

Chapter 13

Technology Considerations

William L. Farwell, Bruce V. Hartley,
Jeff Seymour, and Tony Reid*

- § 13:1 Introduction
- § 13:2 Strategic Issues
 - § 13:2.1 Continual Technology Developments
 - § 13:2.2 Dealing with Metadata
 - § 13:2.3 The Vendor Landscape
- § 13:3 Records and Information Management
 - § 13:3.1 Challenges
 - § 13:3.2 ERM/ECM Software Features
 - § 13:3.3 Questions for ERM/ECM Software Vendors
- § 13:4 Defining the Scope of Electronic Discovery
 - § 13:4.1 Challenges
 - § 13:4.2 Typical Scope Limiters
 - [A] Custodians (ESI Owners)
 - [B] Media Purposes
 - [C] Media Types
 - [D] Data Types
 - [E] Data Ownership
 - [F] Time Frames

* The authors are specialists with Deloitte Financial Advisory Services LLP who advise clients on technology concerns. The opinions expressed in this chapter are those of the authors and not necessarily those of Deloitte Financial Advisory Services LLP or any other Deloitte affiliate. In particular, the reader should not infer any Deloitte endorsement or recommendation of software produced by other parties.

§ 13:5 Preserving Data**§ 13:5.1 Collection As a Means of Preservation****§ 13:5.2 Challenges****§ 13:6 Collecting Data****§ 13:6.1 Challenges****§ 13:6.2 Making Collection Easier****[A] Imaging****[B] Archiving Collected Data****[C] Voice Mail and Video****[D] Chain of Custody****[E] Filters****§ 13:7 Culling Data****§ 13:8 Processing Data****§ 13:8.1 Challenges****§ 13:8.2 Note on Searching****§ 13:9 Reviewing and Analyzing Data****§ 13:9.1 Challenges****§ 13:10 Producing Data****§ 13:10.1 Challenges****Appendix 13A Metadata****§ 13:1 Introduction**

It is possible to manage electronic discovery well, so that it supports your case instead of hindering it. However, aspects of electronic discovery can be extremely technical, and doing an airtight job can require expertise not often taught in law school. Experienced forensic experts can advise lawyers and clients on the level of technology to use at each phase of the discovery process. The cost of guessing wrong can be significant—losing the case, personal sanctions against counsel, or both.

This chapter addresses the technology implications of electronic discovery for parties, their counsel, and experts. It does not presume any particular technical expertise on the reader's part.¹

After addressing some initial strategic issues, the chapter covers the technology issues that affect electronic discovery from start to finish, from the firm's basic management of its electronic records, through the (sometimes overlapping) stages of preserving, collecting, culling, processing, and analyzing data, to the final production of data to other parties or to government entities.

1. This chapter uses some technical terms. Some are in common usage, but others are not. For a glossary of technical terms used in electronic discovery, see *The Sedona Conference Glossary: E-Discovery and Digital Information Management* (The Sedona Conference Working Group Series, 2d ed. 2007), available at www.thosedonaconference.org/content/miscFiles/TSCGlossary_12_07.pdf (viewed June 14, 2008).

§ 13:2 Strategic Issues**§ 13:2.1 Continual Technology Developments**

Technology continues to evolve rapidly, and attorneys and their litigation support experts are obliged to stay current with developments in hardware, software, “middleware,” Internet tools, new media, counterfeiting techniques, artificial intelligence, and data management protocols. While the information in this chapter was current as of the date it was written, this is a fast-moving area of law. Here are just a few factors driving these developments:

- Moore’s Law² continues to operate, expanding storage and shrinking costs exponentially over time. This affects processing speed, memory capacity, and even digital camera resolution. So data requests that may have been unreasonable two years ago may be reasonable now in light of technological developments. What is inaccessible now may become easily accessible someday soon.
- Parties (and their employees and agents) may be storing important data in unlikely places that may still be “active” (and therefore discoverable). Jump drives, iPhones, PDAs, online backup services, GPS tracking devices, radio frequency IDs (RFIDs), implantable chips, and devices yet on the drawing board may contain discoverable data.
- Despite reasonable corporate policies governing regular destruction of corporate records (that is, in the normal course of business, and not under a legal hold), the ability to utterly destroy a record decreases every year. Paradoxically, electronic media make information both less tangible and more permanent.
- Creeping obsolescence is the dark side of technological change. “Legacy” media such as floppy disks and backup tapes deteriorate, and the hardware—and operating systems—necessary to read them also disappear over time. Even properly made backups on current media with state-of-the-art equipment may not be retrievable due to normal degradation of the content or imperfections in the storage media.

2. “Moore’s Law” is the popular name for a 1965 prediction by Intel cofounder Gordon Moore that the number of transistors on a chip will double about every two years. *See* www.intel.com/technology/mooreslaw/index.htm (viewed June 14, 2008).

§ 13:2.2 *Dealing with Metadata*

Metadata is a strategic issue because it can be so important to the overall electronic discovery process and may permeate each phase of the process so thoroughly.

To the question “What is metadata?” the answer is usually “data about data,” which is accurate but somewhat unhelpful. There are many types of metadata, including the following:

- **Document metadata:** Typically accessible by viewing “properties” from within a document; examples include file type, location, author, and statistics such as character and word counts.
- **Application metadata:** Information embedded within a document that may not be visible on the printed page; for example, tracked changes in Microsoft Word and formulas in Microsoft Excel (a/k/a “embedded data”). Application metadata is values or fields, created and stored by a software application, that describe the file.
- **System metadata:** Information stored externally to a file and used by an operating system to track file locations, usage, and other information; some can potentially be useful, such as dates created, modified, and last accessed, but much is digital clutter. Many defendants do not know that the “print spool” system files that a computer creates before sending a file to a printer persist after the file is printed—and may allow forensic experts to recover the file long after it has been otherwise thoroughly deleted from the computer.
- **Email metadata:** For example, dates sent and received, “cc” and “bcc” recipients, message routing information, and identification of attachments. Electronic discovery experts refer to the relationship between an email message and its attachment as “parent-child.”

The amount of metadata that automatically attaches to a document is surprisingly large, and its quality is surprisingly high. There are at least eighty easily accessible application and system metadata fields for each Microsoft Word, Excel, and PowerPoint document, excluding tracked changes. See Appendix 13A for the sorts of metadata that attach to a simple resume in a Word file on a desktop computer.

It is also surprisingly difficult to change the metadata on a file. People trying to hide data often may try to change the date stamp on a file by changing the system time on their computer, manipulating the file and saving it with the backdated time, then reverting to the current time for their computer’s system time. System files and some basic

forensic sleuthing can easily retrace these steps and reveal the real age and provenance of a document.

So metadata can turn out to be pivotal. Its presence or absence—and its careful preservation—pervade the electronic discovery process. It is an issue that should be considered and discussed by all parties early in the preservation process.

§ 13:2.3 **The Vendor Landscape**

Counsel making decisions on how to allocate resources to litigation support in a given matter, whether it is adversarial or investigatory, potentially have an increasing range of choices. The market for electronic discovery support is in flux, with both a proliferation of small niche vendors emerging on one hand and significant expansion of services by a handful of soup-to-nuts vendors on the other. Many smaller vendors focus on one part of the process, such as converting paper records using optical character recognition (OCR) and coding. Other large vendors have invested heavily in infrastructure to provide capacity, reliability, fault tolerance, and geographic availability.

Choosing the right vendor can be essential to a successful discovery outcome. For smaller, localized cases, a specialized vendor may be appropriate. However, for complex matters where discovery may help determine criminal penalties or fines, a larger, integrated provider may be more appropriate.

The following factors should be considered when choosing a vendor:

- **Existing state of company's information management system:** If a company already has a robust, well-oiled document management system in place and the scope of the discovery is small and well-defined, the need for expert help from a large, integrated vendor may be less obvious. However, if organizing, collecting, reviewing, and producing responsive files may be complicated by an ineffective document management system, wide geographic distribution of salient records, or potentially broad scope, the opposite may be true.
- **Time:** If you need to move quickly from scoping the problem to producing relevant files, then larger, more sophisticated vendors who can respond quickly may be a prudent choice. Larger vendors may be more likely to have offshore data processing resources that allow for around-the-clock work on an important case.
- **Chain of custody:** As with physical evidence in a civil or criminal case, it can often be critical to establish and maintain a clear, unbroken chain of custody for each file from every

custodian (discussed in more detail in sections 13:6.1 and 13:6.2 below). Such chains of custody can be easier to maintain and prove with a single integrated vendor than with a series of vendors retained to perform discrete tasks serially, who must pass files to one another (often with varying degrees of care).

- **Gravity:** The importance of the case to the company responding to a suit or agency action is an important factor, and one that attorneys and executive officers take seriously. The potential for criminal penalties can often focus officers' attention and may prove a determining factor in deciding in favor of a full-service vendor. As noted elsewhere in this treatise, attorneys are also subject to fines and sanctions for not responding appropriately to electronic discovery requests; this, too, can be a factor that may argue in favor of a larger vendor.

§ 13:3 Records and Information Management

Records management is itself a massive topic with both legal and technical implications. Generally, the more sophisticated a company's records and information management (RIM) program, the more easily the company can produce electronically stored information (ESI) in the discovery process. Having clearly articulated RIM policies and procedures that are actually followed can be the foundation of an effective records management process. Well before a litigation or investigation, a company needs to evaluate whether its corporate records management policies, especially for electronic records, are current for the specific industry and meet recommended and leading practices generally.

Companies today are increasingly looking toward technology solutions to help support their records management and electronic discovery programs. Traditionally, these programs focused on policies and procedures dealing with the physical management, retention, and disposition of stored records. However, today's complex business environment challenges companies to create legally compliant records management programs. The goal is to design and implement internal processes that are consistent, reproducible, and auditable for managing records through their entire life cycle of creation, active use, storage, and disposition.

§ 13:3.1 Challenges

A major challenge facing records programs today is that business support technologies have potentially made it much easier to create electronic data, but not necessarily easier to manage it. In response, records programs once geared to supporting paper records storage must

now manage an increasing array of electronic files and data being created company-wide.

These developments raise a number of issues:

- The need for consistent standards, policies, and procedures for creating, classifying, storing, and disposing of electronic documents enterprise-wide
- The need to comply with applicable federal and state statutes, rules, and regulations dealing with a record's use, storage, and eventual disposition, including new privacy and electronic discovery requirements
- The need to address a large amount of data that may not be considered a "business record" under current corporate records management policies
- The need to reduce storage and associated IT maintenance costs, including data archiving, migration from older systems, and system backups

In response, new technology solutions have come to market promising to help companies with these and other issues raised by the explosion of electronic data. Products getting the most attention are commonly called electronic records management (ERM) systems, though they are also referred to as enterprise content management (ECM) systems. An ERM system integrates records retention, workflow processes, storage management, email archiving, file versioning control, web content, and compliance and discovery components.

A key component of ERM/ECM systems is the collecting or linking of corporate records and information to a system repository where rules for accessing and retaining the data can be applied. To assist with this effort, an information architecture for both structured data (system data) and unstructured data (such as email, files on shared servers, and other data) should be created.³ Such a files architecture is commonly called a "taxonomy," where each record is classified, usually by subject, by the organization's business unit producing it, or by the function or process responsible for creating it. It is difficult to over-emphasize the importance of devising a clear, accurate, and flexible taxonomy that is easy to understand and implement and meets the needs of the system's users.

An ERM/ECM system's performance can be greatly improved if it is accompanied by the development of an enterprise-wide taxonomy for classifying the content to be stored on the system. The taxonomy provides the organization of information in an ECM system and is a

3. The distinction between structured and unstructured data is addressed in more detail below.

critical aspect of ERM/ECM deployment. It may be simple or extremely detailed, involving several layers of work processes and their resulting records. How records are stored on the system directly affects the ability of users to search, access, and make use of them quickly and efficiently, whether in the ordinary course of business or in an electronic discovery context. A system without a functioning taxonomy can increase search times, cause confusion, and increase costs when one is trying to access records from the repository.

ERM/ECM systems are increasingly adaptable to enterprise-wide applications rather than to the standalone departmental implementations which prevailed in the past. Integration with Microsoft Office products has now become the norm, as more and more records are being created in this work environment. Today's ERM/ECM systems are more compatible with various types of record file formats and systems, from data warehouses to imaging. They may also include features that help manage both paper and electronic records. Specific ERM/ECM products considered by a company should first be vetted to ensure that they are compatible with current business records systems. This can include any records management technology already in place.

§ 13:3.2 ERM/ECM Software Features

ERM/ECM solutions tout a wide range of benefits that can help companies establish and implement a comprehensive records management program, though solutions vary in robustness. Some of the features to consider include the following:

- The ability to manage records and information throughout their life cycle (life-cycle management)
- Advanced filing and searching capabilities that could make it easier for staff to locate needed information quickly and minimize the risk of losing or mistakenly disposing of required business records
- The systemic enforcement of a company's records retention policies and procedures
- An increased ability to meet relevant legal and regulatory requirements
- Potential storage cost savings and increased system performance by reducing duplicate files and unneeded system data
- Potential to reduce litigation risk by helping to identify and reduce draft copies and other transitory records
- Potential to reduce litigation costs by curtailing the volume of data that may need to be identified, preserved, collected, reviewed, and processed

Some software providers may state that their products allow for as much or as little human intervention as needed. For most ERM/ECM systems to function optimally, information must be declared as a record and classified by type before it goes into the repository where retention and other management rules can be applied. This means staff should take the time and effort to accurately classify records they produce or manage before they enter the repository—which, of course, can be tedious and time-consuming.

Auto-categorization is a system feature designed to use rules engines to identify and classify each record rather than depending on staff to do so. However, be aware that auto-classification success rates vary with each provider and product. As part of the provider due diligence process, obtain reports from the provider on success rates from implementations elsewhere. If the provider or product does not offer auto-classification, but you want that feature, check whether the product is compatible with add-on applications that have been developed to auto-classify records before they reach a repository.

§ 13:3.3 Questions for ERM/ECM Software Vendors

Here are questions to ask when evaluating ERM/ECM products:

- **What are our business needs and how well do the product features meet them?** First analyze operations to determine specific business needs the organization is seeking to resolve through the technology. These business needs can then be used in a request for information (RFI) to determine if specific products have the necessary features and functionalities, including electronic discovery, to meet these needs cost-effectively and efficiently.
- **Does the product integrate with other applications and systems already in use?** More ERM/ECM providers are offering “federated” records management. Federation allows for management of content in the system in which it was created, without moving it into a separate system repository. Federated records management allows for more flexibility when laying out the ERM/ECM system. Also, note that some email archiving solutions may work well with Microsoft Exchange or Lotus Notes but not necessarily with both. (You may need to accommodate legacy systems resulting from past or future acquisitions.)
- **What record formats does the product support?** ERM/ECM products support most file types. Determine what file types are currently being created and stored on your systems and check to ensure their compatibility with any product being considered.

- **Does the product meet industry-specific compliance requirements?** Requirements are emerging in discrete industries, such as brokerages and financial services, for managing specific types of electronic records, including the creation of metadata and system security. To show that their products meet these requirements, providers can seek certification under standards such as Department of Defense (DoD) 5015.02-STD, which federal government agencies must use when purchasing document management products. The standard includes highly detailed system features and requirements that must be met before the product can be considered for purchase by the agency. Other companies are not required by any regulations to purchase DoD 5012.02-certified systems. However, a company may want to use this standard as a general guide for purchasing a system in the future.
- **What physical records management capabilities does the product offer?** Today, most but not all records are created electronically. Companies may need to track paper files in offices and boxes sent to off-site storage through a bar code, radio frequency identification (RFID) system, or similar means. Any ERM/ECM product considered should have capabilities to integrate with current paper record storage technologies.
- **How easily are retention schedules and other records policies integrated into the system?** Some products may be more robust than others in this area. Some may permit classifying records into “retention buckets” such as one, two, five, or ten years, while others may allow for specific retentions for record types, including factoring in triggering events, such as when a claim is settled.
- **Does the product have retention management features?** Robust retention management capabilities allow for retention rules to be placed on most content, regardless of whether it has been declared an official record in the system. Current products may offer varying degrees of retention management; generally, cost increases with functionality.
- **What is the product’s records management repository architecture? How are files checked in and out?** Architectures and procedures for accessing and refiling vary significantly in sophistication and ease of use.
- **What are user experiences with the product?** Ask current system users with similarly sized installations to discuss their experiences. Look into established products with a track record for meeting the same types of business needs that you have identified in the RFI process.

- **What training support is offered?** Staff training is essential to any project's success. Trained staff are more comfortable using systems—and may tend to use them more often and more effectively—than staff who do not receive appropriate training. The content and timing of training offered by the product supplier should be appropriate to the user and should include ways to reinforce what is taught, such as individualized follow-up with users.

Firms selling ERM/ECM products with electronic discovery components are proliferating. Conversely, industry consolidation in some sectors has resulted in more sophisticated products with specific features to help implement comprehensive records management and e-discovery programs.

§ 13:4 Defining the Scope of Electronic Discovery

The goal for the scope-definition phase of electronic discovery is to develop an explicit strategy for identifying, locating, and retrieving discoverable data. Ultimately the goal is to conduct a transparent, defensible, auditable, and repeatable process.

Careful, realistic scoping may be equally important to all parties for both legal and technical reasons. When done correctly—and cooperatively—it may have at least three benefits:

- (1) It can be essential to limiting document collection to a manageable size. No one benefits from the production of masses of useless data.
- (2) It can help parties project costs more appropriately, including whether to retain outside experts and what caliber they should be.
- (3) It may constrain the coverage of the legal hold each party may face, limiting it to information that could potentially be discoverable and freeing other resources to be used in the regular conduct of business.

Note that, although defining the scope appears as the first step in the electronic discovery process, it can be iterative. As a case develops—as issues and focus change, and new data and custodians can be identified or removed—the scope of discovery can expand or contract.

§ 13:4.1 Challenges

The major challenge, of course, is the ever-growing volume and variety of data to be considered. Unless your corporate records management system has been designed to anticipate litigation, it can be

extremely difficult to compile an inventory of all data types and how each type is created, secured, stored, retained, deleted, transferred, modified, and archived.

It is difficult to exaggerate the speed at which data creation is expanding, and the extent to which it is subject to discovery. In a perverse corollary of Moore's Law, every eighteen to twenty-four months, companies face a *doubling* of data they produce. Some statistics illustrate the trend:⁴

- Over 90% of all information is now electronic, and 70% of electronic documents are never printed.
- Over 30 billion emails are sent daily, with the total for 2007 estimated at 11 *trillion*.
- Employee email has been subpoenaed at one in five U.S. companies.
- On average, a Fortune 500 company has 125 ongoing cases, with at least 75% requiring e-discovery.

Meanwhile, the typical 40GB hard drive on an individual laptop or desktop can hold:

- 20 million typed, double-spaced pages (without images), or
- 100,000 photos, or
- 8,000 CD-quality songs.⁵

Technology systems and applications can pose a constellation of problems. It is common for a large corporation to have more than 8,000 applications and systems, each with a specific owner and universe of data.

§ 13:4.2 Typical Scope Limiters

The scope of discovery can be limited by applying a number of filters. It may be helpful to classify the potentially relevant data in a matrix applying some of the following filters separately.

[A] Custodians (ESI Owners)

Identify "key players" and "custodians" to prioritize the need to preserve and collect data. Key players are those people who have an

-
4. Rebekah Anderson, *E-Discovery: New Rules Present New Challenges and Opportunities*, MALL NEWSLETTER (Minn. Ass'n. of Law Libraries, St. Paul, Minn.), Jan./Feb. 2007, at 8 (quoting Bruce Hartley of Deloitte Financial Advisory Services LLP), *available at* www.aallnet.org/chapter/mall/news334.pdf (viewed July 15, 2008).
 5. From a calculator on "I'm Afraid I'll Run Out of Disk Space," www.coolnerds.com/Newbies/Fear/hddFear/hddFear.htm (viewed July 17, 2008).

obvious, direct link to the matter being litigated or investigated, while custodians are those who have access to potentially relevant data without necessarily being key players. This could help determine the breadth of the legal hold needed.

[B] Media Purposes

Distinguish among work use, personal use, and disaster recovery. However, note that personal files, including files on home PCs and other media devices, are sometimes very relevant to the case. Generally, disaster recovery files may be deemed “inactive” or “not reasonably accessible” and not normally within the scope of discovery.

[C] Media Types

Classify data sources from hard drives, company servers, and backup tapes. Different media can be backed up in different ways (remotely, in the background, or even surreptitiously). Some media, such as backup tapes, may be of secondary importance if relevant files are still active.

Try to determine early in the discovery process the other media types that may be implicated in the discovery, such as flash drives, memory cards, memory sticks, cell phones, iPods or other MP3 players, personal backup media, GPSs and other handheld devices (including PDAs and BlackBerry devices that may or may not sync with corporate applications such as Outlook or Notes), voice mail, personal email accounts, instant messaging (IM) accounts, home computers, and data held by third parties (including online backup services).

[D] Data Types

It may be useful to make a primary source classification among (1) emails and user files; (2) system files; (3) financial and operational databases; and (4) other types (web, proprietary). (However, note that a proper taxonomy for records management purposes may be content-based, not source-based.)

But you may also want to classify data as to whether it is “unstructured” or “structured,” as that may affect not only how you collect data but what sort of expert help you may need to collect, process, and analyze it.

“Unstructured” data is, generally, data that does not fit into set rows and columns. It may include (1) emails, audio, and video files; and (2) other user-created files such as word processing, spreadsheet, and presentation files residing on network file shares, PC hard drives, or loose media (CDs, DVDs, or USB drives). It has been estimated that more than 85% of all business information exists as unstructured data—commonly appearing in emails, memos, notes from call centers

and support operations, news, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations, and web pages.⁶ Historically, unstructured data comprised primarily hard-copy documents. Typically, attorneys (usually outside counsel) managed the collection, scanning, coding, and review of those documents. But now most documents are electronic, requiring different technologies (and personnel) to collect, analyze, and review them.

“Structured” data describes databases of information. These databases contain the numbers that accountants and analysts work with most frequently; in fact, even when the underlying records (such as personnel files) are stored in hard copy, companies may create a database for them by coding appropriate fields. The transactional detail that comprises systems of record, that describes who paid what to whom and when, who booked what, etc., is stored in these structured databases. Also, importantly, these numbers usually drive damage calculations and financial restatements.

Even though most corporate information may exist as unstructured data, companies can compile or create very large amounts of structured data in databases; in a 2006 white paper, IBM reported that one client had compiled more than 30,000 databases through its collaboration system over ten years. So, managing structured data can become a critical part of the electronic discovery process. Obviously, there is no way to print out the content of 30,000 databases for a meaningful review by counsel.

[E] Data Ownership

Classifying data ownership as either individual or shared can help determine the scope and method of collection, and how to analyze it.

[F] Time Frames

Some data is specific to a period in time (historical), while other data is continually being created or modified (prospective or dynamic).

§ 13:5 Preserving Data

Preservation in the electronic discovery context generally refers to taking steps necessary to ensure that data potentially responsive to a matter is not altered, deleted, or destroyed. This can typically commence with a legal hold—notification of all identified custodians regarding the matter at hand and their duty to preserve data potentially responsive to it. A legal hold notice requesting custodians to

6. Robert Blumberg & Shaku Atre, *The Problem with Unstructured Data*, DM REVIEW MAGAZINE (Feb. 2003), www.dmreview.com/issues/20030201/6287-1.html (registration required) (viewed June 29, 2008).

preserve files and records should be prepared on advice of counsel and should be transmitted quickly, effectively, and as widely as necessary (note the risks of preserving too broadly, discussed below). The legal hold should be based on a sound policy, and in order to be effective, the policy should be in place with supporting procedures before a preservation requirement arises. It can also require supporting IT infrastructure that ensures custodians have the technical means and associated training to comply with the requirements of the legal hold. In some cases, IT should take measures to preserve data so that custodians cannot alter, delete, or destroy data by mistake or even willfully.

The goal of preserving data is to ensure that data is valid and that original data has not been changed, so that it can be used as evidence in a court of law. To a large extent, many of the same concerns and principles that apply to maintaining evidentiary integrity in a criminal case can apply to electronic data preservation. IT personnel, including the system administrators, should not perform preservation without input and guidance from a forensic examiner. Errors (even innocent ones) made at this stage—such as choosing the wrong type of copying tools, or the wrong media—can be extremely expensive, or impossible, to correct.

In addition to the preservation of files existing at the time of the legal hold, it may be necessary to have processes to properly preserve newly created files. This is another area in which advance planning can make all the difference in litigation preparedness.

Properly performed preservation can save money by avoiding sanctions and limiting collection costs. More importantly, it can also reduce downstream processing, hosting, and review costs. It can maintain data integrity, preserving data for later review, appeals, or investigative activities. It can also preserve data for use in other adversarial situations, so counsel may want to consider the length of the preservation stage.

§ 13:5.1 *Collection As a Means of Preservation*

Effectively preserving data can be challenging. Often the simplest way to effectively preserve data with assurance that it cannot be altered, deleted, or destroyed can be to create a copy of it in a forensically sound manner. However, as discussed below, using collection as a means of preservation can increase cost and risk. Technologies designed to address this problem have been evolving, and the problem is getting attention from some enterprise-scale software providers.

Over the past few years, an increasing number of enterprise collection tools have become available. Generally, they allow for more targeted selection of potentially responsive ESI through the use of

filtration or search utilities at the collection source, coupled with an ability to execute the collection process from a central location on remote sources (PCs, file shares), and to copy the resulting data back to a central collection point over the network. The capabilities are now being extended to allow for the locking down of designated data in place (“hold in place”). This capability has great potential for reducing cost and risk by allowing a conservative preservation approach without the negative consequences associated with collection as the means of preservation. Generally, the faster collection can be decoupled from preservation, the better, and technology landscape is improving in this regard.

§ 13:5.2 Challenges

Preservation can be challenging because circumstances often work in favor of preserving more rather than less data. Taking steps to preserve data can require scope definition, which typically means identifying custodians, their associated data sources, and potential indirect data sources such as a financial system or an intranet site. Often, a responding party has limited information, and scope definition can require judgment calls by counsel, technologists, and business people. Moreover, ESI is constantly in flux. If you miss something today, you may not be able to go back tomorrow and capture it (for example, consider email auto-deletion protocols). As a result, it is common to take a conservative approach to preservation, and at the outset cast a fairly wide net.

Moreover, the most common approach to preserving data is to *collect* it—that is, to generate a copy of it. Together with a broad, conservative preservation approach, this factor can increase both cost and risk.

The cost can be fairly obvious—once you collect the data, it gets processed, hosted, and reviewed. The more data, the higher the cost can be, and the cost tends to increase at each step, with review typically much more expensive than any other component.

The risk may not be as obvious. If you over-collect to play it safe for preservation purposes, it can be almost impossible to destroy the collected data during the lifetime of the matter that prompted its collection. Collected data may otherwise have been subject to disposition in the normal course of business, or may not have been subject to any document retention requirements at all. However, now it cannot be destroyed and it becomes a source of ESI that should be considered under future preservation requirements.⁷ Preservation of backup tapes,

7. The vast majority of email may fall into this category: it is not a business record, but as soon as it is collected pursuant to a matter, it has evidentiary value.

which typically contain only a small percentage of ESI relevant to a matter, can be the extreme example of this. Once you hold them past their useful lives for business continuity and disaster recovery purposes, they are by definition archival. In addition, they may present substantial risk, in terms of both their content and the potential cost to restore, process, host, and review the data pursuant to future matters.

There are many other challenges to effective preservation, including the following:

- The preservation duty can vary according to the matter at hand and the relevant rules (the Federal Rules of Civil Procedure, state rules, instructions pursuant to a subpoena by a regulatory body, etc.), although many situations have common elements.
- IT operations in even a small company can be complex and far-reaching. It can be difficult to find a chief information officer who really knows, at a granular level, how every process works, what defaults are used for backup, and what safeguards are in place to protect or preserve data on short notice. Fortunately, the 2006 amendments to Federal Rules of Civil Procedure provide some limited protection against sanctions under certain circumstances where data is lost as result of “routine, good faith operation of an electronic information system.” However, the proper implementation of a legal hold is an essential element of “good faith.”
- Spoliation (withholding, hiding, or destroying relevant evidence) can lead to sanctions even if it is inadvertent.
- Electronic evidence can often be volatile. Data within corporate systems (email, accounting systems, intranet sites, Wiki sites, etc.) is continually modified in the normal course of business. Potentially responsive data can exist in a wide array of locations and on a wide array of devices (PC hard drives, network file shares, USB drives, cell phones, PDAs, iPods, CDs and DVDs, backup tapes, etc.).
- Metadata can be sensitive and can be changed *in the process of collecting it* if proper tools and methods are not employed.
- Legal holds that preserve litigation-related data can interfere significantly with daily business operations, including creating a sudden need for additional storage and alternate means of backup.
- The process can often involve personnel unfamiliar with litigation or forensic procedures, including the custodians (people with relevant data in their control) and IT staff.

- Noncompliance can give opponents access to parts of your information system that could have been kept private had the preservation process been performed correctly.
- Normal automated records, document management, email processes, and backup tape rotations may need to be suspended.

One of the biggest challenges can be balancing a conservative, good faith, well-defined response with the internal and external costs of compliance. Once the duty to preserve has been established, the duty to lock down data can become critical and can be a burden on potential custodians, the IT department, and others in the organization.

§ 13:6 Collecting Data

In this phase, the goal is simple but the task can be complex. The goal is simply to gather data that is potentially responsive to the discovery request. Doing so, however, can be tedious, expensive, confusing, protracted, and overwhelming.

Typical collection is done one custodian at a time, one machine at a time. Advances in technology now allow for multiple collections to be completed simultaneously in a networked environment. Both full forensic images and targeted logical file collections can be performed in a forensically sound manner over a company's network with a variety of tools. However, because of the inconsistency in network bandwidth from company to company, if a large number of full forensic images are needed, they may have to be done physically, rather than over the network. If set up properly, with the right number of collection staff and sufficient equipment, physical collections can be very efficient. The custodian's "home directory" and other shared network drives should not be overlooked during their collection.

In a properly executed collection phase, responsive data should be stored in only a few places, if possible. Chains of custody should be maintained and documented carefully. (Experts have developed sophisticated multi-pocket envelopes that allow for custodial information to be updated without impairing the integrity of the underlying evidence files.)

§ 13:6.1 Challenges

Some of the challenges in the collection phase include the following:

- **Finding the right IT staff:** Turnover in IT departments can be high, duties within the department can be amazingly specialized, and far-flung enterprises that have grown through acquisition may have dozens of incompatible legacy systems that are

imperfectly integrated. Counsel or the experts they retain should interview senior IT staff early, starting with the CIO, and establish a roster of who knows what and who has access to the systems, networks, servers, and backup media that may be affected by the collection process.

- **Data over-collection:** Parties often simply collect too much data in an excess of conservatism. Sometimes this can backfire, as a review of files in response to a specific case can uncover pornographic materials, which (if they include child pornography) can require the reviewer to report the findings to the appropriate authorities for criminal prosecution.
- **Preserving chain of custody and evidence handling:** Again, metadata can be as important as the underlying data, and if the wrong method is used to collect data, its related metadata can be lost or corrupted.
- **Data management:** The volumes can be overwhelming, ranging into terabytes (1,000 GB, equivalent to 50,000 trees reduced to paper) and beyond.
- **Data granularity:** For every medium identified, you may need to collect and track serial numbers, an evidence barcode, an MD5 Hash identifier (a thirty-two-digit cryptographic hexadecimal code), acquisition forms, sector information, a file count by size, and passwords.
- **Media compatibility:** Tape backup systems can vary surprisingly. In addition to four separate types that IBM supports, you may find DLT, SLR, DDS/DAT, Mammoth, IBM 3590, 3570, 3480, 3490E, Travan, AIT, VXA, and LTO Ultrium. Vendors who process tape data should be able to access the data on these various data formats.
- **Nonstandard data:** Structured versus unstructured data, in addition to a growing array of file formats (discussed below), can make classification a challenge. In addition, files can be active, deleted, hidden, encoded, password-protected, or encrypted—so the vendor managing data collection should have the tools to get past passwords and break both general and government-level encryption.
- **International rules:** Many countries, particularly in the EU, have strict rules constraining the transfer of personal data across borders.
- **Numerous file types and extensions:** File extensions are codes of two or more letters that appear at the end of a computer file name that tells the relevant operating system (such as Windows,

Mac, or Linux/Unix) what kind of file it is dealing with.⁸ However, there are literally hundreds of other extensions, with new ones being created continually.⁹ One of the complications in forensic discovery can be that people trying to hide data (or simply to get data through firewalls) can change extensions—such as changing an .exe extension to .doc to fool a firewall into letting through an executable program file that would otherwise be blocked. Fortunately, each file type has a fingerprint that cannot be changed, and the proper forensic software can scan large clusters of data files quickly and identify any that have been renamed from their native format.

§ 13:6.2 Making Collection Easier

As part of the range of collection options, parties and their counsel (and experts) should decide and agree on some collection options to cull the data to a workable size. Some useful options are discussed below.

[A] Imaging

The gold standard for imaging a drive or server for later evidence handling is to do a forensic image, either full or logical. A “full” forensic image copies a drive sector by sector, bit by bit, including erased and overwritten sectors. A “logical” image copies only the active files on a drive. Guidance Software’s EnCase eDiscovery¹⁰ is a widely used software tool for making drive images; in the authors’ experience, it is a simple tool with intuitive GUI, excellent analytics, strong email/Internet support, and a powerful scripting engine. Using it, one forensic examiner can potentially make ten to twenty full forensic hard drive copies at a time in two to five hours, depending on drive size. The image is made with “self-validating” blocks.

-
8. Common file types and extensions for traditional Microsoft Office files are .doc for Word files, .ppt for PowerPoint files, and .xls for Excel files. However, even Microsoft has muddied the waters by changing its own file extensions for Vista/Office 2007 files—they are the same as the ones above, but with an “x” at the end. WordPerfect holdouts will recognize .wpd as the extension for those files, and .rtf and .doc are also extensions for word-based files with limited formatting. PDF files are everywhere, but did you know that the .pdf extension stands for the system-agnostic “Portable Document Format”? There are also the main file extensions for web-based content and images: .htl, .html, and .asp for web pages; .com, .org, .gov, and .net as the main domain types; .jpg, .gif, and .tif for image files; and .mp3 and .wav for video files.
 9. See www.file-extensions.org for a huge library of them, including advice on how to open them, convert them, and back them up.
 10. For details on EnCase eDiscovery software, see www.guidancesoftware.com/ediscovery/index.aspx.

The natural tendency is to choose some imaging standard lower than a full forensic image to save on electronic discovery costs; while sometimes this can be defensible, the initial cost savings can evaporate if, months into the litigation, it becomes clear that hidden content that was never imaged has become material to a case.

[B] Archiving Collected Data

One of the leading practices is to make two encrypted copies of data sets simultaneously so one can be preserved off site while the other is used as the source for making working copies. When third parties are used to handle the archival copies, you may want to confirm their status periodically to ensure that the records match actual inventory and that files to be destroyed are actually destroyed on schedule.

[C] Voice Mail and Video

These media can quickly expand the scope of collection, and, to date, there are few helpful ways of speeding up the scanning of voice mail or video for keywords, dates, or other metadata. Depending on their relevancy to the case, you may consider trying to negotiate with the other party (or the fact-finder) to limit the scope of discovery of voice mail and video files for cost reasons.

[D] Chain of Custody

Try to balance the importance of maintaining strict chains of custody for every bit of data with the cost to the party that needs to maintain them. Just as a forensic expert does with tangible evidence, good forensic evidence practice includes keeping a “bagged and tagged” copy of the data (in an anti-static bag) for six years after the final disposition of a case.

[E] Filters

You should consider applying filters (keyword or Boolean searches, dates, key players, custodians, data types) early to cull files and limit the scope of collection. Increasingly, parties are using concept analytics, keyword searches, and folder analysis (email relegated to project folders by categories or dates) to narrow the scope of discovery and cull irrelevant files.

§ 13:7 Culling Data

Properly collected data can be “culled” for obviously nonresponsive and redundant documents. Parties can filter out system files, “de-dupe” files in batches (removing exact duplicates), and filter files by date or other variables.

Culling can be done as part of the collection, before processing, or during processing.

Done effectively, data culling can dramatically reduce the volume of data that needs to be processed, potentially reducing costs. One of the drawbacks to culling data during collection or before processing can be that, if the scope of discovery later expands, it may be necessary to go back to the original data set to expand the original scope. You should remove system and other known program files from the data collections unless they are relevant to the matter.

De-duping can be done across the population or by custodian. One of the risks of de-duping across a project can be that you may not realize a particular custodian is related to a matter if you include only one copy of the document in the production.

§ 13:8 Processing Data

Once files are collected and culled, they should be processed, searched for privilege, and sorted by type or topic, and in some cases broken into smaller units for review. One vendor estimates that 80% of the time and 80% of the cost of electronic discovery can be spent in processing, review, and analysis, so having powerful workflow tools to manage this part can be critical to managing scarce resources.

Once again, the goal can be deceptively simple—to create a data set that can fully support the attorney review process. Achieving that goal can take concentrated effort and special expertise.

Done effectively, data processing can result in consistent file formats and fully searchable data sets that enable attorneys to find, sort, mark, highlight, and reproduce evidence that may support their case or defense.

Generally, the litigation support industry has moved to “native” file review and production, especially for structured data such as number-intensive databases. There may still be reasons for producing static (TIFF or PDF) images of each file, but generally, a hybrid approach is common, especially for production purposes.

§ 13:8.1 Challenges

The major challenges at the processing stage include:

- **Maintaining parent-child relationships between email messages and their attachments:** It is common for the same email, with one or more attachments, to go to many users—direct recipients, “cc” recipients, and “bcc” (blind copied) recipients. It is also common for recipients to reply and forward the message, with or without the same attachment (or a version of it). Trying to record and view who saw what, or when they saw it, can be daunting.

- **Exploding all archives:** To save space, many senders compress archives of files that have .pst, .zip, or other extensions. Again, determining whether each archive in a set of messages is identical or merely similar can take time and resources.
- **Foreign character sets:** As more companies expand outside the United States or do business in languages other than English, maintaining records that capture information in non-English languages can add a layer of complexity.
- **New media types:** As noted above, an explosion of media types can make it challenging to create data sets that attorneys can review. For example, the need for saving and identifying voice mail messages is becoming more common and can create challenges during the collection and review process.
- **Cost:** Estimating the cost of proper processing of electronic discovery files can be more art than science, and it can be iterative. Once the initial claim and scope of a case is known, a general estimate can be made, but that estimate can shrink or grow, depending on how the early stages of the case progress and whether the parties are successful at limiting the scope of discovery.

§ 13:8.2 Note on Searching

The ability to perform robust searches on data is critical to managing electronic discovery. Tremendous search technology is available. For example, some types of software allow for robust workflows with many users to access data in batches, and other software provides powerful pattern-recognition tools that can apply keywords to limit data sets dramatically. Ideally, searches to cull data should be done early and often. (However, there may be a risk of excluding a file that is actually useful, and having to go back to the well if litigation circumstances change.)

Filters are commonly applied in the collection, processing, and review phases. While filtering is primarily a back-office process, it should meet the needs of the attorneys analyzing the files. Since all parties may have access to filtered data, it may be prudent to negotiate the key terms with the other parties and counsel, and to document the exact nature of the filters applied. You may also want to agree on the best way to access structured data (such as financial and operational databases and reports), as those can not be reduced to “flat” formats such as TIFF and PDF files.

§ 13:9 Reviewing and Analyzing Data

The goal of this phase is to perform legal analysis on the data, including classifying it as to privilege, and using it to understand the facts and circumstances of the case. Attorneys and other forensic

experts should be able to examine a large body of documents, quickly determine their contents by review, categorize each document, or quickly locate key documents of interest. Effective review and analysis can provide powerful tools for understanding what documents contain and how they relate to each other.

Back when most evidence was paper-based, this part of the process was more linear, and more tedious: It required legal staff (attorneys or paralegals) to review, abstract, and synthesize reams of content and sort it by privilege and salience.

Now that most data is electronic, the process may be less linear, but it is also arguably less tedious. The ability to screen out irrelevant information electronically can leave a smaller universe of relevant documents and files that need the direct attention of a legal professional.

A number of vendors provide high-level software to simplify the review and analysis of ESI. Careful consideration should be given to finding the tools best suited to counsel's needs, which may differ from case to case.

If the other phases of the process have been conducted carefully, using the latest software tools to speed up and simplify each phase, the review process can benefit. Attorneys can conduct simple, complex, and ad hoc reviews to locate relevant data. Proper tools can also allow for mass tagging of data as privileged and allow for simple, effective ways to manage documents to streamline attorney review—and for complete, quick reports to manage the process.

§ 13:9.1 **Challenges**

Review-phase challenges may include the following:

- **Variety of review tools:** There is no single best review tool. The competing products generally rely on keyword and Boolean searches, while some have increasingly sophisticated “fuzzy” search capabilities that would, for example, find all references to “cars” when you search for “autos” or “vehicles,” as well as misspellings such as “aotomobiles.” Others may have conceptual search agents with good graphical user interfaces that can help attorneys evaluate a case's merits and potentially simplify complex data sets.
- **Redaction:** Some files need to be redacted pursuant to a claim of privilege. High-quality tools allow for thorough redaction, but occasionally mistakes happen and put privileged content in a public space.
- **Structured data:** When a case's legal issues rely heavily on structured (database-driven) data instead of unstructured, “flat” files, attorneys may rely on other experts to explore, review, and interpret the data for them.

§ 13:10 Producing Data

Generally, the goal at this phase is to be able to deliver data in a useable format to other parties, to a court, or to a regulatory agency.

Final productions are still often image-based, in the form of “load” files—files with fields delimited by a character (such as “comma delimited”) that can be read into another application. A party using Concordance software, for example, generally gets a data file and an image file. Vendors can support more than a dozen litigation-support applications.

“Mixed mode” productions can help save cost and allow a fairer review of structured data. For a mixed mode production, most files (such as Word, PDF, emails, and PowerPoint presentations) may be converted to 300-dpi images (TIFF files), but for other files such as complex Microsoft Excel files, “native” files would be produced. Producing native files allows the other party to see formulas and otherwise hidden content that would not be revealed in TIFF production.

Whatever the production strategy, files are generally delivered on ordinary media such as CDs, DVDs, or hard drives. Most litigation support software allows you to impose electronic Bates stamping or a confidentiality stamp on each page. Parties have been known to wrangle for days on the exact wording of confidentiality legends, which can be as long as five lines. Depending on the number of parties and the review software that they have access to, it may be necessary to produce load or mixed-mode files in more than one proprietary format.

In the production phase, reviewers view, work with, and search documents in batches to filter out the needed information. Information can sometimes go through up to six levels of review before the remaining, processed information is sent to attorneys and subject matter experts for analysis. Approved documents are put into exportable sets, ready to be sent to the client.

The data exchange between parties is normally fluid, continuing in rounds as additional cohorts of data are found, screened, analyzed, and reviewed. With multiple parties, the logistics can get proportionally more complicated.

It can be essential to negotiate early with the other party and its counsel (and experts) to clearly delineate production requirements. For example, an early decision should be reached on whether to stipulate to native or nonstandard production requirements. The experts should leverage the most advanced tools to track productions in complex litigation environments.

§ 13:10.1 Challenges

Below are some of the challenges in the production phase:

- **Whether to produce in native file format:** Producing files in the same format they were created in may be the easiest way to

review them. The “viewing” party can see just what the “creating” party saw, including embedded information, but there are serious drawbacks:

- The reviewing party may need to buy, load, and be able to operate the software that generated the native file.
- Evidentiary integrity may not be preserved if the files are produced in native format. Without hash codes to confirm that no changes are made, one party may not be able to prove that the other party changed a file before presenting it to the finder of fact.
- Redaction and Bates numbering may not be possible with native files.
- Native files may not be searchable across file types, as processed files are.
- **Deciding on a review platform:** Reviewers may know and can be comfortable with industry standards (for example, Concordance, Summation), but the selection process can be complicated because different software products excel in different features. Not all applications can manage complex productions well.