Bloomberg Law®

Digital Discovery & e-Evidence®

Sampling

Analytics, Metrics and Sampling: Tools Needed for Litigating in the Age of eDiscovery

BY PHILIP FAVRO

In the halcyon days of paper documents, discovery involved document productions that often numbered in the hundreds or the thousands. In those days, it was easier to manually sort through documents by hand. If there were concerns that files had been overlooked, they could be quickly reviewed again to better gauge the accuracy of the review.

In the age of eDiscovery, exclusive reliance on manual review is not effective. Nor is it feasible given the sheer volume of electronically stored information (ESI) that must be addressed in most matters. *See Dynamo Holdings Ltd. P'Ship v. Comm'r of Internal Revenue*, No. 2685-11, 2016 BL 235588 (T. C. July 13, 2016). Counsel should explore other methods that satisfy notions of proportionality and that can help clients prepare for the ultimate disposition of a matter.

U.S. Magistrate Judge Patrick Walsh observed as much in 2013 when he urged counsel to use ''21st century computer technology'' to conduct discovery. In so doing, Judge Walsh declared that ''the biggest problem I see with electronic discovery is that lawyers are using 20th century technology—i.e., obtaining all of the documents, organizing them in folders, and trying to read and digest them—to address 21st century issues.'' *See* Hon. Patrick J. Walsh, *Rethinking Civil Litigation in Federal District Court*, 40 ABA LITIG. J. 1 (Fall 2013).

To get up to speed with ''21st century computer technology,'' Judge Walsh recommended that lawyers use analytics. Counsel should also consider using metrics and sampling in connection with analytics. These tools can help lawyers effectively evaluate production quality, among many other uses. This article will provide a brief overview of these concepts and demonstrate their utility in addressing eDiscovery challenges in the twenty-first century.

Philip Favro is a discovery and information governance consultant for Driven, Inc.

Analytics

Look up the word ''analytics'' in the dictionary and it defines the term as referring to a method of logical analysis. This is an excellent characterization of how analytics works in the context of discovery. Analytics provide a logical process for examining substantial amounts of data. They enable counsel to intelligently analyze that data in a fashion superior to the traditional method of manual review. Case studies, court decisions, and legal scholarship confirm that analytics can better identify key documents needed for the disposition of a particular matter, along with expediting the overall search for relevant ESI.

For discovery purposes, analytics include advanced search methodologies such as technology-assisted review (TAR), data clustering, and concept searching. Discovery analytics also encompass email threading and near duplicate identification, tools which can segregate superfluous information from a review set, thereby facilitating and expediting a document review.

Counsel may elect to use one or a combination of analytical search methods or tools to explore a population of documents. See *2016 Guidelines Regarding the Use of Technology-Assisted Review*, COALITION OF TECHNOLOGY RESOURCES FOR LAWYERS 4–5 (2016). This includes conducting searches through an adversary's document production. While concept searching, data clustering, or search terms may be useful in this regard, TAR can be particularly helpful. This is especially the case with more advanced TAR technologies that use active learning training methods, which enable counsel to train a TAR algorithm to identify documents on a specific set of issues. See Hon. John M. Facciola & Philip J. Favro, *Safeguarding the Seed Set: Why Seed Set Documents May Be Entitled to Work Product Protection*, 8 FED. CTS. L. REV. 1, 8–9 (2015).

While analytics offer utility for discovery reviews, they may have limited application in certain scenarios given their different functionalities. For example, TAR is particularly well suited for processing through large numbers of documents. *See Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182, 191 (S.D.N.Y. 2012) (suggest-

 ISSN 1941-3882

ing that TAR is better suited for ''large-data-volume cases''). It may not be the best methodology, though, for exploring a population of fewer than 20,000 documents given training and cost issues. A careful examination of different analytical search methods or tools will help counsel understand which one or a combination of them will be more likely to enhance a client's litigation position in discovery. See *The Sedona Conference Commentary on Proportionality in Electronic Discovery*, 18 SEDONA CONF. J. 141, 174 (2017) (*Proportionality Commentary*) (observing in Principle 6 that parties should have the discretion to select technologies that address their discovery needs).

Metrics

Despite their usefulness, exclusive reliance on analytics may not be sufficient to satisfy the obligations that counsel and client have to certify that productions meet proportionality standards. Without the benefit of benchmarks to demonstrate the quality and nature of a document production, even the most advanced analytics might prove inadequate for establishing that a production satisfies proportionality standards. *See Cargill Meat Solutions Corp. v. Premium Beef Feeders, LLC*, No. 13-cv-1168-EFM-TJJ, 2015 BL 205178 (D. Kan. June 26, 2015) (citing plaintiff's ''unsupported estimate'' of cost as grounds for rejecting its undue burden objection).

To accomplish this objective, counsel need metrics. Metrics are typically defined as a standard of measurement or (as used in business world) a method for evaluating performance. In like manner, metrics offer counsel ways to assess the ''performance'' of a particular document production. Metrics can gauge the extent to which a production contains relevant materials, undisclosed privileged information, and even nonresponsive documents. They can also be used to measure the quality and nature of adversaries' productions. See *Proportionality Commentary* at 167-68 (describing in Principle 4 the role and importance of metrics in substantiating assertions regarding undue burden, among various issues).

Key metrics include prevalence, recall, and precision. Prevalence measures the percentage of responsive information in a particular document population. Recall refers to the percentage of that responsive information produced to an adversary while precision measures the percentage of responsive documents within a particular production.

Although prevalence, recall, and precision are important, different metrics may be essential for demonstrating that a specific document request is disproportionately burdensome. Detailed metrics that reflect the investment of resources (including time, manpower, and costs) that a party will invest to satisfy a discovery request must be developed and then disclosed to adversaries and likely a court in order to invoke proportionality limitations. See *Duffy v. Lawrence Memorial Hosp.*, No. 14-cv-2256, 2017 BL 105975 (D. Kan. Mar. 31, 2017) (modifying its discovery order to allow defendant to produce sampled data given the expense, time, and opportunity costs associated with full compliance with that order).

There are various other metrics that can provide useful benchmarks in discovery. Just as with analytics, counsel should carefully evaluate which metrics will best serve client needs for gauging discovery performance.

Sampling

To obtain these metrics and thereby evaluate particular aspects of a production, counsel often must rely on a third critical tool: sampling. Sampling involves a very simple concept: that ''a few'' can adequately represent ''the many.'' Sampling has been deemed a sufficient measure in a variety of different circumstances such as measuring patient health or determining how voters are leaning on a particular candidate or ballot measure.

A properly designed sample of ESI can likewise be sufficient for discovery purposes. In particular, sampling can be used to depict information from a broader set of data among dozens or hundreds of custodians spanning a range of years. See *Proportionality Commentary* at 165–67 (discussing generally in Principle 4 the benefits of sampling, together with instances where it sampling can be used). Properly sampled data, after it is reviewed, can provide counsel with intelligence on any number of issues relating to a document population. This raw data can be distilled into any number of metrics, which can then be shared with adversaries or the court.

For example, counsel can sample a document population at the outset of a review to determine its prevalence. This will help counsel establish goals for other metrics – particularly recall – relating to its production. Sampling can also help counsel ascertain whether a production contains the level of recall to which the parties agreed in an ESI protocol.

Prior to production, counsel can analyze sampled data to determine if nonresponsive or privileged documents have been mistakenly included. Sampling—as opposed to spot-checking—is essential for accomplishing this task. Spot-checking does not provide a reliable gauge as to the contents of a particular production and could result in the disclosure of nonresponsive or privileged information. This was the case in 2017 when a financial institution inadvertently produced ''a vast trove of confidential information about tens of thousands of [its] wealthiest clients.'' Such a result could have been avoided had the institution's counsel reviewed a properly generated random sample instead of conducting a cursory ''spot-check'' of the client's production.

Finally, sampling can be used to obviate disproportionate production burdens by providing adversaries with representative examples of responsive information. For example, in *Duffy*, the court ordered defendant to produce a random sample of 257 patient records in lieu of searching through 15,574 separate patient files for responsive information. Similarly, *Solo v. United Parcel Service* held that defendant needed to share a sample from six months of certain package shipment information instead of producing five years of such data. *Solo v. United Parcel Service Co.*, No. 14-12719, 2017 BL 6463 (E.D. Mich. Jan. 10, 2017). In each case, a combination of sampling and metrics enabled the responding parties to avoid disproportionate production burdens.

Conclusion

It has been over four years since Judge Walsh urged lawyers to adopt discovery technologies and methods suited to the digital age for pursuing litigation objec-

  DDEE ISSN 1941-3882

tives. While it is understandable that some counsel may have trepidation about learning how to use new innovations, competence in representing clients demands that technological fears be overcome. Becoming proficient in analytics, metrics, and sampling will ultimately help counsel obtain better and more efficient litigation results for clients in the age of eDiscovery.