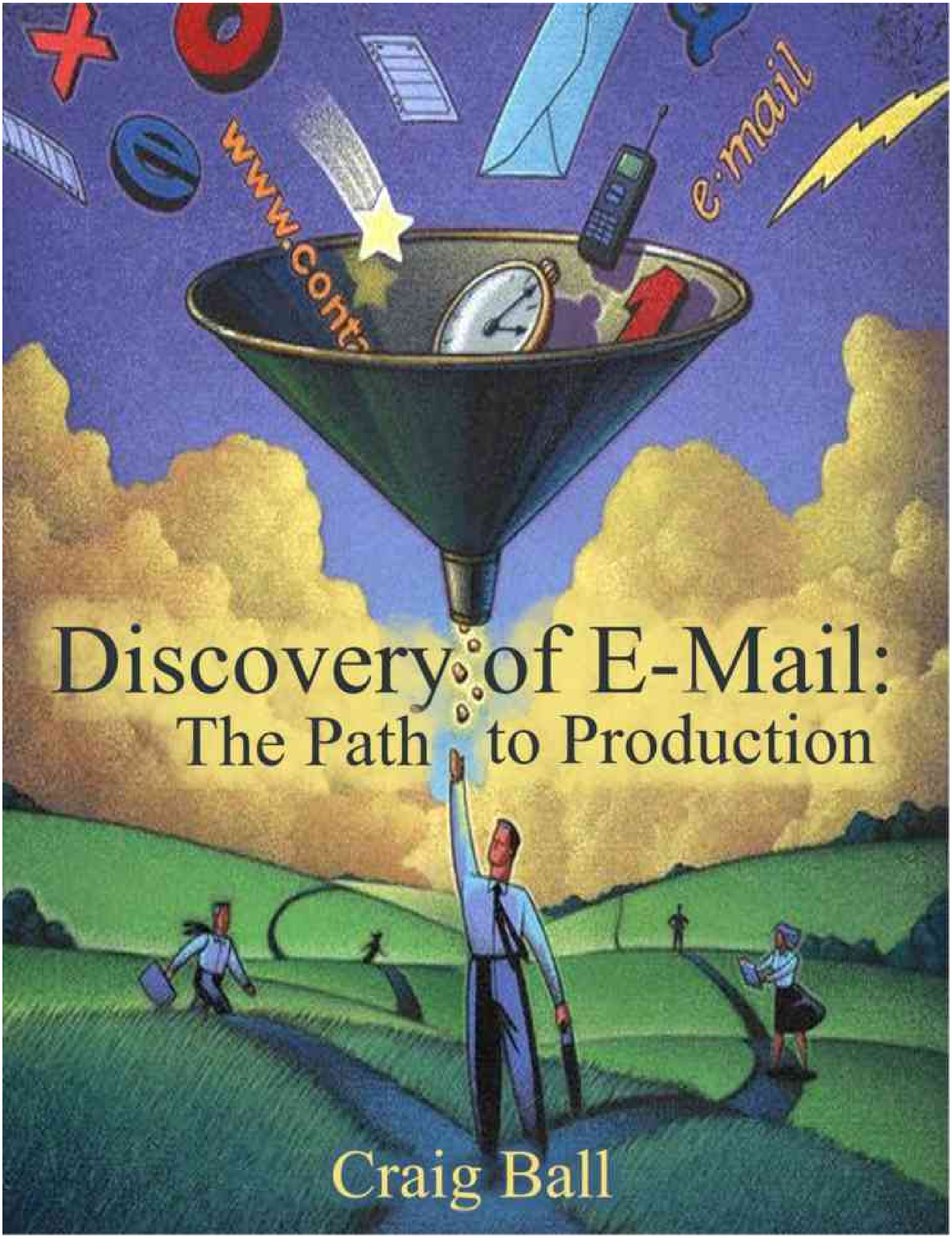

Discovery of E-Mail:
The Path to Production

Craig Ball

Discovery of Electronic Mail: The Path to Production

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.

This article outlines issues and tasks faced in production of electronic mail—certainly the most common and perhaps the trickiest undertaking in electronic discovery. It's a guide to aid attorneys meeting and conferring with opposing counsel, working with e-discovery service providers, drafting production requests and explaining the cost and complexity of e-mail production to clients and the court. It offers no short cuts, but that's not the point. The goal is to keep you from stepping off a cliff. Not every point outlined here 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. One way to avoid re-treading ground is to question expectations from 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?*
- *Be Bates numbered, and by what method?*
- *Be easily authenticable for admission into evidence?*

After attorney review, data harvest is byte-for-byte the costliest phase of electronic discovery. Understandably, producing parties want to search once and be done with it and confine the requesting party to a single list of keywords. From the requesting party's perspective, it's often impossible to frame effective keyword searches absent familiarity with the argot used to describe the events and objects central to the case, resulting in keyword searching missing what well-trained reviewers would find.

Producing parties are often forced to return to the well. Where you anticipate that new keywords will emerge or different search techniques will be used, securing the least costly outcome warrants the most expensive beginning: compiling a comprehensive review set of all potentially relevant e-mail. This entails *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.

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*?

Preservation

The duty to preserve potentially relevant evidence is generally triggered by the anticipation of a claim. Fulfilling a preservation duty with respect to e-mail is made harder by the control reposed in individual users, who establish quirky folder structures, commingle personal and business communications and—most dangerous of all—control deletion and retention of their messages. Although individual users should be directed to retain all potentially relevant messages and *regularly* furnished *sufficient* information to assess relevance *consistently*, the potential for human frailty shouldn't be overlooked. *Don't leave the fox guarding the henhouse.* Act promptly to protect data from spoliation at the hands of users most inclined to sweep it under the rug.

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

- Litigation hold notices to users, including clear, practical and specific retention directives
  - *Notices should remind users of relevant places where their email may reside*
  - *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 users from deleting/altering/corrupting email
  - *Immediate and periodic "snapshots" of relevant user 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 use of wiping software and encryption, with audit and enforcement*

A threshold issue is whether there exists 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 (to name a few) may sufficiently satisfy the preservation duty so as to allow machines to be re-tasked, systems upgraded and back up tape rotation re-initiated. Seeking guidance from the court and working with opposing counsel to craft a preservation order help to insulate a producing party acting in good faith from subsequent claims of spoliation.

Harvest

Knowing what e-mail exists and where, and having taken proper steps to preserve it, it's time to gather potentially relevant messages and attachments into a **comprehensive review** set or select and assemble responsive items into a **preliminary production** set. The difference between the two is that a comprehensive review set is compiled largely without regard to what information will be selected for production. It's a "kitchen sink" assemblage, though ultimately its scope is constrained by the business units, facilities, machines and media selected for examination. By contrast, a preliminary production set is comprised of only those e-mails and attachments that the persons collecting the data from the various files and machines deem responsive to the production requests. When a corporate defendant relies upon each employee to locate and segregate responsive e-mails or when a legal assistant goes from office-to-office selecting e-mails, the resulting collection is a preliminary production set.

The principal advantage of selective harvest is that it cuts the number of messages and attachments subject to attorney review, reducing short run cost. These savings come with attendant risks, among them the need to return to every machine if the initial harvest proves insufficient, the much greater potential for loss or corruption of overlooked evidence and inconsistencies between reviewer judgments. Also, if keyword or concept searches are employed to select e-mail for harvest, be sure to weigh the concerns about such techniques that are discussed later in this article.

The advantage of a comprehensive review set is that despite a larger initial outlay, as new requests and issues arise, the comprehensive collection can be culled again-and-again at little incremental expense. Moreover, by broadly preserving e-mail, a comprehensive review set is a valuable hedge against spoliation claims. For entities subject to ongoing litigation and compliance production, such a comprehensive collection may also be availing in multiple matters.

Whichever method is used, special care must be taken during data harvest to preserve the integrity of the evidence. It's essential to maintain a sound c*hain of custody* for harvested data and be able to establish the *origin* of the e-mail (e.g., system, user account, folder and file from which it was collected) as well as the *custodian* of the e-mail. It's critical to understand that there is more to an e-mail than what a client application like Microsoft Outlook or Lotus Notes displays onscreen. When authenticity is challenged, the unseen header information or encoded attachment data is needed. Accordingly, select a harvest method that preserves *all of the data* in the e-mail.

Another chain of custody requirement is the ability to demonstrate that no one tampered with the data *between* the time of harvest and its use in court. Testimony of the custodians about handling and storage is one solution. Better still, cryptographic hashing, a form of digital "fingerprinting" applied to sections of each e-mail and attachments, generates a alphanumeric value that can be archived with the evidence and used to conclusively establish data integrity, if challenged.

Finally, there is also even more to an e-mail than its contents because, as is true of every file stored on a computer, there is associated *metadata* (data *about* data). Each email must be tracked and indexed by the e-mail client application ("application metadata") and every file containing the e-mail must be tracked and indexed by the file system of the computer storing the data ("system metadata"). E-mail metadata can be important evidence in its own right, helping to establish, e.g., whether or when a message was received, read, forwarded, changed or deleted. System metadata is particularly fragile since most computer users think themselves fully capable of copying a file from one medium to another and fail to appreciate that simply copying a file from a hard drive to a floppy *changes the file's metadata* and potentially destroys critical evidence. Select your methods carefully to insure that the act of harvesting data as evidence doesn't alter the evidence or its metadata. If method chosen alters metadata, *archive the correct metadata before it changes*. Though cumbersome, a spreadsheet reflecting the original metadata is preferable to spoliation. Electronic discovery and computer forensics experts can recommend approaches to resolve these and other data harvest issues.

Population

Your scrupulous e-mail harvest is complete, but what you've reaped is no more ready to be searched for evidence than wheat is fit to be a sandwich. Harvested data arrives in varying incompatible formats on different media. Expect massive database files pulled from Microsoft Exchange and Lotus Domino Servers, .PST and .NSF files copied from local hard drives, HTML pages of browser-based e-mail, paper printouts, .PDF and .TIFF images (some searchable, some not) and all manner of forms and formats described in the Identification section, above. Were you to dump it all on a big hard drive and try to view it or run keyword searches, you'd quickly discover it yields up little information. That's because most of the data isn't stored as text. Some of it is locked up (password protected), some encrypted (e.g., Lotus Notes files) and some compressed, which frustrates text searching as effectively as encryption. The scanned data is a picture, not text, and the e-mail attachments are encoded in a hieroglyphic called "Base 64."

Before search tools and reviewers can do their jobs, the harvested data must be deciphered and reconstituted to be accessible and re-appear as the *words* we see when using e-mail clients and word processors. This is accomplished by, for example,

- Opening password protected files
- Decrypting container files and items (e.g., Lotus Notes .NSF)
- Decompressing email container files (e.g., Outlook .PST, .EBD, .OST)
- Converting attachments to compatible formats (e.g., Base64, MIME)
- Decompressing and decrypting attachments (e.g., .ZIP, .XLS,)
- Optical character recognition of document image attachments (e.g., .TIFF)
- Identifying Unicode-formatted and foreign language attachments and documents (e.g., .DOC)
- Accessing files in obscure or proprietary formats
- Repairing corrupted files

By this point, decisions must be made as to what media and methods will be used to host and review the data. Will counsel for the producing party pore over CDs, DVDs or portable hard drives or wade through network attached storage or online repositories? The assembled data should be organized to make it possible to pair the e-mail with its metadata and to trace messages and attachments back to their origins, by, e.g., custodian, interval, location, business unit or other taxonomy.

De-duplication

You *finally* made it. The e-mails are assembled, accessible and intelligible. You *could* begin your review right away, but unless your client has money to burn, there's one more thing to do before diving in: *de-duplication*. If Jane e-mails Tom, with copies to Dick and Harry and Tom responds with an attachment by clicking "Reply to All," Tom's response is in *both* Tom's Sent Items folder *and* his Inbox, as well as in Jane, Dick and Harry's Inboxes. Save for variations in time of receipt, the messages are functionally identical. Absent de-duplication, Tom's response will be reviewed five times. Not only is this a costly waste of time, it creates the potential for conflicting decisions respecting relevance and privilege issues. The better course would be to use specialized software to remove all but a single instance of Tom's response from the review set.

De-duplication is typically achieved using metadata, cryptographic hashing or a mix of the two. It may be implemented *vertically*, within a single mailbox, folder or custodian, or *horizontally* (also called *globally*) across multiple mailboxes and multiple custodians. It's essential to *track and log all de-duplication* to permit re-population of duplicated items to be produced.

Be careful with horizontal de-duplication as discovery strategies change. An e-mail sent to dozens of recipients may have been de-duplicated from all but one custodian's mailbox in the expectation that the message would be reviewed and a production decision made on review of that single mailbox. If that custodian's e-mail is excluded from review, the de-duplicated e-mail is *never* reviewed, even if all other custodian's mailboxes are examined. Here, de-duplication could result in the failure to produce a discoverable document.

Review

At last, you and your staff are looking at the e-mail to flag:

- Relevant, discoverable and non-privileged items
- Items responsive to particular requests
- Privileged communications (attorney-client, doctor-patient, work product)
- Confidential communications (trade secrets, proprietary data, personal and private)

If the review set is large, counsel may employ keyword or concept search tools to identify privileged or responsive items. Though a cost effective approach and useful when responding to objective requests (e.g., "produce all e-mail between Jane and Tom"), the value of automated search tools is considerably less clear when used to process subjective requests (e.g., "produce all e-mail expressing product safety concerns."). As previously noted, it's often impossible to frame effective keyword searches absent familiarity with the lingo used to describe the events and objects central to the case. Even then, the crucial communiqué, "*Say nothing*" or "*Dump her*" may be overlooked.

Properly used by those who understand their strengths and recognize their limitations, text and concept search tools are an important adjunct to—but an inadequate substitute for—the judgment of a diligent, well-trained reviewer. If you use automated search tools, be prepared to demonstrate to the court and opposing counsel how such tools compare with the efficacy of human reviewers and the basis for such comparison. Know that in the only litigation study comparing the two this author has found, keyword searching fared poorly, finding only about one-fifth of the relevant items identified by human reviewers. The safest approach is to work cooperatively with opposing counsel to select the keywords and frame the searches to be run against the review set. Mailboxes of key witnesses *always* merit careful message-by-message review for relevant intervals.

Re-population

Once it's been decided what to produce and withhold, the production set should be re-populated with all relevant and discoverable non-privileged messages and attachments that were de-duplicated for review. Alternatively, discuss the issue with opposing counsel and determine counsel's preference. Counsel for the requesting party may be satisfied with a log detailing other recipients, if it serves to simplify his review without causing undue confusion. Don't produce de-duplicated e-mail without establishing and memorializing that opposing counsel knows of the de-duplication and waives re-population.

Redaction

When a paper record held discoverable and privileged content, the time-honored solution was to conceal the privileged text with heavy black marking pen and produce a photocopy of the redacted original. Shortsighted efforts to carry that practice into the realm of electronic discovery proved embarrassing when it was discovered that simply obscuring text on the image layer of, e.g., a document file in Adobe

Portable Document Format (.PDF) did nothing to conceal the same text in the file's data layer. Electronic evidence demands different methods to remove privileged and confidential information from discoverable items. Any method employed must eradicate redacted data from all source data including:

- MIME/UU/BASE64 encoded attachments

  All e-mails are plain text file, yet we use them to transport photos, music, programs and all manner of binary files as "attachments". In truth, non-text data aren't "attached" at all. Thanks to an encoding scheme called Base64, binary data hitch a ride, embedded within the body of the e-mail, masquerading as text. If an attachment contains privileged content, know that producing the complete contents of the e-mail (that is, not just the message but the file's headers and footers, too) enables the privileged content to be decoded. Accordingly, Base64 encoded attachments must be redacted from MIME e-mails before their production
- Data layer of document image files (.tiff, .pdf)
- All copied and forwarded counterparts, including :bcc transmittals

Production

Decisions about the medium and format of production, as well as the handling of exceptional attachments, must be made before production of the e-mail can proceed:

- Medium for production: What container will be used for delivery?

  Electronic transmittal (e-mail attachment, FTP transfer)
  External hard drive
  Optical disks
  Online repository
  Hard copies
- Format of Production: In what form will the data files be delivered?

  Native *(.PST, .NSF)*
  Discrete files (.eml)
  Text files (.txt, .rtf)
  Load files (Concordance, Summation)
  Image files without data layer ("naked" .tiff)
  Image files with data layer (.pdf)
  Delimited files
- Protocol for production of exceptional files, for example:

  Databases that must be queried to deliver relevant information
  Spreadsheets and tables containing Z-axis data and embedded formulae
  Voice mail messages and associated metadata
  Data requiring proprietary software
  Data that could not be opened or decrypted; corrupted data
  Other data not lending itself to presentation in a letter size, paper-like format
  Scanned data with handwritten entries and marginalia missed by OCR
- What information will be included in privilege logs?
- What information will be furnished respecting de-duplicated items?

Documentation

Inevitably, something will be overlooked or lost, but sanctions need not follow every failure. Avoid sanctions by documenting diligence at every stage of the discovery effort, to be able to demonstrate why the decision that proved improvident was sound *at the time and place it was made*. Keep a record of where the client looked and what was found, how much time and money was expended and what was sidelined and why.

Conclusion

Responding to electronic discovery is a complex and challenging task--all the more so as we venture beyond the familiar confines of e-mail to the vast and varied sweep of all digital evidence. In the rush to embrace personal computing, businesses got ahead of sensible records management. Empowering individuals with networked PCs delegated responsibility for evidence preservation without adequate guidance or oversight. In short, businesses—and all of us—reaped the benefits of computers at the cost of discovery becoming harder and more expensive.

Some argue that we must make it easier and cheaper to litigate by deeming electronic evidence "out of bounds." Others respond that neither difficulty nor cost can justify curtailing full and fair access to evidence. One fact remains: ***most evidence is electronic***. If we want cases decided on the evidence, *discovery means electronic discovery*, and identifying, preserving, harvesting, managing and presenting digital evidence must be as vital and as accepted as cross-examination or trial by jury.

Electronic discovery is discovery in unfamiliar territory. When you figure out the steps and uncover the traps, it's like any other journey. Here's hoping this article helps you navigate the e-mail trail.

If you feel this outline omits a step or offers incorrect information, please share proposed additions or corrections with me at craig@ball.net. For further information about discovery of electronic mail, please read, "Meeting the Challenge: E-Mail in Civil Discovery" (http://www.ballpoint.org/emailpaper.pdf).