This product is part of the RAND Corporation monograph series. RAND monographs present major research findings that address the challenges facing the public and private sectors. All RAND monographs undergo rigorous peer review to ensure high standards for research quality and objectivity.

Where the Money Goes

Understanding Litigant Expenditures for Producing Electronic Discovery

Nicholas M. Pace, Laura Zakaras

Institute for Civil Justice

A RAND LAW, BUSINESS, AND REGULATION INSTITUTE

This research was conducted by the RAND Institute for Civil Justice, a research institute within RAND Law, Business, and Regulation, a division of the RAND Corporation.

Library of Congress Cataloging-in-Publication Data

Pace, Nicholas M. (Nicholas Michael), 1955-
Where the money goes : understanding litigant expenditures for producing electronic discovery / Nicholas M. Pace, Laura Zakaras.
p. cm.
Includes bibliographical references.
ISBN 978-0-8330-6876-7 (pbk. : alk. paper)
1. Electronic discovery (Law) I. Zakaras, Laura. II. Title.-

K2247.P33 2012
347.73'57—dc23

2012011130

Published 2012 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Summary

Pretrial discovery procedures are designed to encourage an exchange of information that will help narrow the issues being litigated, eliminate surprise at trial, and achieve substantial justice. But, in recent years, claims have been made that the societal shift from paper documents to electronically stored information (ESI) has led to sharper increases in discovery costs than in the overall cost of litigation.

In response, the Federal Rules of Civil Procedure have been amended several times in the past five years, and most states have adopted or amended rules of procedure or evidence to address a range of challenges posed by e-discovery. This evolution in the rules is ongoing: The federal Advisory Committee on Civil Rules is currently exploring issues related to the costs of discovery and may well be on track to propose further amendments to the federal civil rules. Few other issues about the civil justice system in recent years have so focused the attention of policymakers and stakeholders.

Study Purpose and Approach

We hope this monograph will help inform the debate by addressing the following research questions:

- What are the costs associated with different phases of e-discovery production?
- How are these costs distributed across internal and external sources of labor, resources, and services?
- How can these costs be reduced without compromising the quality of the discovery process?
- What do litigants perceive to be the key challenges of preserving electronic information?

We chose a case-study method that identified eight very large companies that were willing, with our assurances of confidentiality, to provide in-depth information about e-discovery production expenses. The companies consisted of one each from the communications, electronics, energy, household care products, and insurance fields, and three from the pharmaceutical/biotechnology/medical device field. We asked participants to choose a minimum of five cases in which they produced data and electronic documents to another party as part of an e-discovery request. In the end, we received at least some reliable e-discovery production cost data for 57 cases, including traditional lawsuits and regulatory investigations.

We also collected information from extensive interviews with key legal personnel from these companies. Our interviews focused on how each company responds to new requests for

e-discovery, what steps it takes in anticipation of those requests, the nature and size of the company's information technology (IT) infrastructure, its document-retention policies and disaster-recovery and archiving practices, its litigation pressure and the types of cases in which it is involved, and what it finds to be the key challenges in this evolving e-discovery environment.

Our analysis is also informed by an extensive review of the legal and technical literature on e-discovery, with emphasis on the intersection of information-retrieval science and the law. We supplemented our data collection with additional interviews with representatives of participating companies, focusing on issues related to the preservation of information in anticipation of discovery demands in current or potential litigation.

Because the participating companies and cases do not constitute a representative sample of corporations and litigation, we cannot draw generalizations from our findings that apply to all corporate litigants or all discovery productions. However, the case-study approach provides a richly detailed account of the resources required by a diverse set of very large companies operating in different industries to comply with what they described as typical e-discovery requests. In what follows, we highlight our key findings.

Costs of Producing Electronic Documents

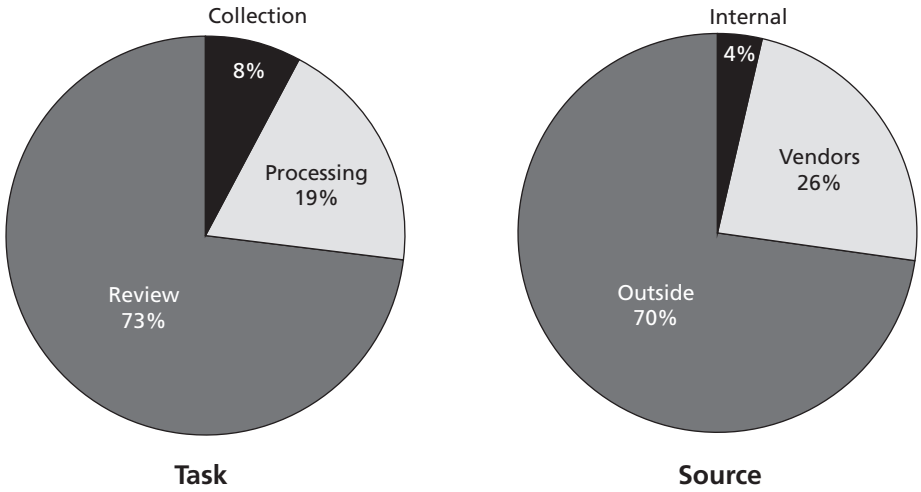
We organized the cost data we received into three tasks:

- *Collection* consists of locating potential sources of ESI following the receipt of a demand to produce electronic documents and data, and gathering ESI for further use in the e-discovery process, such as processing or review.
- *Processing* is reducing the volume of collected ESI through automated processing techniques, modifying it if necessary to forms more suitable for review, analysis, and other tasks.
- *Review* is evaluating digital information to identify relevant and responsive documents to produce, and privileged documents or confidential or sensitive information to withhold.

There were, of course, some gaps in the data. But the data were sufficiently complete to provide interesting insights about relative costs and level of effort across tasks. Figure S.1, for example, shows that the major cost component in our cases was the review of documents for relevance, responsiveness, and privilege (typically about 73 percent). Collection, an area on which policymakers have focused intensely in the past, consumed about 8 percent of expenditures for the cases in our study, while processing costs consumed about 19 percent in typical cases.

We also examined the costs of collection, processing, and review in terms of their sources: *internal*, such as law department counsel and IT department staff; *vendors*; and *outside counsel*. As might be expected because of their historical role in the review process, expenditures for the services of outside counsel consumed about 70 percent of total e-discovery production costs. Internal expenditures, even with adjustments made for underreporting, were generally around 4 percent of the total, while vendor expenditures were around 26 percent (Figure S.1). As Table S.1 shows, vendors played the dominant role in collection and processing, while review was largely the domain of outside counsel. The zero counts for internal processing and review do not mean that corporate resources were not consumed for these tasks, only that none of the

Figure S.1
Relative Costs of Producing Electronic Documents



NOTE: Values reflect median percentages for cases with complete data, adjusted to 100 percent.

RAND *MG1208-S.1*

Table S.1
Case Counts by the Primary Source of Expenditures for E-Discovery Tasks

Task	Internal	Vendor	Outside Counsel	Total Cases Reporting
Collection	6	31	5	42
Processing	0	42	2	44
Review	0	4	45	49

cases reporting complete information had internal expenditures for such activities that were greater than those for external entities, such as vendors or outside counsel.

The task breakdown in the table, however, appears likely to change in the future. Most of the companies whose representatives we interviewed expressed a commitment to taking on more e-discovery tasks themselves and outsourcing those that could be "commoditized." Collection is a good example of this trend. Two of the eight companies were in the process of implementing an automated, cross-network collection tool in order to perform such services without the need for outside vendors, and others were anticipating moving in that direction. Although we found little evidence that the review process was moving in-house, the legal departments in the companies from which we interviewed representatives were taking greater control over at least the "first-pass" review to confirm relevance and responsiveness of documents, choosing vendors and specialized legal service law firms to perform such functions that were formerly delegated to outside counsel.

Reducing the Cost of Review

With more than half of our cases reporting that review consumed at least 70 percent of the total costs of document production, this single area is an obvious target for reducing e-discovery expenditures. We believe that many stakeholder complaints would diminish if expenditures for review were no more burdensome than those for either the collection or processing phase. Because review consumes about $0.73 of every dollar spent on ESI production, while collection and processing consume about $0.08 and $0.19, respectively, *review costs would have to be reduced by about three-quarters* in order to make those costs comparable to processing, the next most costly component of production. Choosing a 75-percent reduction in review expenditures as the desired target is an admittedly arbitrary decision, but more-modest cost savings are not likely to end criticisms from some quarters that the advent of e-discovery has caused an unacceptable increase in the costs of resolving large-scale disputes. To explore possible ways of achieving this target, we synthesized the methods that research on this topic has identified as promising for cutting review costs, both for the traditional approach of an "eyes-on" review of each document and for moving to a new paradigm that relies on computer-categorized review technology to examine documents for relevance, responsiveness, or privilege. We also summarize the literature on the relative quality of traditional review practices and computerized approaches to assess whether moving away from human review would compromise the quality of the process.

Significant Reduction in Current Labor Costs Is Unlikely

Companies are trying a variety of alternatives to the traditional use of outside law firms for most review tasks. In order to reduce the cost of review-related labor, they may hire temporary attorneys or use legal process outsourcing (LPO) companies with stables of contract attorneys. However, the rates currently paid to such project attorneys during large-scale reviews in the United States may well have bottomed out, with further reductions of any significant size unlikely. Another option that has been explored is the use of English-speaking local lawyers in such countries as India and the Philippines. Although such foreign outsourcing uses local attorneys who will work for much less than U.S. counsel, issues related to information security, oversight, maintaining attorney-client privilege, and logistics may limit the utility of offshore approaches for most litigation.

Increasing the Rate of Review Has Its Limits

The most-expansive claims regarding review speed is about 100 documents per hour, and this number assumes that reviewers have the strongest motivations and experience and are examining documents simple enough that a decision on relevance, responsiveness, privilege, or confidential information could be made in an average of 36 seconds. A trained "speed reader" can skim written materials at roughly 1,000 words per minute with about 50-percent comprehension. Therefore, even allocating zero time for bringing up a new document on the screen and zero time for contemplating a decision or the act of clicking the appropriate button to indicate a choice, a maximum of 600 words (about a page and a half) can be read in 36 seconds. Given the trade-off between reading speed and comprehension, especially in light of the complexity of documents subject to discovery in large-scale litigation, it is unrealistic to expect much room for improvement in the rates of unassisted human review.

Techniques for Grouping Documents Are Not the Answer

We describe three techniques that are increasingly used to organize documents—and, in some cases, "bulk-code" like documents—to streamline the review process:

- *Near-duplicate detection* groups together documents that contain mostly identical blocks of text or other information but that nevertheless differ in some minor way (any truly duplicate documents should have been removed during the processing phase).
- *Clustering* identifies the keywords and concepts in each document then groups documents by the degree to which they share keywords or concepts so that documents can be organized by topic rather than in random order to streamline the review.
- *Email threading* groups individual emails into single "conversations," sorting chronologically, and eliminating duplicate material.

These techniques organize material rather than reducing the number of documents in the review set. Commercial vendors of these services claim they can increase the rate of review to 200, 300, or even 500 documents per hour. However, given the physical limitations of reading and comprehension, better organization of the corpus of documents is not likely to account for such astonishing review rates unless decisions about individual documents can be applied to dozens or hundreds of similar items on a routine basis. Although some document sets may lend themselves to bulk coding in this manner, it is unlikely that these techniques would foster sufficiently dramatic improvements in review speed for most large-scale reviews.

Human Reviewers Are Highly Inconsistent

Just how accurate is the traditional approach in these days of computerized review tools flashing documents on screen before a first-year associate or contract lawyer at rates exceeding 50 documents per hour? Some rigorous studies addressing this issue found that human reviewers often disagree with one another when they review the same set of documents for relevance and responsiveness in large-scale reviews. In one study, for example, seven teams of attorneys, all trained in a similar manner and given the same instructions, examined 28,000 documents, clustered into 12,000 families involving similar topics, to judge whether the families were responsive to the facts of the case.[1] The seven teams differed significantly on the percentage of families determined to be responsive, ranging from a low of 23 percent to a high of 54 percent. As indicated by other studies discussed in this monograph, the high level of disagreement, corroborated by other studies discussed in the main text, is caused by human error in applying the criteria for inclusion, not a lack of clarity in the document's meaning or ambiguity in how the scope of the production demand should be interpreted.

Is Predictive Coding an Answer?

We believe that one way to achieve substantial savings in producing massive amounts of electronic information would be to let computers do the heavy lifting for review. Predictive coding is a type of computer-categorized review application that classifies documents according to how well they match the concepts and terms in sample documents. Such machine-learning techniques continually refine the computer's classifications with input from users, just as spam filters self-correct to increase the reliability of their future decisions about new email mes-

[1] Barnett and Godjevac, 2011.

sages, until the ambiguous ratings disappear. With predictive coding, humans (i.e., attorneys) initially examine samples of documents from the review set and make determinations about whether they are relevant, responsive, or privileged. Using those decisions, the software assigns scores to each document in the review set representing the probability that a document matches the desired characteristics. Additional samples of these new decisions are drawn and examined by the attorney reviewers, and the application refines the templates it uses to assign scores. The results of this iterative process are eventually stabilized. At that point, disagreement between the software's decisions and those of human reviewers should be minimized.

Because this is nascent technology, there is little research on how the accuracy of predictive coding compares with that of human review. The few studies that exist, however, generally suggest that predictive coding identifies at least as many documents of interest as traditional eyes-on review with about the same level of inconsistency, and there is some evidence to suggest that it can do better than that.

Not surprisingly, costs of predictive coding, even with the use of relatively experienced counsel for machine-learning tasks, are likely to be substantially lower than the costs of human review. It should be kept in mind that attorney review is still very much in play with predictive coding, but generally only for the smaller subset of documents that the application has judged to be potentially relevant, responsive, or privileged.[2] Because there is scant research on the issue, it is too early to confidently estimate the magnitude of any savings. Evidence, however, suggests the reduction in person-hours required to review a large-scale document production could be considerable. One study, for example, which did not report on cost savings but did report time savings, suggested that predictive coding of a document set previously reviewed in the traditional way would have saved about 80 percent in attorney review hours.[3] Although this estimate did not include the costs of the vendor's services, and the potential reduction in hours would be strongly influenced by the threshold probability scores used for determining potential matches, the savings are still likely to be considerable and meet the goal we set of a three-quarter reduction in review expenditures.

Barriers to the Use of Computer-Categorized Document Review

With such potential to reduce the costs of review without compromising quality, why is it that predictive coding and other computer-categorized document review techniques are not being embraced by litigants? None of the companies in our sample was using predictive coding for review purposes; at the end of 2011, we could find no evidence in the published record that any vendor, law firm, or litigant had used predictive coding in a publicized case that named the parties and court jurisdiction.

Some concerns are likely to pose barriers to the use of predictive coding, including whether it performs well in any of the following:

- identifying *all* potentially responsive documents while avoiding *any* overproduction

---

[2] For example, one potential approach to computer-categorized document review would have the application identify documents likely to be relevant and responsive and then have attorneys examine only the identified set to confirm the decisions and to determine whether those documents contain privileged communications or sensitive information.

[3] Equivio, 2009a.

- identifying privileged or confidential information
- flagging "smoking guns" and other crucial documents
- classifying highly technical documents
- reviewing relatively small document sets.

Another barrier to widespread use could well be resistance to the idea from outside counsel, who would stand to lose a historical revenue stream. Outside counsel may also be reluctant to expose their clients to the risks of adopting an evolving technology. But perhaps most important is the absence of judicial guidance on the matter. At the time we conducted this study, there were simply no judicial decisions that squarely approved or disapproved of the use of predictive coding or similar computer-categorized techniques for review purposes. It is also true that many attorneys would be uncomfortable with the idea of being an early adopter when the potential downside risks appear to be so large. Few lawyers would want to be placed in the uncomfortable position of having to argue that a predictive-coding strategy reflects reasonable precautions taken to prevent inadvertent disclosure, overproduction, or underproduction, especially when no one else seems to be using it.

We propose that the best way to overcome these barriers and bring predictive coding into the mainstream is for innovative, public-spirited litigants to take bold steps by using this technology for large-scale e-discovery efforts and to proclaim its use in an open and transparent manner. The motivation for conducting successful public demonstrations of this promising technology would be to win judicial approvals in a variety of jurisdictions, which, in turn, could lead to the routine use of various computer-categorized techniques in large-scale reviews along with long-term cost savings for the civil justice system as a whole. Without organizational litigants making a contribution in this manner, many millions of dollars in litigation expenditures will be wasted each year until legal tradition catches up with modern technology.

Challenges of Preservation

Some important generalizations emerged from our inquiry into what corporate counsel consider to be the main challenges of preserving electronic information in anticipation of litigation.

Companies Are Not Tracking the Costs of Preservation

Most interviewees did not hesitate to confess that their preservation costs had not been systematically tracked in any way and that they were unclear as to how such tracking might be accomplished, though collecting useful metrics was generally asserted as an important future goal for the company.

Preservation Expenditures Are Said to Be Significant

All interviewees reported that preservation had evolved into a significant portion of their companies' total e-discovery expenditures. Some of them believed that preserving information was now costing them more than producing e-discovery in the aggregate. The way in which organizations perceive the size of preservation expenditures relative to that of production appears to be related to steps taken (or not taken) to move away from ad hoc preservation strategies, the nature of their caseloads, and ongoing impacts on computing services and business practices.

There Are Complaints About the Absence of Clear Legal Authority

A key concern voiced by the interviewees was their uncertainty about what strategies are defensible ones for preservation duties. Determining the reasonable scope for a legal hold in terms of custodians, data locations, and volume was said to be a murky process at best, with strong incentives to overpreserve in the face of the risk for significant sanctions. Similar concerns were voiced about the process itself, with few concrete guideposts said to be available to provide litigants with a level of comfort when deciding not only what to preserve, but how.

The cause for such worries is the absence of controlling legal authority in this area. Although judicial decisions have addressed preservation scope and process, they act as legally binding precedent in only specific jurisdictions, or conflict with decisions rendered by other courts on the same issues. As a result, litigants reported that they were greatly concerned about not making defensible decisions involving preservation and about the looming potential of serious sanctions.

Recommendations

We propose three recommendations to address the complaints of excessive costs and uncertainty that emerged from our interviews.

Adopt Computer Categorization to Reduce the Costs of Review in Large-Scale E-Discovery Efforts

The increasing volume of digital records makes predictive coding and other computer-categorized review techniques not only a cost-effective option to help conduct review but the *only* reasonable way to handle large-scale production. Despite efforts to cull data as much as possible during processing, review sets in some cases may be impossible to examine thoroughly using humans, at least not in time frames that make sense during ongoing litigation. New court rules *might* move the process forward, but the best catalyst for more-widespread use of predictive coding would be well-publicized documentation of cases in which judges examined the results of actual computer-categorized reviews. It will be up to forward-thinking litigants to make that happen.

It should be noted that we believe that computer-categorized review techniques, such as predictive coding, have their greatest utility with production volumes that are at least as large as the cases in our sample.

Improve Tracking of Costs of Production and Preservation

There are many reasons to track discovery costs. Without such data, companies cannot develop strategies for dealing with massive data volumes, such as investing in automated legal-hold–compliance systems or advanced analytic software for early case assessment. A litigant also needs to be able to present a credible argument to a judge that a proposed discovery plan or request will result in unreasonably large expenditures. Finally, the need for better records may be strongest in the context of preservation, in which the absence of publicly reported data in this area frustrates rule-making efforts intended to address litigant complaints.

Bring Certainty to Legal Authority Concerning Preservation

Steps must be taken soon to address litigant concerns about complying with preservation duties. The absence of clear, unambiguous, and transjurisdictional legal authority is thwarting thoughtful preservation efforts, potentially leading to overpreservation at considerable cost; and creating uncertainty about proper scope, defensible processes, and sanctionable behavior.