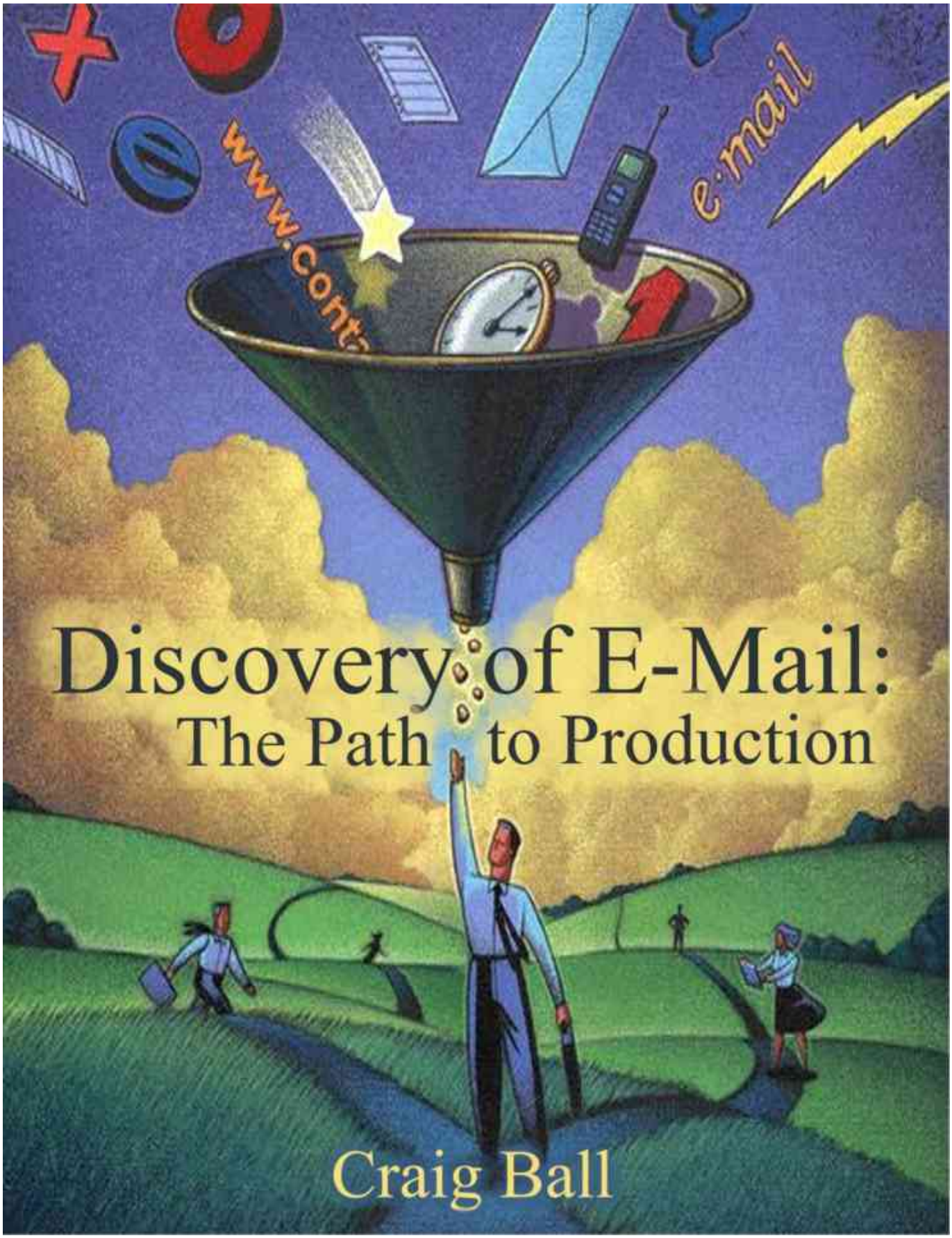

Discovery of E-Mail:
The Path to Production

Craig Ball

Discovery of E-Mail: The Path to Production

Part I:

Asked, "Is sex dirty," Woody Allen quipped, "Only if it's done right." That's electronic discovery: if it's ridiculously expensive, enormously complicated and everyone's lost sight of the merits of the case, you can be pretty sure you're doing it right.

But it doesn't *have* to be that way.

Let's walk a path to production of electronic mail—certainly the most common and perhaps the trickiest undertaking in electronic discovery. The course we take may not be the shortest or the easiest, but that's not the point. We're trying to wend our way to production without stepping off a cliff. Not every point discussed is suited to every production effort, but all deserve *consideration* every time.

Think Ahead

False starts and missteps in electronic discovery are painfully expensive, or even unredeemable if data has been lost. You can avoid re-treading ground by establishing expectations at the outset.
Will the data produced:

- *Integrate paper and electronic evidence?*
- *Be electronically searchable?*
- *Preserve all relevant metadata from the host environment?*
- *Be viewable and searchable using a single application, like a web browser?*
- *Lend itself to Bates numbering?*
- *Be easily authenticable for admission into evidence?*

Like a videogame where players gather keys and tools for later use, meeting these expectations in e-mail production hinges on what you collect along the way through *identification, preservation, harvest* and *population.*

Identification

"Where's the e-mail?" It's a simple question, but one answered too simply—and erroneously--by, "It's on the e-mail server" or "The last sixty days of mail is on the server and the rest is purged." Certainly some of the e-mail will reside on the server, but just as certainly more, even *most*, e-mail is elsewhere, and it's *never all gone* notwithstanding retention policies dictating it disappear. The true location and extent of the e-mail depends on systems configuration, user habits, back up procedures and other hardware, software and behavioral factors. This is true for mom-and-pop shops, for large enterprises and for everything in-between.

Consider a recent case where I was asked to assess whether an associate quitting a law firm stole files and diverted cases. The firm used a Microsoft Exchange e-mail server, so I could have collected or searched the associate's e-mail there. Had I looked only at the server, I would've missed the Hotmail traffic in the Temporary Internet Files folder and the Short Message Service (SMS) exchanges in the PDA synchronization files. Then, there was the Microsoft Outlook archive file (.PST) and offline synchronization file (.OST), both stored on a laptop hard drive and holding thousands more e-mails. Just

looking at the server wouldn't have revealed the stolen data or the diverted business, but searching for e-mail in some of the other places it hides uncovered a treasure trove of damning evidence.

How thorough is your effort to identify e-mail? E-mail resides in some or all of the following venues, grouped according to relative accessibility:

Easily Accessible:

- Online e-mail residing in active files on enterprise servers
  MS Exchange e.g., (.EDB, .STM, .LOG files)
  Lotus Notes (.NSF files)
  Novell GroupWise (.DB files)
- E-mail stored in active files on local or external hard drives and network shares
  User workstation hard drives (e.g., .PST, .OST files for Outlook and .NSF for Lotus Notes)
  Laptops (same as above)
  "Local" e-mail data files stored on networked file servers ("network shares")
  Mobile devices (PDA, "smart" phones, Blackberry)
  Home systems, particularly those with remote access to office networks
- Nearline e-mail
  Optical "juke box" devices
  Back ups of individual users' e-mail folders (i.e., "brick-level" back ups)
- Offline e-mail stored in networked repositories
  e.g., Zantaz EAS®, EMC EmailXtender®, Waterford MailMeter Forensic®

Accessible, but Often Overlooked:

- E-mail residing on remote servers
  ISPs (IMAP, POP, HTTP servers), Gmail, Yahoo Mail, Hotmail, etc.
- E-mail forwarded and carbon copied to third-party systems
  Employee forwards e-mail to self at personal email account
- E-mail threaded behind subsequent exchanges
  Subject and latest contents diverge from earlier exchanges lodged in body of email
- Offline local e-mail stored on removable media
  External hard drives, thumb drives and memory cards
  Optical media: CD-R/RW, DVD-R/RW
  Floppy Drives, Zip Drives
- Archived e-mail
  Auto-archived to additional .PST by Outlook or saved under user-selected filename
- Common user "flubs"
  Users experimenting with export features unwittingly create e-mail archives
- Legacy e-mail
  Users migrate from e-mail clients "abandoning" former e-mail stores
- E-mail saved to other formats
  .pdf, .tiff, .txt, .eml, etc.
- E-mail contained in review sets assembled for other litigation/compliance purposes
- E-mail retained by vendors or third-parties (e.g., former service provider)

- Print outs to paper

More Difficult to Access:

- Offline e-mail on server back up media
  Back up tapes (e.g., DLT, AIT)
- E-mail in forensically accessible areas of local hard drives
  Deleted e-mail
  Internet cache
  Unallocated clusters

The issues in the case, key players, relevant times, agreements between the parties and orders of the court determine the extent to which locations must be examined; however, the failure to *identify* all relevant e-mail carries such peril that caution should be the watchword. Isn't it wiser to invest more to know *exactly* what the client has than concede at the sanctions hearing the client failed to preserve and produce evidence it didn't know it had because *no one bothered to look for it*?

Electronic evidence is fragile and ever changing, so once you've found the e-mail evidence, you must guard against its loss or corruption. Next, we'll walk through preservation thicket.

Part II

Let's continue down the path to production of electronic mail. We've looked beyond the e-mail server to the many other places e-mail hides. Now, having identified the evidence, we're obliged to protect it from deletion, alteration and corruption.

Preservation

Anticipation of a claim is all that's required to trigger a duty to preserve potentially relevant evidence, including fragile, ever-changing electronic data. Preservation allows backtracking on the path to production, but fail to preserve evidence and you've burned your bridges.

Complicating our preservation effort is the autonomy afforded e-mail users. They create quirky folder structures, commingle personal and business communications and—most dangerous of all—control deletion and retention of messages. Best practices dictate that we instruct e-mail custodians to retain potentially relevant messages and that we regularly convey to them sufficient information to assess relevance in a consistent manner. In practice, hold directives alone are insufficient. Users find it irresistibly easy to delete data, so anticipate human frailty and act to protect evidence from spoliation at the hands of those inclined to destroy it. Don't leave the fox guarding the henhouse.

Consider the following as parts of an effective e-mail preservation effort:

- Litigation hold notices to custodians, including clear, practical and specific retention directives

Notices should remind custodians of relevant places where email resides, but not serve as a blueprint for destruction

Be sure to provide for notification to new hires and collection from departing employees

- Suspension of "retention" policies that call for purging email
- Suspension of re-use ("rotation") of back up media containing email
- Suspension of hardware and software changes which make email inaccessible

Replacing back up systems without retaining the means to read older media
Re-tasking or re-imaging systems for new users
Selling, giving away or otherwise disposing of systems and media
- Preventing custodians from deleting/altering/corrupting email

Immediate and periodic "snapshots" of relevant email accounts
Modifying user privileges settings on local systems and networks
Archival by auto-forwarding selected e-mail traffic to protected storage
- Restricting activity—like moving or copying files—tending to irreparably alter file metadata
- Packet capture of Instant Messaging (IM) traffic or effective enforcement of IM prohibition
- Preserve potential for forensic recovery

Imaging of key hard drives or sequestering systems
Suspension of defragmentation
Barring wiping software and encryption, with audit and enforcement

A threshold preservation issue is whether there is a duty of preservation going forward, e.g., with respect to information created during the pendency of the action. If not, timely harvest of data, imaging of drives and culling of relevant back ups from rotation may sufficiently meet the preservation duty so as to allow machines to be re-tasked, systems upgraded and back up tape rotation re-initiated. Securing guidance from the court and cooperating with opposing counsel to fashion practical preservation orders help insulate a producing party from subsequent claims of spoliation.

The Knowledge Hurdle

Thanks to a string of recent, high profile decisions, litigants are gradually awakening to their obligation to preserve electronic evidence. Still, attitudes often range from insufficient ("We'll just stop rotating back up tapes") to incredulous ("Why would we need to preserve voice mail?").

One hurdle is the lack of knowledge on the part of those charged with the responsibility to design and direct preservation efforts: too many don't understand what and how data change or what triggers those changes. They fail to appreciate how the pieces fit together.

For example, in a lawsuit concerning a plant explosion, the defendant, a major oil company, preserved monthly "full" back ups of their e-mail server but failed to hang on to four weeks of incremental back ups immediately preceding the blast. A full back up is a snapshot of the e-mail system at a single point in time. An incremental back up records changes to the e-mail system between snapshots. Did someone think that full back ups were cumulative of the incremental sessions? If so, they missed the fact that any e-mail received and deleted between snapshots might exist on the incremental back ups but be absent from the monthly tapes. They didn't consider how the pieces fit together.

If you've done a good job identifying where e-mail lives, preservation is largely a matter of duplicating the e-mail without metadata corruption or shielding it from subsequent loss or alteration. Both demand technical competence, so you'll need expert help the first time or two. If you ask questions and seek out reasons behind actions, knowledge gained from one effort will guide you through the next.

Adapting Preservation to Minimize Burden and Cost

With digital storage costs at all time lows, it's tempting to minimize spoliation risks by simply keeping everything. Don't. Keeping everything merely postpones and magnifies the cost and complexity of

production. Yet, you can suspend document retention and tape rotation without triggering a costly data logjam, if you adapt your preservation from reflexive to responsive.

Reflexive preservation describes steps you take while figuring out what's relevant and what's not. It's immediate and encompassing action to preserve the status quo while you sift the facts, forge agreements with opponents or seek guidance from the court. Calling a halt to back up tape rotation or suspending retention policies is reflexive preservation.

Reflexive preservation is a triage mechanism and a proper first response; but it's too expensive and disruptive for the long haul. Instead, convert reflexive preservation to responsive preservation by continually tweaking your preservation effort to retain only what's relevant to claims or necessary to meet business and regulatory obligations. Narrow the scope of preservation by agreement, motion practice and sound, defensible judgment.

Having identified the e-mail evidence and preserved it, we need to collect it and make it accessible for review and searching. Next, we hike up harvest hill and perambulate population pass. Wear sensible shoes!

Part III

On the path to production, we've explored e-mail's back alleys and trod the mean streets of the data preservation warehouse district. Now, let's head to the heartland for harvest time--*data* harvest time.

After attorney review, data harvest is byte-for-byte the costliest phase of enterprise e-discovery. Scouring scores of servers, local hard drives and portable media to gather files and metadata is an undertaking no company wants to repeat because of poor planning.

Harvest

Data harvest entails compiling a comprehensive review set or a preliminary production set. The former is a kitchen sink assemblage—unfiltered, though aggregated by business unit, locale, custodian, system or medium. In contrast, a preliminary production set is pre-filtered, comprised of items the data harvester deemed responsive. When a corporate defendant enlists employees to segregate responsive e-mail or a paralegal goes from machine-to-machine or account-to-account selecting messages, the product is a preliminary production set.

Selective harvest holds down cost by reducing the volume for attorney review, attended by increased risk of, e.g., the need to revisit machines, loss or corruption of evidence and inconsistent selections. If keyword or concept searches alone are used to harvest data, the risk of underinclusive production skyrockets.

Strategically, a producing party calls upon an opponent to furnish a list of search terms used to assemble the preliminary production set, eager to impose a "one list, one search" restriction. From the requesting party's perspective, it's hard to frame effective keyword searches without knowing the argot of the opposition. A trained reviewer can "pick up the lingo" One-bite-at-the-apple keyword searches can't. The party seeking discovery either accepts inadequate production or forces the producing party back to the well.

Initially more expensive, comprehensive harvest saves money when new requests and issues arise. Review sets can be searched again-and-again at little incremental expense, and broad preservation serves as a hedge against spoliation sanctions. Companies embroiled in serial litigation or compliance production benefit most from comprehensive collection strategies.

Chain of Custody

Any harvest method must protect evidentiary integrity. A competent chain of custody tracks the origins of e-evidence by, e.g., system, custodian, folder, file and dates. There's more to e-mail than what you see onscreen, so preempt attacks on authenticity by preserving complete header and encoded attachments.

Be prepared to demonstrate that no one tampered with the data between the time of harvest and its use in court. Custodial testimony concerning handling and storage may suffice, but better approaches employ cryptographic hashing of data—"digital fingerprinting"—to prove nothing has changed.

Metadata

As is true of every file on a computer, there's more to an e-mail than its contents: there's metadata, too. Each e-mail is tracked and indexed by the e-mail client ("application metadata") and every file holding e-mail is tracked and indexed by the computer's file system ("system metadata"). E-mail metadata is important evidence in its own right, helping to establish, e.g., whether and when a message was received, read, forwarded, changed or deleted. Metadata's evidentiary significance garnered scant attention until Williams v. Sprint, 2005 W.L. 2401626 (D. Kan. Sept. 29, 2005), a dispute over production of spreadsheets where the court held a party required to produce electronic documents as kept in the ordinary course of business must produce metadata absent objection, agreement or protective order.

System metadata is particularly fragile. Just copying a file from one location to another alters the file's metadata, potentially destroying critical evidence. Ideally, your data harvest shouldn't corrupt metadata, but if it may, archive the metadata beforehand. Though unwieldy, a spreadsheet reflecting original metadata is preferable to spoliation. Electronic discovery and computer forensics experts can recommend approaches to resolve these and other data harvest issues.

Processing and Population

However scrupulous your e-mail harvest, what you've reaped is no more ready to be text searched than a sheaf of wheat is ready to be a sandwich. It's a mish-mash of incompatible formats on different media: database files from Microsoft Exchange or Lotus Domino Servers, .PST and .NSF files copied from local hard drives, HTML fragments of browser-based e-mail and .PDF or .TIFF images. Locked, encrypted and compressed, it isn't stored as text. Image data is a picture, and e-mail attachments are encoded as a hieroglyphic called "Base 64." Run keyword searches on this amalgam and it yields up little information.

Before search tools or reviewers can do their jobs, harvested data must be processed to populate the review set, i.e., deciphered and reconstituted as words by opening password-protected items, decrypting and decompressing container files and running optical character recognition on image files. Searching now will work, but it'll be slow going to slog through duplicate messages in multiple user mailboxes or a single custodian's mailbox at different times. Fortunately, there's a fix for that, too.

Next: de-duplication, deliverables, documentation and the destination on the Path to Production.

Part IV

The e-mail's assembled and accessible. You could begin review immediately, but unless your client has money to burn, there's more to do before diving in: de-duplication. When Marge e-mails Homer, Bart and Lisa, Homer's "Reply to All" goes in both Homer's Sent Items and Inbox folders, and in Marge's, Bart's and Lisa's Inboxes. Reviewing Homer's response five times is wasteful and sets the stage for conflicting relevance and privilege decisions.

Duplication problems compound when e-mail is restored from backup tape. Each tape is a snapshot of e-mail at a moment in time. Because few users purge mailboxes month-to-month, one month's snapshot holds nearly the same e-mail as the next. Restore a year of e-mail from monthly backups, and identical messages multiply like rabbits.

De-Duplication

De-duplication uses metadata, cryptographic hashing or both to exclude identical messages. De-duplication may be implemented vertically, within a single mailbox or custodian, and horizontally, across multiple mailboxes and custodians. When questioning or prepping a witness, you'll want to see all relevant messages in the witness' mailbox, not just unique messages; so track and log de-duplication to facilitate re-population of duplicated items. De-duplication works best when unique messages and de-duplication logs merge in a database, allowing a reviewer to reconstruct mailboxes.

Be wary of "horizontal" de-duplication when discovery strategies change. An e-mail sent to dozens of recipients de-duplicated from all but one custodian's mailbox may be lost forever if the one custodian's e-mail ends up not being produced.

Review Tools

Rather than plow through zillions of e-mails for responsive and privileged items, reviewers often turn to keyword or concept search tools. Automated search tools make short work of objective requests for "all e-mail between Simpson and Burns," but may choke on "all e-mail concerning plant safety." To frame effective keyword searches, you have to know the lingo describing events and objects central to the case. Even then, crucial communiqués like, "My lips are sealed" or "Excellent" may be missed.

Are tireless black box tools an adequate substitute for human review? The jury's still out. In a seminal study, keyword searching fared poorly, finding only about one-fifth of relevant items identified by human reviewers. However, litigation management consultant Anne Kershaw looked at an advanced search tool and found machines performed almost twice as well as humans. The safest course is to arm conscientious, well-trained reviewers with state-of-the-art search tools and work cooperatively with opposing counsel to frame searches. Even then, examine the mailboxes of key witnesses, message-by-message.

Redaction

Paper redaction was easy: We concealed privileged text using a black marker and photocopied. It's trickier to eradicate privileged and confidential information at the data layer of document image files and within encoded attachments and metadata. Run your approach by an expert.

Re-population
For production, should you re-populate to restore relevant, non-privileged items previously de-duplicated, or will the other side accept a de-duplication log? Never produce de-duplicated e-mail without memorializing that opposing counsel knows of the de-duplication and waives re-population.

Deliverables
There isn't just one "right" media or format for deliverables. Options for production media include network transmittal, external hard drives, optical disks, tape, online repositories and hard copies. Formats range from native (.pst), exported (.eml), text (.txt), load files (Concordance, Summation), image files with or without data layers (.pdf, .tiff) and delimited files. Evidence ill-suited to .tiff production (databases, some spreadsheets, etc.), compels native production.

Documentation
Inevitably, something will be overlooked or lost, but sanctions need not follow every failure. Document diligence throughout the discovery effort and be prepared to demonstrate why bad decisions were sound at the time and under the circumstances. Note where the client looked for responsive information, what was found, how much time and money was expended, what was sidelined and why. Avoid sanctions by proving good faith.

Are We There Yet?
The path to production is a long and winding road, but it's heading in the right direction. Knowing how to manage electronic evidence is as vital to trial practice as the ability to draft pleadings or question witnesses. Don't forget what happened on Main Street when they built the Interstate. Paper discovery's the old road. E-discovery's the Interstate.