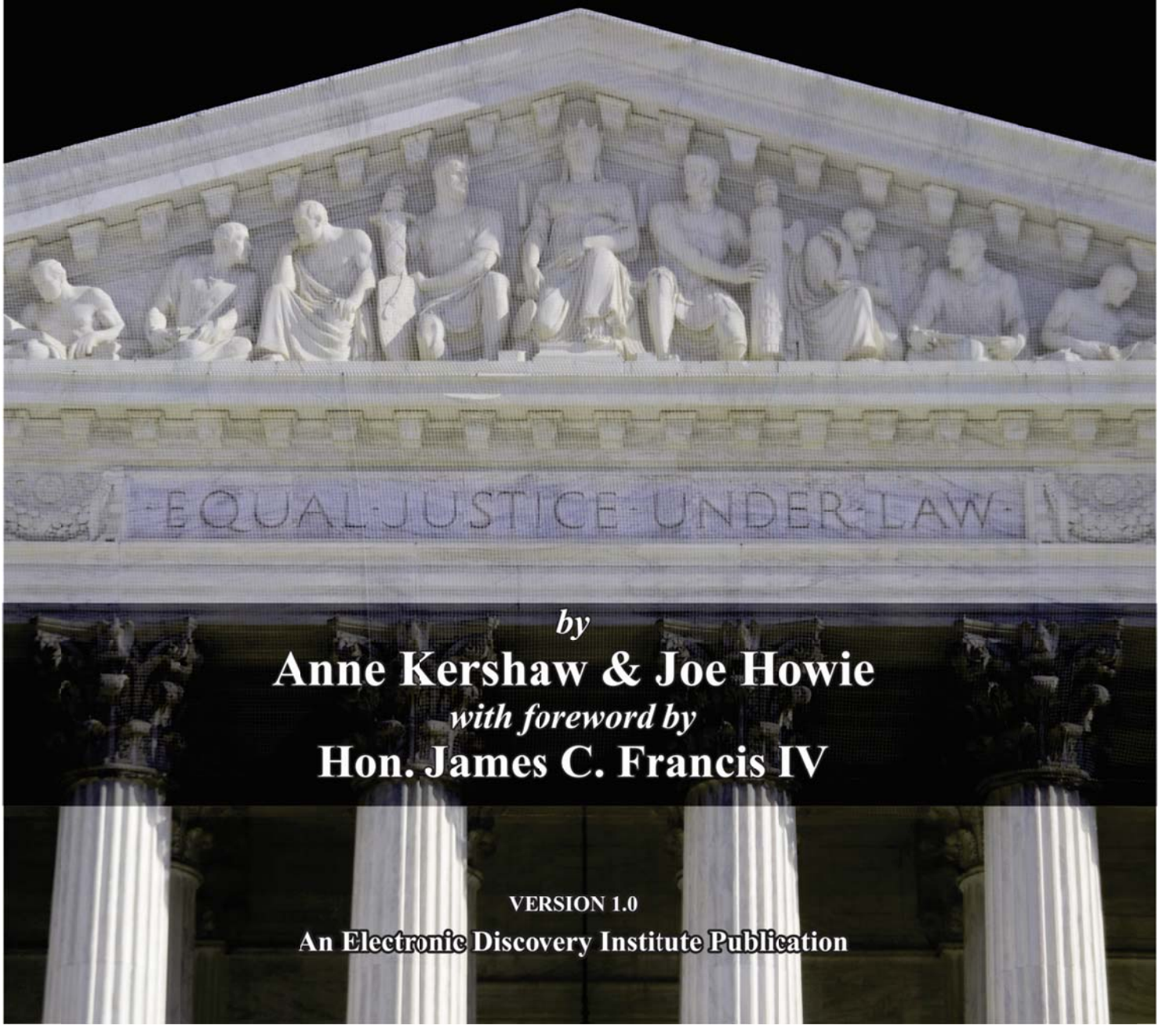
JUDGES' GUIDE TO COST-EFFECTIVE E-DISCOVERY



by

Anne Kershaw & Joe Howie

with foreword by

Hon. James C. Francis IV

VERSION 1.0

An Electronic Discovery Institute Publication

Judges' Guide to Cost-Effective E-Discovery

Foreword by Hon. James C. Francis IV, United States Magistrate Judge, Southern District of New York

The discovery of electronically stored information ("ESI") has become vital in most civil litigation—virtually all business information and much private information can be found only in ESI. At the same time, the costs of gathering, reviewing, and producing ESI have reached staggering proportions. This is caused in good part by the sheer volume of ESI but also by the reluctance or inability of some lawyers to adopt cost-effective strategies. This guide provides an overview of some of the basic processes and technologies that can reduce the costs of processing ESI. Courts may not want to adopt all of the recommendations contained here, but they are worth careful consideration.

For judges who are inclined to be involved directly in managing ESI issues, this guide provides information that can be shared with counsel to help curtail ever-escalating discovery costs. For judges who are more comfortable letting the parties manage the details of e-discovery, it will help separate fact from myth or fiction when lawyers advance conflicting arguments on electronic discovery. At the very least, lawyers for all the parties should be encouraged to be familiar with the principles contained in this guide.

James C. Francis IV
New York
October 1, 2010

Rule Number One:

> "… to secure the just, speedy, and inexpensive determination of every action and proceeding."
>
> FED. R. OF CIV. P. 1

Preface

The *Judges' Guide to Cost-Effective E-Discovery* is an outgrowth of the work by one of the authors (Anne Kershaw) in presenting sessions on e-discovery to workshops for U.S. magistrate judges. It is intended to be used as supporting material in such courses.

The purpose of the *Judges' Guide* is to inform judges of the lessons learned through the work of the eDiscovery Institute in researching cost-effective ways to manage ever-escalating discovery costs. Judges are, after all, the ultimate arbiters of what is reasonable or acceptable in processing electronic discovery. Management expert Peter Drucker has observed that you can't manage what you don't measure, and, where possible, we have tried to provide objective metrics to substantiate the recommendations made in the *Judges' Guide*. Objective data on processing alternatives should always be relevant when deciding issues of proportionality and burden regarding ESI.

The good news is that there are technologies and processes that can dramatically lower the cost of handling e-discovery. The even better news is that many of these technologies and processes not only lower costs, but actually improve several important measures of quality such as shortened time frames, improved confidentiality, and far greater transparency and replicability.

There is no single "correct" way to handle electronic discovery; reasonable people can reach differing conclusions based on the exigencies of each case. Because even individual lawyers may use different processes on different cases, there are probably more ways of gathering, processing, and producing ESI than there are lawyers. We have tried to present a reasonably representative overview of the process, recognizing that some parties may use different processes. However, everyone certainly must agree that controlling litigation costs is vital if we are to maintain America's competitiveness without having to abandon our great system of justice. For that, we are all responsible, and we hope this guide will raise the level of dialog on some of the ways litigation costs can be reduced to more tolerable levels.

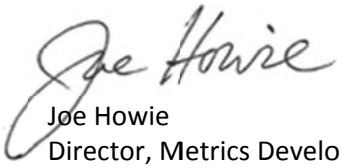
To the extent this guide is being read by counsel, our best advice is start discussions regarding the scope of preservation and collection as early as possible to determine what you're going to preserve and how you are going to collect responsive information. Then get a letter or e-mail off to opposing counsel as soon as you know their names so that any misunderstandings as to expectations can be surfaced as early as possible. In the world of ESI, trying to recover later from an earlier misunderstanding can be expensive and challenging—if not impossible, and waiting for the Rule 16 pretrial conference may well be too late.

If you disagree with the information contained in this guide, or if you would like to see other topics discussed in any future issues of the guide, please contact us by e-mailing info@ediscoveryinstitute.org.

Technology and the law are evolving constantly, and your input will help us keep the *Judges' Guide* current.

We look forward to hearing from you.

Anne Kershaw
Co-Founder and President
eDiscovery Institute



Joe Howie
Director, Metrics Development & Communications
eDiscovery Institute

Acknowledgements

We would like to acknowledge the following people who have contributed significantly to the preparation of this guide:

Hon. James C. Francis IV, United States Magistrate Judge, Southern District of New York, whose review of earlier drafts greatly improved the content and whose support encouraged us to bring the *Judges' Guide* to completion.

Herbert L. Roitblat, Ph.D., co-founder of the eDiscovery Institute and co-founder, Principal, Chief Scientist, and Chief Technology Officer of OrcaTec LLC., who offered helpful comments and advice on statistics and whose earlier writings on reviewer consistency were very helpful in preparing this guide.

Patrick Oot, co-founder and General Counsel of the eDiscovery Institute, who helped develop the outline for the *Judges' Guide* and provided helpful suggestions on content during its development.

Contents

Judges Guide to Cost-Effective E-Discovery

1. Introduction: The Battle for Cost-Effectiveness in E-Discovery

The exchange of electronically stored information ("ESI") in discovery has become essential in most cases because virtually all information now originates electronically. The first ESI discovery challenge, which has been ongoing for several years, has been to get lawyers to preserve and produce relevant ESI. The next challenge, which is just starting, is to get them to do it in a cost-effective manner so that the costs of discovery don't completely upset the scales of justice.

This guide discusses processes and technologies that will reduce discovery costs associated with reviewing and producing e-mail and electronic files, such as word processing documents, spreadsheets, and presentations. In this guide, we recommend specific technologies and processes that have been proven to lower the costs of electronic discovery—often while improving the quality of the results.

Where available, we provide estimates of the savings that could be achieved by the various technologies and processes. In the appendices, we include a technical competency quiz, which can be used for self-assessment.

We have tried not to re-plow old ground in this guide—we have not restated all the statistics in support of the notion that ESI has become burdensome, and we've assumed the reader is, for the most part, familiar with the basic terminology of electronic discovery. For readers who are interested in some of this background material, we recommend the following publications of The Sedona Conference® available at www.theSedonaConference.org:

- *The Sedona Conference® Glossary: E-Discovery & Digital Information Management (Third Edition)*
- *The Sedona Principles after the Federal Amendments, August, 2007*

This guide does not address imaging (copying) entire hard drives for forensic investigation, although we endorse Sedona Principle 5 that "it is unreasonable to expect parties to take every conceivable step to preserve all potentially relevant electronically stored information."

Similarly, while we touch on the use of databases by attorneys for housing and managing documents for attorney review and production to adversaries, we do not go into great detail on preserving and collecting ESI from databases. We would note that, apart from data that may have been archived or retired from a database, databases should not present cost issues when it comes to discovery – they are designed to efficiently organize and present information. The question is generally how best to share that information—a question that can be resolved with the execution of appropriate privilege non-waiver agreements, claw-back agreements, and the generation of reports or exporting subsets of data from databases.

2. Overview of the E-Discovery Process for E-mails and Electronic Documents[1]

One of the challenges of e-mail is that it can be found in many different places, e.g., on company "e-mail servers"[2] that run software designed to manage multiple e-mail accounts for company employees (e.g. Microsoft Exchange or IBM's Lotus Notes), on "web e-mail servers" such as Yahoo, GMail, MSN, or AOL that provide e-mail for virtually anyone, or even on a local computer, such as a desktop or laptop (i.e. a "client") that has a program installed that captures or synchronizes with an e-mail account on a server and houses copies of the e-mail on the local machine's hard drive so that the e-mail can be viewed when the local

[1] As noted above, this guide does not address forensic investigative discovery from hard drives or provide detail on discovery from databases.

[2] A "server" is a computer running a service or program for other computers, which are often called users or "clients."

machine is not connected to a server. Increasingly, e-mail may also be stored on cell phones or PDA's.

In addition, e-mails can be saved electronically or moved to areas outside the e-mail program and saved in other electronic formats, such as ".txt" or ".msg."

Electronic documents or files, such as Word or WordPerfect documents, PowerPoint presentations, or Excel spreadsheets, can be saved and located on servers, local machines, or on external drives that may be attached to a server or local machine for additional storage. In addition, electronic files exist as attachments to e-mail.

Both e-mail and electronic files consist of a number of electronic components. They have "text," which is the substantive content of the communication that is visible to normal users and "metadata," which are additional pieces of information about the file or communication that is supplied by the computer, the software, or a human. Metadata for e-mail, for example, will include the e-mail address and/or names of the senders and recipients, the subject line, the date and time, and information regarding the e-mail's Internet journey if it originated outside the organization. Metadata for electronic files may include its location on a computer drive (referred to as the "file path"), the size and type of the file, the author, and the creation and last modified dates.[3] In civil cases, the metadata associated with e-mail are important mostly because it allows for easier organization and management of the e-mail.

The process for collecting, reviewing, and producing e-mails and electronic documents for discovery in litigation usually involves: (1) finding the relevant e-mail and electronic files and copying them from their original electronic home in a way that preserves relevant metadata, (2) converting the e-mail and electronic files to a format and placing them where people outside the original computer system can review them without the ability to modify them (known also as "processing"), and (3) reviewing and then producing or turning over the data to the requesting party.

Step 1 – Finding and Collecting Relevant Information

Ideally, knowledgeable client personnel are available to identify and segregate relevant information with their attorney's assistance and instruction. This type of targeted collection is the most effective method for containing e-discovery costs because it limits the volume of ESI that is processed. Rarely, however, do attorneys have confidence that all people with information are available or that all relevant information has been identified. This uncertainty may prompt them to take all e-mail or files associated with a particular employee (known also as a "custodian") or department. It is this "wholesale" rather than targeted collection of data that results in massive volumes duplicative data and massive costs.

Collecting e-mail is often done by using the e-mail program, e.g., Microsoft Outlook or Lotus Notes, to make a copy of the e-mail account or folder. In the case of electronic files, collection is often accomplished by using a program such as WinZip or RoboCopy to put the files in an electronic envelope or folder to ensure preservation of the metadata and then moving the envelope to an external drive that can be delivered to another location.

A note about collecting source information: while on many levels, it is preferable to have employees of a party segregate relevant e-mails and electronic files (with guidance and instruction) and place them in "collection" folders to obtain a targeted collection, it is important to understand that, in so doing, you will not be collecting the original folder or "file path" information. This information would, however, be preserved in the data's original environment, assuming preservation orders are followed properly. Accordingly, it is important that this process be discussed in early discovery planning sessions and also made clear to the requesting party.

Step 2 – Processing

E-mail and electronic files are essentially just assemblies of electronic elements (the text and the metadata) that are pieced together and presented by the software in a way that allows users to read and manipulate the information presented. In other words, business users typically need the original software

[3] The "author" information for electronic documents is often carried over electronically by the computer from copied files and is often incorrect. For example, an accounting department employee may create and distribute an expense report template that will list that employee as the "author," even though many different employees may actually fill-in the template in the process of submitting expense statements.

application that created the file in order to read and manipulate the information.

For discovery purposes, however, lawyers want reviewers to be able to read the information but *not* be able to manipulate it; otherwise the integrity of the data is suspect, and its evidentiary value is lost. To do this, the electronic elements of the file are commonly "processed" or fed into a different software application that will present the data so that reviewers can read it, but not change it, thus maintaining the data's integrity. This process involves running a computer application that parses out and separates the electronic elements of the original dataand converts them to a common format. This process is often referred to as "normalizing the data." The common or normalized format can then be imported into software designed for reviewing documents that will present the documents to reviewers without allowing changes to them. Note that, at this point, decisions will need to be made as to which data elements will be imported because many are unnecessary.

Most standard processes include the rendering of a "TIFF" or "PDF"[4] image of the document to reside alongside the data elements in the review repository. TIFF or PDF versions of documents involve essentially taking a picture of the documents as they appeared in their original software application. Processing documents without taking a picture or image is often referred to as a "native" production. When working with native files, reviewers will either have the original applications or, more commonly, the review platform will have viewing software, such as Stellant, that can view—but not necessarily edit—files created by a wide variety of software applications. Other review platforms will convert the native files to an HTML representation, although that approach may result in a less than perfect representation of how the document looked originally.

Historically, the normalized data was put in a "load file," often located on an external hard drive, that was used to load the data into the review software. However, there are now some review applications that ingest and process data directly. Pricing for processing is usually by the gigabyte, ranging from $45 to $1,500 per gigabyte, depending on the provider and services included.

Note that not all electronic files can be processed. For example, some files are encrypted, password-protected, or corrupted, and it is important that the processing vendor provide an "exception report" that lists any e-mail or files that were not processed.

Most of the cost-saving measures discussed in this guide involve technologies used in the processing phase.

Step 3 – Review and Production

After the data are processed it is imported into a document review application or software, also referred to as a review platform, repository, or database. The document repository can be made available to multiple reviewers through a secure Internet connection using a URL or Domain Name, which lawyers can access using a web browser, such as Internet Explorer or Firefox. Alternatively, the repository can be made available through a server on the organization's network or, if only one person at a time needs to use the repository, it can be housed on the hard drive of a laptop or desktop.

The software application presenting the documents in the repository for review (the "review application") will have various customizable features for tagging or designating documents for discovery purposes, such as "relevant," "not relevant," "privileged," "confidential," and so on. Most review applications also allow the document reviewers to "redact" or block out information within a document that might be privileged or confidential.

One extremely important feature in a review application is that it allow for group tagging or designation so that when groups of documents are determined to be relevant, such as all drafts of a contract, the group can be designated without having to tag each document individually.

In addition, many law firms continue to create and add "issue" or "subject" codes to documents, a process that is extremely time consuming and fraught with error.[5] While limited issue coding, may be appropriate

[4] "PDF" indicates that a document file meets the Portable Document Format specification, which was created originally by Adobe, but is now maintained as an open standard by the International Standards Organization (ISO). There are several applications that can view PDF documents. "TIF" indicates a file meets the specifications for a tagged image file, an image format that can be read by many software applications.

[5] See the later discussion on the issue of the lack of agreement on relevance coding between two or more review

in some cases, e.g. "hot" or "key" document codes, the relative value of applying issue codes across a document set is usually not worth the substantial additional time and cost, particularly because the majority of the documents are text searchable and can be retrieved later with word searches rather than codes.[6]

After the documents have been reviewed, those designated for production can be provided to the requesting party in various forms; the format of production is certainly something the parties should have discussed and agreed upon during or before a FED. R. CIV. P. 26 conference.

If the data are reviewed on a web-based review application, the most inexpensive and efficient way to produce the documents is to provide the requesting party with access to a secure and designated section or folder on the application. In that case, production simply involves providing the requesting party with a URL, user names, and passwords. More traditionally, methods for producing electronic documents involve creating on an external hard drive a load file containing the various data elements that comprise the documents, and sending the hard drive to the requesting party. Such load files should be made according to specifications agreed to by the parties. CDs or DVDs may also be used, but these media are not preferred because they are unreliable, often failing or skipping for no apparent reason.

3. No Single Silver Bullet: The Successive Reductions Concept

There is no single silver bullet that solves all problems associated with escalating discovery costs and delays. As noted above, the single most effective cost reduction method is the focused collection of records most likely to contain relevant information.[7] Some argue that e-discovery is best accomplished by taking large amounts of data from clients and then applying keyword or other searches or filters. While, in some rare cases, this method might be the only option, it is also apt to be the most expensive. In fact, keyword searching against large volumes of data to find relevant information is a challenging, costly, and imperfect process. A much better approach is to ask key client contacts to help you locate core relevant information and then, by reading that information, determine other sources of relevant information.[8]

Once records are collected, cost reduction is achieved by applying different technologies and processes in a succession of steps, each of which: (a) reduces the volume of data under consideration by electronically eliminating duplicates and irrelevant documents, and/or (b) leverages the time of the reviewers by grouping closely related items in a way that makes it reasonable to make relevance or privilege decisions for all records in a group based on reviewing one item or a small subset of the group or by making review decisions on a subset of the records and then using software to determine which other records are sufficiently similar to the sample set's responsive records to be deemed responsive.

In this way, the volume is reduced in successive discrete stages. This guide will present technologies or concepts in the order in which they are usually applied in the ESI discovery process. A 20 percent savings in a process that costs tens or hundreds of thousands of dollars can be appreciable in its own right, but the cumulative effect of applying a series of reductions can be even more dramatic. Consider the following table, which demonstrates the cumulative effect of applying

teams. Issue codes are even more complex, and consistency or agreement between reviewers is even harder to achieve.

[6] Issue coding (sometimes referred to as "subjective" coding) was first used as a way to locate and retrieve documents regarding certain relevant topics or subject matters, before documents were maintained electronically and text searchable.

[7] Patrick Oot, Anne Kershaw, and Herbert L. Roitblat, *Practitioners' View: Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. U. L. REV. 533 (2010).

[8] Although keyword searching presents many challenges when applied against large volumes of heterogeneous data from many sources, it is considerably less problematic and therefore very useful when applied against targeted, largely relevant data populations. For example, the word "bat" is used to describe a ledge or shelf in a ceramics kiln, but is also used in a baseball context, can reference a nocturnal flying mammal, or can be used in idioms, e.g. to bat around some ideas, to go to bat for someone, or to be crazy as a bat. Targeted collection can avoid sweeping in sports-related e-mail and other clearly irrelevant items. By having a controlled context for searching, documents containing the term, "bat," are more likely to be responsive. In other words, just because a term might result in large numbers of irrelevant documents in a Google search, doesn't mean that search terms are without value in a more controlled document collection.

the various processes described in more detail in the following sections of this guide:

Table 1, Example of Cumulative Effect of Successive Reductions in Volume			
Process	Savings		Volume To Examine (GB)
	%	GB	
Original Volume	--	--	500
Less: De-NISTing (removing software & software support files)	20%	100	400
Less: Duplicate Consolidation	40%	160	240
Less: E-mail Threading	30%	72	168
Less: Domain Name Analysis	60%	101	67
Less: Predictive Coding	50%	38	39
Overall Reduction	**92.2%**	**453**	**39 (just 7.8% of original)**

4. Hashing and DeNISTing

Identifying Identical Files: Hash Values

To be able to discuss some of the basic technologies or processes for reducing the volume of electronic discovery, we need to define a very important concept: calculating "hash" values for files. (Note: we return to the topic of hash values in Chapter 5, with just a few words on basics for now).

The processes on how to calculate hash values are specified as standards in publically available documents. Hashing standards specify the sequence in which mathematical operations are performed using literally every bit of a file as input in order to calculate a fixed-length value.

The "hash" value of a file is a kind of digital fingerprint for the file that uniquely identifies it. The principle is simple: no two non-identical files should produce the same hash values. Hashing can be immensely useful throughout the e-discovery process, including verifying data transfers and identifying duplicate files. The advantage of having standards is that if different software packages calculate the hash values for the same content using the same standard, they will produce the same hash values for the same files.

Hashing is used widely outside of the electronic discovery world. For example, it has been used since the early days of the Internet to confirm that complete messages were sent and received. Both the sending and receiving parties calculate hash values, and if—both are the same—there is a high level of confidence that no data have been dropped or added. Hashing is also used to confirm that data encryption is successful. Hash values are calculated for files before they are encrypted and then again after they're decrypted. If the before and after hash values are the same, there is a high level of confidence that nothing was lost or added during the encryption process.

Two hashing standards encountered quite frequently in electronic discovery are the MD5 (Message Digest 5) standard, which creates a 128-bit hash value, and the SHA-1 standard (Secure Hash Algorithm), which creates a 160-bit hash value. The chances that two different files having the same MD5 hash values are one in 340,282,366,920,938,000,000,000,000,000,000,000 and chances of two different files having the same SHA-1 hash values are about 4 billion times less likely—in other words, a virtual impossibility[9] Adding, deleting, or changing even a single character in a file results in the creation of a different hash value by the hashing software. Even changing the type of font used for a single character or adding an extra space results in a different hash value.

While computer scientists have been able to create two dissimilar files that have the same MD5 hash value, we are unaware of a "collision" (two different files creating the same hash values) in real life discovery. There is a far greater chance of something else going wrong in the e-discovery process than there is that some evidence will slip thought the cracks because of a collision in hash values.

Focus on User-Generated Data: "DeNISTing"

The focus of discovery is to find information that is relevant to the case at hand. This will virtually never include files distributed by software companies as part of their program files, unless one of those companies is a party to the law suit and the software is at issue.

Considering that individual software programs such

[9]The specification for the MD5 algorithm has been an Internet standard since 1992: *See* http://www.faqs.org/rfcs/rfc1321.html. The SHA-1 standard, promulgated by the National Institute of Standards and Technology, and often used in encryption, is described at: http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf. There are other SHA standards that have longer hash values and are hence even more secure.

as Microsoft Word are often distributed as hundreds or thousands of files requiring hundreds of gigabytes of drive space, it should be obvious that just the simple expedient of not collecting or processing commercial software-related files can significantly reduce volume.

One simple—but fallible—way to avoid this problem would be to skip or pass over files that are in software folders or subdirectories or that have file extensions that make them look like software, e.g. "Word.exe." However, this would enable criminals and other wrongdoers to escape detection by keeping their files in such subdirectories or by saving them with file extensions that make them look like software, such as with *.exe or *.dll extensions.

Fortunately for litigants and law enforcement agencies, the National Institute of Standards and Technology ("NIST") publishes what it calls the National Software Reference Library ("NSRL").[10] The NSRL is a comprehensive listing of files known to be distributed with a large number of software packages, e.g., Microsoft Office. The listing includes not only the names of the files and their file sizes, but also the hash values for those files. The so-called "NIST" list can be used to safely reduce the volume of electronic data that is collected. This process of eliminating from consideration those files that are listed is called "deNISTing" the files.

When lawyers are talking about all the terabytes of data they will need to examine, it is useful to ask them if that's after deNISTing–particularly if they have not used targeted collection to limit the volume of data that was collected.

5. Metadata

"Metadata" are data that describes a particular characteristic or property about a computer drive, folder, file, or e-mail. For example, one important characteristic of a computer file is its name. Other important characteristics are the file's location, size, or perhaps its MD5 hash value. E-mail metadata could include, among other things, the From, To, CC, Date Sent, and Subject data.

Information about electronic files can be generated by many sources, including the computer operating system and the software application that created or modified a file. Some metadata, such as who had physical control of the data storage device on which the file was located, may have to be provided from information external to the computer system, e.g. from the person who gathered the data.

Types of Computer Metadata

Generally speaking, there are two types of computer metadata. One type is stored externally from the objects or files being managed. For example, the computer operating system usually tracks the date and time each file was created and the date it was last accessed. Such metadata can be useful in limiting the volume of data that must be considered. e.g., requiring only the production of documents created or modified during a given time period.

Another type of metadata is stored internally in a file. For example, Microsoft Word permits whoever creates a document to indicate a document's "author," "title," and the "keywords" associated with the document.[11] This type of metadata is not usually displayed when the Word document is printed to paper or if the contents of the file are used to create a PDF file (the type of file read usually with Adobe Acrobat).

As a practical matter, user-supplied metadata are "problematic" for e-discovery collection purposes because users typically do not actually add this information to their documents or if they do add them, they may do so inconsistently. Sometimes programs are set up to provide user information, and that information may never get updated as needed. For example, one specific problem is that the "author" name assigned by the computer is often incorrect.[12]

Sometimes the term, "metadata," is used to refer to information stored in a file that may be displayed only when the file is viewed in a particular way. For example, many software packages permit users to make comments on content in the file and to then either display or hide those comments. In addition, the file

[10] For more information on the National Software Reference Library, see http://www.nsrl.nist.gov/.

[11] To see how to create or change such metadata values in Word, search in Word Help for "properties," which is how Microsoft refers to certain of those metadata values. You can also see some of the metadata values when using Windows Explorer. Simply point to a file, right-click on it and select "Properties."

[12] See footnote 3 for an example of an incorrect author on expense report template.

might contain a history of how the document or content had been revised and by whom. This type of metadata is often referred to as "embedded data."

Managing Metadata in Attorney Work Product Databases or Repositories

Databases are often used by lawyers to manage the e-discovery process. It is sometimes useful to think of databases as similar to large spreadsheets with columns or "fields" for certain types of data, e.g., name of file, date created, author, etc., with a row for each document or file being tracked. Metadata are often represented in databases with specific fields holding certain types of information, including the text of the document. For example, one field might hold the file name, another the date created, and another field could hold the date last accessed or hash value for the file.

There are several advantages of having certain types of information in specified fields. One is that search specifications can be very specific and focus on one property or type of information. Having certain types of information in certain fields also permits users to select which fields to display on the screen or in a report. This lets them disregard irrelevant data and focus on the relevant. Another advantage is that records can be sorted by different types of data. For instance, one sort could present documents alphabetically by author, another by subject, and perhaps a third sort by date.

The document review systems work in some ways like e-mail management systems. For example, when people use e-mail software, they usually select which e-mails to read by picking from a list of selected metadata values presented in a columnar form, e.g. from, subject, and sent. The same e-mails could be represented in a different view using different fields or the same fields in a different order, for example. sent, from, to, subject.

Once an e-mail is selected to be read, the software usually presents the e-mail in a separate window that will contain a selective presentation of standard metadata with the text or body of the e-mail underneath. Document management systems typically operate in comparable fashion, presenting search results in columnar format while permitting more complete views of individual records for examination.

6. Minimizing Déjà Vu—Duplicate Consolidation or "Deduping"

People conducting reviews of ESI for production purposes, such as selecting relevant documents or screening for privileged records, can be left with a strong sense of déjà vu about having seen certain records before. The reason is that they may have, in fact, seen the same files many times before because electronic discovery is shot through with duplicate files, that is, files that are bit-for-bit exactly the same as each other.

Electronic documents and files are so easy to copy and distribute that every person can have his or her own personal "library" of files that gets swept up in discovery. With e-discovery, the challenge is to identify duplicate records and avoid repeatedly examining the same content. Attorneys who don't do this properly may be wasting their clients' money and presenting exaggerated claims of burden and time needed for reviews.[13]

Duplicate files can make up as much as 60-80 percent or more of the records in a collection.[14] It obviously makes sense to avoid repeatedly examining exactly the same files for exactly the same purposes. Consolidating duplicates permits faster reviews and reduces the risk that the same records are treated inconsistently. Consolidating duplicates is a basic way of controlling the escalating costs of discovery and litigation.

---

[13] Under FED. R. OF CIV. P. 26(b)(2)(C)(i), a court *must* limit the frequency or extent of discovery if it determines "the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive."

[14] The information on volume reduction with duplicate consolidation and the extent of its use is taken from the "Report on Kershaw-Howie Survey of E-Discovery Providers Pertaining to Deduping Strategies," published by the eDiscovery Institute, ("Deduping Survey"). That study formed the basis for the articles, Anne Kershaw and Joe Howie , *EDD Showcase: Highway Robbery?*, L. TECH. NEWS, Aug. 2009, at 30, and Patrick Oot, Anne Kershaw, and Joe Howie, *Ethics and Ediscovery Review*, ACC DOCKET , Jan-Feb, 2010, at 46.

The companies that participated in the Deduping Survey included ACT, BIA, CaseCentral, Clearwell, Daticon, Encore, Fios, FTI Consulting, GGO, Iris Data Services, Kroll Ontrack, LDM, LDSI, Rational Retention, Recommind, StoredIQ, Trilantic, and Valora.

People performing production reviews on computers usually have views of the records they are assigned that are similar in many ways to how they read e-mails. One view is a columnar view of selected metadata about the records with column headings like:

	Doc#	Date	From	To	Subject	Custodian
Row 1						
Row 2						
…						

Each horizontal row will represent one document or file. Reviewers can browse the columnar view and click on a row to view and analyze the original record, much like the e-mail example earlier

When there is no duplicate consolidation (or what some call "deduping"), reviewers see a row in the columnar view for each *copy* of a file. However, when a reviewer examines the full record of a copy the reviewer sees exactly the same full display for each copy of a file–reviewers are making exactly the same decisions on virtually exact copies; they are, in essence, double-billing for the second and subsequent copies.

Lawyers will sometimes have custodian-level[15] deduping performed, meaning that only one copy of a record will be presented for each person or custodian who had a copy—even if that person had more than one copy. That is a step in the right direction, but there can be many people who received copies and when reviewers look at the full record view for each copy, each of them will look exactly alike.

The most efficient way to consolidate duplicates is to show reviewers only one copy of a record regardless of how many people originally had or produced copies. We use the term "duplicate consolidation" to indicate that the information about who had copies and where they stored them can all be consolidated and presented in one field associated with just one record. For example, if an e-mail was found in the records of the author Sam Jones, as well as in the e-mail systems of Sheila Schwartz and Mike Mitchell, all of that information can be consolidated and presented in a single database record in a "Sources" field as shown below:

DocNo	Doc Type	Author	Recipient	Sources
ABC1001	E-mail	Sam Jones	Sheila Shwartz; Mike Mitchell; Brenda Xu	Sam Jones; Sheila Schwartz: Mike Mitchell; John Cox

Considering that there can be ten or more copies of some e-mails, it should be clear that presenting the information in one field for one record gives reviewers a much better comprehension of everyone who had copies of a file. It can also help spot trends such as particular custodians who may not have produced what might have been expected of them – in the above example, where's Brenda Xu's copy, why isn't she listed as a source? If this ends up being a consistent pattern, was there a problem collecting data from her? Was she somehow missed in the collection process or did her files get overlooked in processing?[16]

This type of information is particularly helpful when making decisions pertaining to privilege or privilege waiver. In the above example, if DocNo ABC1001 is privileged, how did John Cox come to have a copy of the e-mail? Does his possession of the e-mail create a waiver of privilege argument?

Consolidating information about duplicate files *across a collection* has been shown to significantly lower the volume of information that has to be reviewed.[17] Parties who consolidate duplicates just within the records of *individual* custodians will review, *on average*, 27 percent more records than a party that consolidates

[15] As noted previously, a "custodian" is the individual from whom the documents were collected.

[16] Putting all sources of information in a single field can be so overwhelming that it is seldom actually used. An alternative to achieving this overall audit-type function performed by putting all sources in a single field in the consolidated database entry is to have production reports that list custodians—and the number of different types of records produced by them—and to cross-check that against the number of times that person's name or e-mail address appears as a sender or recipient. There can be many reasons why the numbers don't match completely, e.g., hardware failures, varying compliance with record retention policies, etc., but counsel should be aware of this sort of overall process management information.

[17] The process of identifying duplicates across an entire collection is sometimes called "horizontal" deduping; the process of examining only the contents of an individual custodian is called "vertical" deduping. Because "deduping" may imply that some copies are discarded, we prefer the use of the term "duplicate consolidation" to reflect that the information is simply consolidated in one place for faster, better reviews, not discarded.

duplicates across a collection, sometimes substantially more.

Furthermore, failing to consolidate duplicates across a collection:

- Exposes parties to risks of inconsistent review decisions, e.g. producing one copy of a record where another record had been redacted or listed as privileged,
- Causes lawyers to inflate estimates given to adversaries and courts about the costs and time required to review productions.
- Exposes trade secrets, attorney-client information and other company-confidential information to far more reviewers than necessary.
- Results in reviewers making uninformed decisions about records because they cannot readily determine everyone who had copies of the records.

Duplicate consolidation is the "low-hanging fruit" of e-discovery cost control in that it can be implemented in virtually any review platform. The identification of duplicates and the consolidation of information about the various copies can be accomplished prior to data being loaded in the review system, and that is not necessarily so with some of the other technologies. For example, the availability of e-mail threading (discussed in chapter 6) or concept clustering technologies (see chapter 8) may depend on the type of software used by the review database or repository.

Despite the obvious advantages of consolidating duplicate information across custodians, the EDI Survey showed that *about half the time parties do not use this readily available and proven technology*. By inflating the volume of records to be analyzed for production purposes, the failure to use this technology also delays the dispute resolution process.

One important point about duplicate consolidation: The producing party can still preserve all of the data it examined in making the consolidation decisions, and in those rare instances where a forensic examination is needed of all the metadata pertaining to a specific e-mail, that preserved data could still be made available as an exception rather than as the rule.

Inconsistent Review Decisions

A study by the eDiscovery Institute[18] and studies by the Legal Track of the Text Retrieval Conference ("TREC"), [19] which is sponsored by NIST, have shown that different teams of reviewers examining the same records seldom agree with each other more than 80 percent of the time, and that includes records considered non-responsive. When consistency measures look at how many documents originally deemed responsive were selected by a second review team, only about half of the records deemed responsive originally are selected by the second team.[20]

Inconsistent review decisions are to be expected when multiple reviewers review exactly the same content—this shouldn't be a surprise, it's a virtual certainty. When large numbers of reviewers examine large numbers of identical records, they will invariably make different decisions on the same records—some reviewers will decide to produce some copies while others will decide to redact other identical copies or place them on the privileged list, and still others may decide they are not responsive. In fact, even the same reviewers can make different decisions when presented with the same records due to changes in understanding

[18] Herbert Roitblat, Anne Kershaw, and Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61(1) J. AM. SOC'Y INFO. SCI. & TECH. 70 (2010) (hereinafter "EDI Document Categorization Study") available at:

http://www.eDiscoveryInstitute.com/pubs/DocumentCategorization.pdf

[19] Herbert Roitblat, "TREC Legal Track 2008," http://orcatec.blogspot.com/2009/04/trec-legal-track-2008.html.

[20] Basically, the percentage of records that are deemed unresponsive can make the overall agreement rate look higher. For example, if the first team ("TEAM A") performing review selects 10 percent of the records as being responsive and a second team ("TEAM B") selects 10 percent of the records at random as being responsive, the second team will agree with the first team on 82 percent of the records -- despite the fact that the second team only selected 10 percent of the same responsive records as the first team!

	TEAM A Responsive	TEAM A NonRespons.	Total Records
Team A	10	90	100
Team B Responsive	*1*	9	10
Team B NonRespons.	9	*81*	90
Agreement	1	81	82

The percentage of documents originally selected as responsive that are selected as responsive by the second team is obviously a key metric. More on the topic of inter-reviewer agreement rates in Chapter 9.

of the case, fatigue, or a variety of factors.

Recommendation: In cases where a party produces a record that is identical to another copy that was withheld as privileged, the court should consider whether the inadvertently producing party used reasonable efforts to guard against such production, given the long availability of hashing technology and its reliability. One could argue that if the produced and the withheld records both produce the same hash value the inadvertent production was avoidable.

Attachment Context

There is one situation where parties typically do not consolidate all the multiple identical copies of documents, and that is where the same document has been used as attachments to different e-mails. Because reviewers may want to see documents in the context of the e-mails to which they were attached, parties typically include a copy each time the record is attached to a different e-mail. Thus, if a document is attached to two different e-mails, it will be represented twice in the review database.

E-mail. There are two special comments to make about consolidating duplicates for e-mail. The first is that there are very minor differences in the metadata associated with the copy held by the sender and the various copies in the files of the recipients. For example, the time may be different by a few seconds or the hidden data can include how a message was routed over the Internet to a specific recipient. This information can be different for each recipient.

As a result, a hash value calculation that includes all possible metadata will not match e-mails that are—for all practical purposes—identical. This can lead to a great deal of waste and duplicative reviews because, *in point of fact, none of the first or second level reviewers ever look at or consider those metadata values when making privilege or responsiveness decisions*. The solution is to exclude such metadata from the hash value calculations so that minor differences in these values do not result in different hash values. *In other words, reasonably sophisticated processes can determine which metadata are handed to the hashing software for it to use in calculating hash values.* In fact, different parties may select different methods of calculating hash values for e-mails.

A second point is more technical in nature, but could, in the right circumstance, result in a major reduction in discovery volume. In circumstances where a producing party maintains more than one e-mail system, e.g., a Lotus Notes system and a Microsoft Outlook system, minor differences in how data are stored by those systems can cause them to produce different hash values for the same e-mails. For example, one e-mail system may put brackets around an e-mail address, e.g., "Joe Smith <Joe@Smith.com>" and another may put parentheses, e.g. "Joe Smith (Joe@Smith.com)." That difference will cause different hash values to be calculated—and this difference can occur even between different versions of the same e-mail software.

One answer is to "normalize" the e-mails by essentially putting them in the format that was used to transmit them as specified in an Internet technical document.[21] In other words, base the hashing comparison on what the e-mail looked like while it was being transmitted, not on how it was stored once received. It isn't important that, as a judge you be able to perform the normalization, only that you know that it can be done.

7. The Rest of the Story: E-mail Threading

E-mail constitutes the lion's share of the volume of ESI. One feature of e-mail presenting special challenges and opportunities is that e-mail conversations take place in threads, which are created when a party to an e-mail either forwards it or replies to it.

Typically, but not invariably, the text of the earlier e-mail is contained in the reply or the forward, although users may edit or change the earlier text. The special challenge when deciding responsiveness or privilege is that an individual e-mail may contain only a piece of an overall conversation or, in the best case, all the conversation up to but not beyond that point in time. Thus, the reviewer may be unaware of subsequent replies to or forwards of a particular e-mail.

A subsequent e-mail in the thread might affect the designation of the earlier segment as responsive or privileged. Alternatively, the e-mail may have been

[21] See the RFC2822 standard at: http://www.faqs.org/rfcs/rfc2822.html.
This topic is discussed in John Martin, *Deduping Across Different E-Mail Platforms, A Faceted View of Duplicate Consolidation*, ALSP UPDATE, Sept. 2009.

forwarded to someone outside the zone of the privilege or a reply to a subset of the original distribution list might have gone to someone outside the privilege. For example, an e-mail between outside counsel and three executives might be forwarded by one of the executives to a friend of his who was not a lawyer, spouse, psychotherapist, or clergyman, thereby waiving the privilege on the earlier e-mails.

From a review standpoint, presenting all the e-mails in a thread or group to the same reviewer at the same time has huge advantages in terms of speed of comprehension and review. A survey by the eDiscovery Institute[22] suggests that, on average, using e-mail threading can save 36 percent compared to not using e-mail threading, with individual respondents reporting instances of savings on specific cases of 50 percent or more. Furthermore, e-mail threading helps ensure consistency of decision-making for all the e-mails in the thread.

Some e-mail threading systems that group e-mails in threads for review purposes will confirm that a later e-mail contained all the content from earlier e-mails in the thread or branch of the thread, enabling the reviewer to focus attention on the last e-mail in the thread or branch. This has obvious advantages when deciding relevance or privilege because the reviewer can see the name of everyone who received copies of the content and can make review decisions in light of the context of the overall conversation.

Different e-mail threading technologies have varying abilities to create threads from e-mails from different e-mail systems and/or to include scanned copies of e-mail printouts in threads.

Recommendation: When faced with lawyers complaining about the large volume of e-mail they have to review, ask them if they are using e-mail threading techniques. Also, consider what threading systems, if any, was used by a party when deciding fee applications for reviewing e-mails.

[22] Anne Kershaw and Joe Howie, *Report on Kershaw-Howie Survey of Leading E-Discovery Providers Pertaining to Email Threading*, , published by the eDiscovery Institute, Jan. 5, 2010 available at www.eDiscoveryInstitute.org. The companies participating in the survey were Anacomp, Capital Legal Solutions, Clearwell, Daticon EED, Equivio, InterLegis, Kroll Ontrack, Logik, OrcaTec, Recommind, TCDI, Trilantic, and Valora.

Privilege Lists

Some courts require that a party seeking to claim privilege on an e-mail must list the author and recipients of each e-mail that was embedded in a particular e-mail. For example, if an e-mail were a reply to an earlier e-mail, there would have to be, in essence, two entries on the privilege log for that e-mail, one showing the parties to the earlier or initiating e-mail and another showing the parties to the reply. Having a review system that tracks which e-mails are associated in threads can greatly speed the creation of such privilege lists and permit users to confirm that the text shown in later e-mails accurately reflects the information contained in the earlier e-mails, i.e., that no one edited or changed that information in the later e-mail.

Suggested Terminology

Court opinions addressing how to log e-mails on privilege lists have used inconsistent terminology, e.g. the terms "chains," "strings," and "strands" have all been used somewhat interchangeably. For example, sometimes a single e-mail is described as being a chain[23] or a string as opposed to containing some of the text of earlier e-mails in the thread. At least one opinion referred to earlier e-mails in a string as being "attached" to a subsequent e-mail when in fact the subsequent e-mail was most likely a reply or a forward of an earlier e-mail.[24]

There are more than semantic differences in

[23] *See, e.g.*, Rhoads Indus., Inc. v. Building Materials Corp. of Am., 254 F.R.D. 238 (E.D. Pa. 2008) defines "so-called 'e-mail chains' or 'e-mail strings,' i.e., *a series of e-mail messages by and between various individuals*" (emphasis added), which misses the essential point that a thread or string involves replies or forwards, not just having common senders or recipients; also used terminology of "top" e-mail to refer to the content added to earlier e-mails; Baxter Healthcare Corp. v. Fresenius Med. Care Holding, Inc., 2008 WL 4547190 (N.D. Cal. 2008) (Defendant to create a privilege log entry for each e-mail in a "string."); Muro v. Target Corp., 243 F.R.D. 301, 307 (N.D. Ill. 2007) uses the term "string" to refer to an e-mail that contains the e-mail messages contained in the string; In re Universal Serv. Fund Tel. Billing Practices Litig., 232 F.R.D. 669 (D. Kan. 2005) defines a "strand" as follows: "an e-mail strand (or string) occurs where the printed e-mail actually consists of more than one message, usually formatted with the most recent message first;"

[24] United States v. ChevronTexaco Corp., 241 F.Supp. 2d 1065 (N.D. Cal. 2002)

forwarding, replying, and attaching. Consider the example where Attorney Bob e-mails CEO Carol, attaching a Microsoft Word file containing a formal legal opinion for the use of Carol's company. Bob does not discuss the contents of the opinion letter, but does mention some news about Alice, a mutual friend who works for the *Wall Street Journal*. If Carol wants to make a comment to Bob about the news and provide a copy to Alice, how she does that could determine whether the attorney-client privilege that applies to the legal advice in the Word document is waived.

If Carol does a "reply" to Bob and copies Alice, the Word document will not be sent to Alice, and there is no waiver situation. If Carol does a "forward" of Bob's e-mail to Alice with a copy to Bob, the Word document will be sent to Alice, and there is a potential waiver situation. If Carol creates a whole new thread by creating a new e-mail to both Bob and Alice, but "attaches" Bob's e-mail, in many e-mail systems that will also cause the attachment to Bob's e-mail to be sent to Alice, thereby also creating an arguable waiver situation.

Adopting a common terminology and being technically correct in describing replies and forwards vs. attaching e-mails will improve analysis and understanding. Here are some suggested definitions:

- **"Initiating" e-mail**. An e-mail that was sent, but which was not a reply to, or a forward of, an earlier e-mail.
- **"Draft" e-mail**. An e-mail that was drafted but not sent.
- An **e-mail "thread" or "string"** is the initiating e-mail plus all of various e-mails that resulted from people forwarding or replying to that initiating e-mail or subsequent replies or forwards. Note that when a person replies to or forwards an initiating e-mail, the "To," "From," "Subject," and "Date" metadata of the initiating e-mail often become part of the text of the replied or forwarded e-mail. Other metadata, such as the "Internet Header" information basically disappear.

  For most cases, this doesn't matter, but it is important to understand that—other than a single-e-mail thread, i.e., an e-mail that was sent but to which there was no reply and which was never forwarded, no individual e-mail will have all of the data and/or metadata associated with all of the earlier e-mails in the thread. Therefore, while an individual e-mail may be said to contain *evidence* of the thread, or to contain a possibly imperfect snapshot of the thread (recognizing that the text of the e-mail that was forwarded or to which a reply was sent may have been edited or changed by the person forwarding or replying), it does not contain all of the information pertaining to the whole thread.
- An **"embedded" e-mail** is one whose content is included in the body of a later e-mail.
- An **"orphan" e-mail** is one that cannot be associated with one or more of the earlier e-mails in the thread.
- A **conversation "branch"** occurs when there is more than one reply or forward to an individual e-mail. The key point here is that a reviewer could not determine by looking at the last e-mail in one branch what was said or happened in a different branch.
- **BCC**'s. The term, "BCC," is derived from "Blind Carbon Copy," which is a linguistic artifact from the earlier era of typewriters. It indicates that some people received copies of an e-mail without others being aware of everyone who received copies. Typically only the sender sees who all received BCC copies.
- **Attached e-mails** are e-mails that are attached to another e-mail much as a Word document could be attached to an e-mail. The significance of "attaching" an e-mail to another is that an e-mail from one thread can be associated with an e-mail in a different thread or threads whereas if a "reply" or a "forward" is used, the e-mail remains in its original thread.

E-mail Threading Systems

There are several e-mail threading systems that will identify which e-mails were in a thread. Each of them involves some combination of the following techniques for associating individual e-mails into threads:

1. **Hidden Metadata**. E-mails can contain information not normally viewed by typical users. Such metadata can include fields such as "MessageID" and "Reply To Message ID" that can enable the association of individual e-mails into threads with a high degree of confidence, particularly if all of the e-mails in the thread are

available. Not all e-mail systems employ the same metadata values in the same way, sometimes making it problematic to associate e-mails from different systems into consolidated threads.

2. **Visible Metadata**. Some metadata that are viewed usually by typical users can be used to assist in associating e-mails into threads, e.g. information in the From, To, CC, Subject, and Date fields. Given that different threads between the same people might use the same Subject line; this method by itself is perhaps the least reliable.
3. **Textual Analysis**. Some e-mail threading systems will compare the text in the body of the e-mail to confirm whether it, in fact, contains the text of e-mails that occurred earlier in the thread.

When evaluating claims of the use of e-mail threading to speed review, judges should be inclined to approve such uses.

Scanned Paper E-mail. Normally when people are discussing e-mail threading they are referring to the process of associating native electronic e-mails into threads. However, there are also some systems that claim to be able to associate scanned paper copies of e-mails into threads.

8. Who Sent That?—Domain Name Analysis

In the old days of paper files, if all correspondence was filed by the sender's company, a party responding to a discovery request could safely disregard folders where the sending company would not be reasonably suspected of having had anything relevant to do with the litigation. In the brave new era of ESI, the same logic should enable producing parties to disregard e-mails from domain names not reasonably suspected of having any relevance to the litigation, e.g. CNNSports.com or LocalSoccerClub.com.

Experience and anecdotal accounts from conversations with others suggest that reductions in the volume of e-mail in a collection on the order of 30-80 percent are achievable by gathering only e-mail from domain names associated with organizations that are reasonably likely to have generated relevant content or content that could lead to relevant content.

One exception to this sort of high-level treatment of e-mails according to sender's domain is the various domain names used by providers of ostensibly personal or private e-mail services, e.g., AOL.com, Hotmail.com, Yahoo.com, Gmail.com, etc. People sometimes use such private e-mail addresses as a convenience when traveling or out of the office, but other times they can be used as a way of attempting to avoid having some illicit activity detected. A more granular approach to such domain names might involve making decisions based on who the individual users were.

It is reasonable to expect producing parties to maintain a list of the domain names it used to exclude records from consideration.

9. Getting It Together: Clustering, Near-Duping, and Grouping

There are various technologies available that cluster or group like records ("Grouping Technology") so that review decisions regarding relevance and/or privilege can be made for a group of records based on an examination of one or a few of the records in the group. The use of such Grouping Technologies will normally represent a good faith effort to take reasonable steps to identify relevant material and should be encouraged as a way to lower the costs of litigation.

Recommendations:

1. To the extent the costs of the additional software or processing used to accomplish clustering are included in fee petitions, the court should be inclined to approve them as having saved in attorney fees.
2. Courts should consider what cost-effective processes have been used when evaluating attorney fee requests. One could reasonably expect efficient processes from lawyers with high billing rates.

10. Predictive Coding/Automated Categorization

The cost of having lawyers conduct pre-production relevance and privilege review is often the single largest e-discovery cost element. In the traditional "linear

review" model, lawyers visually examine every record that was potentially responsive or privileged and make a decision on each document. Deduping cuts down those costs by examining only one of several identical copies. E-mail threading groups e-mails by conversations for faster, more consistent decision making, often enabling reviewers to skip earlier e-mails whose content is repeated later in the thread. Clustering can group records that are very similar so that decisions can be made on groups of records without having to examine each record. As clustering is often implemented, reviewers will examine at least one record in each group or cluster.

"Predictive coding" and "automated categorization" refer to another approach to pre-production document review that minimizes the number of documents that have to be examined visually. Attorneys review a subset of records at the outset of the case, and software is used to propagate those decisions to the other documents that had not been examined visually.

A recent survey showed that, on average, predictive coding reduced review costs by 45 percent, with several respondents reporting much higher savings in individual cases.[25] That same survey showed that *the largest single obstacle to more widespread adoption of predictive coding was the uncertainty over judicial acceptance of this approach.*

In the predictive coding survey, the respondents used several variations in their approaches. Most involved the selection of an initial review set by queries or random selection, propagation of the review decisions to unreviewed documents using some form of clustering technology—with follow-on sampling of selected, and/or unselected records to determine if further iterations were in order.

A recent study[26] demonstrated that predictive coding or document categorization produces quality comparable to traditional human linear review. In the study, two review teams examined a sample set of 5,000 documents taken from a document population that had been reviewed previously for purposes of responding to a DOJ investigation. Two e-discovery service providers used their versions of predictive coding to review the entire collection. The resulting agreement rates are shown below:

Comparison	Overall Agreement (includes Responsive and Non-Responsive)	% of Documents Selected As Responsive in Original Review Later Identified as Responsive in Second Review
Team A vs. Original	75.6%	48.8%
Team B vs. Original	72.0%	53.9%
System C vs. Original	83.2%	45.8%
System D vs. Original	83.6%	52.7%

One of the biggest problems with these types of studies is establishing a base line with which to compare the performance of different review teams or methodologies. In this study, the original decisions were taken as the base line. As shown above, the automated document categorization generally performed as well as the human review teams in matching the decisions of the original review team.

To the extent predictive coding is compared to manual review, judges should be aware that studies focusing on results of reviews by different teams of reviewers on the same sets of records have shown that the second team may select a little more than half the records as responsive as the first review team did.[27] In other words, manual review is itself heavily flawed, costs significantly more than predictive coding, and exposes sensitive or confidential corporate information to far more reviewers than is necessary—all factors that should be considered in any consideration of whether a responding party made reasonable efforts to select records for production.

Predictive coding or automated document categorization could greatly improve the replicability of original decisions. All survey respondents indicated they

[25] See "*eDiscovery Institute Survey on Predictive Coding*," available at www.eDiscoveryInstitute.org. The companies that participated in the survey were Capital Legal Solutions, Catalyst Repository Systems, Equivio, FTI Technology, Gallivan Gallivan & O'Melia, Hot Neuron, InterLegis, Kroll Ontrack, Recommind, Valora Technologies, and Xerox Litigation Services.

[26] EDI Document Categorization Study, supra, note 18.

[27] The EDI Document Categorization Study summarizes reviewer consistency studies by the Text Retrieval Conference ("TREC") sponsored by the National Institute of Standards and Technology. In a 2008 study, second reviewers selected 62 percent of the records initially selected, and in a 2006 study, the second reviewers selected 58 percent of the records initially selected.

maintained an audit trail of how documents were processed, and most indicated that, by reapplying those decisions, their systems would produce the same results as the original processes. One indicated that the order in which documents were loaded, sampled, and reviewed could have a minor impact on the results.

Note that, with manual linear review, there is virtually no transparency. In many cases, individual reviewers do not document their rationale for deeming a particular document as being responsive or not, only the decision remains. By contrast with the predictive coding systems, the audit trail permits an examination of why something was included or excluded.

Recommendation: Parties should be encouraged to use predictive coding or automated classification processes to lower the costs of producing electronic evidence.

11. The Case for Focused Sampling

When people talk about sampling, they usually mean random sampling in which each record in a population has an equal chance of being selected. In litigation and e-discovery, sampling is often used in the context of sampling archived records such as backup tapes to see if they contain relevant records or by sampling discovery records to see if those that had not been selected for production actually contained relevant information. *In discussing sampling in this chapter, we will use the term, "responsiveness rate," ("RR") to indicate the percentage of records that are responsive to a discovery request.*

Entire books and courses have been devoted to statistics and sampling, but here are a few observations on the use of sampling and statistics in e-discovery. First an overview, then the observations:

Overview

Measurements taken from samples are used as estimates of what an examination of the whole population would yield. For example, if 14 percent of a sample is responsive, 14 percent of the population is also estimated to be responsive. A confidence level in conjunction with a confidence interval describes how likely it is that the sample measurement accurately describes the population, e.g., a 95 percent confidence level with a 4 percent margin of error means that, if repeated samples of the same size were drawn, 19 out of 20 times, the results would be within the confidence interval or margin of error, that is, the observed rate plus or minus a margin of error, e.g., 14 percent plus or minus 4 percent.

The larger the sample size, the higher the confidence level and/or the smaller the margin of error.

Observations

1. When evaluating the validity of conclusions drawn from random sampling, consider the number of records you would expect to be found given the responsiveness rate and the sample size. In general, be wary when the estimated responsiveness rate is lower than the margin of error.

For example, let's assume that, in an employment discrimination case, there are 200 e-mails that are directly relevant to how the plaintiff was perceived or discussed by her co-workers or her supervisor, distributed within the organization as follows—noting that all responsive e-mails are within the plaintiff's work section:

Collection Source	Responsive Records	Total Records	Resp. Rate
Section Supervisor	50	10,000	.005000
Section Co-Workers (8 employees)	150	80,000	.001875
Section Total	200	90,000	.00222
Department (50 employees)	200	500,000	.000400
Division (500 employees)	200	5,000,000	.000040
Company (5,000 employees)	200	50,000,000	.000004

Suppose we use the general advice that a sample size of 500 will provide an estimate of the population with a 95 percent confidence level with a margin of error of ± 4.4 percent.[28] If 500 of the supervisor's

[28] *See, e.g., Sample Size Table*, RESEARCH ADVISORS, http://research-advisors.com/tools/SampleSize.htm (last visited Oct. 31, 2010).

"Professional researchers typically set a sample size level of about 500 to optimally estimate a single population parameter, e.g., the proportion of likely voters who will vote for a particular candidate. This method will construct a 95 percent confidence interval with a Margin of Error of about ±4.4 percent (for large populations)."

records were sampled, we would expect, on average, to find 500 x .005 or 2.5 records, meaning either two or three relevant records. On the other hand, sampling 500 records from the overall department would yield, on average, 0.2 responsive records, meaning that one of five times we sampled, we would expect to find one record. Sampling 500 records from the entire company would average .002 responsive records, meaning that every 500th time we sampled, we would expect to find one record, which is not surprising, considering the responsiveness rates for those document populations are much smaller than the margin of error.

The point, of course, is that restricting sampling populations to the records associated with people most likely to have been in a position to have known about or taken action on the relevant subject matter during the relevant time period—a "people-centric" record gathering process—will improve the chances that "random" sampling will yield useful results. There is little point to randomly sampling everything.

> *2. It is simplistic to think of one overall "responsiveness" rate. A document that is "responsive" may satisfy any of a number of criteria, each of which will have an individual responsiveness rate that is generally lower than the overall responsiveness rate. Those issue-specific responsiveness rates need to be considered when evaluating the completeness of productions.*

A typical Rule 34 request will include a number of specifications for the documents or ESI the requesting party wants produced. For example, in our hypothetical employment discrimination case, the plaintiff may want records that reflect pay rates for all employees in a department over time and records that contain instances where the plaintiff's supervisor or colleagues referred to her in a derogatory or derisive fashion. There may be literally thousands of pay rate records that are largely duplicative, and relatively few records indicating derogatory or derisive communications.

Sampling done to corroborate the completeness of the production ought to take into consideration whether any new information was found versus redundant or duplicative information. Sampling 500 records and finding 10 more records on pay rates that are already well-established should not result in having to search for more records. On the other hand, sampling on the issue of derogatory or derisive communications may require a sample size larger than 500, given the relatively low responsiveness rate of records pertaining to this one issue or request.

12. Special Cases

The topics covered in the earlier part of this guide will apply in most cases involving ESI. This chapter covers a few topics, which—although not occurring in as many cases—can represent significant dollar expenditures when they do occur:

A. Foreign Language Translation Costs

Human translation services or even machine translation can have significant costs. Parties having large volumes of records translated should first perform deduping, e-mail threading, or clustering to reduce the number of records that must be translated so that duplicative translation services are used.

Recommendation: When presented with legal bills for translation costs, inquire as to whether the party that ordered or performed the translation used some of the basic technology described earlier in this guide, such as across-custodian deduping, e-mail threading, or deduping/near-duping to minimize the number of times the same content was translated.

B. Searching Audio Records

Some methods of searching audio recordings can result in large expenditures of time and money, e.g., transcribing tapes to full text and searching the full text or having humans listen to the actual recordings. Technology is available to search the actual recordings to look for the phonemes or sounds that are recorded without having to convert those sound files to text files. Furthermore, there are usually metadata associated with the recordings for such things as date and time, person making the recording, the person for whom it was recorded, etc.

Recommendation: When presented with arguments about cost, burden, or time required to search audio tapes, ask what measures have been taken to use metadata to narrow the scope of the content that would have to be searched, and ask whether the parties have considered phonetic searching of the native files. Don't assume that human review or transcribing to full text will yield better results.

C. Searching Tape Backups

When the Federal Rules of Civil Procedure were amended to explicitly incorporate ESI, the drafters wisely refrained from referencing specific storage devices in the definition of what is "not reasonably accessible."[29] Many lawyers do not understand the distinction between "in-system" backup data and backup data that have been removed from the backup system's standard rotation process.

Generally speaking, backup data that are in-system and part of the ongoing rotation for disaster recovery is accessible but it is also completely redundant of the live data that it are backing up (as it should be).[30] Collection from these in-system backups is therefore unnecessary if data are otherwise preserved and collected from the live data sources although in some situations, it may be more convenient to collect data from such backup systems.

For various reasons, backup tapes or hard drives may be removed from the in-system rotation process, but parties should understand that the tapes will preserve only what was in the active files when the tapes were created—for example, if someone gets an e-mail at 1:00 PM and deletes it at 1:15 PM the tape backup that is made at 1:30 PM will typically not contain that e-mail. In other words, just because there are tapes doesn't mean that everything that was ever relevant was preserved.

When tapes or drives have been removed from the in-system rotation process, the accessibility of the data will depend on a number of factors, but particularly the length of time that has transpired since the tapes were removed. In addition, advancements in tape indexing technology may make it easier and less costly to index the content of older backup tapes, but these technologies should be tested before placing any reliance on them. If the tape indexing technology works, it may be reasonable to at least sample some tapes where earlier it might have been prohibitively expensive.

Recommendation: When lawyers are discussing the high costs of searching out-of-system backup tapes, ask them what alternatives they have considered.

13. Worst Practices

There are some practices that will invariably be associated with inefficient or error-prone review of e-discovery. They are:

A. Wholesale-printing of ESI to paper for review

While we hope this practice is not widespread, there are still computer-phobic lawyers who order electronic files printed to paper for review, despite the fact that the people who originally created, sent, received, and used them did so almost exclusively in electronic format. Printing to paper causes delays, and the incurring of more clerical staffing costs. It can also cause unnecessary delays and errors in associating review decisions with the correct review database record.[31]

Recommendation: Courts should deny printing fees for wholesale print-to-paper for review and should cut hours or hourly rates for those involved in such reviews. Printing to paper is also an indicator of lack of sophistication, and courts should examine whether the cost-effective techniques outlined in this guide have been used. For example, the chances that lawyers are not consolidating duplicates across custodians are much

[29] F. R. CIV. P. 26(b)(2)(B). *See* http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/EDiscovery_w_Notes.pdf.

[30] A word about rotation: usual backup disaster recovery procedures involve creating successive backups over a period of time, such as a week, in case the most recent backup and any before it are defective. These tapes are then rotated and overwritten the following week.

[31] The focus of this guide is on cost effectiveness, but courts should also be mindful of the environmental impact of this practice. For example:

- A pound of paper contains about 100 sheets of paper. A stack of paper a foot high contains about 3,000 pages http://eetd.lbl.gov/paper/ideas/html/copyfactsA.htm. One tree yields about 8-9,000 pages of paper.
- http://www.examiner.com/life-the-cubicle-in-national/cubicle-101-how-to-identify-the-office-tree-killer-who-wastes-paper
- http://wiki.answers.com/Q/How_much_paper_can_one_tree_produce
- http://wiki.answers.com/Q/How_many_sheets_of_paper_are_made_from_a_tree

One gigabyte of electronic data typically contains about 70,000 pages or about 8-9 trees' worth of paper. When printed, the paper will weigh about 700 pounds and when stacked will be about 21 feet high.

higher if they aren't even sophisticated enough to review electronic evidence on computers. They are also far less likely to use threading or clustering technology.

B. Wholesale printing to paper, then scanning to TIF or PDF

To the extent parties elect to work with TIF or PDF images of electronic discovery, those TIF or PDF copies can be made directly from the electronic files without printing to paper. Fees related to wholesale printing to paper—followed by scanning those same records—should be disallowed. Costs associated with OCRing those scanned records to create searchable text should also be disallowed as the searchable text was present in the original native file.

14. Ethics and E-Discovery[32]

The ethical duty to be technically competent in processing electronic discovery is rooted in various rules of the *ABA Model Rules of Professional Conduct*:

- **Rule 1.1. Competence:** How can a lawyer who doesn't understand ESI represent a client in litigation when the predominant form of evidence is electronic?
- **Rule 1.3. Diligence:** A threshold level of diligence would be to make an effort to understand how the principal sources of discovery can be processed.
- **Rule 1.5. Fees:** The obligation to refrain from double-billing implies a duty to take reasonable steps to identify duplicative work.
- **Rule 3.2. Expediting Litigation:** The courts should not have to tolerate the delays caused by inefficient or unknowledgeable attorneys.
- **Rule 3.3. Candor toward the Tribunal:** Estimates of how much effort and time are required to process electronic records ought to be based on reasonably competent processing.
- **Rule 4.1**. **Truthfulness in Statements to Others:** Same as above.

Unless lawyers possess or retain others who possess such competency, the objective of Rule 1 of the Federal Rules of Civil Procedure cannot be obtained, namely, "to secure the just, speedy, and inexpensive determination of every action and proceeding."

Judges can help protect parties from technically incompetent representation for the sake of the clients and the integrity of the judicial process.

Recommendations:

1. As suggested throughout this guide, consider the effectiveness of the ESI processing tools employed by a lawyer when deciding cost-shifting applications, sanctions, and privilege waiver questions.
2. Start reporting flagrant offenders to state disciplinary boards or agencies. The Federal Rules of Civil Procedure e-discovery amendments have been in place for almost four years. The technical incompetence of lawyers may not end when judges report the incompetent practitioners, but it will diminish it drastically.

15. About the Authors

Anne Kershaw is the founder of A. Kershaw, P.C. // Attorneys & Consultants (www.AKershaw.com), a litigation management consulting firm providing independent analysis and innovative recommendations for the management of all aspects of litigation, including the management of data, records, and documents for discovery in litigation and investigations. Anne Kershaw is an expert on reducing costs associated with accumulated legacy data and electronic discovery.

Ms. Kershaw is also the President and co-founder of the E-Discovery Institute

Ms. Kershaw has been involved with high tech litigation management since 1993, taking management roles on national coordinating and trial counsel teams defending multi-state product liability claims.

In addition, Ms. Kershaw is a Faculty Member of Columbia University's Executive Master of Science in Technology Management Program, teaches at the Georgetown University E-Discovery Academy and is an Advisory Board Member of the Georgetown University Law Center's Advanced E-Discovery Institute.

She is a member of The Sedona Conference®

[32] Patrick Oot, Joe Howie, and Anne Kershaw, *Ethics and EDiscovery Review*, 28(1) ACC DOCKET 46 (2010). Available at www.eDiscoveryInstitute.org.

Working Group on Best Practices for Electronic Document Retention and Production (www.thesedonaconference.org), the Counsel on Litigation Management (www.litmgmt.com), the International Legal Technology Association (www.iltanet.org), the Association of Records Management (www.arma.org), and the National Association of Women Lawyers (www.nawl.org). Ms. Kershaw also co-chairs the ABA's E-Discovery and Litigation Technology Committee for the Corporate Counsel Litigation Section Committee (www.abanet.org/litigation/committees/corporate/home.html). She is also the author of numerous published articles, all posted on www.AKershaw.com.

Ms. Kershaw is a 1987 graduate, cum laude, of New York Law School (evening division) and a 1982 graduate of Bard College. She is admitted to practice in the United States District Courts for the Eastern and Southern Districts of New York, United States District Court for the District of New Jersey, United States Court of Appeals for the Second, Third, and Ninth Circuits, United States Supreme Court, and all New York State Courts.

Joe Howie is Director, Metrics Development and Communications, for the eDiscovery Institute and is principal in Howie Consulting. Mr. Howie practiced law with the Ohio Attorney General's Antitrust Section for three years and was in-house counsel for Phillips Petroleum for six years. He has written articles for a number of publications and organizations, including the ABA, ACC, ILTA, ALSP, *Law Technology News*, *InsideCounsel*, and others. His 1989 article for *California Lawyer*, "The Litigation Weapon of the '90s, Electronic Media Discovery," was one of the first heralding the advent of what we now call e-discovery or ESI.

Mr. Howie is a member of the Editorial Advisory Board for *Law Technology News* and writes a column for *InsideCounsel*.

For the last 20 years he has been involved in developing and marketing software and technology for the practice of law, including one of the first PC- and image-based litigation support applications, autocoding, linguistic pattern matching, and near duplicate detection software as well as legal publishing.

Mr. Howie holds a JD, cum laude, from Ohio State University, a masters in systems management from the University of Southern California and a business degree from the University of Michigan.

16. About the Electronic Discovery Institute

The Electronic Discovery Institute (or "eDiscovery Institute") is a 501(c)(3) nonprofit research organization. It conducts research and surveys to identify and provide metrics for technologies and processes that can reduce the cost of or improve the quality of processing ESI. Its research and survey results are available free of charge on its website, www.eDiscoveryInstitute.org.

If you have an idea on other surveys or studies that would help measure or promote cost-effective processing of electronic discovery or if you would simply like more information on the Institute, please contact us at info@eDiscoveryInstitute.org.

Appendix A: E-Discovery Technical Competency Quiz[33]

There is an ethical obligation to be competent in our representation of clients. If you are responsible for managing e-discovery as part of your corporate job responsibilities, you should be able to score at least 90 on this quiz. If your outside counsel control how your records are processed, they should also be able to score at least 90 on this quiz. (There are 11 questions worth 10 points each that a reasonably competent e-discovery practitioner should know and a more technical bonus question worth 20 points. Answers are shown following the quiz.)

Select the single most correct answer, 10 points for each correct answer:

1. One gigabyte of data
    a. may contain 25,000-500,000 pages, depending on the file types.
    b. always contains at least 1,000 pages regardless of data types.
    c. never contains more than 200,000 pages regardless of data types.
2. A *.exe file
    a. can be ignored safely for e-discovery purposes.
    b. could contain data.
    c. should be examined only if found in the "My Documents" folder.
3. DeNISTing
    a. refers to the use of data from the National Software Reference Library.
    b. should only be done if there is a concern that an investigation may have criminal consequences.
    c. is a seldom-used NASDAQ investigative procedure.
4. Deduplication
    a. means producing the minimum number of copies of an item if there are duplicates.
    b. means reviewing only the minimum number of copies of an item if there are duplicates.
    c. means having record custodians only give you one copy of their files if they have duplicates.
5. Vertical deduplication
    a. refers to consolidating duplicates within reporting relationships, e.g., within those custodians who report to the same person.
    b. refers to consolidating duplicates from within the records of a single custodian.
    c. refers consolidating duplicates from within individual branches or folders of drives.
6. Hashing
    a. refers to a process of scrambling a file for encryption purposes.
    b. refers to the process of cracking a password on an encrypted file.
    c. refers to a process of calculating a message ID for a file.
7. MD5 values
    a. will always be different for different files.
    b. ignore capitalization.

[33] This quiz appeared as Sidebar 2 to Patrick Oot, Joe Howie, and Anne Kershaw, *Ethics and EDiscovery Review*, 28(1) ACC DOCKET 46 (2010), published by the Association of Corporate Counsel. The quiz and the answers are © 2010 the Association of Corporate Counsel. *Reprinted with permission from the Association of Corporate Counsel 2010. All Rights Reserved.* www.acc.com

   c. are highly unlikely to be duplicated if calculated from two different files.
8. SHA values
   a. are more reliable than MD5
   b. are less reliable than MD5.
   c. are on a par with MD5.
9. Deduping ESI across custodians for review
   a. can, on average, lower review costs by 20 percent.
   b. decreases risks of inconsistent review decisions.
   c. lowers the number of people who are likely to have access to the records.
   d. speeds review.
   e. none of the above.
   f. Answers a, b, c and d.
10. Digital forensics
   a. typically involves copying a drive or device on a bit-by-bit basis.
   b. is required in all high-dollar litigation.
   c. requires the disclosure of the identity of the forensic examiner to opposing counsel.
11. Metadata
   a. are critical in most non-criminal proceedings
   b. generally require the services of a forensic examiner
   c. are sometimes used to refer to certain data within a file as well as data about a file maintained externally to the file.
12. (20 point bonus) RFC2822
   a. is an Internet standard that can be used to consolidate e-mails that were produced from different e-mail systems.
   b. Is an internet communications security standard used by the DOD and leading ESI providers for secure data hosting.
   c. Is an internet standard providing for secure VPN access.

Appendix B: Answers to E-Discovery Competency Quiz

1. A. Actual number of pages can vary widely depending on the file type. Some image files can contain many megabytes each. Simple text files may be very small—only one or two KB, i.e., about .000001 or .000002 GB each.
2. B. Users can save Word documents or other files with an ".exe" extension. To open them, they just change the extension back to the appropriate one, e.g. ".doc." The processing vendor should provide an exception report, listing all files with executable or unknown extensions, including their original file location.
3. A. The National Institute of Standards and Technology (NIST) is the umbrella organization for the National Software Reference Library, which contains file names and hash values for files distributed with many software packages.
4. B. Parties can dedupe for review purposes, but still produce multiple copies of each selected record.
5. B.
6. C. For example, the abbreviation, "MD5," was formed from the words, "Message Digest."
7. C. It is theoretically possible for two different files to have the same hash value, but it is highly unlikely.
8. A.
9. F.
10. A.
11. C.
12. A.

Appendix C: Summary of Deduping Survey Report

The deduping survey was conducted in September/October of 2009. It asked respondents to quantify the reduction in volume achieved by consolidating duplicate e-mails and electronic files within the records of individual custodians ("custodian-level" deduping or "vertical" deduping), and the savings that could be achieved if duplicates were consolidated across all the custodians ("project level" deduping or "horizontal" deduping). The results were:

Company	Reductions Achieved by De-Duping as a Percentage of Number of Original Documents in Culled Data Collections								
	Q. 9A Single-Custodian Deduping			Q. 9B Across-Custodian Deduping			Q. 9C Across-Custodian Compared to Single-Custodian		
	Avg.	Min.	Max.	Avg.	Min.	Max	Avg.	Min.	Max.
1. Act	15	10	20	25	20	35	10	10	15
2. BIA	25	10	50	40	15	75	15**	5**	25**
3. CaseCentral	19	3	21	51	44	56	32**	41**	36**
4. Clearwell	25	15	35	55	45	65	30	30	30
5. Daticon	36.8	0	74	46.8	0	79	10	0	5**
6. Encore	20	10	50	25	10	60	20	10	50
7. Fios	12	4	23	30	24	41	18	20	18
8. FTI Consulting*1									
9. GGO	20	10	50	40	10	65	20	0	15
10. Iris Data Services	13	4	58	33	20	90	23	16	76
11. Kroll Ontrack*2									
12. LDM Global	20			35			15		
13. LDSI – no b/u tapes	1-2		10-15	1-15	2-25	40	6.5**		27.5**
LDSI – w/b/u tapes	60	50	75	70	60	85	10**	10	10**
14. Rational Retention*3									
15. Recommind	5	0	20	30	10	60	25	10	40
16. StoredIQ	35	10	60	50	20	80	15	10	20
17. Trilantic	20	5	30	40	5	70	20	0	40
18. Valora	15	5	25	30	20	40	17.5	5	30
Totals	342	136	603	608	316	941	287	167	437
Number Responses	16	14	15	16	14	15	16	14	15
Average of Responses	21.4	9.7	40.2	38.1	22.6	62.7	17.9	11.9	29.2

Note: Double asterisks indicate the value was calculated from 9A and 9B as survey respondent did not provide the values for 9C

For footnotes, see full report.

Despite the clear savings from consolidating duplicates across custodians, it was only used in approximately half the cases (51.9 percent).

Appendix D: Further EDI Resources

This guide and the following resources are available in PDF format for free downloading at eDiscoveryInstitute.org:

Articles

Herbert Roitblat, Anne Kershaw, and Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61(1) J. AM. SOC'Y INFO. SCI. & TECH. 70 (2010).

Patrick Oot, Anne Kershaw, and Herbert L. Roitblat, *Practitioners' View: Mandating Reasonableness in a Reasonable Inquiry*, 87 DENV. U. L. REV. 533 (2010).

Patrick Oot, Joe Howie, and Anne Kershaw, *Ethics and EDiscovery Review*, 28(1) ACC DOCKET 46 (2010) (Considers ethical implications of failing to consolidate duplicate electronic records from the standpoint of obligations to clients, the courts, and opposing parties.)

Anne Kershaw and Joe Howie, *EDD Showcase: Crash or Soar?*, L. TECH NEWS, Oct. 2010, at 1 (About predictive coding.)

Anne Kershaw and Joe Howie, *Exposing Context: E-mail Threads Reveal the Fabric of Conversations*, L. TECH. NEWS, Jan. 2010, at 32.

Anne Kershaw and Joe Howie , *EDD Showcase: Highway Robbery?*, L. TECH. NEWS, Aug. 2009, at 30 (Discussing the ethics of what is, in essence, double-billing for duplicative attorney review of e-discovery that can take place without deduping across custodians.)

Survey Reports

Report on Kershaw-Howie Survey of E-Discovery Providers Pertaining to Deduping Strategies

Report on Kershaw-Howie Survey of E-Discovery Providers Pertaining to Email Threading

EDiscovery Institute Survey on Predictive Coding

a 501(c)(3) nonprofit research organization
303 South Broadway, Suite 430,
Tarrytown, NY 10591
866-471-3977
www.eDiscoveryInstitute.org
For information, email Info@eDiscoveryInstitute.org