F.R.D. 331, 333 (D.D.C. 2008) (Facciola, M.J.); *Victor Stanley, Inc. v. Creative Pipe, Inc.,* 250 F.R.D. 251, 260, 262 (D. Md. 2008) (Grimm, M.J.); *see also NDLON v. Immigration and Customs Enforcement, Dep't of Homeland Security,* 811F. Supp. 2d 713 (S.D.N.Y. 2011) (Scheindlin, J.); see *generally* Baron, *Law in the Age of Exabytes, supra* note 23.

34 *William A. Gross Construction Associates, Inc. v. American Manufacturers Mutual Insurance Co.,* 256 F.R.D. 134, 136 (S.D.N.Y. 2009) (Peck, M.J.); *see also* Hon. Andrew Peck, *Search, Forward*, Law Tech. News (Oct. 1, 2011) at 1-2, *available at* http://www.recommind.com/sites/default/files/LTN_Search_Forward_Peck_Recommind.pdf.

35 *See Disability Rights Council of Greater Washington v. Washington Metro. Transit Auth.,* 242 F.R.D. 139, 148 (D.D.C. 2007); *see generally Mazza, et al., supra* note 12, at [54] (discussing concept search).

Another type of search tool relies on mathematical probabilities that a certain text is associated with a particular conceptual category. These types of machine learning tools, which include "clustering" and "latent semantic indexing," are potentially helpful in addressing cultural biases of taxonomies because they do not depend on linguistic analysis, but on mathematical probabilities. They can also help identify communications hidden in code language and neologisms. For example, if the labor lawyer were searching for evidence that management was targeting neophytes in the union, she might miss the term "n00b" (a neologism for "newbie"). This technology, first used in government intelligence, and now increasingly used as part of computer- or technology-assisted review in e-discovery, is particularly apt in helping lawyers find information when they do not know exactly what to look for. For example, when a lawyer is looking for evidence that key players conspired to violate the labor union laws, she will usually not know the "code words" or expressions the players may have used to disguise their communications. For a discussion of recent developments in computer- or technology-assisted review, see below, text accompanying note 49.

With so many different search methods currently available, it is important to choose the most appropriate search strategy for any particular case. The choice of a search method will always depend heavily on the particular context. Practitioners should be aware of the strengths and limitations of varying approaches. Sometimes the most appropriate search method is obvious from the outset of a case; in other situations, the best method(s) only become evident after experimentation and use. But practitioners must recognize if a particular search method is ineffective and must be willing to modify their approach based on the results. See also Practice Pointers, below.

Resistance by the Legal Profession

Some litigators continue to primarily rely upon manual review of information as part of their review process.[36] Principal rationales are: (1) concerns that computers cannot be trusted to replace the human intelligence required to make complex determinations on relevance and privilege; (2) the perception that there is a lack of scientific validity of search technologies necessary to defend against a court challenge; and (3) widespread lack of knowledge (and confusion) about the capabilities of automated search tools.

Other parties and litigators may accept simple keyword search, yet be reluctant to use alternative search techniques. They may not be convinced that the chosen method would be defensible if confronted with a court challenge. They may perceive a risk that problem documents will not be found despite the additional effort, or that documents might be missed which would otherwise be picked up in a straight keyword search. Moreover, acknowledging that there is no one solution for all situations, they may opt for an accepted, lowest common denominator approach. Finally, litigators often lack the time and resources to sort out these highly complex technical issues on a case-by-case basis.[37]

---

36 *But see In re Instinet Group, Inc.,* 2005 WL 3501708 at *3 (Del. Ch. Dec. 14, 2005). The court reduced plaintiffs' attorneys' fee claim by $1 million (75% of the total claim) for "obvious" inefficiencies in plaintiffs' counsel's review of paper printouts ("blowbacks") from digital files. The court stated that plaintiffs' counsel's decision to "blow back" the digital documents to paper "both added unnecessary expense and greatly increased the number of hours required to search and review the document production."

37 *See*, *e.g.*, Ron Friedmann, *A Future Beyond Hammers*, Strategic Legal Technology Blog (Feb. 4, 2005, 1:38PM), http://prismlegal.com/a-future-beyond-hammers/ (suggesting that not one solution fits all cases); *see also* Ron Friedmann, *Thoughts on Full Text Retrieval (A KM and Litigation Support Topic)*, Strategic Legal Technology Blog (July 30, 2003) (questioning the incremental value of sophisticated searching over simple searching because of the costs of implementation and need to build taxonomies and to test methodologies).

But the legal landscape is changing rapidly. The year 2012 saw the first judicial opinions approving the use of the alternative search method of computer- or technology-assisted review (as described in Section V, below).[38] In the *Moore* opinion, the magistrate judge noted that "[c]ounsel no longer have to worry about being the "first" or "guinea pig" for judicial acceptance of computer-assisted review. ... Computer-assisted review now can be considered judicially-approved for use in appropriate cases."[39] This and future precedent will hasten acceptance by lawyers (and their clients) of the use of such alternate methods and techniques.

Challenging the Choice of Search Method

Challenge to the choice of a search methodology used in a review prior to production can arise in one of two contexts: (1) a requesting party's objection to the unilateral selection of a search method by a responding party; or (2) a court's *sua sponte* review of the use of a method or technology. Accordingly, the preferable method to preempt challenges—advocated by the proponents of the 2006 Federal Rules Amendments and some practitioners—is for a full and transparent discussion among counsel of the search methodology. Where the parties are in agreement on the method and a reasonable explanation can be provided, it is unlikely that a court will second-guess the process.

Absent agreement, a party has the presumption, under Sedona Principle 6, that it is in the best position to choose an appropriate method of searching and culling its data. However, a unilateral choice of a search methodology may risk challenge if an opponent can show that the results of the search are not accurate, complete, or reliable. As a practical matter, those who might object to a particular search and retrieval technology may face several challenges. First, the legal system has, for decades, blessed the use of keyword search tools and databases for discovery review. And second, if human review or even keyword searching is the benchmark for accuracy and completeness, it arguably should not be difficult to measure new technologies against keyword search or human review, especially when guided by a reasonable process. The discovery standard is, after all, reasonableness, not perfection.

Given the continued exponential growth in information, a large body of precedent will likely develop over time that critically analyzes new and alternative search methods in use in particular legal contexts. Indeed, the first case in which a court held an evidentiary hearing on a challenge to the use of keywords in favor of one type of alternative computer-assisted review method occurred in 2012.[40]

---

38 *See Moore v. Publicis Groupe SA*, 287 F.R.D. 182 (S.D.N.Y. 2012) (Peck, M.J.), *aff'd*, *Moore v. Publicis Groupe SA*, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012) (Carter, J.) (approving joint search protocol for technology-assisted review); *see also In re Actos (Pioglitazone) Products*, 2012 WL 3899669 (W.D. La. July 27, 2012) (same); *Global Aerospace Inc. v. Landow Aviation LP*, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012) (Chamblin, J.) (granting responding party's request to use technology-assisted review); *EORHB, Inc. v HOA Holdings, LLC*, #1 2012 WL 5399073 (Del. Ch. Oct. 15, 2012) (ordering, *sua sponte*, parties to use technology-assisted review with same vendor).

39 *Moore*, 287 F.R.D. at 195.

40 *See Kleen Prod., LLC v. Packaging Corp. of Am.,* 2012 WL 4498465 (N.D. Ill. Sept. 28, 2012) (Nolan, M.J.) (discussing evidentiary hearings held prior to parties reaching agreement on choice of method).

IV. Some Key Terms, Concepts, and History in Information Retrieval Technology

The evaluation of Information Retrieval ("IR") systems has, at least until recently, largely been of interest to computer scientists and graduate students in information and library science. Unlike performance benchmarking for computer hardware, there are no accepted, objective criteria for evaluating the performance of IR systems. That is, for IR systems, the notion of effectiveness is subjective. Human judgment is ultimately the criterion for evaluating whether an IR system returns the relevant information in the correct manner. Two users may have differing needs when using an IR system. For example, one may want to find all potentially relevant documents. Another may want to correctly sort information by priority. In addition, subject matter and information type may impact a user's IR requirements.

Over the past 50 years, a large body of research has emerged concerning the evaluation of IR systems. The study of IR metrics helps quantify and compare the benefits of various search and IR systems. In 1966, C.W. Cleverdon listed various "metrics" which have become the standard for evaluating IR systems within what has become known as the "Cranfield tradition."[41] Two of the metrics, *precision* and *recall,* are based on binary relationships. That is, either a document is relevant or it is not, and either a document is retrieved or it is not. Several modifications and additional metrics have been added in the IR literature since then, as the scientific field continues to add and refine techniques for measuring the efficiency of IR systems—both in terms of retrieval and also in user access to relevant information.

Measuring the Effectiveness of Information Retrieval Methods

Recall, by definition, is "an information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved."[42] That is, out of the total number of relevant documents in the document collection, how many were retrieved correctly?

Precision is defined as "an information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant."[43] Put another way, how much of the returned result set is "on target"?

41 *See* Cyril W. Cleverdon, et al., Aslib Cranfield Research Project: Factors Determining The Performance of Indexing Systems (Vol. I,, 1966), *available at* http://www.sigir.org/museum/pdfs/Factors%20Determining%20the%20Performace%20of%20Indexing%20Systems%20Vol%201%20-%20Part%201%20Text/pdfs/frontmatter.pdf; Cyril W. Cleverdon, et al., Aslib Cranfield Research Project: Report of Cranfield II (Vol II, 1966), *available at* http://www.sigir.org/museum/pdfs/Factors_Determining_the_Performace_of_Indexing_Systems_Vol_2/pdfs/frontmatter.pdf.; *see generally* C.J. Rjiisbergen, Information Retrieval (2d ed. 1979), *available at* http://www.dcs.gla.ac.uk/Keith/Preface.html.

42 *See* Ricardo Baeza Yates & Berthier Ribeiro Neto, Modern Information Retrieval 437, 455 (1999) (glossary), *available at* http://www.sims.berkeley.edu/~hearst/irbook/glossary.html; *see also The Grossman-Cormack Glossary*, *supra* n.7.

43 R. Yates & B. Neto, *supra* n.42, at 455.

Recall and precision can be expressed by simple ratios:

$$\textbf{Recall} = \frac{\textbf{Number of responsive documents retrieved}}{\textbf{Number of responsive documents overall}}$$

$$\textbf{Precision} = \frac{\textbf{Number of responsive documents retrieved}}{\textbf{Number of documents retrieved}}$$

If a collection of documents contains, for example, 1,000 documents, 100 of which are relevant to a particular topic and 900 of which are not, then a system that returned only these 100 documents in response to a query would have a precision of 1.0, and a recall of 1.0.

If the system returned all 100 of these documents, but also returned 50 of the irrelevant documents, then it would have a precision 100/150 = .667, and still have a recall of 100/100 = 1.0.

If it returned only 90 of the relevant documents along with 50 irrelevant documents, then it would have a precision of 90/140 = 0.64, and a recall of 90/100 = 0.9.

Importantly for the practitioner, there is typically a trade-off between precision and recall. One can often adjust a system to retrieve more documents—increasing recall—but the system achieves this result at the expense of retrieving more irrelevant documents—decreasing precision. Effectively, one can cast either a narrow net and retrieve fewer relevant documents, along with fewer irrelevant documents, or cast a broader net and retrieve more relevant documents, but at the expense of retrieving more irrelevant documents.[44]

Measuring the Efficiency of Information Retrieval Methods

Efficiency is important to the success of an IR system, but it does not affect the quality of the results. Efficiency is measured in three ways. The first measurement is the mean time for returning search results (this can be measured by the average time it takes to return results, or the computational complexity of the search). The second measurement is the mean time it takes a user to complete a search. This measurement is more subjective and is a function of the ease of use of the IR system. A third method involves the number of documents that must be reviewed to achieve a particular level of recall or precision.

44 There are many other common metrics that are considered in information retrieval literature, including F-measure, average precision, and average search length. F-measure is an approximation of the "crossover point" between precision and recall, where both are maximized. Average precision determines the precision level for each retrieved relevant item. Average search length is the average position of a relevant retrieved item. Still other terms include "fallout" (the ratio of the number of non-relevant items retrieved to the total number of items retrieved") and "elusion" (the proportion of responsive documents that have been missed by the search). *See generally The Grossman-Cormack Glossary*, *supra* n.7.

The Blair and Maron Study

A well-known study testing recall and precision in a legal setting was conducted by David Blair and M.E. Maron in 1985.[45] The Blair and Maron study demonstrated the problems caused by the rich use of human language among the many people that can be involved in a dispute and how difficult it is to take such richness into account in a search for information. Indeed, Blair and Maron found that attorneys were only about 20% effective at identifying all of the different ways that document authors could refer to words, ideas, or issues in their case.

For the purposes of their study, Blair and Maron evaluated a case involving an accident on the San Francisco Bay Area Rapid Transit (BART) in which a computerized BART train failed to stop at the end of the line. There were about 40,000 documents, totaling about 350,000 pages, in the discovery database. The attorneys worked with experienced paralegal search specialists in an effort to find all of the documents that were relevant to the issues. The attorneys estimated that they had found more than 75% of the relevant documents, but more detailed analysis found that the number was actually only about 20%. The authors found that the different parties in the case used different words to describe the same thing, depending on their role in the case. The parties on the BART side of the case referred to "the unfortunate incident," but parties on the victim's side called it a "disaster." Other documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were often not mentioned.

As Roitblat notes, *supra*, note 45, Blair and Maron found:

> that the terms used to discuss one of the potentially faulty parts varied greatly depending on where in the country the document was written. Some people called it an 'air truck,' a 'trap correction,' 'wire warp,' or 'Roman circle method.' After 40 hours of following a 'trail of linguistic creativity' and finding many more examples, Blair and Maron gave up trying to identify all of the different ways in which the document authors had identified this particular item. They did not run out of alternatives, they only ran out of time.

More Recent Studies on Precision and Recall in E-Discovery

In the years since publication of the 2007 Version of this Commentary, a variety of other efforts have been made to study the precision/recall issues in a legal discovery context. Many of these have been initiated by members of The Sedona Conference. Numerous studies emanating from the TREC Legal Track (see discussion, below) have confirmed relatively low rates of recall obtained from basic keyword search.[46] Moreover, two widely-cited recent studies have provided the foundation for lawyers making the claim that some of the more advanced computer- or technology-assisted review methods yield more accurate results than reliance solely on human

45 David L. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness For a Full-Text Document-Retrieval System*, 28 Comm. ACM 289 (1985). The discussion that follows of the Blair and Maron study is drawn directly from Herbert L. Roitblat, *Search and Information Retrieval Science*, 8 Sedona Conf. J. 225, 231 (2007).

46 *See*, *e.g.*, Stephen Tomlinson, et al., *Overview of The TREC 2007 Legal Track*, http://trec.nist.gov/pubs/trec16/t16_proceedings.html (baseline Boolean search captured only 22% of universe of all relevant documents found by all combined search methods); *see generally*, *TREC Legal Track Overview Papers,* 2006-2011, http://trec-legal.umiacs.umd.edu/.

review (as measured by precision, recall, and F_1 measures).[47] Both the 2010 study by Roitblat et al., and the 2011 study by Grossman and Cormack, used secondary data to compare the effectiveness of human review to that of certain computer- or technology-assisted review methods. Roitblat et al. used a collection of 1.6 million documents reviewed in response to a Department of Justice Second Request. The 1.6 million documents were re-reviewed by two teams using (unspecified) technology-assisted review methods. A statistical sample of 5,000 documents from the 1.6 million was also re-reviewed by two teams using manual review. The two technology-assisted review efforts both yielded better agreement with the original production than the two human review efforts, supporting the conclusion that technology-assisted review can be at least as effective as human review. Grossman and Cormack used the TREC 2009 collection of 836,135 documents captured from Enron by the Federal Energy Regulatory Commission. During the course of the TREC 2009 Legal Track Interactive Task, these documents were reviewed for responsiveness to seven "topics" (requests for production composed by TREC) by several participating teams using various technology-assisted review methods, which are detailed in the TREC proceedings.[48] Two teams showed superior effectiveness over five of the topics, according to F_1 (the harmonic mean of recall and precision). Grossman and Cormack compared these results to those of a human review of a statistical sample conducted (for three topics) by professional contract reviewers and (for two topics) by third-year law students. For all topics and all measures (recall, precision, and F_1) either the technology-assisted review was superior, or the measured difference was not statistically significant. On average, the technology-assisted reviews achieved recall, precision, and F_1 scores of 76.7%, 84.7%, and 80.0%, respectively, while the human reviews achieved 59.3%, 31.7%, and 36.0%, respectively. Overall, these studies indicate that one should not presume human review to be the most effective approach, that certain technology-assisted review methods can improve on human review, and that no review method is perfect. On the other hand, these studies do not indicate that all technology-assisted review methods are more effective than human review in all circumstances.

The limitations of keyword approaches to search and retrieval first exposed in the Blair and Maron study, and validated in subsequent research, have not faulted the ability of computers to locate documents meeting the attorneys' search criteria—but rather the inability of the attorneys and paralegals to anticipate all of the possible ways that people might refer to the issues in the case. The richness and ambiguity of human language causes severe challenges in identifying relevant information.

As Blair and Maron (and subsequent studies) demonstrate, human language is highly ambiguous and full of variation. In the years since Blair and Maron, and with increasing attention focused on the e-discovery space, the IR community has been engaged in research and the development of methods, tools, and techniques that compensate for endemic ambiguity and variation in human language, thereby improving the recall and precision of searches.

---

47 *See* Maura R. Grossman & Gordon 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), *available at* http://jolt.richmond.edu/v17i3/article11.pdf; 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://jolt.richmond.edu/v17i3/article11.pdf; Patrick Oot, et al., M*andating Reasonableness In A Reasonable Inquiry*, 87 Denv. U. L. Rev. 533 (2011), *available at* http://law.du.edu/documents/denver-university-law-review/v87-2/Oot_PDF.pdf; *see also* Howard Turtle, *Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance*, 1994 Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 212-220 (using structured case law in Westlaw databases), *see generally*, Rand 2012 Study, at 59-69 (summarizing results from these and other recent studies).

48 *See* Bruce Hedin, et al., *Overview of the TREC 2009 Legal Track*, http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf.

V. Boolean and Beyond: A World of Search Methods, Tools, and Techniques

In the decades since the Blair and Maron study, a variety of new search tools and techniques have been introduced to help find relevant information and weed out irrelevant information. Understanding these various tools and methods is critical. All automated methods are not created equal and do not perform the same functions and tasks. It is important to know what each methodology does when it is used alone or in conjunction with other methods and which tools are most effective for which purposes.

Clearly, different search methods have different functions and values in different circumstances. There is no one best system for all situations, an important fact for practitioners learning the techniques of search and retrieval technology to understand.

A more detailed description of search methods and techniques is set out in the Appendix. These methods can be grouped into four broad categories, but there are hybrid and crosscutting approaches that defy easy placement in any particular "box."

Keywords and Boolean Operators

First, there are keyword-based methods, ranging from the simple use of keywords alone, to the use of strings of keywords with what are known as "Boolean operators" (including AND, OR, "AND NOT," or "BUT NOT").

Categorizations of Data Sets

Second, there are other techniques relying on categorizations of the entire data set with various methodologies heavily reliant on deriving (i.e., coming to a consensus on) a thesaurus, taxonomy, or "ontology" of related words or terms. These techniques can be used to categorize the entire data set into specified categories all at once or as more data is added to the data set.

However, data sets generally need to be indexed to use any of the latter methodologies—where the indexing will take more time depending on what one indexes (e.g., indexing all of the data will take substantially longer than indexing selected fields).

There are a variety of indexing tools, some of which are available as open source tools. Indexing structured data may take less time than indexing data in an unstructured form, if only designated fields are indexed. Indexing an unstructured data set is time consuming because of the need to index all the words (except for "and," "a," "the," or other common "noise" or "stop" words). Knowing what is being indexed will be important to set expectations in terms of timing and making the data useful for querying or review.

Alternative search methods to keywords can, in some instances, free the user from having to guess, for every document, what word the author might have used. For example, there are more

than 120 words that could be used in place of the word "think" (e.g., guess, surmise, anticipate and so on). As the Blair and Maron study shows, people are actually very poor at guessing all of the right words to use as search terms to find the documents they are looking for without overwhelming the retrieval with irrelevant documents. In light of this fact, alternative search methods help organize large collections of documents which humans would otherwise be unable to organize.

Using a thesaurus, taxonomy, or ontology generally provides the results one would expect, because these systems explicitly incorporate one's expectations about what is related to what. They are most useful when one has (or can buy) a good idea of the conceptual relations to be found in one's documents—or one has the time and resources needed to develop them. Clustering, Bayesian classifiers, and other types of systems have the power to discover potentially unanticipated relationships in the text. This means that one gets unexpected results from time to time, which can be of great value, but can also be somewhat over-inclusive (or even wrong). An example: after training on a collection of medical documents, one of these systems learned that Elavil and Klonopin were related (they are both anti-anxiety drugs). A search for Elavil turned up all the documents that contained that word, along with "false positive" documents containing only the word "Klonopin" (without the word "Elavil").

Such systems can discover the meaning of at least some acronyms, jargon, and code words appropriate to the context of the specific document collection. No one has to anticipate their usage in all possible contexts; the systems, however, can help to derive them directly from the documents.

Computer- or Technology-Assisted Review Methods

Third, there are a variety of methods that combine both technological and human inputs in an iterative search design. These techniques stem from various research directions in Artificial Intelligence and can be categorized under the general rubric of computer- or technology-assisted review. When drawing on machine-learning techniques, these tools are sometimes referred to as predictive coding; however, the more inclusive nomenclature for the entire spectrum of advanced methods remains computer- or technology-assisted review. Generally put, computer- or technology-assisted approaches are based on iterative processes where one (or more) attorneys or IR experts train the software, using document exemplars, to differentiate between relevant and non-relevant documents. In most cases, these technologies are combined with statistical and quality assurance features that assess the quality of the results. The research cited above has demonstrated such techniques superior, in most cases, to traditional keyword-based search, and, even, in some cases, to human review.[49]

The computer- or technology-assisted review paradigm is the joint product of human expertise (usually an attorney or IR expert working in concert with case attorneys) and technology. The quality of the application's output, which is an assessment or ranking of the relevance of each document in the collection, is highly dependent on the quality of the input, that is, the human training. Best practices focus on the utilization of informed, experienced, and reliable individuals

49 *See supra,* notes 47-48.

training the system. These individuals work in close consultation with the legal team handling the matter, for engineering the application. Similarly, as explained below, the defensibility and usability of computer- or technology-assisted review tools require the application of sound approaches to selection of a "seed" or "training" set of documents, monitoring of the training process, sampling, and quantification and verification of the results.[50]

Well-thought-out techniques are needed in order to ensure that the set of documents used to train the system provides thorough coverage of the entire document population and particularly, of the relevant material (both expected and unexpected) contained therein. Sampling will be an essential component of any effective quality control regimen; it is important that samples drawn for quality control are separate from those used for training in order to ensure independence of statistical measurement.

Iterative monitoring of the training process ensures that the system is adequately trained, (meaning that additional exemplars could not substantially enhance review effectiveness), while avoiding wasteful use of expensive "expert" resources through excessive training. In some cases, active learning techniques are applied to accelerate the training process, reducing to a minimum the number of training documents required, and avoiding the inefficiencies of random sampling, especially in low prevalence populations. Active learning systems select new exemplars for training based on knowledge of the population that the application has generated from previous training examples.

Statistically valid *measurement techniques,* based on precision and recall, as described above, are expressed within confidence intervals and at certain confidence levels to estimate results. Quality assurance techniques are critical in order to verify outcomes; for example, sampling of the documents culled as irrelevant to verify that they contain the expected low prevalence of relevant documents.

None of these systems is magical. Language is sometimes shared between two people who invent a shorthand or code. And all tools require a healthy dose of good legal judgment, based on a well-thought-out approach. Some techniques may be difficult to understand to those without technical backgrounds, but they need not be mysterious. If a vendor cannot explain how a system works, then the buyer should beware and either require an explanation or consider an alternate approach.

There is no magic to the science of search and retrieval; it is comprised of mathematics, linguistics, computer science, and hard work. If lawyers do not become conversant in this area, they risk surrendering the field's intellectual jurisdiction to other disciplines, as well as risk poor quality and costly e-discovery outcomes.

50 *See generally*, *The Sedona Conference Commentary on Achieving Quality in E-Discovery*, *supra* note 32.

VI. Practical guidance for Evaluating the Use of Automated Search and Retrieval Methods

Practice Point 1. ***In many settings involving large amounts of relevant electronically stored information (ESI), relying solely on a manual search process for the purpose of finding responsive documents may be feasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary under certain circumstances.***

For the reasons articulated in prior sections, the demands placed on practitioners and parties in litigation and elsewhere increasingly dictate that practitioners must evaluate the use of automated search and retrieval methods in a wide variety of cases and contexts. Particularly (but not exclusively) in large and complex litigation, where discovery is expected to encompass hundreds of thousands to hundreds of millions of potentially responsive electronic records; there is no reasonable possibility of marshalling the human labor required to undertake a document-by-document, manual review of the potential universe of discoverable materials. This is increasingly true both for parties responding to a discovery request and for parties who propound discovery (and receive a massive amount of material in response). Where the infeasibility of undertaking manual review is acknowledged, utilizing automated search methods may not only be reasonable and valuable, but also necessary.

Even in less complex settings, overreliance on manual review may be an inefficient use of scarce resources. This is especially the case where automated search tools used on the front end of discovery could prove useful in a variety of ways, including early case assessment for settlement or other purposes, or for prioritizing or grouping documents to allocate resources or facilitate later manual review.

Of course, the use of automated search methods is not intended to entirely eliminate the need for manual review; indeed, in many cases, both automated and manual searches will be conducted, with initial automated searches used for culling down a universe of material to more manageable size (or prioritizing documents), followed by a secondary manual review process. So too, while automated search methods may help identify privileged documents within a larger set, the majority of practitioners may still rely on largely manual review processes to identify the basis for the assertion of privilege.

Practice Point 2. ***The successful use of any automated search method or technology will be enhanced by a well-thought-out process with substantial human input on the front end.***

As discussed above, the decision to employ an automated search method or technology cannot be made in a vacuum, on the assumption that the latest "tool" will solve an attorney's discovery obligations. Rather, to maximize the chances of success in terms of finding responsive documents, a well-thought-out strategy capitalizing on "human knowledge" available to a party should be implemented at the earliest opportunity. This knowledge can take many forms.

First, a party must evaluate the specific legal setting, since the nature of the lawsuit or investigation, the field of law involved, and the specific causes of action under which a discovery obligation may arise must all be taken into account. For example, keyword searches alone in highly technical patent cases may prove highly efficacious. But in other types of cases, including those with broad causes of action involving subjective states of intent, a practitioner should consider alternative search methods.

Second, in any legal setting involving consideration of automated methods for conducting searches, counsel and client should perform an analysis to define the target universe of documents that is central to the relevant causes of action. This would include not only assessing relevant subject areas and "drilling down" with as much specificity as possible, but also analyzing the custodians (or others) who would be in possession of such relevant data. Time and cost considerations must also be factored in, including budgeting for human review time. These practice points apply whether a party is a defendant and holds a universe of potentially discoverable data, or a plaintiff who is expecting to receive a massive data set in response to requests for production.

Practice Point 3. ***The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed. Parties and their counsel must match the use case with the tools and best practices appropriate to address it and must incorporate proportionality considerations involving the overall cost and the stakes of the litigation.***

The choice of a search and retrieval method for a given situation depends upon a number of factors. Two of the biggest decisions to make are the acceptable level of false positive "noise" (i.e. achieving higher "precision"), and the acceptable level of false negatives (i.e., maximizing "recall"). There are a number of overarching factors that lawyers should consider in evaluating the use of particular search and retrieval methods in particular settings.

First, the "heterogeneity" of the relevant universe of ESI is a significant factor. ESI that is potentially relevant may be found in multiple locations, and in a variety of forms, including structured and unstructured active computer environments, removable media, backup tapes, and a variety of email applications and file formats. In some cases, information that provides historical, contextual, tracking, or managerial insight (such as metadata) may be relevant to a specific matter and demand specialized data mining search tools. But in other cases, the very same data will be irrelevant.

Second, the volume, prevalence, and condition of the likely relevant ESI, and the extent to which ESI is contained within static or dynamic electronic applications, are all relevant to the party's or its counsel's decisions.

Third, for any particular search and IR method, the time it takes and its cost (compared to other automated methods or human review) must also be considered.

Fourth, the goals of the search are a factor (e.g., capturing or finding as many responsive documents as possible regardless of time and cost versus finding responsive documents as

efficiently as possible, i.e., with the least number of non-responsive documents). In other words, the practitioner must consider the desired trade-off between precision and recall. Given the particular setting, the party seeking to employ one or more search methods should assess the relative importance in that setting of finding responsive ESI versus the importance of eliminating non-responsive data. Depending on this assessment, one or more alternative search methodologies may prove to be a better match in the context of a particular task.

Fifth, one must consider the skills, experience, financial, and practical constraints of the representatives of the party making the selection (e.g., the attorneys, litigation support staff, vendors, the Special Master, etc.).

Sixth, the current status of electronic discovery in the matter, including the extent to which activities including preservation and collection are occurring in addition to processing and/or attorney review.

Seventh, one must investigate empirical research supporting the reliability of the search and IR method for particular types of data, or in particular settings.

* * *

Although not the focus of this Commentary, practitioners may also wish to consider the use of advanced search methods at earlier stages of the e-discovery process before traditional document review.

- Early Case Assessment: A key objective of early case assessment is to assess case risk and cost, with the goal of avoiding futile and wasteful litigation activity. Tools with sophisticated metadata search and computer- or technology-assisted review capabilities can be helpful in identifying a small subset of highly probative documents within the population with high precision. Users can then focus on this "rich" subset of data to make rapid and informed decisions on case strategy.

- Culling: For proper use in culling strategies, search and retrieval technologies should support a valid statistical framework able to estimate characteristics such as prevalence, precision, and recall. This facilitates a user's quantifiable assessment of the cost and risk involved in culling decisions: for example, whether culling a certain set of documents will potentially yield 5% or 95% of the relevant documents.

- Prioritized review: Prioritized review structures the review process to start with the documents most likely to be relevant, progressing to the documents least likely to be relevant. Stratified review, a variation of prioritized review, matches document importance and reviewer quality, such that more skilled reviewers can review documents requiring greater expertise. Prioritized and stratified reviews require computer- or technology-assisted review tools that are able to classify documents as "more likely to be relevant" through "less likely to be relevant."

Finally, in adopting proportionality, parties need to balance the costs of using alternative search methods with the perceived benefits and risks in a given litigation context. Costs and time

will vary depending on the desired rate of recall (how many documents need to be found) and precision (how many documents need to be reviewed to yield relevant documents). Risk also is reflected in recall, i.e., how many documents are found versus how many are "left on the table." The proportionate costs versus benefits and risks that a user is willing to bear are a function of what technology is reasonably available and what is at stake in the particular matter, taking into account what e-discovery phase is being addressed (document review versus other aspects of discovery, at earlier stages, per the above).

Practice Point 4. ***Parties and their counsel should perform due diligence when choosing a particular information retrieval product or vendor service.***

The prudent practitioner should ask questions regarding search and retrieval features and the specific processing and searching rules that are applied to such features. Some tools are fully integrated into a vendor's search and review system, whereas others are "stand alone" tools that may be used separately from the particular review platform. It is essential not only to understand how the various tools function, but also to understand how the tools fit within the overall workflow planned for discovery. A practitioner should inquire as to which category or categories the specific tool fits into, how it functions, and what third-party technology lies behind the tool.

It is also essential that specific methods or tools be made understandable to the court, opposing parties, and the attorney's client. How data is captured and indexed (and how long it takes to build an index) also may affect a decision on use; it is therefore important to understand how a particular system deals with rolling input and output over time, considering its flexibility and scalability. The ability to perform searches across metadata, to search across multiple indices or stores of data, to search embedded data, to refine search results (nested searches), to save queries, to capture duplicates and perform deduplication, to trace email threads, and to provide listings of related terms or synonyms are all examples of specific functional requirements that should be clarified depending on case needs.

Other types of due diligence may involve administrative matters (e.g., understanding maintenance and upkeep, additional charges, system upgrades, availability of consultants or technicians to address problems, system performance), quality control issues (e.g., prior testing of the method or tool in question; how databases and dictionaries supporting concept search were populated; the strength of the provider's application development group), and, finally, any relevant licensing issues which could involve proprietary software or escrow agreements with third parties.

Practice Point 5. ***Because of the characteristics of human language, no search and information retrieval tool can guarantee the identification of all responsive documents in large data collections. Moreover, different search methods may produce different results, subject to a measure of statistical variation inherent in the science of information retrieval.***

Just as with past practice involving manual search through traditional paper document collections, there is no requirement that "perfect" searches will occur—only that parties and

their counsel act reasonably and in good faith in the performance of their discovery obligations. From decades of IR research, it is clear that a 100% rate of recall, i.e., the ability to retrieve all responsive documents from a given universe of electronic data, is an unachievable goal. As discussed in prior sections, the richness of human language, with its attendant elasticity, causes all present day automated search methods to fall short of perfection.

Moreover, there will always be a measure of statistical variation associated with alternative search methods, i.e., some responsive documents will be found by one search method while being missed by others. Even the same search method may return different results if new documents are added to the searched universe. Particularly in the context of a large data set, a search method should be judged by its overall results (such as using measures of recall and precision), rather than being judged by whether it produces the identical document set as compared with a different technique.

However, it is important not to compare "apples with oranges." Given the present state of information science, it would be a mistake to assume that one search method will work optimally across all types of possible inquiries or data sets (e.g., what works well in finding word processing documents in a given proprietary format may not be as optimal for finding information in structured databases, or in a collection of scanned images). This is another area where, consistent with the above principles, a good deal of thought should be given at the outset to the precise problem, in terms of its scope and relevancy considerations, before committing to a particular search method.

Practice Point 6. ***Parties and their counsel should make a good faith attempt to cooperate when determining the use of particular search and information retrieval methods, tools, and protocols (including keywords, concepts, computer- and technology-assisted review and other types of search parameters and quality control measures).***

The body of case law that has emerged since 2006[51] indicates that courts are becoming more comfortable with addressing search and retrieval issues, particularly in the context of approving protocols, or ordering parties to share information that would lead to the development of more refined search protocols. The fact that some courts have waded into these issues demonstrates how rapidly the law has been evolving since the 2006 amendments to the Federal Rules of Civil Procedure.[52]

Under Rule 26(f), the parties' initial planning should address "[a]ny issues relating to disclosure or discovery of electronically stored information," as well as "[a]ny issues relating to preserving discoverable information." These initial discussions on preservation and production should include a specific discussion on search methods and protocols to be employed by one or both

51 *See* Baron, *Law in the Age of Exabytes*, *supra* note 23.

52 *See* Kenneth J. Withers, *Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure*, 4 Nw. J. Tech. & Intell. Prop. 172 (2006) (what "probably strikes the reader [of Treppel, *supra* note 26] as matter of fact, sensible, and routine, would have been extraordinary a scant six years ago, when the last major revision of the discovery rules went into effect [in 2000])."

parties.[53] While disclosure of these methods and protocols is not mandated or legally required under this rule, the advantages of collaborating should strongly be considered. In many cases, reaching an early consensus on the scope of searches can minimize the overall time, cost, and resources spent on such efforts, as well as minimize the risk of collateral litigation challenging the reasonableness of the search method employed.[54]

The Sedona Conference *Cooperation Proclamation*, published in 2008 underscores Practice Point 6, here, by including among the methods for accomplishing cooperation in e-discovery: (i) "Exchanging information on relevant data sources, including those not being searched;" and (ii) "Jointly developing automated search and retrieval methodologies to cull relevant information."[55]

Practice Point 7. ***Parties and their counsel should expect that their choice of search methodology (and any validation of it) will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and at trial).***

Counsel should be prepared to explain what keywords, search protocols, and alternative search methods were used to generate their production set, including ESI made subject to a subsequent manual search for responsiveness and privilege. This explanation may need to come from a technical "IT" expert, a statistician, or an expert in search and retrieval technology. Parties should anticipate that in contested matters, an opposing party may request the justification of particular search methods used; this may require a demonstration of the recall and precision (or other measures) for the output of a chosen search method.[56] Counsel must be prepared to answer questions and even prove the reasonableness and good faith of their methods.

Practice Point 8. ***Parties, counsel, and the courts should be alert to new and rapidly evolving search and information retrieval methods. Moreover, parties and their counsel should recognize that information retrieval is a distinct field of study that includes expertise in such areas as computer science, statistics, and linguistics, and that consultation with or utilization of experts in information retrieval may improve the quality of search results in complex cases involving large volumes of ESI.***

Given the rapid evolution of technology, what constitutes a reasonable search and IR method is subject to change. The legal community needs to be vigilant and must examine new and emerging techniques and methods which can yield better search results. In particular settings, lawyers

---

53 *See*, *e.g.*, *In re* Pilot Project Regarding Case Management Techniques for Complex Civil Cases in the Southern District of New York (Standing Order, Nov 1. 2011), http://www.nysd.uscourts.gov/cases/show.php?db=notice_bar&id=261.

54 *See* Paul and Baron, *Information Inflation*, *supra* note 13, at [50-55] (discussing an iterative collaboration process that includes adoption of multiple "meet and confers" to discuss and refine preliminary search results).

55 The Sedona Conference *Cooperation Proclamation*, *available at* https://thesedonaconference.org/download-pub/1703. A companion piece, *The Case for Cooperation*, published in *The Sedona Conference Journal* goes on to say:

> [W]orking cooperatively with opposing counsel to identify a reasonable search protocol, rather than making boilerplate objections to the breadth of a requested protocol or unilaterally selecting the keywords used without disclosure to opposing counsel, may help avoid sanctions or allegations of intentional suppression. Indeed, 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.

10 Sedona Conf. J. 339, 344 (2009 Supp.).

56 *See*, *e.g.*, *Moore*, 287 F.R.D. at 182; *Kleen Prod.,* 2012 WL 4498465 at *1.

should endeavor to incorporate evolving technological progress at the earliest opportunity in the planning stages of discovery or other legal setting involving search and retrieval issues.

Successful search of large amounts of ESI is increasingly dependent on the expertise of those doing the searches. Attorneys who lack expertise in IR or statistics should consider consulting or collaborating with IR experts, when appropriate, in complex cases. In general, the bar would do well to understand that a greater appreciation of IR and scientific methods will result in better overall search. This is analogous to the situation where a tool, such as a scalpel, could in theory be used by anyone, regardless of their expertise in medicine or surgery. But when that tool is used in the hands of a trained surgeon (as opposed to someone who lacks that expertise), common sense dictates that the results will likely be better.

VII. Future Directions in Search and Retrieval Science

What prospects exist for improving present day search and retrieval methodologies? And how can lawyers play a greater role in working with the IR research community to improve the accuracy and efficiency of search and review technology?

A. Harnessing the Power of Artificial Intelligence (AI)

A statement from page 36 of The Sedona Conference, *Navigating The Vendor Proposal Process* (2007 ed.), under the general heading "Advanced Search and Retrieval Technology," bears repetition here: "Technology is developing that will allow for electronic relevancy assessments and subject matter, or issue coding. These technologies have the potential to dramatically change the way electronic discovery is handled in litigation, and could save litigants millions of dollars in document review costs. Hand-in-hand with electronic relevancy assessment and issue coding, it is anticipated that advanced searching and retrieval technologies may allow for targeted collections and productions, thus reducing the volume of information involved in the discovery process."

The growing enormity of data stores, the inherent elasticity of human language, and the unfulfilled goal of computational thinking to approximate the ability and subtlety of human language all present steep challenges to the IR and AI communities in their quest to develop optimal search and retrieval techniques.

But the future holds promise. Not only is there available technology to apply sophisticated computer algorithms to data mine traditional text, but new and better approaches to image and voice pattern recognition are looming on the horizon. Indeed, there is already rudimentary technology available for searching audio by search terms, and some processes are confronting—and succeeding—in searching by image or picture.[57] Clearly, at some point, all forms of data stored in corporations and institutions will be within the scope of future information demands in legal settings, and likely within the ambit of future automated search processes.

Finding information on the Web sometimes is easier than finding documents on one's own hard drive. The post-Google interest in building better search engines for the Web can only lead to new and better search techniques applied to more well-defined contexts, such as corporate and institutional intranets and data stores.

A "2020 Science" report issued by Microsoft in March 2006 anticipated the near-term development of "novel data mining technologies and novel analysis techniques," including "active learning" in the form of "autonomous experimentation" and "artificial scientists," in replacement of "traditional machine learning techniques [that] have failed to bring back the knowledge out of the data."[58] With the emergence of computer- or technology-assisted review

57 *See*, *e.g.*, Idée, Inc., *TinEye: Reverse Image Searc*h, http://www.tineye.com/ (last visited November 25, 2013).

58 *See* Microsoft, *2020 Science*, at 15, *available at* http://research.microsoft.com/towards2020science/downloads.htm (last visited November 25, 2013).

methods in e-discovery, we are beginning to see such techniques played out in litigation. Beyond the short-term horizon, scientists are expected to embrace emergent technologies including the use of genetic algorithms, nanotechnology, quantum computing, and a host of other advanced means of information processing. And future AI research in the specific domain of search and retrieval is unbounded and, at least in part, unpredictable.

B. The Role of Process in the Search and Retrieval Challenge

Every search and retrieval technology has its own methodology to ensure the technology works properly, relying on a set of instructions that outline the workflow for the tool. How well these methods are applied significantly impacts the performance, and therefore the results achieved by the technology. This is where process comes in to play. Process provides order and structure by setting guidelines and procedures designed to ensure that a technology performs as intended. Effectively applied, process drives the consistent and predictable application of the search and retrieval technology. The results derived from the consistent and predictable application of search and retrieval tools establish the technology's credibility and value.

The Importance of Process

A process is a considered series of events, acts, or operations leading to a predictable result or effect. A process, like a technology, is a "tool" that can be used to assist in completing a task. The use of a well-defined and controlled process promotes consistency, reliability, and predictability of the results and ensures the efficient use of the resources required to produce them. As such, a process does not find the answer to or attain the objective of a task on its own. Process, no matter how well designed and executed, cannot replace the exercise of judgment; however, process promotes the exercise of judgment by ensuring that the most accurate and reliable information is available when making decisions. In the search and retrieval context, this means the availability of consistent and reliable information to assist the parties in making informed decisions.

The use of process promotes consistency by establishing a defined approach to a task. The resulting consistency promotes reliability and predictability. Reliability and predictability allow for better planning, performance, and cost management. Altogether, risk is reduced and confidence is promoted.

One can visualize search and retrieval as a process enabling a party to distinguish potentially discoverable information from among a broader set of electronic data collected for the purposes of production. It consists of several steps that take place in the context of a particular search and retrieval technology. Because the application of process is flexible, it can be used to address unique conditions that might be associated with a technology, such as where the use of a search and retrieval technology itself creates issues. For example, the use of search and retrieval technologies to address significant volumes of information may not address all problems: as review volumes increase (even with carefully crafted and tested search criteria) the likelihood of being swamped by false positives or missing false negatives increases greatly. By developing and implementing process steps which consistently address these issues, their impact can be diminished and the reasonableness and good faith of the technology can be established.

"Process" as a Measure of Reasonableness and Good Faith

Search and retrieval in this new era requires the establishment and recognition of a new standard. A standard of absolute perfection is and always has been unrealistic; but now, with quantitative data available, we know perfection is not only unrealistic but also quite simply unachievable.

Rather than perfection, which would demand the identification and production of every relevant, non-privileged document, the standard against which to measure these new technologies and processes should incorporate the same principles that have traditionally governed all discovery: reasonableness, good faith, and proportionality.[59] Although these terms raise concerns about ambiguity and uncertainty, they can actually represent a well-defined set of expectations in the context of the discovery process.

A process that emphasizes reasonableness, good faith, and proportionality is fully consistent with what is required under the discovery process. Discovery of information relevant to a dispute gathered by an opponent is often central to a fair and efficient resolution.[60] A party need only identify and produce that which is relevant, as defined by the rules, with the degree of diligence expected and available by experienced practitioners acting reasonably.[61] As noted in Sedona Principles 6 and 11, a party may choose to implement this approach in a reasonable manner.

Sound process applied to the use of search and retrieval technology can readily establish a measurable means for conducting discovery that satisfies the Rules. Reasonableness, good faith, and proportionality can be defined and measured by identifying performance criteria based on their attributes. Accordingly, the unreasonable and unattainable goal of "perfection" should not hinder the attainable and measurable goal of reasonableness.

As search and retrieval technologies and associated processes are developed, parties will no doubt want to use them to achieve defensible and credible results. If a party fails to adhere to appropriate performance guidelines, it will be subject to scrutiny and criticism, and perhaps even sanctions. Therefore, an established process—in conjunction with sound technology—can serve as a benchmark for conducting future discovery. Further, defensibility with opposing counsel and the court will likely depend implementing and adhering to processes developed for use with search and retrieval technologies.

Implementing Process

Using a search and retrieval technology in conjunction with an implementing process in will involve iterative activity. This will incorporate feedback loops at appropriate decision points to allow integration of what a case team learns after each step of the process. This, in turn, will calibrate and maximize the technology's ability to identify relevant information. It is through this feedback that case teams will acquire sound information to use in making both strategic and tactical decisions.

---

59 *See* Fed. R. Civ. P. 26(g) & 26(b)(2)(C).

60 *See Hickman,* 329 U.S.at 507.

61 Under Fed. R. Civ. P. Rule 26(g)(1), an attorney of record is expected to certify that to the best of his or her "knowledge, information, and belief, formed after a reasonable inquiry," that disclosures are "complete and correct" as of the time they were made. Similarly, under Rule 26(g)(2), an attorney must certify that to the best of his or her "knowledge, information, and belief, formed after a reasonable inquiry," that discovery "requests, responses, and objections" are made "consistent with these rules." *See also* Fed. R. Civ. P. 26(b)(2)(C).

The initial search and retrieval process should be designed as a "pilot" process that can be evaluated and modified as the team learns more about the corpus of information to be reviewed. One useful approach initiates the process by focusing on the information collected from a few key custodians at the center of the facts at issue in the litigation or investigation. Focusing on information collected from core custodians (information with a higher likelihood of relevance) will help the team efficiently understand the issues and language used by the custodians, and enable them to more efficiently develop and implement an appropriate search and retrieval process for subsequent custodians and ESI.

The initial selection and refinement of search terms can also benefit from sampling techniques used to rank the effectiveness of various terms or concepts. Reviewing samples of information that include selected search terms or concepts and ranking their relative value based on their efficacy in retrieving relevant information (recall) and their efficiency in excluding non-relevant information (precision) can help focus the selection of appropriate search terms.[62]

The development of process control logs and improved second-level review techniques can also help the review team consistently apply the designed process to all of the information to be reviewed. Additionally, a second-level review process based on statistical sampling techniques can ensure acceptable levels of quality. While these techniques are relatively unknown in the typical review processes in use today, their widespread adoption in businesses of all types should drive their implementation in large document review projects in the near future.[63]

C. How The Legal Community Can Contribute to The Growth of Knowledge

Human review of documents in discovery is expensive, time consuming, and error-prone. The application of linguistic and statistically-based content analysis, search and retrieval technologies such as computer- or technology-assisted review, and other tools, techniques, and process in support of the review function can effectively reduce cost, time, and error rates.[64]

Recommendations

1. ***The legal community should support collaborative research with the scientific and academic sectors aimed at establishing the efficacy and efficiency of a range of automated search and information retrieval methods.***

2. ***The legal community should encourage the establishment of objective benchmarking criteria, for use in assisting lawyers in evaluating the competitive legal and regulatory search and retrieval services market.***

Up until recently, in the years since the 1985 Blair and Maron study, there was little in the way of peer-reviewed research establishing the efficacy of various methods of automated content

62 *See supra*, text at Part IV.

63 *See generally The Sedona Conference Commentary on Achieving Quality in E-Discovery*, *supra* note 32.

64 *See supra* note 47 and accompanying text.

analysis, search, and retrieval as applied to a legal discovery context. Research into the relative efficacy of search and retrieval methods should acknowledge that each alternative should be viewed in the context of its suitability to specific document review tasks. Different technologies, tools, and techniques obviously have different strengths and weaknesses. Moreover, the outcomes of the application of advanced content analysis, search, and retrieval methods may be significantly different based on expertise of the operator. Ideally, research should advance the goal of setting a minimum or baseline standard for what constitutes an adequate information retrieval process, as well as reaching agreement on how to benchmark competing methods against agreed-upon objective evaluation measures.

Since 2006, The Sedona Conference has supported the TREC Legal Track (part of the TREC research program run by the National Institute of Standards and Technology). NIST is a federal agency that collaborates with industry and academia to develop and apply technology, measurements, and standards. TREC is designed "to encourage research in information retrieval from large text collections."[65] The TREC legal track has involved evaluation of a set of search methodologies based on lawyer relevancy assessments on topics drawn from large publicly available document databases. The results that have come out of the TREC Legal Track represent the type of objective research into the relative efficacy of Boolean and other search methods that the legal community should further encourage.[66]

However, a need exists to expand upon TREC research to accommodate the potential retrieval of tens or hundreds of millions of arguably relevant documents among a greater universe of terabytes, petabytes, exabytes, and beyond, and to study new and emerging forms of ESI, including text messaging and all forms of online social media. Members of The Sedona Conference community have and will continue to participate in collaborative workshops and other fora focused on issues involving IR.[67] How best to leverage the work of the IR community to date is an enterprise beyond the scope of this paper. The Sedona Conference intends to remain in the forefront of the efforts of the legal community in seeking out centers of excellence in this area, including the possibility of fostering private-public partnerships aimed at focused research.

65 The Text Retrieval Conference (TREC) was started in 1992. Its purpose is to support research within the IR community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC is overseen by a program committee consisting of representatives from government, industry, and academia. Each TREC track involves a test database of documents and topics. Participants run their own search and retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST generally pools the individual results, judges the retrieved documents for correctness, and evaluates the results. The TREC cycle ends with a workshop forum for participants to share their experiences. The TREC test collections and evaluation software are available to the IR research community at large, so organizations can evaluate their own retrieval systems at any time. TREC has successfully met its dual goals of improving the state-of-the-art in IR and of facilitating technology transfer, and some of today's commercial search engines include technology first developed at TREC. For further information, *see* http://trec.nist.gov (last visited November 25, 2013).

66 *See* TREC Legal Track Overview Papers, 2006-2011, http://trec-legal.umiacs.umd.edu/.

67 *See*, *e.g.*, Discovery of Electronically Stored Information ("DESI") Workshops, held at the Eleventh, Twelfth, Thirteenth, and Fourteenth International Conferences on Artificial Intelligence and Law (ICAIL 2007, ICAIL 2009, ICAIL 2011, ICAIL 2013) (links available at http://www.umiacs.umd.edu/~oard/desi5/); Information Retrieval for E-Discovery (SIRE) Workshop, SIGIR 2011 (link available at http://www.umiacs.umd.edu/~oard/sire11/).

Appendix A: Types of Search Methods

This Appendix is a "survey" of some of the different search methods found in the information retrieval literature, which form the basis of offerings by vendors in the legal marketplace. The list is not exhaustive. Indeed, as the main body of the Commentary makes clear, rapid technological progress will inevitably affect how methods are described, implemented, and subsequently replaced with new ways of performing search and retrieval.

Three further notes on this survey are in order. First, the following search methods are not intended to be mutually exclusive. Indeed, many products encourage the use of hybrid, combined, or cumulative approaches to search.

Second, the choice of method is condition- and objective-specific. The method best suited to one circumstance may not be the best suited to another circumstance. The methods reviewed in this survey may be viewed as distinct tools in the search "toolbox," each with its appropriate application or applications.

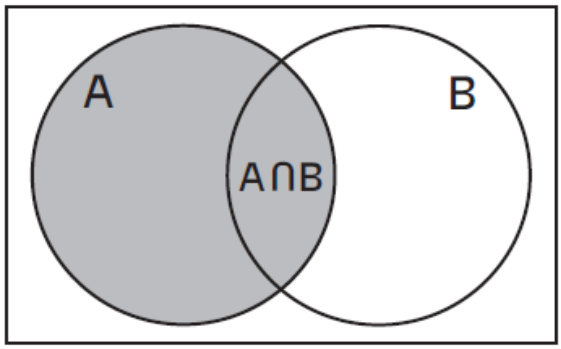
Third, potential users of the technologies reviewed in this survey are reminded that the tools alone cannot guarantee an effective search, any more than a well-made scalpel guarantees a successful surgery. Search tools will be effective only if applied with sound methodological principles and with appropriate expertise.

A. Boolean Search

A "Boolean search" utilizes the principles of Boolean logic named for George Boole, a British born mathematician. Boolean logic is a method for describing a "set" of objects or ideas. Boolean logic was applied to IR as computers became more widely accepted. Boolean searches can easily be applied to large sets of unstructured data and return results which exactly match the search terms and logical connectors applied by the operators.

As used in set theory, a Boolean notation demonstrates the relationship between sets or groups of information—in our example, two sets of information, "A" and "B"—and, in effect, creates a new set of information.

If a search seeks information contained within either original set "A" *or* original set "B" (essentially any area within either set in the Venn diagram below), the searcher is creating a "Union" of A and B (denoted "A∪B"). If the search seeks information which would be found within both set "A" *and* set "B," if each set was searched separately (the area within the overlap of the Venn diagram below), the searcher is creating an "Intersection" of A and B (denoted "A∩B"). A Venn diagram picture easily depicts these relationships (see below).



The "**OR**" Boolean operator directs that the set may contain any, some, or all of the keywords searched. The purpose of this command is to encompass alternative vocabulary terms. **OR** is represented by the union of the sets ("A∪B") (the entire area within both the circles above). The use of **OR** *expands* the resulting Boolean set.

The "**AND**" Boolean operator identifies the intersection of two sets or two keywords. The purpose of this command is to help construct more complex concepts from more simple vocabulary word "building blocks." **AND** is represented by intersection of the sets ("A∩B") (the shaded area within the intersection of the two circles). The use of **AND** *restricts* the resulting Boolean set.

The "**NOT**" Boolean operator eliminates unwanted terms. The purpose of this command (often preceded by either "**AND**" or "**BUT**") helps suppress multiple meanings of the same term; in other words, eliminating ambiguity. **NOT** would be represented by the area within the rectangle surrounding both circles, or the "empty" set ("ø").

Different search engines or tools may provide additional Boolean operators or connectors to create more complex search statements. These may include:

- **Parenthesis:** A Boolean search may include the use of parentheses to force a particular order to the execution of the search, as well as to create more refined and flexible criteria. Any number of logical ANDs (or any number of logical ORs) may be chained together without ambiguity; however, the combination of ANDs and ORs and AND NOTs or BUT NOTs can lead to ambiguous directions. In such cases, parentheses may be used to clarify the order of operations. The operations within the innermost pair of parentheses are performed first, followed by the next pair out, etc., until all operations are completed.

- **Proximity or NEAR/WITHIN operator:** This technique checks the location of terms and only matches those within the specified distance. This is a useful method for establishing relevancy between search criteria, as well as paring down irrelevant matches and obtaining better results (improving precision). Some search engines permit the user to define the order, in addition to the distance of the search terms. For example: budget w/10 deficit would mean "deficit within the 10 words of word budget."

- **Phrase searching:** Some search engines provide an option to search a set of words as an exact phrase, either by typing the phrase in quotation marks (" ") or by using a command. When they receive this kind of instruction, the search engine will locate all words that precisely match the search terms, and then discard those which are not next to each other in the correct order. To perform this task efficiently, the index typically will store the position of the word in the document, so the search engine can tell where the words are located.

- **Wildcard operators** (also sometimes referred to as "truncation" or "stemming"). This search capability allows the user to widen the search by searching a word stem or incomplete term. Such a search is typically reflected by a symbol such as a question mark (?), asterisk (*), or exclamation point (!). The search engine may also allow the user to restrict the truncation to a certain number of letters by adding additional truncation symbols. For example: "Teach??" would find "teaches" and "teacher," but would not find "teaching." In addition, some engines will allow for internal truncation such as "wom?n," which would find "women" or "woman." The "*" and "!" terms have broader application: for example, hous* would find house, housemate, Houston, household, or other words with the stem "hous."

B. Probabilistic Search Models

Probability theories are used in IR to make decisions regarding relevant documents. A probabilistic search system is based on a formula that places a value on words, their interrelationships, proximity, and frequency. By computing these values, a relevancy ranking can be determined for each document in a search result. This weighting may be based on a variety of factors:

- Frequency of terms within a document—the more times the term appears, the more weight it carries;

- Location of terms within a document—terms in titles and closer to the top of documents are more heavily weighted;

- Adjacency or proximity—the closer the terms are to each other, the heavier the weighting;

- Explicit or implicit feedback on relevance, in which the top-ranked documents are examined and used to refine the probabilistic model.[68]

Examples of probabilistic search models include Okapi BM25, Bayesian networks, and language models.[69]

C. "Fuzzy" Search

Boolean and probabilistic search models rely on exact word matches to form the results of a query. Exact matching is very strict: either a word matches or it doesn't. "Fuzzy" search is an attempt to improve recall by matching more than the exact word: fuzzy matching techniques

68 In explicit relevance feedback, the user codes the top-ranked documents and only those coded relevant are used to refine the model. In implicit relevance feedback, the top-ranked documents are simply assumed to be relevant.

69 *See* Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, Introduction to Information Retrieval, 219-252 (2008)*, available* at http://www-nlp.stanford.edu/IR-book/.

try to reduce words to their core and then match all forms of the word. The method is similar to stemming in Boolean classifiers, discussed above.

Some algorithms for fuzzy matching rely on the understanding that the beginning and end of English words are more likely to change than the center, so they count matching letters and give more weight to words with matching letters in the center than at the edges. Unfortunately, this can sometimes yield results that make little sense (a search for "Tivoli" might bring up "ravioli").

Many systems allow the user to assign a degree of "fuzziness" based on the percentage of characters that are different. Fuzzy search, or matching, has at least two different variations: finding one or more matching strings of a text, and finding similar strings within a fixed string set often referred to as a "dictionary." Fuzzy search has many applications in legal IR including: spellchecking, auto-filling of email addresses, and OCR cleanup.

D. Dimensionality Reduction Systems

Bayesian classifiers are often considered "naïve" because they assume that every word in a document is independent of every other word in the document. In contrast, there is a class of concept-learning technologies that rely on the notion that words are often correlated with one another, and that there is value in that correlation. These methods are also referred to as "dimensionality-reduction techniques" or "dimension-reduction systems."

These systems recognize there is redundancy among word usage and take advantage of that redundancy to find "simpler" representations of text. For example, a document that mentions "lawsuits" is also likely to mention "lawyers," "judges," "attorneys," etc. These words are not synonyms, but they do share certain meaning characteristics. The presence of any one of these words would be suggestive of a common theme. Documents that mentioned any of these terms would likely be about law. Conversely, in searching for one of these words, one might be interested in finding a document that did not contain that exact word, but did contain one of these related words.

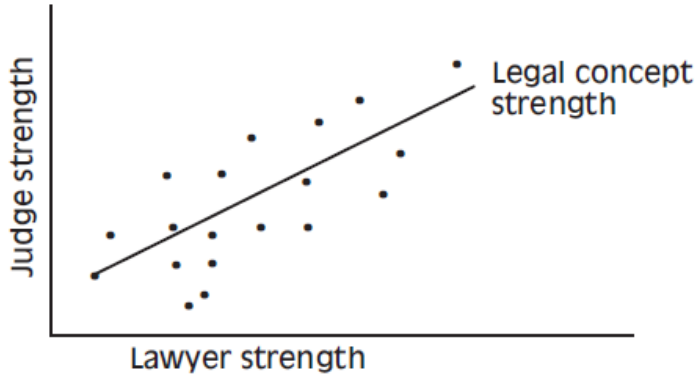


Figure 1. Dimension reduction – the original dimensions of "lawyer" and "judge" are combined into a single dimension. Each point in the graph represents a document. Its location in the graph shows how much the document is related to each dimension.

The figure above illustrates the kind of relationships identified by such systems. The word "lawyer" tends to occur in the same context as the word "judge." Each document has a certain strength along the "lawyer" dimension, related, for example, to how many times the word "lawyer" appears. Similarly, documents have strength along the "judge" dimension, related, for example, to how many times the word "judge" appears. These systems find a new dimension that summarizes the relationship between "lawyer" and "judge." In this example, we are reducing the dimensions from two to one.

Mathematically, we can then describe documents by how much strength they have along this dimension and not concern ourselves with its strength along either the original "lawyer" or "judge" dimensions. The new dimension is a summary of the original dimensions, and the same thing can be done for all words in all documents. We can locate documents along these new, reduced, dimensions or we can represent words along these dimensions in a similar way.

Similarly, multiple words can be represented along dimensions; instead of having just one summary dimension, we can have many of them. Instead of describing a document by how it relates to each of the words it contains, as is done with Vector Space Models,[70] we can describe the document by how it relates to each of these reduced dimensions. Latent Semantic Indexing ("LSI," also called "Latent Semantic Analysis") is the most well-known of these dimension-reducing techniques, but there are others, including neural networks and other kinds of statistical language modeling.

These techniques are similar to one another in that they "learn" the representations of the words in the documents from the documents themselves. Their power comes from reducing the dimensionality of the documents. They simplify representation, and make recognizing meaning easier.

For example, a collection of a million documents might contain 70,000 or more unique words. Each document in this collection can be represented as a list of 70,000 numbers, where each number stands for each word (i.e., the frequency with which that word occurs in that document). Using these techniques, one can represent each document by its strength along each of the reduced dimensions.

One can think of these strengths as a "meaning signature," where similar words will have similar meaning signatures. Documents with similar meanings will have similar meaning signatures. As a result, the system can recognize documents that are related, even if they have different words, because they have similar meaning signatures.

E. Machine-Learning Approaches

There are two main types of machine learning: Unsupervised and Supervised. Unsupervised learning is performed using a large set of examples, without any additional human input. In Supervised Learning methods, the learning examples are tagged individually by a user, and the learning process relies heavily on these examples. Both Supervised and Unsupervised learning may use dimension-reduction techniques as described in Section D.

70 *See* Roitblat, *supra* note 47.

1. Unsupervised Machine Learning (Statistical Clustering)

Systems may use statistics or other unsupervised machine-learning tools to recognize the category to which certain information belongs. The simplest of these is the use of "statistical clustering." Clustering is the process of grouping together documents with similar content. There are a variety of ways to define similarity, but one way is to count the number of words that overlap between each pair of documents. The more words they have in common, the more likely they are to be about the same thing.

Many clustering tools build hierarchical clusters of documents. Some organize the documents into a fixed number of clusters. The quality or "purity" of clustering (i.e., the degree to which the cluster contains only what it should contain) is rarely as high as that obtained using custom built taxonomies or ontologies, but since they require no human intervention to construct, clustering is often an economical and effective first-pass at organizing the documents in a collection. One of the major hurdles when deploying clustering is that there are no objective measures of "clustering quality." Some systems improve the quality of clusters that are produced by starting with a selected number of clusters, each containing selected related documents. These selected documents then function as "seeds" for the clusters. Other related documents are then joined to them to form clusters that correspond to their designer's interests. Then, additional documents are added to these clusters if they are sufficiently similar.

2. Supervised Machine Learning

In the context of text categorization, the general objective of machine learning is to generate a classifier that can automatically classify new untagged documents accurately and efficiently based on a small set of tagged exemplar documents. There are many learning algorithms that can be used in the classifier paradigm, including, but not limited to, Naïve Bayesian, Artificial Neural Networks, Support Vector Machines, and Logistic Regression. The choice of which algorithm to use depends on the nature of the task at hand, the type of data, the characteristics of the learning process, etc.

Selecting the set of exemplar documents that will be used to build a classifier is a key challenge of supervised machine learning—it must ensure comprehensive coverage of the population of documents while minimizing the size of this training set to control costs.

One common approach to selecting the training set is the use of active learning. Active learning is a widely researched field, within which several methods and technologies have been developed and tested.[71]

The premise of active learning, and the main differentiator between this approach and other types of machine learning, is that the learning process is conducted in a number of iterations. The algorithm selects a small sample of documents for each iteration. Each sample is tagged by an expert user, and the tags are fed back to the algorithm as a training set. The algorithm learns from these new tags and generates another sample for the next iteration. The process continues until

71 For a general discussion and comparison of various active learning techniques, see S. B. Kotsiantis, *Supervised Machine Learning: A Review of Classification Techniques*, 31 Informatica Journal 249 (2007), *available at* http://www.informatica.si/PDF/31-3/11_Kotsiantis%20-%20Supervised%20Machine%20Learning%20-%20A%20Review%20of...pdf; G. Cormack and M. Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, 18th Text Retrieval Conference (TREC 2009) Proceedings (2010), http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf.

sufficient learning has occurred and the algorithm can accurately predict the user's classification decisions.

The main advantage of the active-learning technique is that it enables the algorithm to make an informed decision as to which documents are to be included in the next sample. The basis for this decision becomes more and more informed as the number of iterations increases. The objective of active learning is to optimize learning performance by choosing sample documents that provide the maximum contribution to the training of the classifier. In comparison to random sampling, active learning dramatically reduces the number of documents needed for the training stage.

F. Concept and Categorization Tools: Thesauri, Taxonomies, and Ontologies

To deal with the problem of synonymy, some systems rely on a thesaurus, which lists alternative ways of expressing the same or similar ideas. When a term is used in a query, the system uses a thesaurus to automatically search for all similar terms. The combination of query term and the additional terms identified by the thesaurus can be said to constitute a "concept."

The quality of the results obtained with a thesaurus depends on the quality of the thesaurus, which, in turn, depends on the effort expended to match the vocabulary and usage of the organization using it. Generic thesauri, which may attempt to represent the English language or are specialized for particular industries, are sometimes available to provide a starting point, but each group or organization has its own jargon and own way of talking that require adjustment for effective categorization. In America, for example, the noun "jumper" is a child's one-piece garment. In Australia, the noun "jumper" is a sweater. In America, a 3.5 inch removable disk device was called a "floppy" during its heyday. But in Australia, it was called a "stiffy."

Taxonomies and ontologies are also used to provide conceptual categorization. Taxonomy is a hierarchical scheme for representing classes and subclasses of concepts. The figure below shows a part of a taxonomy for legal personnel. Attorneys, lawyers, etc., are all types of legal personnel. The only relations typically included in a taxonomy are hierarchical or inclusion relations. Items lower in the taxonomy are subclasses of items higher in the taxonomy. For example, the NAICS (North American Industry Classification System) is one generally available taxonomy that is used to categorize businesses. In this taxonomy, the category "Information" has subclasses of "Publishing," "Motion Picture and Sound Recording Industries," and "Broadcasting."

One can use this kind of taxonomy to recognize conceptual relationship among these different types of personnel. If a category includes law personnel, then any document that mentions attorney, lawyer, paralegal, etc., should be included in that category. Like thesauri, there are a number of commercially available taxonomies for various industries.

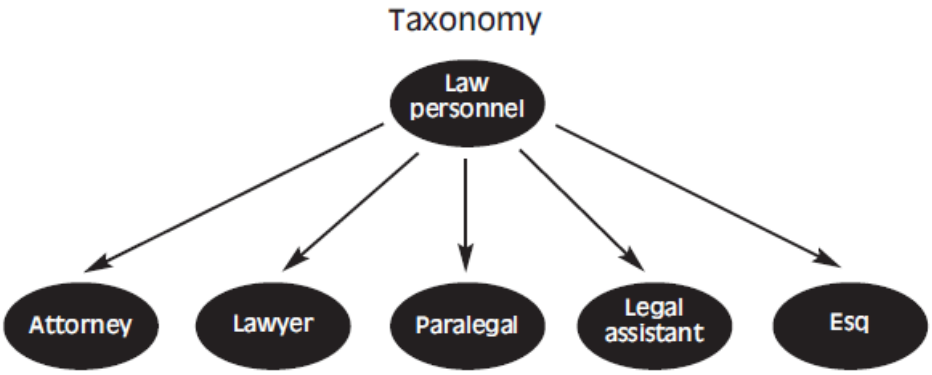


Figure 2. A simple taxonomy for law personnel.

Predefined taxonomies exist for major business functions and specific industries. It may be necessary to adapt these taxonomies to one's particular organization or matter.

An ontology is a more generic species of taxonomy, often including a wider variety of relationship types than are found in the typical taxonomy. An ontology specifies the relevant set of conceptual categories and how they are related to one another. The figure below shows part of an ontology covering subject matter similar to that described in the preceding taxonomy. For clarity, only a subset of the connections between categories is shown. According to this ontology, if the category includes attorneys, the user may also be interested in documents that use words such as "lawyer," "paralegal," or "Esq." Like taxonomies, ontologies are most useful when they are adapted to the specific information characteristics of the organization.

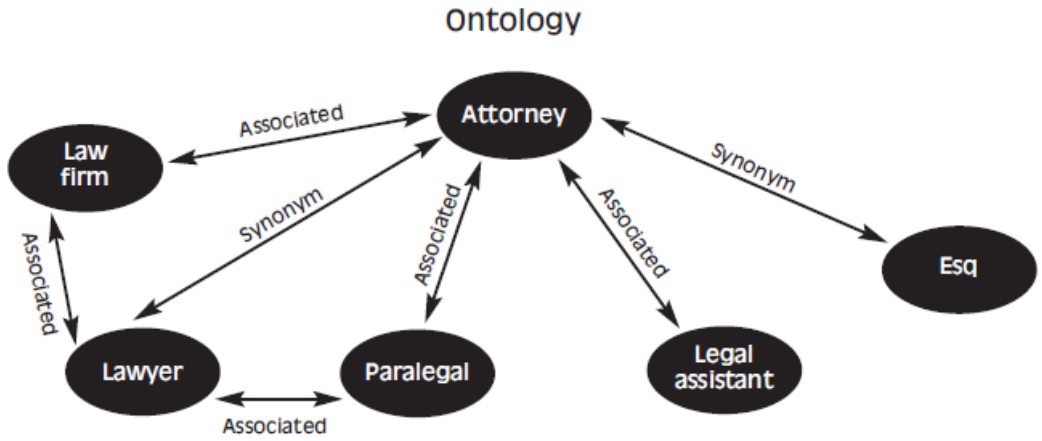


Figure 3. A section of an ontology of legal personnel.

Taxonomies, ontologies, and thesauri are all knowledge structures. They represent explicit knowledge about some subject. An expert writes down the specific relations she knows about. Although there are tools that help the expert create these structures, they still tend to represent only the information the expert can explicitly describe as important.

The structure of the thesaurus, taxonomy, or ontology can be used as the organizing principle for a collection of documents. Rules are derived that specify how documents with specific words in them are related to each of these categories, and the computer can then be used to organize the documents into the corresponding categories.

These rules can be created explicitly, or they can be created using machine learning techniques. Explicit rules are created by knowledge engineers. For example, one rule might include a Boolean statement like this: (acquir* or acquisition or divest* or joint venture or alliance or merg*) and (compet* or content or program*) that specifies the critical words that must appear for a document to be assigned to the "merger" category. The effectiveness of rules like these depends critically on the ability of the knowledge engineers to guess the specific words that document authors actually used. Syntactic rules may also be employed by some systems. For example, a system may only look for specific words when they are part of the noun phrase of a sentence.

G. Presentation/Visualization and Social-Networking Tools

Presentation and visualization software technologies may incorporate search and retrieval functionality that may be found to have useful applications. These technologies can organize information (e.g., emails) so that a searcher can more efficiently study the search topic (including finding relevant emails). They also are good at highlighting patterns of "social networks" within an organization that would not necessarily be apparent by more traditional searches. Subject to some exceptions, the results of any search and retrieval query can be presented in a variety of forms, including as a:

1. List—items in sequence, for example messages ordered by sent date
2. Sort—sortable items aggregated into rows by columns, for example messages by sender
3. Group—items categorized or totaled, for example count of messages by sender
4. Cluster—items in groups organized by spatial proximity, for example relevant groups spiraling out to less relevant groups
5. Tree—items in parent/child hierarchy, for example, folder and subfolder(s)
6. Timeline—items arrayed by a time element, for example a list/group of items arrayed by sent date
7. Thread—items grouped by conversation
8. Network—items arrayed by person, for example a diagram of message traffic between sender(s) and recipient(s)
9. Map—items plotted by geography, for example items plotted by city and state of origin
10. Cube—items in a multidimensional pivot table; including, table, group, timeline, and tree functionality

In practice, a searcher can load search results into a presentation technology for an organized view and then drill down to access discrete items of particular interest or concern. This often iterative process may help a searcher to learn more about, act on, and manage search results.

The Sedona Conference Working Group Series & WGS Membership Program

"**Dialogue designed to move the law forward in a reasoned and just way.**"

The Sedona Conference Working Group Series ("WGS") represents the evolution of The Sedona Conference from a forum for advanced dialogue to an open think-tank confronting some of the most challenging issues faced by our legal system today.

The WGS begins with the same high caliber of participants as our regular season conferences. The total, active group, however, is limited to 30-35 instead of 60. Further, in lieu of finished papers being posted on the website in advance of the Conference, thought pieces and other ideas are exchanged ahead of time, and the Working Group meeting becomes the opportunity to create a set of recommendations, guidelines or other position piece designed to be of immediate benefit to the bench and bar, and to move the law forward in a reasoned and just way. Working Group output, when complete, is then put through a peer review process, including where possible critique at one of our regular season conferences, hopefully resulting in authoritative, meaningful and balanced final papers for publication and distribution.

The first Working Group was convened in October 2002, and was dedicated to the development of guidelines for electronic document retention and production. The impact of its first (draft) publication—*The Sedona Principles; Best Practices Recommendations and Principles Addressing Electronic Document Production* (March 2003 version)—was immediate and substantial. *The Principles* was cited in the Judicial Conference of the United State Advisory Committee on Civil Rules Discovery Subcommittee Report on Electronic Discovery less than a month after the publication of the "public comment" draft, and was cited in a seminal e-discovery decision of the Federal District Court in New York less than a month after that. As noted in the June 2003 issue of Pike & Fischer's *Digital Discovery and E-Evidence, "The Principles*...influence is already becoming evident."

The WGS Membership Program was established to provide a vehicle to allow any interested jurist, attorney, academic or consultant to participate in Working Group activities. Membership provides access to advance drafts of Working Group output with the opportunity for early input, and to Forums where reference materials are posted and current news and other matters of interest can be discussed. Members may also indicate their willingness to volunteer for Brainstorming Groups and Drafting Teams. See the "Working Group Series" area of our website: www.thesedonaconference.org for further details on our Working Group Series and the Membership Program.

Visit www.thesedonaconference.org