SAMPLING: MAKING ELECTRONIC DISCOVERY MORE COST EFFECTIVE

Milton Luoma
Metropolitan State University
700 East Seventh Street
St. Paul, Minnesota 55337
651 793-1246 (fax)
651 793-1481
Milt.Luoma@metrostate.edu

Vicki Luoma
Minnesota State University
145 Morris Hall
Mankato, Minnesota 56001
507 389-5497
507 389-1916
Vicki.Luoma@mnsu.edu

ABSTRACT

With the huge volumes of electronic data subject to discovery in virtually every instance of litigation, time and costs of conducting discovery have become exceedingly important when litigants plan their discovery strategies. Rather than incurring the costs of having lawyers review every document produced in response to a discovery request in search of relevant evidence, a cost effective strategy for document review planning is to use statistical sampling of the database of documents to determine the likelihood of finding relevant evidence by reviewing additional documents. This paper reviews and discusses how sampling can be used to make document review more cost effective by considering issues such as an appropriate sample size, how to develop a sampling strategy, and taking into account the potential value of the litigation in relation to the costs of additional discovery efforts.

Keywords: sampling, statistical sampling, electronic discovery

1. INTRODUCTION

Litigation has always been about the adversarial relationship and zealous representation of one's clients and their interests, but with the rapid expansion in the volume of electronically stored information (ESI) lawyers have found themselves having to become non-adversarial in the discovery phase of litigation now that electronic discovery is the norm. With the relatively low cost of data storage and the seemingly limitless amount of ESI to search, the new Federal

Rules of Civil Procedure were amended to require lawyers and the parties to fully cooperate in the management of the discovery process.

At the onset of litigation, a party must comply with Federal 26(a)(1)(B), which requires full disclosure of a great deal of basic information as described below.

> Rule 26. Duty to Disclose; General Provisions Governing Discovery
>
> (a) Required Disclosures.
>
> (1) Initial Disclosures.
>
> (A) *In General*. Except as exempted by Rule 26(a)(1)(B) or as otherwise stipulated or ordered by the court, a party must, without awaiting a discovery request, provide to the other parties:
>
> (i) the name and, if known, the address and telephone number of each individual likely to have discoverable information — along with the subjects of that information — that the disclosing party may use to support its claims or defenses, unless the use would be solely for impeachment;
>
> (ii) a copy — or a description by category and location — of all documents, electronically stored information, and tangible things that the disclosing party has in its possession, custody, or control and may use to support its claims or defenses, unless the use would be solely for impeachment; (Federal Rules of Civil Procedure, 2007)

The rule requires adversaries to exchange either a copy of or a description of all electronically stored information by category and location that may be used in their legal claim or defense against the claim. This requirement is tantamount to asking a poker player to show his or her hand before bets are placed. However, it really is not as simple as showing your hand in a poker game because most of the time the party has no idea what they have, where it is located and how to produce it.

2. HOW MUCH TRUTH CAN YOU AFFORD?

There is simply too much information to produce all of one's ESI or even to list everything one has or even to know what one has. In the oft cited case of Zubukake v. UBS Warburg, Judge Shira Scheindlin wrote: "Discovery is not just about uncovering the truth, but also about how much of the truth the parties can

afford to disinter." (Zubulake v. UBS Warburg LLC, 2003)

Laura Zubulake sued her former employer UBS Warburg over gender discrimination. The case became a catalyst for development of the new discovery rules and procedures. Zubulake requested documents stored or produced in electronic format, which were primarily emails. UBS Warburg claimed either the data could not be found or it had been lost. In a series of five pre-trial rulings the judge examined cost shifting, discovery obligations, and responsibilities of maintaining and retrieving data. Judge Scheindlin ultimately found that the defendant had a duty to preserve data that it knew or should have known were relevant to the litigation. To determine the issue of cost shifting the judge ordered the defendant to restore and review information from five backup tapes out of a total of 94 available tapes. The court allowed Zubulake to select five tapes out of the 94 for sampling. Defendant Warburg was ordered to submit an affidavit with the results of the sampling along with costs. (Zubulake, 2004)

Zubulake chose five tapes with emails from her former supervisor. After the five sample tapes were restored, the defendant revealed there were 6,203 unique emails contained in the sample data. In the next step in the recovery process keyword searches were used to find emails that made reference to Zubulake, reducing the messages to 1,075 unique messages and claimed that of those 1,075 only 600 were subject to Zubulake's document request. This process cost Warbug over $19.000.00. Warburg estimated the cost to restore and produce the remaining tapes to be approximately $273,649.39. (Zubulake v. UBS Warburg LLC, 2003). There are numerous important rulings in this case but what is remarkable is the use of sampling as a method of reducing costs and narrowing search requirements. This case used sampling to determine whether more searching should be conducted and whether costs should be shifted to the party seeking the information. (Zubulake, 2004)

In 2007 the new Federal Rules of Civil Procedure were adopted and Rule 34 included a provision for sampling. Rule 34 reads as follows:

> 1.1 Rule 34. **Producing Documents, Electronically Stored Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes**
>
> *1.1.1.1 (a) In General*
>
> *1.1.1.2 A party may serve on any other party a request within the scope of Rule 26 (b)*
>
> (1) to produce and permit the requesting party or its representative to inspect, copy, test, or sample the following items in the responding party's possession, custody, or control:

> (A) any designated documents or electronically stored information — including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations — stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form; or
>
> (B) any designated tangible things; or
>
> (2) to permit entry onto designated land or other property possessed or controlled by the responding party, so that the requesting party may inspect, measure, survey, photograph, test, or sample the property or any designated object or operation on it. (FRCP 26(b)

The important point is that a party may serve a request to sample data. Sampling should be used routinely in cases with large amounts of electronically stored information to find the data needed whether in producing the data or in defending a search methodology. It also permits a lawyer to be both cooperative and adversarial at the same time. This procedure places a greater responsibility on the requesting party to apply the reasonableness standard to determine what should be sampled. On the other hand, sampling permits the party providing the data to verify to the court that he or she has made a reasonable effort to comply with discovery by checking the results.

3. RECENT CASES FOLLOWING THE ZUBULAKE GUIDELINES

In 2010 in Makrakis v. Demelis, the plaintiff sought damages from the defendant nurse Demelis and her employer, Brigham & Women's Hospital, for damages when the nurse improperly administered a toxic dose of a drug to the plaintiff. The plaintiff asked the court for an order requiring the hospital to restore all electronic backup tapes containing emails originating from thirteen employees or former employees of the hospital from 1987 to 2010. The plaintiffs sought an order requiring the hospital to hire a third-party vendor to search the restored email archives using the keywords "Makrakis," "DeMelis," "pancuronium," and "Pavulon." Further, plaintiffs sought a court order compelling production of all emails sent or received by DeMelis at any time. The defendants opposed the request on the grounds that the search would be unduly burdensome, prohibitively expensive and not add anything relevant to the information they already had. The court, citing Zubulake, ordered the defendants to sample a small number of backup tapes, at the expense of the requesting party. (Makrakis v. Demelis, 2010)

In another 2010 case the court ruled that a phased approach to ESI discovery is

appropriate and reasonable approach. In this case through sampling the discovery costs were reduced from the estimated $60,000 to $13,000 (Barrera v. Boughton, 2009)

In a 2009 case the court found that, that "sampling to test both the cost and the yield is now part of the mainstream approach to electronic discovery." (S.E.C v. Collins & Aikman Corp, 2009)

Courts now require litigants to be even more responsible for the success and accuracy of the discovery process by requiring them to defend their chosen search methods and to show how they have verified or validated their results. In other words, what tests were done to establish that the methodologies used were efficacious?

Judge Grimm found:

> Additionally, the defendants do not assert that any sampling was done of the text searchable ESI files that were determined not to contain privileged information on the basis of the keyword search to see if the search results were reliable. Common sense suggests that even a properly designed and executed keyword search may prove to be over-inclusive or under-inclusive, resulting in the identification of documents as privileged which are not, and non-privileged which, in fact, are. The only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents determined to be privileged and those determined not to be in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive. There is no evidence on the record that the Defendants did so in this case. (Victor Stanley, Inc. v. Creative Pipe, Inc., 2010)

Based on these and other cases with similar rulings, sampling must be considered by all parties to litigation in order to reduce discovery costs.

4. SAMPLING – HOW IT WORKS

Sampling can assist one both in finding the data required to strengthen one's case, but it can also be used to certify one's ESI discovery results. In sampling one must be able to show precision, confidence, and the expected deviations. The Electronic Discovery Reference Model (EDRM) Search Group outlines a strategy for using sampling. (The Electronic Discovery Reference Model, 2005) As the EDRM Search Guide states, sampling can only be done by the one who has the data, which may not always be in line with the requesting parties' demands. The Sedona Conference, Working Group Commentary, Achieving Quality in the E-Discovery Process discusses several sampling methods and their purposes. (Working Group 1, 2009)

Essentially, sampling a set of electronic documents is a tradeoff between

obtaining every possible relevant document, which will invariably result in a very high cost, versus reviewing a smaller set of the documents at a lower cost, but running the risk of missing relevant documents that may be critical to the case. For large scale litigation or in cases where limited resources may be available for the discovery process, sampling is an intelligent alternative to attempting to review every possible document that is available.

There are several types of sampling that can be used in a sampling procedure. For example, the Sedona Conference identified five quality measures including judgment sampling as very helpful (p12) even though one cannot make generalized statements about the entire population of documents. This form of sampling can be used in a quality control context where a small sample of documents can be selected from a set of documents that have been reviewed by junior counsel to determine whether or not the document reviewer has exercised proper judgment regarding how the document was classified, that is, as relevant or not. However, not just any sample will do. The very best kind of sample is one that is representative of the entire population of electronic documents.

While several other sampling methods exist, but the most important of these is statistical sampling that permits one to generalize about the entire population of documents based on a random sample of documents. The question that must be answered for anyone designing a sampling procedure is how large must the sample be? The answer to that question depends on how confident one wants to be that the sample size is truly representative of the population and what range of the estimate of the proportion of relevant documents is required.

To determine the sample size when one wishes to determine the proportion of documents in the population of documents that are relevant for discovery purposes, one must determine or estimate five items: 1) the desired interval range within which the population proportion is expected, 2) the confidence level for estimating the interval within which to expect the population proportion, 3) the standard error of the proportion, 4) an estimate of the proportion of the population which contains relevant documents, and 5) calculate the sample size.

First, the desired interval range within which the population proportion is expected is a wholly subjective decision. For example, if one wants the resulting interval range to be within 10 percent of the population's true proportion of relevant documents, then this figure will be plus or minus 0.10. If one wants a tighter limit on the interval range, such as five percent, then this figure will be plus or minus 0.05. So, if one wants to be able to say the population of electronic documents contains X% relevant documents plus or minus 10%, then the sample size will be determined with this requirement in mind as shown below.

Second, the confidence level desired for the final estimate of the population range is a subjective choice where the calculations are based on the Normal distribution, or classical bell curve, and incorporates values based on the standard deviation or standard error of the Normal distribution. For example, a common choice for

confidence level is 90% or 95%, so ultimately one will be able to say something like, "I am 95% certain that the population of electronic documents contains 70% + or - 10% documents relevant to the litigation at hand."

Third, one must estimate the standard error of the proportion of relevant documents. This figure is obtained by dividing the result of step 1 by 1.65 if one desires a confidence level of 90% or dividing the result of step 1 by 1.96 if one desires a confidence level of 95%; or dividing the result of step 1 by 3.00 if one desires a confidence level of 99%. The following table shows the results of using interval ranges of 10%, 5%, and 1% and confidence levels of 90%, 95%, and 99%.

Estimate of the Standard Error of the Proportion of Relevant Documents			
Proportion of Relevant Documents + or - %	90% confidence	95% confidence	99% confidence
10%	0.06061	0.05102	0.03333
5%	0.03030	0.02551	0.01667
1%	0.00606	0.00255	0.001667

Fourth, to determine sample size one first needs to estimate the proportion of the documents in the population that are relevant. Since that is not generally known beforehand, one must estimate that proportion before calculating the sample size. The best way to estimate that proportion is to complete some preliminary testing or pilot sampling by randomly selecting several documents and determining the proportion of this sample that contains relevant documents. Usually, about 30 documents per pilot sample are sufficient. This preliminary testing or pilot sampling can be repeated several times. If the selection of documents for each pilot sample is random, then the average proportion of relevant documents contained in the samples should be close to the population's proportion of relevant documents. The product of the proportion of relevant documents multiplied by the proportion of non-relevant documents is referred to as the dispersion of the sample.

Finally, the sample size is calculated by dividing the sample dispersion by the estimate of the standard error of the proportion multiplied by itself. For example, suppose we wish to be 95% confident that the proportion of relevant documents in the population is 20% plus or minus 5% and the proportion of relevant documents in the pilot sampling procedure was 20%, then our sample size is (0.20)(0.80) / 0.02551 =245.866 rounded off to 246, which is a reasonable number to review.

On the other hand, consider the situation if one desires to be 99% confident that the proportion of relevant documents in the population is 20% plus or minus 1% and the proportion of relevant documents in the pilot sampling procedure was 20%, then our sample size is (0.20)(0.80) / 0.001667 = 57,577! Clearly, the tighter the interval and the higher the confidence level desired increases the sample size – in some cases quite dramatically.

5. CONCLUSION

In conclusion, it is clear that ESI sampling has become an important aspect of electronic discovery. It has been used as a means of validating search methodologies as well as a means of containing discovery costs and maintaining quality control over the discovery process. While several sampling methods are available, statistical sampling can be an effective way of describing the characteristics of an entire population of ESI documents based on a relatively small sample of documents randomly selected from the population. It further permits one to establish the confidence level of the sampling results and the range of accuracy of the results. It therefore behooves lawyers to educate themselves on the procedures involved in the development of statistical sampling methodologies, which may at the very least satisfy the safe harbor provisions of the Federal Rules of Civil Procedure.

REFERENCES

Barrera v. Boughton, 256 F.R.D. 403, 418 (S.D.N.Y. 2009).

Cooper, D. R. and P.S. Schindler (2003), Business Research Methods, McGraw-Hill, Boston.

Federal Rules of Civil Procedure, 26(a)(1)(B) (December 2007).

Makrakis v. Demelis, 2010 WL 3004337 (09-706-C July 13, 2010).

S.E.C v. Collins & Aikman Corp, 256 F.R.D. 403, 418 (S.D.N.Y. 2009).

The Electronic Discovery Reference Model. (2005). Retrieved December 16, 2010, from The Electronic Discovery Reference Model: http://www.law.com/jsp/legaltechnology/eDiscoveryRoadmap.jsp

Victor Stanley, Inc. v. Creative Pipe, Inc., 2010 WL 3530097 (D.MD 2010).

Working Group 1. (2009). "Achieving Quality in the E-Discovery Process.". *Commentary by the Working Group 1 of The Sedona Conference®*. Sedona: Sedona Conference.

Zubulake v. UBS Warburg LLC, 217 F.R.D. 309, 312 (S.D.N.Y 2003).