Richmond Journal of Law and Technology

Volume 20 | Issue 2 Article 4

2014

Defensible Data Deletion: A Practical Approach to Reducing Cost and Managing Risk Associated with Expanding Enterprise Data

Dennis R. Kiker

DEFENSIBLE DATA DELETION: A PRACTICAL APPROACH TO REDUCING COST AND MANAGING RISK ASSOCIATED WITH EXPANDING ENTERPRISE DATA

Dennis R. Kiker*

Cite as: Dennis R. Kiker, *Defensible Data Deletion: A Practical Approach to Reducing Cost and Managing Risk Associated with Expanding Enterprise Data*, 20 RICH. J.L. & TECH. 6 (2014), http://jolt.richmond.edu/v20i2/article6.pdf.

I. INTRODUCTION

[1] Modern businesses are hosts to steadily increasing volumes of data, creating significant cost and risk while potentially compromising the current and future performance and stability of the information systems in which the data reside. To mitigate these costs and risks, many companies are considering initiatives to identify and eliminate information that is not needed for any business or legal purpose (a process referred to herein as "data remediation"). There are several challenges for any such initiative, the most significant of which may be the fear that information subject to a legal preservation obligation might be destroyed. Given the volumes of information and the practical limitations of search technology, it is simply impossible to eliminate all risk that such information might be overlooked during the identification or remediation process. However, the law does not require that corporations eliminate "all risk." The law requires that

* Dennis Kiker has been a partner in a national law firm, director of professional services at a major e-Discovery company, and a founding shareholder of his own law firm. He has served as national discovery counsel for one of the largest manufacturing companies in the country, and counseled many others on discovery and information governance-related issues. He is a Martindale-Hubbell AV-rated attorney admitted at various times to practice in Virginia, Arizona and Florida, and holds a J.D., *magna cum laude* & Order of the Coif from the University of Michigan Law School. Dennis is currently a consultant at Granite Legal Systems, Inc. in Houston, Texas.

corporations act reasonably and in good faith,[1] and it is entirely possible to design and execute a data remediation program that demonstrates both. Moreover, executing a reasonable data remediation program yields more than just economic and operational benefits. Eliminating information that has no legal or business value enables more effective and efficient identification, preservation, and production of information requested in discovery.[2]

[2] This Article will review the legal requirements governing data preservation in the litigation context, and will demonstrate that a company can conduct data remediation programs while complying with those legal requirements. First, we will examine the magnitude of the information management challenge faced by companies today. Then we will outline the legal principles associated with the preservation and disposition of information. Finally, with that background, we will propose a framework for an effective data remediation program that demonstrates reasonableness and good faith while achieving the important business objectives of lowering cost and risk.

II. THE PROBLEM: MORE DATA THAN WE WANT OR NEED

[3] Companies generate an enormous amount of information in the ordinary course of business. More than a decade ago, researchers at the University of California at Berkeley School of Information Management

[1] *See* THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: SECOND EDITION BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION 28 (Jonathan M. Redgrave et al. eds., 2007) [hereinafter "THE SEDONA PRINCIPLES"], *available at* http://www.sos.mt.gov/Records/committees/erim_resources/A%20-%20Sedona%20Principles%20Second%20Edition.pdf (last visited Jan. 30, 2014); *see also* Louis R. Pepe & Jared Cohane, *Document Retention, Electronic Discovery, E-Discovery Cost Allocation, and Spoliation Evidence: The Four Horsemen of the Apocalypse of Litigation Today*, 80 CONN. B. J. 331, 348 (2006) (explaining how proposed Rule 37(f) addresses the routine alteration and deletion of electronically stored information during ordinary use).

[2] *See* THE SEDONA PRINCIPLES, *supra* note 1, at 12.

and Systems undertook a study to estimate the amount of new information generated each year.[3] Even ten years ago, the results were nearly beyond comprehension. The study estimated that the worldwide production of original information as of 2002 was roughly five exabytes of data, and that the storage of new information was growing at a rate of up to 30% per year.[4] Put in perspective, the same study estimates that five exabytes is approximately equal to all of the words ever spoken by human beings.[5] Regardless of the precision of the study, there is little question that the volume of information, particularly electronically stored information ("ESI") is enormous and growing at a frantic pace. Much of that information is created by and resides in the computer and storage systems of companies. And the timeworn adage that "storage is cheap" is simply not true when applied to large volumes of information. Indeed, the cost of storage can be great and come from a number of different sources.

[4] First, there is the cost of the storage media and infrastructure itself, as well as the personnel required to maintain them. Analysts estimate the total cost to store one petabyte of information to be almost five million dollars per year.[6] The significance of these costs is even greater when one realizes that the vast majority of the storage for which companies are currently paying is not being used for any productive purpose. At least one survey indicates that companies could defensibly dispose of up to 70% of the electronic data currently retained.[7]

---

[3] *See* Peter Lyman & Hal R. Varian, *How Much Information 2003?*, http://www.sims.berkeley.edu/research/projects/how-much-info-2003/ (last visited Feb. 9, 2014).

[4] *Id.*

[5] *See id.*

[6] Jake Frazier, *Hoarders: The Corporate Edition*, BUSINESS COMPUTING WORLD (Sept. 25, 2013), http://www.businesscomputingworld.co.uk/hoarders-the-corporate-edition/.

[7] *Id.*

[5] Second, there is a cost associated with keeping information that currently serves no productive business purpose. The existence of large volumes of valueless information makes it more difficult to find information that is of use. Numerous analysts and experts have recognized the tremendous challenge of identifying, preserving, and producing relevant information in large, unorganized data stores.[8] As data stores increase in size, identifying particular records relevant to a specific issue becomes progressively more challenging. One of the best things a company can do to improve its ability to preserve potentially relevant information, while also conserving corporate resources, is to eliminate information from its data stores that has no business value and is not subject to a current preservation obligation.

[6] Eliminating information can be extremely challenging, however, due to the potential cost and complexity associated with identifying information that must be preserved to comply with existing legal obligations. When dealing with large volumes of information, manual, item-by-item review by humans is both impractical and ineffective. From the practical perspective, large volumes of information simply cannot be reviewed in a timely fashion with reasonable cost. For example, consider an enterprise system containing 500 million items. Even assuming a very aggressive review rate of 100 documents per hour, 500 million items would require five million man-hours to review. At any hourly rate, the cost associated with such a review would be prohibitive.

[7] Even when leveraging commonly used methods of data culling to reduce the volume required for review, such as deduplication, date culling, and key word filtering, the anticipated volume would still be unwieldy

[8] *See* JAMES DERTOUZOS ET. AL, RAND INST. FOR CIVIL JUSTICE, THE LEGAL AND ECONOMIC IMPLICATIONS OF E-DISCOVERY: OPTIONS FOR FUTURE RESEARCH ix (2008), *available at* http://www.rand.org/content/dam/rand/pubs/occasional_papers/2008/RAND_OP183.pdf; *see also* Robert Blumberg & Shaku Atre, *The Problem with Unstructured Data*, INFO. MGMT. (Feb. 1, 2003, 1:00 AM), http://soquelgroup.com/Articles/dmreview_0203_problem.pdf; THE RADICATI GROUP, TAMING THE GROWTH OF EMAIL: AN ROI ANALYSIS 3-4 (2005), *available at* http://www.radicati.com/wp/wp-content/uploads/2008/09/hp_whitepaper.pdf

when even a 90% reduction in volume would require review of 50 million items. Moreover, studies have long demonstrated that human reviewers are often quite inconsistent with respect to identifying "relevant" information, even when assisted by key word searches.[9]

[8] Current scholarship also shows that human reviewers do not consistently apply the concept of relevance and that the overlap, or the measure of the percentage of agreement on the relevancy of a particular document between reviewers, can be extremely low.[10] Counter-intuitively, the result is the same even when more "senior" review attorneys set the "gold standard" for determining relevance.[11] Recent studies comparing technology-assisted processes with traditional human review conclude that the former can and will yield better results. Technology can improve both recall (the percentage of the total number of relevant documents in the general population that are retrieved through search) and precision (percentage of retrieved documents that are, in fact, relevant) than humans can achieve using traditional methods.[12]

---

[9] *See* David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval* System, COMM. ACM, March 1985, at 289-90, 295-96.

[10] *See* Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697, 701 (2000), *available at* http://www.cs.cornell.edu/courses/cs430/2006fa/cache/Trec_8.pdf (finding that relevance is not a consistently applied concept between independent reviewers). *See generally* Hebert 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, 77 (2010).

[11] *See* Voorhees, *supra* note 10, at 701 (finding that the "overlap" between even senior reviewers shows that they disagree as often as they agree on relevance).

[12] *See generally* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11 ¶ 2 (2011), http://jolt.richmond.edu/v17i3/article11.pdf (analyzing data from the TREC 2009 Legal Track Interactive Task Initiative).

[9] There is also growing judicial acceptance of parties' use of technology to help reduce the substantial burdens and costs associated with identifying, collecting, and reviewing ESI. Recently, the U.S. District Court for the Southern District of New York affirmed Magistrate Judge Andrew Peck's order approving the parties' agreement to use "predictive coding," a method of using specialized software to identify potentially relevant information.[13]

[10] Likewise, a Loudon County, Virginia Circuit Court judge recently granted a defendant's motion for protective order allowing the use of predictive coding for document review.[14] The defendant had a data population of 250 GB of reviewable ESI comprising as many as two million documents, which, it argued, would require 20,000 man-hours to review using traditional human review.[15] The defendant explained that traditional methods of linear human review likely "misses on average 40% of the relevant documents, and the documents pulled by human reviewers are nearly 70% irrelevant."[16]

[11] Similarly, commentary included with recent revisions to Rule 502 of the Federal Rules of Evidence indicate that using computer-assisted tools may demonstrate reasonableness in the context of privilege review: "Depending on the circumstances, a party that uses advanced analytical software applications and linguistic tools in screening for privilege may be

[13] *See* Moore v. Publicis Groupe SA, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534, at *1-3 (S.D.N.Y. Apr. 26, 2012).

[14] *See* Global Aerospace, Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 Va. Cir. LEXIS 50, at *2 (Va. Cir. Ct. Apr. 23, 2012).

[15] *See* Mem. in Supp. of Mot. for Protective Order Approving the Use of Predictive Coding at 4-5, Global Aerospace, Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 Va. Cir. LEXIS 50 (Va. Cir. Ct. Apr. 9, 2012).

[16] *Id.* at 6-7.

found to have taken 'reasonable steps' to prevent inadvertent disclosure."[17]

[12] Simply put, dealing with the volume of information in most business information systems is beyond what would be humanly possible without leveraging technology. Because such systems contain hundreds of millions of records, companies effectively have three choices for searching for data subject to a preservation obligation: they can rely on the search capabilities of the application or native operating system, they can invest in and employ third-party technology to index and search the data in its native environment, or they can export all of the data to a third-party application for processing and analysis.

III. THE SOLUTION: DEFENSIBLE DATA REMEDIATION

[13] Simply adding storage and retaining the ever-increasing volume of information is not a tenable option for businesses given the cost and risk involved. However, there are risks associated with data disposition as well, specifically that information necessary to the business or required for legal or regulatory reasons will be destroyed. Thus, the first stage of a defensible data remediation program requires an understanding of the business and legal retention requirements applicable to the data in question. Once these are understood, it is possible to construct a remediation framework appropriate to the repository that reflects those requirements.

A. Retention and Preservation Requirements

[14] The U.S. Supreme Court has recognized that "'[d]ocument retention policies,' which are created in part to keep certain information from getting into the hands of others, including the Government, are common in business."[18] The Court noted that compliance with a valid

[17] FED. R. EVID. 502(b) Advisory Committee's Notes, Subdivision (b) (2007).

[18] Arthur Anderson LLP v. United States, 544 U.S. 696, 704 (2005).

document retention policy is not wrongful under ordinary circumstances.[19] Document retention policies are intended to facilitate retention of information that companies need for ongoing or historical business purposes, or as mandated by some regulatory or similar legal requirement. Before attempting remediation of a data repository, the company must first understand and document the applicable retention and preservation requirements.

[15] It is beyond the scope of this Article to outline all of the potential business and regulatory retention requirements.[20] Ideally, these would be reflected in the company's record retention schedules. However, even when a company does not have current, up-to-date retention schedules, embarking on a data remediation exercise affords the opportunity to develop or update such schedules in the context of a specific data repository. Most data repositories contain limited types of data. For example, an order processing system would not contain engineering documents. Thus, a company is generally focused on a limited number of retention requirements for any given repository. There are exceptions to this rule, such as with e-mail systems, shared-use repositories (e.g., Microsoft SharePoint), and shared network drives. Even then, focusing on

[19] *Id.*; *see* Managed Care Solutions, Inc. v. Essent Healthcare, 736 F. Supp. 2d 1317, 1326 (S.D. Fla. 2010) (rejecting plaintiffs' argument that a company policy that e-mail data be deleted after 13 months was unreasonable) (citing Wilson v. Wal-Mart Stores, Inc., No. 5:07-cv-394-Oc-10GRJ, 2008 WL 4642596, at *2 (M.D. Fla. Oct. 17, 2008); Floeter v. City of Orlando, No. 6:05-CV-400-Orl-22KRS, 2007 WL 486633, at *7 (M.D. Fla. Feb. 9, 2007)). *But see* Day v. LSI Corp., No. CIV 11–186–TUC–CKJ, 2012 WL 6674434, at *16 (D. Ariz. Dec. 20, 2012) (finding evidence of defendant's failure to follow its own document policy was a factor in entering default judgment sanction for spoliation).

[20] For purposes of this article, such laws and regulations are treated as retention requirements with which a business must comply in the ordinary course of business. This article focuses on the requirement to exempt records from "ordinary course" retention requirements due to a duty to preserve the records when litigation is reasonably anticipated. In short, this article relies on the distinction between *retention* of information and *preservation* of information, focusing on the latter. *See infra* text accompanying note 23.

the specific repository will enable the company to likewise focus on some limited subset of its overall record retention requirements. Once a company has identified the business and regulatory retention requirements applicable to a given data repository, information in the repository that is not subject to those requirements is eligible for deletion unless it is subject to the duty to preserve evidence.

[16] The modern duty to preserve derives from the common law duty to preserve evidence and is not explicitly addressed in the Federal Rules of Civil Procedure.[21] The duty does not arise until litigation is "reasonably anticipated."[22] Litigation is "reasonably anticipated" when a party "knows" or "should have known" that the evidence may be relevant to current or future litigation.[23] Once litigation is reasonably anticipated, a company should establish and follow a reasonable preservation plan.[24] Although there are no specific court-sanctioned processes for complying with the preservation duty, courts generally measure the parties' conduct in a given case against the standards of reasonableness and good faith.[25]

---

[21] *See* Sylvestri v. Gen. Motors, Inc., 271 F.3d 583, 590 (4th Cir. 2001); *see also* Stanley, Inc. v. Creative Pipe, Inc., 269 F.R.D. 497, 519 (4th Cir. 2010).

[22] *See* Cache la Poudre Feeds v. Land O'Lakes, 244 F.R.D. 614, 621, 623 (D. Colo. 2007); *see also* THE SEDONA PRINCIPLES, *supra* note 1, at 14.

[23] *See* Pension Comm. of the Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC, 685 F. Supp. 2d 456, 466 (S.D.N.Y. Jan. 15, 2010 *as amended* May 28, 2010); Rimkus Consulting Grp., Inc. v. Cammarata, 688 F. Supp. 2d 598, 612-13 (S.D. Tex. 2010); Zubulake v. UBS Warburg LLC, 220 F.R.D. 212, 216 (S.D.N.Y. 2003) (*Zubulake IV*); *see also* The Sedona Conference, *Commentary on Legal Holds: The Trigger & The Process*, 11 SEDONA CONF. J. 265, 269 (2010) [hereinafter "*Commentary on Legal Holds*"].

[24] *Commentary on Legal Holds*, *supra* note 23, at 269 ("Adopting and consistently following a policy or practice governing an organization's preservation obligations are factors that may demonstrate reasonableness and good faith."); *see* THE SEDONA PRINCIPLES, *supra* note 1, at 12.

[25] *Commentary on Legal Holds*, *supra* note 23, at 270 (evaluating an organization's preservation decisions should be based on good faith and reasonable evaluation of relevant facts and circumstances).

In this context, a "defined policy and memorialized evidence of compliance should provide strong support if the organization is called up on to prove the reasonableness of the decision-making process."[26]

[17] The duty to preserve is not without limits: "[e]lectronic discovery burdens should be proportional to the amount in controversy and the nature of the case" so the high cost of electronic discovery does not "overwhelm the ability to resolve disputes fairly in litigation."[27] Moreover, courts do not equate reasonableness with "perfection."[28] Nor does the law require a party to take "extraordinary" measures to preserve "every e-mail" even if it is technically feasible to do so.[29] "Rather, in accordance with existing records and information management principles, it is more rational to establish a procedure by which selected items of value can be identified and maintained as necessary to meet the organization's legal and business needs[.]"[30]

[18] Critical tasks in a preservation plan are the identification and documentation of key custodians and other sources of potentially relevant information.[31] Custodians identified as having potentially relevant information should generally receive a written litigation hold notice.[32]

---

[26] *Id.* at 274.

[27] *Rimkus Consulting*, 688 F. Supp. 2d at 613 n.8 (quoting THE SEDONA PRINCIPLES, *supra* note 1, at 17); *see also* Stanley v. Creative Pipe, Inc., 269 F.R.D. 497, 523 (D. Md., 2010); *Commentary on Legal Holds*, *supra* note 23, at 270.

[28] *Pension Comm.*, 685 F. Supp. 2d at 461 ("Courts cannot and do not expect that any party can meet a standard of perfection.").

[29] *See* THE SEDONA PRINCIPLES, *supra* note 1, at 28, 30 (citing Concord Boat Corp. v. Brunswick Corp., No. LR-C-95-781, 1997 WL 33352759, at *4 (E.D. Ark. Aug. 29, 1997)).

[30] THE SEDONA PRINCIPLES, *supra* note 1, at 15.

[31] *See Commentary on Legal Holds*, *supra* note 23, at 270; *id.* at 28.

[32] *See Pension Comm.* 685 F. Supp. 2d at 465; s*ee also Commentary on Legal Holds*,

The notice should be sent by someone occupying a position of authority within the organization to increase the likelihood of compliance.[33] The Sedona Guidelines also suggest that a hold notice is most effective when it:

1) Identifies the persons likely to have relevant information and communicates a preservation notice to those persons;
2) Communicates the preservation notice in a manner that ensures the recipients will receive actual, comprehensible and effective notice of the requirement to preserve information;
3) Is in written form;
4) Clearly defines what information is to be preserved and how the preservation is to be undertaken; and
5) Is regularly reviewed and reissued in either its original form or an amended form when necessary.[34]

[19] The legal hold should also include a mechanism for confirming that recipients received and understood the notice, for following up with custodians who do not acknowledge receipt, and for escalating the issue until it is resolved.[35] To be effective, the legal hold should be periodically reissued to remind custodians of their obligation and to apprise them of changes required by the facts and circumstances in the litigation.[36]

[20] Experience has also shown that legal holds that are not properly managed and ultimately released are less likely to receive the appropriate level of attention by employees. Thus, the legal hold process should also include a means for determining when litigation is no longer reasonably

supra note 23, at 270.

[33] THE SEDONA PRINCIPLES, *supra*, note 1, at 32.

[34] *Commentary on Legal Holds*, *supra* note 23, at 270.

[35] *Id.* at 283-85.

[36] *See id.* at 285.

anticipated and the hold can be released, while ensuring that information relevant to another active matter is preserved.[37]

B. The Remediation Framework

[21] Against this backdrop, it is possible to outline a framework for data remediation that is compliant with legal preservation requirements. The following describes a high-level data remediation process that can be applied to virtually any data environment and any risk tolerance profile. The general process is described in Figure 1 below:

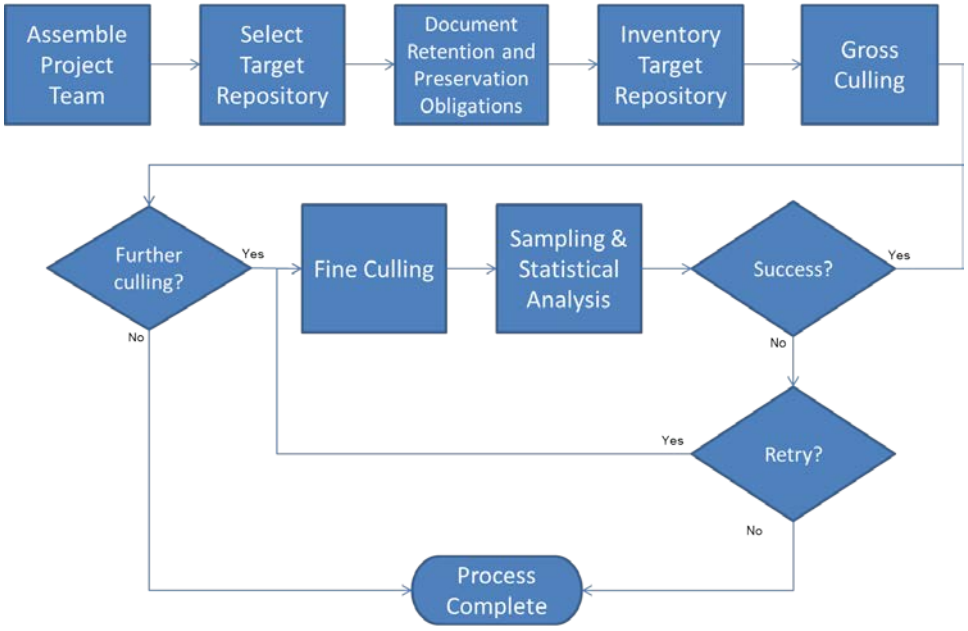

Figure 1: Data Remediation Framework

[37] *Id.* at 287.

1. Assemble the Team

[22] A successful data remediation project depends on invested participation by at least three constituents in the organization: legal, information technology ("IT"), and records and information management ("RIM"). In addition, the project may require additional support from experts experienced in information search and retrieval and statistical analysis. In-house and/or outside counsel provides legal oversight and risk assessment for the project team, as well as guidance on legal preservation obligations. IT provides the technological expertise necessary to understand the structure and capabilities of the target data repository. RIM professionals provide guidance on business and regulatory retention obligations. The need for information search and retrieval experts and statisticians depends on the complexity of the data remediation effort as described below. Finally, including business users of the information may be necessary as required to fully document retention requirements applicable to a particular repository if not adequately documented in the organization's document retention policy and schedule.

2. Select Target Data Repository

[23] Selecting the target data repository requires consideration of the costs and benefits of the data remediation exercise. Each type of repository presents unique opportunities and challenges. For example, e-mail systems, whether traditional or archived, are notorious for containing vast amounts of information that is not needed for any business or legal purpose. Similarly, shared network drives tend to contain large volumes of unused and unneeded information. Backup tapes, legacy systems, and even structured databases are other possible targets. IT and RIM resources are invaluable in identifying a suitable target repository. For example, IT can often run reports identifying directories and files that have not been accessed recently.

3. Document Retention and Preservation Obligations

[24] As discussed above, it is critical to understand the retention and preservation obligations that are applicable to the data contained in the target repository. Retention obligations include the business information needs as well as any regulatory requirements mandating the preservation of data. Ideally, these are incorporated into the document retention policy and schedule for the organization. If not, it will be important to document those requirements applicable to the target repository.

[25] Preservation obligations are driven by existing and reasonably anticipated litigation.[38] In some cases this may be the most challenging part of the project, particularly for highly litigious companies, because, unlike business needs and regulatory requirements, preservation obligations are constantly changing as new matters arise and circumstances evolve in existing matters. Successful completion of the remediation project will require a detailed understanding of, and constant attention to, the preservation obligations applicable to the target repository. As discussed below, some of the risk associated with this aspect of the project can be ameliorated through selection of the appropriate repository and culling criteria. Nevertheless, the scope and timing of the project will be driven in large part by the preservation obligations applicable to the target repository.

4. Inventory Target Data Repository

[26] After selecting the target data repository, the team must inventory the information within that repository. This does not involve creating an exhaustive list or catalog of every item within the repository. Rather, inventorying the repository involves developing a good understanding of the types of information that are contained there, the date ranges of the information, and other criteria that will enable identifying information that must be retained and that which can be deleted. The details of the inventory will vary by data repository. For example, for an e-mail server,

[38] *See supra* ¶ 16.

the pertinent criteria may include only date ranges and custodians, whereas for a shared network drive, the pertinent criteria may include departments and individuals with access, date ranges, and file types.

5. Gross Culling

[27] The next step is to determine the "gross culling" criteria for the data repository. In this context, "gross culling" refers to an initial phase of data culling based on broad criteria as opposed to fine or detailed culling criteria that may be used in a later phase of the exercise.[39] The nature of the information contained within the repository will determine the specific criteria to be used, but the objective is to locate the "low-hanging fruit," the items within the repository that can be readily identified as not falling within any retention or preservation obligation. These are black-and-white decisions where the remediation team can definitively determine without further analysis that the items identified can be deleted.

[28] For example, in most cases, dates are effective gross culling criteria. Quite often, large volumes of e-mail and loose files (data retained in shared network drives or other unstructured storage) predate any existing retention or preservation obligation for such items. Similarly, in repositories that are subject to short or no retention guidelines, the business need for the data can be evaluated in terms of the date last accessed. In the case of shared network drives, for example, it is not uncommon to find large volumes of information that has not been accessed by any user in many years.[40] Such information can be disposed of with very little risk.

[39] *See* Alex Vorro, *How to Reduce Worthless Data*, INSIDECOUNSEL (Mar. 1, 2012), http://www.insidecounsel.com/2012/03/01/how-to-reduce-worthless-data?t=technology.

[40] *See, e.g.*, Anne Kershaw, *Hoarding Data Wastes Money*, BASELINE (Apr. 16, 2012), http://www.baselinemag.com/storage/Hoarding-Data-Wastes-Money/ (80% of the data on shared network and local hard drives has not been accessed in three to five years).

6. Fine Culling

[29] Sometimes, the process need go no further than the gross culling stage. Depending on the volume of data deleted and the volume and nature of the data remaining, the remediation team may determine that the cost and benefit of attempting further culling of the data are not worth the effort and risk. In some cases, however, gross culling techniques will not identify sufficient volumes of unneeded data and more sophisticated culling strategies must be employed.

[30] The precise culling technique and strategy will depend on the specific data repository, its native search capabilities, and the availability of other search tools. For example, many modern e-mail archiving systems have fairly sophisticated native search capabilities that can locate with a high degree of accuracy content pertinent to selected criteria. Other systems will require the use of third-party technology. In either case, the fine culling process will require selection of culling criteria that will uniquely identify items not subject to a retention or preservation obligation and be susceptible to verification. Depending on the nature of the data and the complexity of the necessary search criteria, the remediation team may need to engage an expert in information search and retrieval.

7. Sampling and Statistical Analysis

[31] Regardless of the specific fine culling strategy employed, the remediation team should validate the results by sampling and analysis to ensure defensibility. Generally, it will be advisable to engage a statistician to direct the sampling effort and perform the analysis because both can be quite complex and rife with opportunity for error.[41] Moreover, in the

[41] Statistical sampling results can be as valid using a small random sample size as they are for using a larger sample size because, in a simple random sample of any given size, all items are given an equal probability of being selected for the statistical assessment. In fact, to achieve a confidence interval of 95% with a margin of error of 5%, a sample size of 384 would be sufficient for the population of 300 million. *See Sample Size Table*, RESEARCH ADVISORS, http://research-advisors.com/tools/SampleSize.htm (last visited on Jan. 12, 2014) (citing Robert V. Krejcie & Daryle W. Morgan, *Determining Sample Size for Research Activities, Educational and Psychological Measurement* 30 EDUC. &

event that the company's process is ever challenged, validation by an independent expert is compelling evidence of good faith. It is important to realize that the statistical analysis cannot demonstrate that no items subject to a preservation obligation are included in the data to be destroyed. It can only identify the probability that this is the case, but it can do so with remarkable precision when properly performed.[42]

8. Iteration

[32] Fine culling and validation should continue until the remediation team achieves results that meet its expectations regarding the volume of data identified for deletion and the probability that only data not subject to a preservation obligation are included in the result set.

IV. CONCLUSION

[33] The enormity of the challenge that expanding volumes of unneeded information creates for businesses is difficult to understate. Companies literally spend millions of dollars annually to store and maintain information that serves no useful purpose, funds that could be directed to productive uses such as hiring, research, and investment. Facing this challenge, on the other hand, is a challenge of its own, perhaps due more to the fear of adverse consequences in litigation than any other factor. It is possible, however, to develop a defensible data remediation process that enables a company to demonstrate good faith and reasonableness while eliminating the cost, waste, and risk of this unnecessary data.

PSYCHOL. MEASUREMENT 607, 607-610 (1970). However, samples can be vulnerable to discrete "sampling error" because the randomness of the selection may result in a sample that does not reflect the makeup of the overall population. For instance, a simple random sample of messages will on average produce five with attachments and five with no attachments, but any given test may over-represent one message type (e.g., those with attachments) and under-represent the other (e.g., those without).

[42] *See, e.g.*, *Statistics*, WIKIPEDIA, http://en.wikipedia.org/wiki/Statistics (last visited on Feb. 9, 2014).