Competency and Strategy in Electronic Discovery

Craig Ball

December 2017 Ver. 8.1222
© Craig Ball, All Rights Reserved

Contents

Goals for this Collection 2

E-Discovery Update 2017 4

A Dozen E-Discovery Strategies for Requesting and Producing Parties 21

Introduction to Discovery in U.S. Civil Litigation 22

The "E-Discovery Rules" (1,16,26,34 & 45) of the Federal Rules of Civil Procedure With Committee Notes accompanying 2006 and 2015 Amendments 27

What Every Lawyer Should Know About E-Discovery 72

Introduction to Digital Computers, Servers and Storage 76

Getting your Arms around the ESI Elephant 94

The Internet of Things Meets the Four Stages of Attorney E-Grief 99

Custodial Hold: Trust but Verify 106

Elements of an Effective Legal Hold Notice 108

Opportunities and Obstacles: E-Discovery from Mobile Devices 111

Introduction to Metadata 138

Deep Diving into Deduplication 164

Mastering E-Mail in Discovery 172

Luddite Lawyer's Guide to Computer Backup Systems 211

Databases in E-Discovery 230

Search is a Science 254

Forms that Function 266

Exercise: Forms of Production and Cost 286

Preparing for Meet and Confer 293

About the Author 305

Goals for this Collection

The goal of this collection of articles is nothing less than to completely change the way you think about electronically stored information.

In a world where less than one-in-one-hundred cases are tried, discovery strategy, particularly e-discovery strategy, is more often vital than trial strategy. Strategy isn't simply doing what the rules require, and the law allows. Strategy requires we explore our opponent's fears, goals and pain points … and our own. *Is it just about the money? Can we deflect, distract or, deplete the other side's attention, energy or resources? How can they save face while we get what we want?*

Yet, strategic use of e-discovery garners little attention, perhaps because the fundamentals demand so much focus, there's little room for flourishes. As lawyers, we tend to cleave to one way of approaching e-discovery and distrust any way not our own. ***If you only know one way of doing things, how do you act strategically?***

Strategic discovery is the domain of those who've mastered the tools, techniques and nuances of efficient, effective discovery. That level of engagement, facility and flexibility is rare; but, you can be *more* strategic in e-discovery even if you've got a lot to learn. These readings are designed to get you thinking about the fantastic journey data takes from its simple, seamless existence as an endless stream of ones and zeroes to the seemingly-endless variety of documents, communications, records and formats that confound us in e-discovery. More, the goal is that you learn to use e-discovery strategically, making wise choices because you understand the sources and processes of ESI well enough to stand firm or compromise.

Craig Ball, December 22, 2017

E-DISCOVERY UPDATE 2017

Craig Ball

E-Discovery Update 2017

Never have lawyers enjoyed more ways to answer the questions, "what happened and why?" The world teems with sensor-laden, networked devices informing abundant apps. Once-ephemeral actions and communications are routinely recorded, ready to illuminate intent and serve as Boswell to behavior. Interaction and information on demand have changed us. We stand astride physical and virtual worlds, often more engaged with distal persons than with those at our table. Instant information gratification renders no question too trivial to Google and no attitude or experience insufficiently trenchant to share on Facebook.

Some despair that privacy is gone, the President tweets, and there's no "ducking and covering" from a cyberattack. But, as lawyers doggedly pursuing facts, we can rejoice. The digital universe is paying attention and stands ready to clue us in. All we must do is know where to look, ask the right questions and be tenacious seeking answers.

If you've paid close attention to e-discovery, then the landscape of e-discovery at the midpoint of 2017 looks much like it did a year ago, when the amended federal rules that kicked in at the close of 2015 were a source of uncertainty, particularly as to proportionality and sanctions. With a longer view, it's clear that proportionality is a blunt instrument, and not all courts are bowing to limits on their power to sanction spoliation of electronically stored information (ESI).

Proportionality

Proportionality describes the sensible proposition that the burdens of discovery shouldn't outweigh its benefits vis-à-vis the needs of the case. The 2015 amendments to Rule 26 of the Federal Rules of Civil Procedure shifted the elements of proportionate discovery—residing elsewhere in the rule for 30+ years—into the scope of discovery; *viz.*:

Unless otherwise limited by court order, the scope of discovery is as follows: Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense **and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable.**

FRCP Rule 26(b)(1), amended language in bold.

Proportionality is routinely (and inarguably) advocated as, "a $50,000 case shouldn't prompt discovery costing $100,000.00." Of course, it shouldn't; but, the parties rarely hold the same view of a case's value or their exposure. As well, the significance of a case cannot always be measured in monetary terms. Consequently, proportionality has manifested after the amendments as

(improperly) a boilerplate objection and as (usefully) an analytical framework by which courts issue protective orders according to their sound sense of fairness and discretion. The wise practitioner must couch objections and responses in the elements of the amended Rule, recognizing that courts will be prone to treat those elements as a checklist.

Texas' Take: Calling proportionality the "pole star" informing the exercise of discretion over electronic-discovery disputes, the Texas Supreme Court recently laid out the Texas proportionality factors and pronounced them "in line" with federal counterparts, stating, "[A]ll discovery is subject to the proportionality overlay embedded in our discovery rules and inherent in the reasonableness standard to which our electronic-discovery rule is tethered." *In Re State Farm Lloyds, Relator*, Nos. 15-0903, 15-0905 (Tex. Sup. Ct. May 26, 2017).

The Texas proportionality factors read a bit differently than the federal factors and are "certainly not exclusive." Per *In Re State Farm Lloyds,* Texas looks at:

1. Likely benefit of the requested discovery;
2. The needs of the case;
3. The amount in controversy;
4. The parties' resources;
5. Importance of the issues at stake in the litigation;
6. The importance of the proposed discovery in resolving the litigation; and
7. Any other articulable factor bearing on proportionality.

Spoliation Sanctions

Lawyers approach e-discovery with less enthusiasm than one brings to a root canal. Only the stick of sanctions has served to force litigators to preserve and produce ESI. Courts are loathe to issue sanctions and have done so in only the most egregious circumstances involving the intentional destruction of relevant ESI. Still, parties and counsel unskilled in e-discovery worried that their negligent destruction of evidence might serve as the basis for serious sanctions, like summary dismissal or an adverse inference instruction to the jury. A split between the federal circuits arose over whether serious sanctions could be grounded on negligence or required proof of prejudice and/or malevolent intent, *e.g.*, the Second Circuit required proof of negligence and prejudice where the Fifth Circuit required a showing of bad faith to underpin serious sanctions.

In 2015, the committee charged with drafting the Federal Rules of Civil Procedure sought to resolve the split by amending Rule 37 to limit the ability of judges to sanction the loss and destruction of electronic evidence unless specific requirements are met. FRCP Rule 37(e) now states:

If electronically stored information that should have been preserved in the anticipation or conduct of litigation is lost because a party failed to take reasonable steps to preserve it, and it cannot be restored or replaced through additional discovery, the court:

(1) upon finding prejudice to another party from loss of the information, may order measures no greater than necessary to cure the prejudice; or

(2) only upon finding that the party acted with the intent to deprive another party of the information's use in the litigation may:

(A) presume that the lost information was unfavorable to the party;

(B) instruct the jury that it may or must presume the information was unfavorable to the party; or

(C) dismiss the action or enter a default judgment.

FRCP Rule 37(e), as amended 2015.

Note the threshold inquiries:

a. Was ESI lost? The amended rule doesn't change anything for the loss of non-electronic items, like paper records or tangible evidence.
b. Should the lost ESI have been preserved for the litigation?
c. Was the ESI lost because reasonable steps weren't taken to preserve it?
d. Can the lost ESI be restored or replaced?

When all these criteria are met, the Rule lays out two exclusive paths:

1. If the lost ESI prompts prejudice to "another" party (presumably the *requesting* party), the Court may order curative measures minimally necessary to offset the prejudice,

OR

2. If it is determined that the spoliator "acted with the intent to deprive" another party of the use of the ESI in the litigation, the Court may impose serious sanctions *(i.e.*, adverse presumption, adverse inference or dismissal/default).

The amended Rule was intended to occupy the field in terms of ESI spoliation sanctions; but not all judges accept that their inherent, discretionary power to sanction spoliation has been curtailed. *Cf., Cat3 LLC v. Black Lineage, Inc.,* No. 14 Civ. 5511 (AT) (JCF) (S.D.N.Y. January 12, 2016) and *Hsueh v. N.Y. State Dep't of Fin. Servs.*, No. 15 Civ. 3401 (PAC), 2017 WL 1194706 (S.D.N.Y. Mar. 31, 2017).

Texas' Take: The Texas Supreme Court lately weighed in on standards governing spoliation in *Brookshire Bros., Ltd. v. Aldridge*, –S.W.3d–, 2014 WL 2994435 (Tex. July 3, 2013), holding that an adverse inference instruction for spoliation may only be given to a jury when the destruction of evidence was intentional or deprived the opposing party of "any meaningful ability to present a claim or defense." The court added that "[s]poliation findings—and their related sanctions—are to be determined by the trial judge, outside the presence of the jury, in order to avoid unfairly prejudicing the jury by the presentation of evidence that is unrelated to the facts underlying the lawsuit" and that "evidence bearing directly upon whether a party has spoliated evidence is not to be presented to the jury except insofar as it relates to the substance of the lawsuit."

Forms of Production

Lawyers continue to long for the days of paper records and memoranda in red rope folders, and why not? Litigation was simple when you could carry the case file in a briefcase. But, while the legal profession adapted to the demise of typewriters and carbon paper, it clings to the delusion that discovery can be printed out as pixels or ink.

Twenty-first century evidence is principally data, not documents. Accordingly, the forms in which we receive ESI determines if it's utile and complete. Strikingly, lower cost and recognition of native production's superior utility and completeness have driven a slow, sure move away from conversion of ESI to so-called "static" forms of production. If diminished utility and completeness were not sufficient justification to make smart designations of forms of production, the markedly increased per-gigabyte cost paid vendors to ingest and host flabby static formats called "TIFF images" should give any lawyer pause. Poorly-chosen forms of production are not the biggest contributors to the high cost of e-discovery as inefficient approaches to review are most costly, but waste occasioned by the failure to designate, obtain and utilize native and near-native forms of production is still substantial and one of the easiest to fix.[1]

Texas' Take: In federal practice, squabbles over forms of production have become rarer as counsel are less prone to squander energy and goodwill seeking to convert spreadsheets, presentations and other rich formats into static TIFF images over an opponent's objection. Unfortunately, the

[1] Native forms of production are the same forms the data occupies in the ordinary course of business. It's the form that information takes when the witnesses create and use it. Near-native forms are those which preserve those elements of functionality and completeness as can reasonably be achieved when it's infeasible to produce in native forms. That is, an e-mail may need to be converted from a native container format to a near-native single message format. What makes the latter format near-native is that the form selected retains the essential elements that allow an e-mail application to process the data as e-mail.

trend toward efficiency and lower cost has been set back in Texas by the Supreme Court's decision in *In Re State Farm Lloyds, Relator*, Nos. 15-0903, 15-0905 (Tex. Sup. Ct. May 26, 2017). [2]

In a mandamus action seeking to overturn a court's order requiring native forms of production be employed, and despite the plain language of Texas Rule of Civil Procedure 196.4, the Texas Supreme Court held that "neither party may dictate the form of electronic discovery. The requesting party must specify the desired form of production, but all discovery is subject to the proportionality overlay embedded in our discovery rules and inherent in the reasonableness standard to which our electronic-discovery rule is tethered. The taproot of this discovery dispute is whether production in native format is reasonable given the circumstances of [the] case. Reasonableness and its bedfellow, proportionality, require a case-by-case balancing of jurisprudential considerations, which is informed by factors the discovery rules identify as limiting the scope of discovery…." *Id.*

The Court could have recognized that native formats are those used in the ordinary course and, accordingly, are the original evidence as used every day by the parties. Production in native (or, when infeasible, near-native format) is inherently reasonable absent a showing of undue burden or cost because native format is, by definition, the form in which the data is found, as it ordinarily exists in the producing party's systems. Requiring that forms of production be litigated on proportionality grounds according to the circumstances of each case will serve to slow resolution and increase the cost of litigation for all, versus a default rule that parties produce in the forms in which they ordinarily hold the responsive data absent an agreement or order to supply alternate forms.

Cross Border Discovery

If you've lawfully engaged in e-discovery from persons and companies residing within the European Union, you've surely bumped up against the EU's 1995 Data Protection Directive (Directive 95/46/EC) regulating the "processing" of personal data of EU citizens. "Processing" includes collection, retrieval, transmission, use and disclosure—essentially, every action attendant to e-discovery. Moving data to the United States once implicated a regulatory regime of self-certification called the Safe Harbor Principles. In October 2015, the European Court of Justice ruled that the Safe Harbor regime provided an inadequate level of data protection, and

[2] DISCLOSURE: I served as an expert witness for the homeowners in the case. The homeowners prevailed in terms of resisting mandamus; however, the Texas Supreme Court lost an opportunity to point the way toward lower cost and more efficient e-discovery for all litigants, instead grafting a ponderous analytical framework onto what should be one of the simplest processes in e-discovery. Requiring requesting parties to show cause why evidence should *not* be degraded from the forms used in the ordinary course of business to static forms places the burden on the wrong party. As well, requiring a special showing to demand metadata integral to the original evidence is akin to requiring production of the consonants in a document but demanding good cause be shown to obtain the vowels.

one year ago, the European Commission adopted the EU-US Privacy Shield framework (effective July 12, 2016) to enable U.S. companies to more easily receive personal data from EU entities. That said, the viability of the Privacy Shield has been thrown in doubt by President Donald Trump's issuance of an Executive Order on January 25, 2017, requiring that U.S. privacy protections extend only to citizens and permanent residents of the U.S.

As if there weren't enough confusion attendant to cross-border discovery, effective May 25, 2018, the 1995 Data Protection Directive will be supplanted by a new set of data privacy standards called the GDPR for "General Data Protection Regulation" (Regulation 2016/679). The GDPR broadens privacy protections for EU citizens, including a right of explicit consent to processing of personal data and a right to request erasure of personal data. Notwithstanding the optimism of some commentators, the GDPR seems certain to make it more difficult and, accordingly, more expensive to conduct e-discovery from sources based in the European Union. Of course, the EU is just one of several regions around the world that place widely-varying and onerous hurdles in the path of U.S. e-discovery. It won't be as simple as getting the Court to order production when to do so subjects a party to criminal or civil penalties in other jurisdictions.

Cybersecurity and Privacy

Cybersecurity and personal privacy are real and compelling concerns. Whether we know it or not, virtually everyone has been victimized by data breach. Lawyers are tempting targets to hackers because lawyers and law firms hold petabytes of sensitive and confidential data. Lawyers bear this heady responsibility despite being far behind the curve of information technology and arrogant in dismissing their need to be more technically astute. Cloaked in privilege and the arcana of law, litigators have proven obstinate when it comes to adapting discovery practice to changing times and threats, rendering them easy prey for hackers and data thieves.

Corporate clients better appreciate the operational, regulatory and reputational risks posed by lackluster cybersecurity. Big companies have been burned to the point that when we hear names like Sony, Target or Anthem, we may think "data breach" before "electronics," "retail" or "health care." The largest corporations operate worldwide, so are subject to stricter data privacy laws. In the United States, we assume if a company owns the system, it owns the data. Not so abroad, where people have a right to dictate how and when their personal information is shared.

Headlines have forced corporate clients to clean up their acts respecting data protection, and they've begun dragging their lawyers along, demanding that outside counsel do more than pay lip service to protecting, *e.g.*, personally-identifiable information (PII), protected health information (PHI), privileged information and, above all, information lending support to those who would sue the company for malfeasance or regulators who would impose fines or penalties.

Corporate clients are making outside counsel undergo security audits and institute operational and technical measures to protect company confidential information. These measures include encryption in transit, encryption at rest, access controls, extensive physical security, incident response capabilities, cyber liability insurance, industry (*i.e.*, ISO) certifications and compulsory breach reporting. For examples of emerging 'standards,' look at the Model Information Protection and Security Controls for Outside Counsel Possessing Company Confidential Information lately promulgated by the Association of Corporate Counsel.

Forcing outside counsel to harden their data bulwarks is important and overdue; but, it's also disruptive and costly. Many small firms will find it more difficult to compete with legal behemoths. Savvier small firms, nimbler in their ability to embrace cybersecurity, will frame it as a market differentiator. At the end of the day, firms big and small must up their game in terms of protecting sensitive data.

Enhanced cybersecurity is a rising tide that floats all boats.

Well, maybe not *all* boats. Let me share who's likely to get swamped by this rising tide: *requesting parties (*or, as corporations call them *"plaintiffs' lawyers"),* and their *experts* and *litigation support providers*. Requesting parties and others in the same boat will find themselves grossly unprepared to supply the rigorous cybersecurity and privacy protection made a condition of e-discovery.

Again, *cybersecurity and personal privacy are real and compelling concerns*, but these security concerns will also be used tactically to deflect and defer discovery. They will serve as hurdles and pitfalls tending to make plaintiffs' lawyers think twice before pursuing meritorious cases. If you haven't run into this, you soon will, and your instinct may be to resist. *Don't.*

Fighting to be cavalier about data security is a battle that requesting parties cannot win and should not fight. Requesting parties must instead be ready to put genuine protections in place and articulate them when challenged.

I know some will say, "all we have to do is sign a protective order." But they don't see the trap set by executing protective orders without the ability (and sometimes without the intention) to meet the obligations of the order. High profile gaffes will follow, and the failure of a few will be the undoing of many.

A protective order isn't the answer if it's an empty promise. Requesting parties can't agree to employ stringent data protection and then go about business as usual: e-mailing confidential data, storing it on unencrypted media and failing to ensure that all who receive confidential data from counsel handle it with requisite caution.

Here's how it will go down for some prominent plaintiffs' lawyer:

1. Producing parties will demand protective orders imposing stringent, but appropriate, data protection practices and breach reporting requirements.

2. Requesting parties will sign these orders because—let's be frank—requesting parties will agree to almost anything if they believe it will get them "the smoking gun." Plus, how do you persuade a judge that she shouldn't issue a protective order when all the other side wants are sensible measures like access controls, encryption and breach reporting to protect sensitive data and PII?

3. Requesting parties will treat information produced in discovery with the same care they bring to their own confidential information, which is to say, not much and less than that protective orders typically require.

4. Confidential data will be mishandled, probably with so little actual prejudice as to prompt requesting counsel to ignore the breach reporting obligation in the order, reasoning "no harm, no foul."

5. The breach will ultimately come to light, opening counsel's mishandling of produced data to scrutiny and prompting discovery about discovery. The failure to set up secure systems, establish policies, train employees, test and audit processes and require contractors and experts to do the same will be gleefully dissected in court.

6. The producing party will beat its chest in lamentations of irreparable harm. The legal press will have a field day. The judge will be wrathful. The requesting party's counsel will look like a clown and might lose his ability to serve on plaintiffs' steering committees.

7. Producing parties will ceaselessly argue the now-proven hazard of e-disclosure, and requesting parties everywhere will be tarred with the same brush, challenged to prove they aren't going to be the next ugly breach. Judges will be less willing to grant full and fair discovery and more willing to impose arduous conditions for access.

A cynical and dystopian prediction? Perhaps. But don't imagine it won't happen. It's happening now.

The way to keep this in check is for requesting parties to act now to prepare to receive and protect confidential data sought in discovery.

Requesting parties cannot expect to be held to a lesser standard of cybersecurity than the producing parties compelled to surrender confidential data to them. A grizzled trial lawyer once

warned me, "*Defendants are forgiven several lies. Plaintiffs get none*." So, a party can be incautious with its own data because it's theirs; but counsel who fail to protect an opposing party's confidential data will be harshly judged. They don't just hurt their clients and opponents; they undermine the very foundations of discovery.

So, what must counsel for requesting parties do? Here are a dozen suggestions:

1. Take cybersecurity duties seriously. It's not someone else's job. It's your job. You are the gatekeeper. This is Rule One, not by accident.

2. Don't just treat an opponent's confidential data with the care you afford your own; treat it better. It's like money in your trust account. You don't treat client monies/data like your own. You don't commingle client monies/data with yours, and you don't use that money/data for anything but permissible purposes with careful recordkeeping.

3. If there's a protective order, read it closely and be sure you fully understand what it obliges you to do in terms of the day-to-day conduct of any who access confidential information.

4. A proper chain of custody is essential. You must be ready to establish who received confidential data and the justification for its disclosure. You must be able to prove you had a good faith basis to believe that the person receiving confidential data understood the need to protect the data and possessed the resources, training and skill to do so. This obligation encompasses anyone who gets the data from you, including experts, clerical staff, associated counsel and service providers. Anyone with access to confidential data must be well-prepared to protect the data because their failure is your failure.

5. Proceed with caution when disclosing confidential data to experts. Industry experts serve multiple masters and may seek to exploit confidential data obtained in one matter in other engagements. Secure the expert's written commitment not to do so, and enforce it. Additionally, don't supply confidential data to an expert without first obtaining the expert's consent to receive and protect it. People who appreciate the burden of protecting other people's sensitive data want to hold as little of it as possible.

6. Recognize that you don't get to decide what data warrants protection. The designation rules. If you think something isn't properly designated as confidential or sensitive, challenge the designation; but, until the other side concedes or the Court rules, the designation sets the duty.

7. Confidential data should be encrypted in transit and at rest. This means that none of the confidential data gets attached to an e-mail, moved to portable media (*e.g.,* a thumb drive

or a portable hard drive) or uploaded to the cloud *unless it is encrypted.* No exceptions. No excuses. BTW, if you store or transmit the decryption keys alongside the encrypted data, it's doesn't count as encrypted.

8. Perimeter protection isn't enough. The biggest risks to confidential data are internal threats, that is, from a craven or careless member of your own team. Trust but verify. Access to confidential data should be afforded only on an as-needed/when-needed basis.

9. Access to confidential data must be monitored and logged, as feasible. Remote access and after-hours access should be audited. Safeguard the other side's confidential data in much the same manner as banks protect the contents of safety deposit boxes: There is physical security (walls, doors, alarm systems and guards) and monitoring of the perimeter (cameras and key cards). There's a vault to keep all contents safe when the perimeter is breached, and access controls to make contents available only to authorized persons (dual-keyed boxes and ID/signature scrutiny). Data protection also incorporates elements of perimeter security (limiting physical access to the devices and systems), monitoring (logging and auditing), a vault (strong encryption with sound key management) and access controls (two-factor login credentials and user privilege management).

10. Have a written data security and incident response policy and protocol in place and *conform your practice to it.* Be sure all employees with access to sensitive and confidential data agree to be bound by the policy, and train everyone in proper cybersecurity. You must first recognize a risk to be prepared to meet it. "No one told me to do that" is not the testimony you want to hear when your staff take the stand.

11. Be wary of oppressive obligations to destroy or "return" data when a case concludes. Confidential case data tends to seep into mail servers, litigation databases, document management tools and backup systems. Are you prepared to shut down your firm's e-mail and destroy its backup media because you failed to consider what an obligation to eradicate data would really entail? Have you budgeted for the cost of eradication and certification when the case concludes?

12. Consider cloud-based storage and review tools that integrate encryption, two-factor authentication and access logging. The cloud's key advantage lies in a user's ability to shift many of the physical and operational burdens of cybersecurity to a third-party. It's not a complete solution, but it serves to put a secure environment for confidential data within reach of firms of all sizes.

If this sounds like a big, costly pain, you're paying attention. It's a headache. It slows you down, and the risks grow and change as fast as the technology. But if requesting parties don't put adequate protections in place on their own, courts will allow producing parties to dictate what hoops requesting parties must jump through to obtain discovery–if, indeed, courts don't deem the risk so disproportionate that they deny access altogether.

E-discovery is hard enough. Don't make it harder by giving opponents the ability to claim you can't be trusted to protect their information.

Technology-Assisted Review

Technology-Assisted Review or TAR is the use of computers trained by lawyers to distinguish between responsive and non-responsive ESI. Properly implemented and tasked to the right sort of ESI, it works more quickly, affordably and reliably than an army of human reviewers looking at every potentially relevant item. It is an existential threat to the costly, customary and wildly error-prone approach firms typically take to large-scale document review.
Even as I write that, I know you won't believe it. Yet, it's true. The devil is in the details.

In the last few years, the use of TAR has grown markedly, but quietly. TAR still has the aura of a science experiment. Many who have used TAR tools to speed review are reluctant to disclose same lest their methodology be scrutinized. That's the catch-22 with TAR in 2017: lawyers trust it enough to use it, but not enough to stand behind it. Perhaps because they don't understand TAR well enough to defend it, or perhaps because they just don't trust it themselves. Likely, they would claim that having to defend TAR would be so costly and time-consuming that it would defeat the point of using it. So, they clam up or claim it's "work product" and refuse to confirm or deny its use.

Recent efforts by the Duke EDRM to set standards for TAR deployments are likely to embolden lawyers and courts to use TAR to speed e-discovery and lower costs. Several courts have approved the use of TAR, but none have required its use…yet. Inevitably, the merits of TAR will prompt a court to require its use when alternate methods are be shown to be too slow, costly or unreliable.

Mobile Goes Mainstream

Can anyone doubt the changes wrought by the modern "smart" cellphone? My current home in New Orleans sits at the corner of one-way streets, my porch a few feet from motorists. At my former NOLA home, my porch faced cars stopped for a street light. From both vantage points, I've seen drivers looking at their phones, some so engrossed they failed to move when they could. Phones impact how traffic progresses through controlled intersections in every community. We are slow-moving zombies in cars.

Distracted driving has eclipsed speeding and drunken driving as the leading cause of motor vehicle collisions. Walking into fixed objects while texting is reportedly the most common reason young people visit emergency rooms today. Instances of "distracted walking" injury have doubled every year since 2006. Doing the math, 250 ER visits in 2006 are over half a million ER visits today, *because we walk into poles, doors and parked cars while texting*.

Look around you. CAUTION: *This will entail looking up from your phone.* How many are using their phones? At a concert, how many are experiencing it through the lens of their cell phone cameras? How many selfies? How many texts? How many apps?

Lately I've begun asking CLE attendees how many are never more than an arm's length from their phones 24/7. A majority raise their hands. These are tech-wary lawyers, and most are Boomers, not Millennials.

Smart phones have changed us. Litigants are at a turning point in meeting e-discovery duties, and lawyers ignore this sea change at peril. The "legal industry" has chosen self-deception when it comes to mobile devices. It's a lie in line with corporate bottom lines, and it once found support in the e-discovery case law and rules of procedure. But, no more.

Today, if you fail to advise clients to preserve relevant and unique mobile data when under a preservation duty, you're committing malpractice.

Yes, I used the "M" word, and not lightly.

I wouldn't have called it malpractice a few years ago. But two things have changed, and we can't hide our heads in the sand. These are paradigm shifts.

The two things are, first, the data on phones and tablets are *not* just copies of information held elsewhere. Mobile data is unique, and often relevant, probative evidence. Second, the locking down of phone content has driven the preservation of mobile content from the esoteric realm of computer forensics to the readily accessible world of apps and backups. These developments mean that, notwithstanding the outdated rationales lawyers trot out for ignoring mobile, the time has come to accept that mobile is routinely within the scope of preservation obligations.

Too, lawyers need to stop treating mobile devices like biohazards and realize that there are easy, low-cost ways to preserve relevant mobile content *without taking phones away from users.* Because it's easy and cheap to preserve it, mobile content is accessible, and its preservation, when potentially relevant, is proportionate under the Rules.

That's a strong stand, and one some will angrily reject. I get where they're coming from. It was wonderful to be able to ignore mobile in e-discovery. Mobile was a black hole. It wasn't just that you had to hire technical experts to use expensive tools to preserve the contents of phones; it was like pulling teeth to get users to let loose of their devices for the hours or days it took to collect them. Even when they did hand them over, more than a few users claimed to have entered the wrong password too many times and "accidentally" wiped the contents of the phone. "Oops. My bad."

If that never happened to one of your clients, it may be because your client wasn't preserving phone data, indulging in the assumption that whatever they'd glean from the phone would be collected elsewhere. They deemed mobile redundant.

Lecturing about mobile and IoT in D.C. last year, an associate from a megafirm confided to me that his firm routinely advised all its litigation clients that they need not preserve the content of mobile devices because "all the relevant content would be duplicated on the servers." I asked if the firm had ever tested its advice against the relevant data to determine if there was truth in what they were telling clients. He admitted they never had and offered that they'd never do so. The firm didn't want to know the facts because the fairy tale of "replicated elsewhere" was what the client wanted to hear.

Is it a fairy tale? I have my own views based on my own comparisons of mobile content versus other collected sources. What I see demonstrates that the claim that what's relevant on a phone is preserved elsewhere is a whopper. I am routinely finding examples of relevant data stored on mobile devices that is not found among the other sources of data routinely preserved in e-discovery. The replication fairy tale is a relic of a bygone era of Blackberry Enterprise Servers and phones with lower IQs than the brilliant devices now our constant companions and confidantes.

But, I'm not asking you (or courts) to take my word for it. *Test it yourself.*

If you're going to tell the tale, then get some metrics to make it plausible. Use sampling. Process the phones of a few key custodians and compare all the potentially relevant items collected from their mobile devices against the other sources collected for the sampled custodians. What's the differential? Is the unique evidence from the mobile device probative and material?
I've done that, and so I know replication is a fairy tale. If you want to claim it's true *for your client in your case*, how about putting some facts to work? Bear the burden of proof, or start bearing the onus of truth. When you have the facts, you'll have to let loose of the legend and preserve relevant mobile content.

That's the bad news for those who would prefer to ignore mobile. But take heart, as that will seem like great news compared to the next development. Yet, there's a silver lining. Mobile preservation's become quick, cheap and easy.

A few years ago, mobile phones shared some of the characteristics of personal computers in that they held latent data that could be recovered using specialized tools sold for princely sums by a couple of shadowy tech companies. So, the preservation of mobile devices slipped into the shadows, too. Phones and tablets were *forensic* evidence, and only forensic examiners could collect their contents.

Although users used mobile devices all day, the contents of mobile devices were dubbed "not reasonably accessible." It was too costly and burdensome to preserve a phone. Good thing, because users were holding onto their phones tighter than Willie Nelson clutches a bong. Users protested, *"My mobile phone is the only way the kids' school can reach me in an emergency, and I can't use another phone because everyone texts now, and WHO REMEMBERS PHONE NUMBERS ANYMORE?"*

So, the next altered paradigm: In e-discovery today, the forensic-level preservation of phones—the sort geared to deleted content and forensic artifacts—is a fool's errand. As the public learned from the FBI's tussle with Apple over unlocking the iPhones of the San Bernardino terrorists, modern smart phones are locked down hard. Content is encrypted, and even the keys to access the encrypted content are themselves encrypted. Phone forensics isn't what it used to be. More and more, we can't get to that cornucopia of recoverable forensically-significant data.

At the same time, it's quick, easy and free for a user to generate a full, unencrypted backup of a phone without surrendering possession. The user can even place the backup in a designated location for safekeeping by counsel or IT. Will this be a "forensic image" of the contents? Strictly speaking, no. But as the phone manufacturers tighten their security, "forensic imaging" becomes less and less likely to yield up content of the sort encompassed by a routine e-discovery preservation obligation. Not every case is a job for C.S.I.—and I say that as someone who makes a living through computer forensics.

I grant that a full unencrypted backup of an iPhone isn't going to encompass all the data that might be gleaned by a pull-out-all-stops forensic preservation of the phone. But so what? As my corporate colleagues love to say, "*the standard for ESI preservation isn't perfect.*" I always agree adding, "*but it isn't lousy either.*" Preserving by backup isn't perfect; but it isn't lousy. I've come to regard it as sufficient and proportionate. It's good *enough*, and in most cases, darn good.

I think this is important. It's a game changer for what most litigants are doing today. In a view I hope will come to be shared by all who think it through—preservation of mobile device content must become a standard component of a competent preservation effort except where the mobile content can be shown to be beyond scope. Mobile content has become so relevant and unique, and the ability to preserve it so undemanding, that the standard must be preservation.

Automated and Hosted Processing and Review

The accepted e-discovery workflow has long involved the collection of data by technical personnel and its delivery into the hands of an e-discovery service provider who would process the data into images and generate load files holding extracted text and metadata. These images and extractions would then be loaded into the law firm's "review platform," a tool that mated the extracted text with its corresponding page image and facilitated search and tagging of the collection by multiple reviewers. This approach made it hard to quickly assess a case (because it took a lot of time and money to get the ESI in front of reviewers) and rendered e-discovery too complex or costly for small- and mid-size firms (who would have to make a significant capital investment in review software, servers and workstations.

Lately, the cloud and the development of automated workflows in cloud-based Software-as-a-Service or "SaaS" tools has made it possible for lawyers and support personnel with little technical savvy and no capital investment to upload, process, review and create production sets on a pay-as-you-go basis. Typically charging per gigabyte of data, the cloud service provider processes the data to, *inter alia,* extract its contents and eliminate duplicate items. All processing is done in the cloud, and users pay the host provider monthly (again, typically, on a per gigabyte basis) to rent storage space and access the hosted data. Automated systems allow users to upload data and initiate processing themselves, at any time of day or night.

By standardizing processes, automating workflows and eliminating personnel costs, hosted discovery service providers can offer sophisticated e-discovery services at historically low prices. Though not the optimum approach for very large data sets or unconventional file types, automated and hosted processing and review promises to make e-discovery feasible and affordable in more matters.

Consolidation

Another trend that shows no sign of abating in 2017 is the consolidation of e-discovery software and service providers as companies gobble each other up. A decade ago, fear, intimidation and technical incompetence made lawyers and clients easy prey for e-discovery vendors charging premium prices. Everyone charged a fortune, so everyone—particularly the service providers themselves—assumed that's what e-discovery costs (*i.e.*, a bundle) and the gravy train would run

forever. Price gouging was aided by systematic pricing obfuscation, making apples-to-apples comparisons difficult. Over time, buyers of e-discovery services came to see how much those offerings were merely commodities, and the bottom fell out of the market as sellers embarked on a death race to the bottom on pricing. The result has been that providers—including some of the biggest names in the industry--had to fold their tents and sold out to their competition in dozens of face-saving mergers.

Fortunately for consumers, consolidation has yet to prompt price increases; however, slim margins and commoditization still plague the survivors, who continue to collapse into one another at a rate of attrition not offset by startups. It's a buyers' market for e-discovery; but, it behooves the buyer to understand what they are getting.

Attorney Competence

After years trying to persuade lawyers to acquire the barest technical fundamentals of e-discovery, I never cease to marvel at the ingenuity and compelling arguments my trial lawyer colleagues use to explain why they shouldn't need to know this "e-stuff." But, no thanks to me, the battleship is turning in other states, and a conversation has started about the need to equip the next generation of lawyers with the technical knowledge they need to thrive in an era when all information is digital and all evidence electronic. In 2015, California issued a formal ethics opinion requiring that counsel involved in matters involving electronically-stored information to either 'learn it, get help or get out.' The opinion sets out nine skill sets that lawyers dealing with e-discovery must possess or be obliged to decline the representation. The Opinion notes that an attorney handling e-discovery matters, either by themselves or in association with competent co-counsel or expert consultants, should be able to:

- Initially assess e-discovery needs and issues, if any;
- Implement/cause to implement appropriate ESI preservation procedures;
- Analyze and understand a client's ESI systems and storage;
- Advise the client on available options for collection and preservation of ESI;
- Identify custodians of potentially relevant ESI;
- Engage in competent and meaningful meet and confer with opposing counsel concerning an e-discovery plan;
- Perform data searches;
- Collect responsive ESI in a manner that preserves the integrity of that ESI; and
- Produce responsive non-privileged ESI in a recognized and appropriate manner.

The State Bar of California Standing Committee on Professional Responsibility and Conduct, Formal Opinion No. 2015-193.

Likewise, the state of Florida now mandates that that its lawyers obtain three hours of technical training each year, in addition to its existing MCLE requirements. Most states have nothing of this nature and don't even offer MCLE credit for information technology training.

A Dozen E-Discovery Strategies for Requesting and Producing Parties
E-Discovery Strategy for Requesting Parties

1. Anticipate sources: Just because you don't know all sources of potentially relevant information held by your opponent doesn't mean you can't anticipate some of them.
2. Be specific in your preservation demand. Use it to inform and close doors.
3. Lose the boilerplate discovery request. ESI isn't just another flavor of "document."
4. Supply a written agenda for meet and confer and enough time and guidance to respond.
5. *Always* specify forms of production, spec the load file and seek native forms when useful.
6. Be ready to articulate the objective behind any request for data and metadata.
7. Gear the timing of e-discovery to insure readiness for depositions.
8. Scrutinize the capabilities and limits of your opponent's electronic search methodology.
9. Know what you want most: *discovery or sanctions.*
10. E-discovery is a marathon, not a sprint. Tenacity pays dividends. Build your record.
11. Come to court armed with metrics. One good example is better than a slew of suspicion.
12. Always be prepared to address proportionality objections.

E-Discovery Strategy for Producing Parties

1. Initiate a legal hold immediately, and draft the hold notice with its discovery in mind.
2. Never state anything is gone without verification, especially when dealing with IT staff.
3. Respond to preservation demands with a written notice of what you will and won't do.
4. Be proactive. Present a reasonable e-discovery plan and a responsive proposal.
5. Requesting parties so want to get *something* they will agree to almost *anything*.
6. Seek to shift costs whenever feasible, even when you will not prevail.
7. Come to court armed with metrics. Quantify cost. Use real numbers, not extrapolations.
8. Promote use of highly precise keyword searches as these are least helpful to opponents.
9. Test to insure your searches pick up known responsive and privileged items.
10. Avoid categorical representations about ESI as they rarely survive scrutiny.
11. Set reasonable parameters limiting collection and search (custodian, interval, file types)
12. As rational, demand reciprocity in preservation, collection, search and production.

Introduction to Discovery in U.S. Civil Litigation

Until the mid-20th century, the trial of a civil lawsuit was an exercise in ambush. Parties to litigation knew little about an opponent's claims or defenses until aired in open court. A lawyer's only means to know what witnesses would say was to somehow find them before trial and persuade them to talk about the case. Witnesses weren't obliged to speak with counsel, and even when they did so, what they volunteered outside of court might change markedly when under oath on the stand. Too, at law, there was no right to see documentary evidence before trial.

John Henry Wigmore, nicely summed up the situation in his seminal, *A Treatise on the System of Evidence in Trial at Common Law* (1904). Citing the Latin maxim, *nemo tenetur armare adversarium suum contra se* ("no one is bound to arm his adversary against himself"), Wigmore explained:

> To require the disclosure to an adversary of the evidence that is to be produced would be repugnant to all sportsmanlike instincts. Rather permit you to preserve the secret of your tactics, to lock up your documents in the vault, to send your witness to board in some obscure village, and then, reserving your evidential resources until the final moment, to marshal them at the trial before your surprised and dismayed antagonist, and thus overwhelm him. Such was the spirit of the common law; and such in part it still is. It did not defend or condone trickery and deception; but it did regard the concealment of one's evidential resources and the preservation of the opponent's defenseless ignorance as a fair and irreproachable accompaniment of the game of litigation.
> Id. At Vol. III, §1845, p. 2402.

Our forebears at common law[3] feared that disclosure of evidence would facilitate unscrupulous efforts to tamper with witnesses and promote the forging of false evidence. The element of surprise was thought to promote integrity of process.

Legal reformers hated "trial by ambush" and, in the late-1930's, they sought to eliminate surprise and chicanery in U.S. courts by letting litigants obtain information about an opponent's case before trial in a process dubbed "discovery."[4] The reformer's goal was to streamline the trial process and enable litigants to better assess the merits of the dispute and settle their differences without need of a trial.

---

[3] "Common law" refers to the law as declared by judges in judicial decisions ("precedent") rather than rules established in statutes enacted by legislative bodies.

[4] That is not to say that discovery was unknown. Many jurisdictions offered a mechanism for a Bill of Discovery, essentially a separate suit in equity geared to obtaining testimony or documents in support of one's own position. However, Bills of Discovery typically made no provision for obtaining information about *an opponent's* claims, defenses or evidence—which is, of course, what one would most desire. As well, some states experimented with procedural codes that allowed for discovery of documents and taking of testimony (*e.g.*, David Dudley Field II's model code). For a comprehensive treatment of the topic, see, Ragland, George, Jr., Discovery Before Trial, 1932.

After three years of drafting and debate, the first Federal Rules of Civil Procedure went into effect on September 16, 1938. Though amended many times since, the tools of discovery contained in those nascent Rules endure to this day:

- Oral and written depositions (Rules 30 and 31);
- Interrogatories (Rule 33);
- Requests to inspect and copy documents and to inspect tangible and real property (Rule 34);
- Physical and mental examinations of persons (Rule 35);
- Requests for admissions (Rule 36);
- Subpoena of witnesses and records (Rule 45).

Tools of Discovery Defined

Depositions

A deposition is an interrogation of a party or witness ("deponent") under oath, where both the questions and responses are recorded for later use in hearings or at trial. Testimony may be elicited face-to-face ("oral deposition") or by presenting a list of questions to be posed to the witness ("written deposition"). Deposition testimony may be used in lieu of a witness' testimony when a witness is not present or to impeach the witness in a proceeding when a witness offers inconsistent testimony. Deposition testimony is typically memorialized as a "transcript" made by an official court reporter, but may also be a video obtained by a videographer.

Interrogatories

Interrogatories are written questions posed by one party to another to be answered under oath. Although the responses bind the responding party much like a deposition on written questions, there is no testimony elicited nor any court reporter or videographer involved.

Requests for Production

Parties use Requests for Production to demand to inspect or obtain copies of tangible evidence and documents, and are the chief means by which parties pursue electronically stored information (ESI). Requests may also seek access to places and things.

Requests for Physical and Mental Examination

When the physical or mental status of a party is in issue (such as when damages are sought for personal injury or disability), an opposing party may seek to compel the claimant to submit to examination by a physician or other qualified examiner.

Requests for Admission

These are used to require parties to concede, under oath, that particular facts and matters are true or that a document is genuine.

Subpoena

A subpoena is a directive in the nature of a court order requiring the recipient to take some action, typically to appear and give testimony or hand over or permit inspection of specified documents or tangible evidence. Subpoenas are most commonly used to obtain evidence from persons and entities who are not parties to the lawsuit.

Strictly speaking, the Federal Rules of Civil Procedure do not characterize subpoenas as a discovery mechanism because their use is ancillary to depositions and proceedings. Still, they are employed so frequently and powerfully in discovery as to warrant mention.

Scope of Discovery Defined

Rule 26(b)(1) of the Federal Rules of Civil Procedure defines the scope of discovery this way:

> Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable.

The Federal Rules don't define what is "relevant," but the generally accepted definition is that matter is deemed relevant when it has any tendency to make a fact more or less probable. Information may be relevant even when not admissible as competent evidence, such as hearsay or documents of questionable authenticity.

The requirement that the scope of discovery be proportional to the needs of the case was added to the Rules effective December 1, 2015, although it has long been feasible for a party to object to discovery efforts as being disproportionate and seek protection from the Court.

Certain matters are deemed beyond the proper scope of discovery because they enjoy a privilege from disclosure. The most common examples of these privileged matters are confidential attorney-client communications and attorney trial preparation materials (also called "attorney work product"). Other privileged communications include confidential communications between spouses, between priest and penitent and communications protected by the Fifth Amendment of the U.S. Constitution.

Protection from Abuse and Oppression

The discovery provisions of the Federal Rules of Civil Procedure are both sword and shield. They contain tools by which litigants may resist abusive or oppressive discovery efforts. Parties have the right to object to requests and refrain from production on the strength of those objections. Parties may also seek **Protective Orders** from the court. Rule 26(c) provides:

> The court may, for good cause, issue an order to protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense, including one or more of the following:
> (A) forbidding the disclosure or discovery;
> (B) specifying terms, including time and place or the allocation of expenses, for the disclosure or discovery;
> (C) prescribing a discovery method other than the one selected by the party seeking discovery;
> (D) forbidding inquiry into certain matters, or limiting the scope of disclosure or discovery to certain matters;
> (E) designating the persons who may be present while the discovery is conducted;
> (F) requiring that a deposition be sealed and opened only on court order;
> (G) requiring that a trade secret or other confidential research, development, or commercial information not be revealed or be revealed only in a specified way; and
> (H) requiring that the parties simultaneously file specified documents or information in sealed envelopes, to be opened as the court directs.

Character and Competence in Discovery

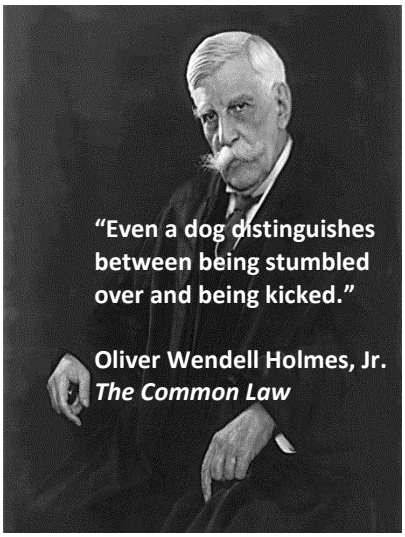
Discovery is much-maligned today as a too costly, too burdensome and too intrusive fishing expedition.[5] Certainly, its use is tainted by frequent instances of abuse and obstruction; yet, the fault for this stems from the architects of discovery--principally lawyers--and not the mechanics. Discovery is effective and even affordable when deployed with character and competence.

But, what's feasible is often at odds with what's done. There is a sufficient dearth of character and competence among segments of the bar as to ensure that discovery abuse and obstruction are commonplace; so much so that many lawyers frequently rationalize fighting fire with fire in a race to the bottom.

Character is hard to instill and harder still to measure; but, competence is not. We can require that lawyers master the ends and means of discovery—particularly of electronic discovery, where so many lag—and we can objectively assess their ken. When you can establish competence, you

[5] Such concerns are not new. Well before the original Rules went into effect, the Chairman of the Rules Advisory Committee exclaimed, "We are going to have an outburst against this discovery business unless we can hedge it about with some appearance of safety against fishing expeditions." Proceedings of the Advisory Committee (Feb. 22, 1935), at CI-209-60-0-209.61. Many still curse "this discovery business," particularly those most likely to benefit from the return of trial by ambush and those who would more-or-less do away with trials altogether.

can more easily discern character or, as Oliver Wendell Holmes, Jr. aptly observed, you can know what any dog knows; that is, the difference between being stumbled over and being kicked.



To leap the competence chasm for e-discovery, lawyers must first recognize the value and necessity of acquiring a solid foundation in the technical and legal aspects of electronic evidence, and bar associations, law schools and continuing education providers must supply the accessible and affordable educational opportunities and resources needed to help lawyers across.

The "E-Discovery Rules" (1,16,26,34 & 45)
of the Federal Rules of Civil Procedure
With Committee Notes accompanying 2006 and 2015 Amendments

Rule 1. Scope and Purpose

These rules govern the procedure in all civil actions and proceedings in the United States district courts, except as stated in Rule 81. They should be construed, administered, and employed by the court and the parties to secure the just, speedy, and inexpensive determination of every action and proceeding.

Notes

(As amended Dec. 29, 1948, eff. Oct. 20, 1949; Feb. 28, 1966, eff. July 1, 1966; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2015 Amendment

Rule 1 is amended to emphasize that just as the court should construe and administer these rules to secure the just, speedy, and inexpensive determination of every action, so the parties share the responsibility to employ the rules in the same way. Most lawyers and parties cooperate to achieve these ends. But discussions of ways to improve the administration of civil justice regularly include pleas to discourage over-use, misuse, and abuse of procedural tools that increase cost and result in delay. Effective advocacy is consistent with — and indeed depends upon — cooperative and proportional use of procedure.

This amendment does not create a new or independent source of sanctions. Neither does it abridge the scope of any other of these rules.

***

Rule 16. Pretrial Conferences; Scheduling; Management

(a) Purposes of a Pretrial Conference. In any action, the court may order the attorneys and any unrepresented parties to appear for one or more pretrial conferences for such purposes as:

(1) expediting disposition of the action;

(2) establishing early and continuing control so that the case will not be protracted because of lack of management;

(3) discouraging wasteful pretrial activities;

(4) improving the quality of the trial through more thorough preparation; and

(5) facilitating settlement.

(b) Scheduling.

(1) *Scheduling Order.* Except in categories of actions exempted by local rule, the district judge—or a magistrate judge when authorized by local rule—must issue a scheduling order:

(A) after receiving the parties' report under Rule 26(f); or

(B) after consulting with the parties' attorneys and any unrepresented parties at a scheduling conference.

(2) *Time to Issue.* The judge must issue the scheduling order as soon as practicable, but unless the judge finds good cause for delay, the judge must issue it within the earlier of 90 days after any defendant has been served with the complaint or 60 days after any defendant has appeared.

(3) *Contents of the Order.*

(A) *Required Contents.* The scheduling order must limit the time to join other parties, amend the pleadings, complete discovery, and file motions.

(B) *Permitted Contents.* The scheduling order may:

(i) modify the timing of disclosures under Rules 26(a) and 26(e)(1);

(ii) modify the extent of discovery;

(iii) provide for disclosure, discovery, or preservation of electronically stored information;

(iv) include any agreements the parties reach for asserting claims of privilege or of protection as trial-preparation material after information is produced, including agreements reached under Federal Rule of Evidence 502;

(v) direct that before moving for an order relating to discovery, the movant must request a conference with the court;

(vi) set dates for pretrial conferences and for trial; and

(vii) include other appropriate matters.

(4) *Modifying a Schedule.* A schedule may be modified only for good cause and with the judge's consent.

(c) Attendance and Matters for Consideration at a Pretrial Conference.

(1) *Attendance.* A represented party must authorize at least one of its attorneys to make stipulations and admissions about all matters that can reasonably be anticipated for discussion at a pretrial conference. If appropriate, the court may require that a party or its representative be present or reasonably available by other means to consider possible settlement.

(2) *Matters for Consideration.* At any pretrial conference, the court may consider and take appropriate action on the following matters:

(A) formulating and simplifying the issues, and eliminating frivolous claims or defenses;

(B) amending the pleadings if necessary or desirable;

(C) obtaining admissions and stipulations about facts and documents to avoid unnecessary proof, and ruling in advance on the admissibility of evidence;

(D) avoiding unnecessary proof and cumulative evidence, and limiting the use of testimony under Federal Rule of Evidence 702;

(E) determining the appropriateness and timing of summary adjudication under Rule 56;

(F) controlling and scheduling discovery, including orders affecting disclosures and discovery under Rule 26 and Rules 29 through 37;

(G) identifying witnesses and documents, scheduling the filing and exchange of any pretrial briefs, and setting dates for further conferences and for trial;

(H) referring matters to a magistrate judge or a master;

(I) settling the case and using special procedures to assist in resolving the dispute when authorized by statute or local rule;

(J) determining the form and content of the pretrial order;

(K) disposing of pending motions;

(L) adopting special procedures for managing potentially difficult or protracted actions that may involve complex issues, multiple parties, difficult legal questions, or unusual proof problems;

(M) ordering a separate trial under Rule 42(b) of a claim, counterclaim, crossclaim, third-party claim, or particular issue;

(N) ordering the presentation of evidence early in the trial on a manageable issue that might, on the evidence, be the basis for a judgment as a matter of law under Rule 50(a) or a judgment on partial findings under Rule 52(c);

(O) establishing a reasonable limit on the time allowed to present evidence; and

(P) facilitating in other ways the just, speedy, and inexpensive disposition of the action.

(d) Pretrial Orders. After any conference under this rule, the court should issue an order reciting the action taken. This order controls the course of the action unless the court modifies it.

(e) Final Pretrial Conference and Orders. The court may hold a final pretrial conference to formulate a trial plan, including a plan to facilitate the admission of evidence. The conference must be held as close to the start of trial as is reasonable, and must be attended by at least one attorney who will conduct the trial for each party and by any unrepresented party. The court may modify the order issued after a final pretrial conference only to prevent manifest injustice.

(f) Sanctions.

(1) *In General.* On motion or on its own, the court may issue any just orders, including those authorized by Rule 37(b)(2)(A)(ii)–(vii), if a party or its attorney:

(A) fails to appear at a scheduling or other pretrial conference;

(B) is substantially unprepared to participate—or does not participate in good faith—in the conference; or

(C) fails to obey a scheduling or other pretrial order.

(2) *Imposing Fees and Costs.* Instead of or in addition to any other sanction, the court must order the party, its attorney, or both to pay the reasonable expenses—including attorney's fees—incurred because of any noncompliance with this rule, unless the noncompliance was substantially justified or other circumstances make an award of expenses unjust.

Notes

(As amended Apr. 28, 1983, eff. Aug. 1, 1983; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2006 Amendment

The amendment to Rule 16(b) is designed to alert the court to the possible need to address the handling of discovery of electronically stored information early in the litigation if such discovery is expected to occur. Rule 26(f) is amended to direct the parties to discuss discovery of electronically stored information if such discovery is contemplated in the action. Form 35 is amended to call for a report to the court about the results of this discussion. In many instances, the court's involvement early in the litigation will help avoid difficulties that might otherwise arise.

Rule 16(b) is also amended to include among the topics that may be addressed in the scheduling order any agreements that the parties reach to facilitate discovery by minimizing the risk of waiver of privilege or work-product protection. Rule 26(f) is amended to add to the discovery plan the parties' proposal for the court to enter a case-management or other order adopting such an agreement. The parties may agree to various arrangements. For example, they may agree to initial provision of requested materials without waiver of privilege or protection to enable the party seeking production to designate the materials desired or protection for actual production, with the privilege review of only those materials to follow. Alternatively, they may agree that if privileged or protected information is inadvertently produced, the producing party may by timely notice assert the privilege or protection and obtain return of the materials without waiver. Other arrangements are possible. In most circumstances, a party who receives information under such an arrangement cannot assert that production of the information waived a claim of privilege or of protection as trial-preparation material.

An order that includes the parties' agreement may be helpful in avoiding delay and excessive cost in discovery. *See Manual for Complex Litigation*(4th) §11.446. Rule 16(b)(6) recognizes the propriety of including such agreements in the court's order. The rule does not provide the court with authority to enter such a case-management or other order without party agreement, or limit the court's authority to act on motion.

Committee Notes on Rules—2015 Amendment

The provision for consulting at a scheduling conference by "telephone, mail, or other means" is deleted. A scheduling conference is more effective if the court and parties engage in direct simultaneous communication. The conference may be held in person, by telephone, or by more sophisticated electronic means.

The time to issue the scheduling order is reduced to the earlier of 90 days (not 120 days) after any defendant has been served, or 60 days (not 90 days) after any defendant has appeared. This change, together with the shortened time for making service under Rule 4(m), will reduce delay at the beginning of litigation. At the same time, a new provision recognizes that the court may find good cause to extend the time to issue the scheduling order. In some cases it may be that the

parties cannot prepare adequately for a meaningful Rule 26(f) conference and then a scheduling conference in the time allowed. Litigation involving complex issues, multiple parties, and large organizations, public or private, may be more likely to need extra time to establish meaningful collaboration between counsel and the people who can supply the information needed to participate in a useful way. Because the time for the Rule 26(f) conference is geared to the time for the scheduling conference or order, an order extending the time for the scheduling conference will also extend the time for the Rule 26(f) conference. But in most cases it will be desirable to hold at least a first scheduling conference in the time set by the rule.

Three items are added to the list of permitted contents in Rule 16(b)(3)(B).

The order may provide for preservation of electronically stored information, a topic also added to the provisions of a discovery plan under Rule 26(f)(3)(C). Parallel amendments of Rule 37(e) recognize that a duty to preserve discoverable information may arise before an action is filed.

The order also may include agreements incorporated in a court order under Evidence Rule 502 controlling the effects of disclosure of information covered by attorney-client privilege or work-product protection, a topic also added to the provisions of a discovery plan under Rule 26(f)(3)(D).

Finally, the order may direct that before filing a motion for an order relating to discovery the movant must request a conference with the court. Many judges who hold such conferences find them an efficient way to resolve most discovery disputes without the delay and burdens attending a formal motion, but the decision whether to require such conferences is left to the discretion of the judge in each case.

***

Rule 26. Duty to Disclose; General Provisions Governing Discovery

(a) Required Disclosures.

(1) *Initial Disclosure.*

(A) *In General.* Except as exempted by Rule 26(a)(1)(B) or as otherwise stipulated or ordered by the court, a party must, without awaiting a discovery request, provide to the other parties:

(i) the name and, if known, the address and telephone number of each individual likely to have discoverable information—along with the subjects of that information—that the disclosing party may use to support its claims or defenses, unless the use would be solely for impeachment;

(ii) a copy—or a description by category and location—of all documents, electronically stored information, and tangible things that the disclosing party has in its possession, custody, or control and may use to support its claims or defenses, unless the use would be solely for impeachment;

(iii) a computation of each category of damages claimed by the disclosing party—who must also make available for inspection and copying as under Rule 34 the documents or other evidentiary material, unless privileged or protected from disclosure, on which each computation is based, including materials bearing on the nature and extent of injuries suffered; and

(iv) for inspection and copying as under Rule 34, any insurance agreement under which an insurance business may be liable to satisfy all or part of a possible judgment in the action or to indemnify or reimburse for payments made to satisfy the judgment.

(B) *Proceedings Exempt from Initial Disclosure.* The following proceedings are exempt from initial disclosure:

(i) an action for review on an administrative record;

(ii) a forfeiture action in rem arising from a federal statute;

(iii) a petition for habeas corpus or any other proceeding to challenge a criminal conviction or sentence;

(iv) an action brought without an attorney by a person in the custody of the United States, a state, or a state subdivision;

(v) an action to enforce or quash an administrative summons or subpoena;

(vi) an action by the United States to recover benefit payments;

(vii) an action by the United States to collect on a student loan guaranteed by the United States;

(viii) a proceeding ancillary to a proceeding in another court; and

(ix) an action to enforce an arbitration award.

(C) *Time for Initial Disclosures—In General.* A party must make the initial disclosures at or within 14 days after the parties' Rule 26(f)conference unless a different time is set by stipulation or court order, or unless a party objects during the conference that initial disclosures are not appropriate in this action and states the objection in the proposed discovery plan. In ruling on the objection, the court must determine what disclosures, if any, are to be made and must set the time for disclosure.

(D) *Time for Initial Disclosures—For Parties Served or Joined Later.* A party that is first served or otherwise joined after the Rule 26(f)conference must make the initial disclosures within 30 days after being served or joined, unless a different time is set by stipulation or court order.

(E) *Basis for Initial Disclosure; Unacceptable Excuses.* A party must make its initial disclosures based on the information then reasonably available to it. A party is not excused from making its disclosures because it has not fully investigated the case or because it challenges the sufficiency of another party's disclosures or because another party has not made its disclosures.

(2) *Disclosure of Expert Testimony.*

(A) *In General.* In addition to the disclosures required by Rule 26(a)(1), a party must disclose to the other parties the identity of any witness it may use at trial to present evidence under Federal Rule of Evidence 702, 703, or 705.

(B) *Witnesses Who Must Provide a Written Report.* Unless otherwise stipulated or ordered by the court, this disclosure must be accompanied by a written report—prepared and signed by the witness—if the witness is one retained or specially employed to provide expert testimony in the case or one whose duties as the party's employee regularly involve giving expert testimony. The report must contain:

(i) a complete statement of all opinions the witness will express and the basis and reasons for them;

(ii) the facts or data considered by the witness in forming them;

(iii) any exhibits that will be used to summarize or support them;

(iv) the witness's qualifications, including a list of all publications authored in the previous 10 years;

(v) a list of all other cases in which, during the previous 4 years, the witness testified as an expert at trial or by deposition; and

(vi) a statement of the compensation to be paid for the study and testimony in the case.

(C) *Witnesses Who Do Not Provide a Written Report.* Unless otherwise stipulated or ordered by the court, if the witness is not required to provide a written report, this disclosure must state:

(i) the subject matter on which the witness is expected to present evidence under Federal Rule of Evidence 702, 703, or 705; and

(ii) a summary of the facts and opinions to which the witness is expected to testify.

(D) Time to Disclose Expert Testimony. A party must make these disclosures at the times and in the sequence that the court orders. Absent a stipulation or a court order, the disclosures must be made:

(i) at least 90 days before the date set for trial or for the case to be ready for trial; or

(ii) if the evidence is intended solely to contradict or rebut evidence on the same subject matter identified by another party under Rule 26(a)(2)(B) or (C), within 30 days after the other party's disclosure.

(E) *Supplementing the Disclosure.* The parties must supplement these disclosures when required under Rule 26(e).

(3) *Pretrial Disclosures.*

(A) *In General.* In addition to the disclosures required by Rule 26(a)(1) and (2), a party must provide to the other parties and promptly file the following information about the evidence that it may present at trial other than solely for impeachment:

(i) the name and, if not previously provided, the address and telephone number of each witness—separately identifying those the party expects to present and those it may call if the need arises;

(ii) the designation of those witnesses whose testimony the party expects to present by deposition and, if not taken stenographically, a transcript of the pertinent parts of the deposition; and

(iii) an identification of each document or other exhibit, including summaries of other evidence—separately identifying those items the party expects to offer and those it may offer if the need arises.

(B) *Time for Pretrial Disclosures; Objections.* Unless the court orders otherwise, these disclosures must be made at least 30 days before trial. Within 14 days after they are made, unless the court sets a different time, a party may serve and promptly file a list of the following objections: any objections to the use under Rule 32(a) of a deposition designated by another party under Rule 26(a)(3)(A)(ii); and any objection, together with the grounds for it, that may be made to the admissibility of materials identified under Rule 26(a)(3)(A)(iii). An objection not so made—except for one under Federal Rule of Evidence 402 or 403—is waived unless excused by the court for good cause.

(4) *Form of Disclosures.* Unless the court orders otherwise, all disclosures under Rule 26(a) must be in writing, signed, and served.

(b) Discovery Scope and Limits.

(1) *Scope in General.* Unless otherwise limited by court order, the scope of discovery is as follows: Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable..

(2) *Limitations on Frequency and Extent.*

(A) *When Permitted.* By order, the court may alter the limits in these rules on the number of depositions and interrogatories or on the length of depositions under Rule 30. By order or local rule, the court may also limit the number of requests under Rule 36.

(B) *Specific Limitations on Electronically Stored Information.* A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.

(C) *When Required.* On motion or on its own, the court must limit the frequency or extent of discovery otherwise allowed by these rules or by local rule if it determines that:

(i) the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive;

(ii) the party seeking discovery has had ample opportunity to obtain the information by discovery in the action; or

(iii) the proposed discovery is outside the scope permitted by Rule 26(b)(1).

(3) *Trial Preparation: Materials.*

(A) *Documents and Tangible Things.* Ordinarily, a party may not discover documents and tangible things that are prepared in anticipation of litigation or for trial by or for another party or its representative (including the other party's attorney, consultant, surety, indemnitor, insurer, or agent). But, subject to Rule 26(b)(4), those materials may be discovered if:

(i) they are otherwise discoverable under Rule 26(b)(1); and

(ii) the party shows that it has substantial need for the materials to prepare its case and cannot, without undue hardship, obtain their substantial equivalent by other means.

(B) *Protection Against Disclosure.* If the court orders discovery of those materials, it must protect against disclosure of the mental impressions, conclusions, opinions, or legal theories of a party's attorney or other representative concerning the litigation.

(C) *Previous Statement.* Any party or other person may, on request and without the required showing, obtain the person's own previous statement about the action or its subject matter. If the request is refused, the person may move for a court order, and Rule 37(a)(5) applies to the award of expenses. A previous statement is either:

(i) a written statement that the person has signed or otherwise adopted or approved; or

(ii) a contemporaneous stenographic, mechanical, electrical, or other recording—or a transcription of it—that recites substantially verbatim the person's oral statement.

(4) *Trial Preparation: Experts.*

(A) *Deposition of an Expert Who May Testify.* A party may depose any person who has been identified as an expert whose opinions may be presented at trial. If Rule 26(a)(2)(B) requires a report from the expert, the deposition may be conducted only after the report is provided.

(B) *Trial-Preparation Protection for Draft Reports or Disclosures.* Rules 26(b)(3)(A) and (B) protect drafts of any report or disclosure required under Rule 26(a)(2), regardless of the form in which the draft is recorded.

(C) *Trial-Preparation Protection for Communications Between a Party's Attorney and Expert Witnesses.* Rules 26(b)(3)(A) and (B) protect communications between the party's attorney and any witness required to provide a report under Rule 26(a)(2)(B), regardless of the form of the communications, except to the extent that the communications:

(i) relate to compensation for the expert's study or testimony;

(ii) identify facts or data that the party's attorney provided and that the expert considered in forming the opinions to be expressed; or

(iii) identify assumptions that the party's attorney provided and that the expert relied on in forming the opinions to be expressed.

(D) *Expert Employed Only for Trial Preparation.* Ordinarily, a party may not, by interrogatories or deposition, discover facts known or opinions held by an expert who has been retained or specially employed by another party in anticipation of litigation or to prepare for trial and who is not expected to be called as a witness at trial. But a party may do so only:

(i) as provided in Rule 35(b); or

(ii) on showing exceptional circumstances under which it is impracticable for the party to obtain facts or opinions on the same subject by other means.

(E) *Payment.* Unless manifest injustice would result, the court must require that the party seeking discovery:

(i) pay the expert a reasonable fee for time spent in responding to discovery under Rule 26(b)(4)(A) or (D); and

(ii) for discovery under (D), also pay the other party a fair portion of the fees and expenses it reasonably incurred in obtaining the expert's facts and opinions.

(5) *Claiming Privilege or Protecting Trial-Preparation Materials.*

(A) *Information Withheld.* When a party withholds information otherwise discoverable by claiming that the information is privileged or subject to protection as trial-preparation material, the party must:

(i) expressly make the claim; and

(ii) describe the nature of the documents, communications, or tangible things not produced or disclosed—and do so in a manner that, without revealing information itself privileged or protected, will enable other parties to assess the claim.

(B) *Information Produced.* If information produced in discovery is subject to a claim of privilege or of protection as trial-preparation material, the party making the claim may notify any party that received the information of the claim and the basis for it. After being notified, a party must promptly return, sequester, or destroy the specified information and any copies it has; must not use or disclose the information until the claim is resolved; must take reasonable steps to retrieve

the information if the party disclosed it before being notified; and may promptly present the information to the court under seal for a determination of the claim. The producing party must preserve the information until the claim is resolved.

(c) Protective Orders.

(1) *In General.* A party or any person from whom discovery is sought may move for a protective order in the court where the action is pending—or as an alternative on matters relating to a deposition, in the court for the district where the deposition will be taken. The motion must include a certification that the movant has in good faith conferred or attempted to confer with other affected parties in an effort to resolve the dispute without court action. The court may, for good cause, issue an order to protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense, including one or more of the following:

(A) forbidding the disclosure or discovery;

(B) specifying terms, including time and place or the allocation of expenses, for the disclosure or discovery;

(C) prescribing a discovery method other than the one selected by the party seeking discovery;

(D) forbidding inquiry into certain matters, or limiting the scope of disclosure or discovery to certain matters;

(E) designating the persons who may be present while the discovery is conducted;

(F) requiring that a deposition be sealed and opened only on court order;

(G) requiring that a trade secret or other confidential research, development, or commercial information not be revealed or be revealed only in a specified way; and

(H) requiring that the parties simultaneously file specified documents or information in sealed envelopes, to be opened as the court directs.

(2) *Ordering Discovery.* If a motion for a protective order is wholly or partly denied, the court may, on just terms, order that any party or person provide or permit discovery.

(3) *Awarding Expenses.* Rule 37(a)(5) applies to the award of expenses.

(d) Timing and Sequence of Discovery.

(1) *Timing.* A party may not seek discovery from any source before the parties have conferred as required by Rule 26(f), except in a proceeding exempted from initial disclosure under Rule 26(a)(1)(B), or when authorized by these rules, by stipulation, or by court order.

(2) Early Rule 34 Requests.

Time to Deliver. More than 21 days after the summons and complaint are served on a party, a request under Rule 34 may be delivered:

(i) to that party by any other party, and

(ii) by that party to any plaintiff or to any other party that has been served.

(B) When Considered Served. The request is considered to have been served at the first Rule 26(f) conference.

(3) *Sequence.* Unless the parties stipulate or the court orders otherwise for the parties' and witnesses' convenience and in the interests of justice:

(A) methods of discovery may be used in any sequence; and

(B) discovery by one party does not require any other party to delay its discovery.

(e) Supplementing Disclosures and Responses.

(1) *In General.* A party who has made a disclosure under Rule 26(a)—or who has responded to an interrogatory, request for production, or request for admission—must supplement or correct its disclosure or response:

(A) in a timely manner if the party learns that in some material respect the disclosure or response is incomplete or incorrect, and if the additional or corrective information has not otherwise been made known to the other parties during the discovery process or in writing; or

(B) as ordered by the court.

(2) *Expert Witness.* For an expert whose report must be disclosed under Rule 26(a)(2)(B), the party's duty to supplement extends both to information included in the report and to information given during the expert's deposition. Any additions or changes to this information must be disclosed by the time the party's pretrial disclosures under Rule 26(a)(3) are due.

(f) Conference of the Parties; Planning for Discovery.

(1) *Conference Timing.* Except in a proceeding exempted from initial disclosure under Rule 26(a)(1)(B) or when the court orders otherwise, the parties must confer as soon as practicable—and in any event at least 21 days before a scheduling conference is to be held or a scheduling order is due under Rule 16(b).

(2) *Conference Content; Parties' Responsibilities.* In conferring, the parties must consider the nature and basis of their claims and defenses and the possibilities for promptly settling or resolving the case; make or arrange for the disclosures required by Rule 26(a)(1); discuss any issues about preserving discoverable information; and develop a proposed discovery plan. The attorneys of record and all unrepresented parties that have appeared in the case are jointly responsible for arranging the conference, for attempting in good faith to agree on the proposed discovery plan, and for submitting to the court within 14 days after the conference a written report outlining the plan. The court may order the parties or attorneys to attend the conference in person.

(3) *Discovery Plan.* A discovery plan must state the parties' views and proposals on:

(A) what changes should be made in the timing, form, or requirement for disclosures under Rule 26(a), including a statement of when initial disclosures were made or will be made;

(B) the subjects on which discovery may be needed, when discovery should be completed, and whether discovery should be conducted in phases or be limited to or focused on particular issues;

(C) any issues about disclosure, discovery, or preservation of electronically stored information, including the form or forms in which it should be produced;

(D) any issues about claims of privilege or of protection as trial-preparation materials, including—if the parties agree on a procedure to assert these claims after production—whether to ask the court to include their agreement in an order under Federal Rule of Evidence 502;

(E) what changes should be made in the limitations on discovery imposed under these rules or by local rule, and what other limitations should be imposed; and

(F) any other orders that the court should issue under Rule 26(c) or under Rule 16(b) and (c).

(4) *Expedited Schedule.* If necessary to comply with its expedited schedule for Rule 16(b) conferences, a court may by local rule:

(A) require the parties' conference to occur less than 21 days before the scheduling conference is held or a scheduling order is due under Rule 16(b); and

(B) require the written report outlining the discovery plan to be filed less than 14 days after the parties' conference, or excuse the parties from submitting a written report and permit them to report orally on their discovery plan at the Rule 16(b) conference.

(g) Signing Disclosures and Discovery Requests, Responses, and Objections.

(1) *Signature Required; Effect of Signature.* Every disclosure under Rule 26(a)(1) or (a)(3) and every discovery request, response, or objection must be signed by at least one attorney of record in the attorney's own name—or by the party personally, if unrepresented—and must state the signer's address, e-mail address, and telephone number. By signing, an attorney or party certifies that to the best of the person's knowledge, information, and belief formed after a reasonable inquiry:

(A) with respect to a disclosure, it is complete and correct as of the time it is made; and

(B) with respect to a discovery request, response, or objection, it is:

(i) consistent with these rules and warranted by existing law or by a nonfrivolous argument for extending, modifying, or reversing existing law, or for establishing new law;

(ii) not interposed for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation; and

(iii) neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action.

(2) *Failure to Sign.* Other parties have no duty to act on an unsigned disclosure, request, response, or objection until it is signed, and the court must strike it unless a signature is promptly supplied after the omission is called to the attorney's or party's attention.

(3) *Sanction for Improper Certification.* If a certification violates this rule without substantial justification, the court, on motion or on its own, must impose an appropriate sanction on the signer, the party on whose behalf the signer was acting, or both. The sanction may include an order to pay the reasonable expenses, including attorney's fees, caused by the violation.

Notes

(As amended Dec. 27, 1946, eff. Mar. 19, 1948; Jan. 21, 1963, eff. July 1, 1963; Feb. 28, 1966, eff. July 1, 1966; Mar. 30, 1970, eff. July 1, 1970; Apr. 29, 1980, eff. Aug. 1, 1980; Apr. 28, 1983, eff. Aug. 1, 1983; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 17, 2000, eff.

Dec. 1, 2000; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 28, 2010, eff. Dec. 1, 2010; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2006 Amendment

Subdivision (a). Rule 26(a)(1)(B) is amended to parallel Rule 34(a) by recognizing that a party must disclose electronically stored information as well as documents that it may use to support its claims or defenses. The term "electronically stored information" has the same broad meaning in Rule 26(a)(1) as in Rule 34(a). This amendment is consistent with the 1993 addition of Rule 26(a)(1)(B). The term "data compilations" is deleted as unnecessary because it is a subset of both documents and electronically stored information.

Changes Made After Publication and Comment. As noted in the introduction [omitted], this provision was not included in the published rule. It is included as a conforming amendment, to make Rule 26(a)(1) consistent with the changes that were included in the published proposals.

[*Subdivision (a)(1)(E)*.] Civil forfeiture actions are added to the list of exemptions from Rule 26(a)(1) disclosure requirements. These actions are governed by new Supplemental Rule G. Disclosure is not likely to be useful.

Subdivision (b)(2). The amendment to Rule 26(b)(2) is designed to address issues raised by difficulties in locating, retrieving, and providing discovery of some electronically stored information. Electronic storage systems often make it easier to locate and retrieve information. These advantages are properly taken into account in determining the reasonable scope of discovery in a particular case. But some sources of electronically stored information can be accessed only with substantial burden and cost. In a particular case, these burdens and costs may make the information on such sources not reasonably accessible.

It is not possible to define in a rule the different types of technological features that may affect the burdens and costs of accessing electronically stored information. Information systems are designed to provide ready access to information used in regular ongoing activities. They also may be designed so as to provide ready access to information that is not regularly used. But a system may retain information on sources that are accessible only by incurring substantial burdens or costs. Subparagraph (B) is added to regulate discovery from such sources.

Under this rule, a responding party should produce electronically stored information that is relevant, not privileged, and reasonably accessible, subject to the (b)(2)(C) limitations that apply to all discovery. The responding party must also identify, by category or type, the sources containing potentially responsive information that it is neither searching nor producing. The identification should, to the extent possible, provide enough detail to enable the requesting party

to evaluate the burdens and costs of providing the discovery and the likelihood of finding responsive information on the identified sources.

A party's identification of sources of electronically stored information as not reasonably accessible does not relieve the party of its common-law or statutory duties to preserve evidence. Whether a responding party is required to preserve unsearched sources of potentially responsive information that it believes are not reasonably accessible depends on the circumstances of each case. It is often useful for the parties to discuss this issue early in discovery.

The volume of—and the ability to search—much electronically stored information means that in many cases the responding party will be able to produce information from reasonably accessible sources that will fully satisfy the parties' discovery needs. In many circumstances the requesting party should obtain and evaluate the information from such sources before insisting that the responding party search and produce information contained on sources that are not reasonably accessible. If the requesting party continues to seek discovery of information from sources identified as not reasonably accessible, the parties should discuss the burdens and costs of accessing and retrieving the information, the needs that may establish good cause for requiring all or part of the requested discovery even if the information sought is not reasonably accessible, and conditions on obtaining and producing the information that may be appropriate.

If the parties cannot agree whether, or on what terms, sources identified as not reasonably accessible should be searched and discoverable information produced, the issue may be raised either by a motion to compel discovery or by a motion for a protective order. The parties must confer before bringing either motion. If the parties do not resolve the issue and the court must decide, the responding party must show that the identified sources of information are not reasonably accessible because of undue burden or cost. The requesting party may need discovery to test this assertion. Such discovery might take the form of requiring the responding party to conduct a sampling of information contained on the sources identified as not reasonably accessible; allowing some form of inspection of such sources; or taking depositions of witnesses knowledgeable about the responding party's information systems.

Once it is shown that a source of electronically stored information is not reasonably accessible, the requesting party may still obtain discovery by showing good cause, considering the limitations of Rule 26(b)(2)(C) that balance the costs and potential benefits of discovery. The decision whether to require a responding party to search for and produce information that is not reasonably accessible depends not only on the burdens and costs of doing so, but also on whether those burdens and costs can be justified in the circumstances of the case. Appropriate considerations may include: (1) the specificity of the discovery request; (2) the quantity of information available from other and more easily accessed sources; (3) the failure to produce

relevant information that seems likely to have existed but is no longer available on more easily accessed sources; (4) the likelihood of finding relevant, responsive information that cannot be obtained from other, more easily accessed sources; (5) predictions as to the importance and usefulness of the further information; (6) the importance of the issues at stake in the litigation; and (7) the parties' resources.

The responding party has the burden as to one aspect of the inquiry—whether the identified sources are not reasonably accessible in light of the burdens and costs required to search for, retrieve, and produce whatever responsive information may be found. The requesting party has the burden of showing that its need for the discovery outweighs the burdens and costs of locating, retrieving, and producing the information. In some cases, the court will be able to determine whether the identified sources are not reasonably accessible and whether the requesting party has shown good cause for some or all of the discovery, consistent with the limitations of Rule 26(b)(2)(C), through a single proceeding or presentation. The good-cause determination, however, may be complicated because the court and parties may know little about what information the sources identified as not reasonably accessible might contain, whether it is relevant, or how valuable it may be to the litigation. In such cases, the parties may need some focused discovery, which may include sampling of the sources, to learn more about what burdens and costs are involved in accessing the information, what the information consists of, and how valuable it is for the litigation in light of information that can be obtained by exhausting other opportunities for discovery.

The good-cause inquiry and consideration of the Rule 26(b)(2)(C) limitations are coupled with the authority to set conditions for discovery. The conditions may take the form of limits on the amount, type, or sources of information required to be accessed and produced. The conditions may also include payment by the requesting party of part or all of the reasonable costs of obtaining information from sources that are not reasonably accessible. A requesting party's willingness to share or bear the access costs may be weighed by the court in determining whether there is good cause. But the producing party's burdens in reviewing the information for relevance and privilege may weigh against permitting the requested discovery.

The limitations of Rule 26(b)(2)(C) continue to apply to all discovery of electronically stored information, including that stored on reasonably accessible electronic sources.

Changes Made after Publication and Comment. This recommendation modifies the version of the proposed rule amendment as published. Responding to comments that the published proposal seemed to require identification of information that cannot be identified because it is not reasonably accessible, the rule text was clarified by requiring identification of sources that are not

reasonably accessible. The test of reasonable accessibility was clarified by adding "because of undue burden or cost."

The published proposal referred only to a motion by the requesting party to compel discovery. The rule text has been changed to recognize that the responding party may wish to determine its search and potential preservation obligations by moving for a protective order.

The provision that the court may for good cause order discovery from sources that are not reasonably accessible is expanded in two ways. It now states specifically that the requesting party is the one who must show good cause, and it refers to consideration of the limitations on discovery set out in present Rule 26(b)(2)(i), (ii), and (iii).

The published proposal was added at the end of present Rule 26(b)(2). It has been relocated to become a new subparagraph (B), allocating present Rule 26(b)(2) to new subparagraphs (A) and (C). The Committee Note was changed to reflect the rule text revisions. It also was shortened. The shortening was accomplished in part by deleting references to problems that are likely to become antique as technology continues to evolve, and in part by deleting passages that were at a level of detail better suited for a practice manual than a Committee Note.

The changes from the published proposed amendment to Rule 26(b)(2) are set out below. [Omitted]

Subdivision (b)(5). The Committee has repeatedly been advised that the risk of privilege waiver, and the work necessary to avoid it, add to the costs and delay of discovery. When the review is of electronically stored information, the risk of waiver, and the time and effort required to avoid it, can increase substantially because of the volume of electronically stored information and the difficulty in ensuring that all information to be produced has in fact been reviewed. Rule 26(b)(5)(A) provides a procedure for a party that has withheld information on the basis of privilege or protection as trial-preparation material to make the claim so that the requesting party can decide whether to contest the claim and the court can resolve the dispute. Rule 26(b)(5)(B) is added to provide a procedure for a party to assert a claim of privilege or trial-preparation material protection after information is produced in discovery in the action and, if the claim is contested, permit any party that received the information to present the matter to the court for resolution.

Rule 26(b)(5)(B) does not address whether the privilege or protection that is asserted after production was waived by the production. The courts have developed principles to determine whether, and under what circumstances, waiver results from inadvertent production of privileged or protected information. Rule 26(b)(5)(B) provides a procedure for presenting and addressing these issues. Rule 26(b)(5)(B) works in tandem with Rule 26(f), which is amended to direct the

parties to discuss privilege issues in preparing their discovery plan, and which, with amended Rule 16(b), allows the parties to ask the court to include in an order any agreements the parties reach regarding issues of privilege or trial-preparation material protection. Agreements reached under Rule 26(f)(4) and orders including such agreements entered under Rule 16(b)(6) may be considered when a court determines whether a waiver has occurred. Such agreements and orders ordinarily control if they adopt procedures different from those in Rule 26(b)(5)(B).

A party asserting a claim of privilege or protection after production must give notice to the receiving party. That notice should be in writing unless the circumstances preclude it. Such circumstances could include the assertion of the claim during a deposition. The notice should be as specific as possible in identifying the information and stating the basis for the claim. Because the receiving party must decide whether to challenge the claim and may sequester the information and submit it to the court for a ruling on whether the claimed privilege or protection applies and whether it has been waived, the notice should be sufficiently detailed so as to enable the receiving party and the court to understand the basis for the claim and to determine whether waiver has occurred. Courts will continue to examine whether a claim of privilege or protection was made at a reasonable time when delay is part of the waiver determination under the governing law.

After receiving notice, each party that received the information must promptly return, sequester, or destroy the information and any copies it has. The option of sequestering or destroying the information is included in part because the receiving party may have incorporated the information in protected trial-preparation materials. No receiving party may use or disclose the information pending resolution of the privilege claim. The receiving party may present to the court the questions whether the information is privileged or protected as trial-preparation material, and whether the privilege or protection has been waived. If it does so, it must provide the court with the grounds for the privilege or protection specified in the producing party's notice, and serve all parties. In presenting the question, the party may use the content of the information only to the extent permitted by the applicable law of privilege, protection for trial-preparation material, and professional responsibility.

If a party disclosed the information to nonparties before receiving notice of a claim of privilege or protection as trial-preparation material, it must take reasonable steps to retrieve the information and to return it, sequester it until the claim is resolved, or destroy it.

Whether the information is returned or not, the producing party must preserve the information pending the court's ruling on whether the claim of privilege or of protection is properly asserted and whether it was waived. As with claims made under Rule 26(b)(5)(A), there may be no ruling if the other parties do not contest the claim.

Changes Made After Publication and Comment. The rule recommended for approval is modified from the published proposal. The rule is expanded to include trial-preparation protection claims in addition to privilege claims.

The published proposal referred to production "without intending to waive a claim of privilege." This reference to intent was deleted because many courts include intent in the factors that determine whether production waives privilege.

The published proposal required that the producing party give notice "within a reasonable time." The time requirement was deleted because it seemed to implicate the question whether production effected a waiver, a question not addressed by the rule, and also because a receiving party cannot practicably ignore a notice that it believes was unreasonably delayed. The notice procedure was further changed to require that the producing party state the basis for the claim.

Two statements in the published Note have been brought into the rule text. The first provides that the receiving party may not use or disclose the information until the claim is resolved. The second provides that if the receiving party disclosed the information before being notified, it must take reasonable steps to retrieve it. [1]

The rule text was expanded by adding a provision that the receiving party may promptly present the information to the court under seal for a determination of the claim.

The published proposal provided that the producing party must comply with Rule 26(b)(5)(A) after making the claim. This provision was deleted as unnecessary.

Changes are made in the Committee Note to reflect the changes in the rule text.

The changes from the published rule are shown below. [Omitted]

Subdivision (f). Rule 26(f) is amended to direct the parties to discuss discovery of electronically stored information during their discovery-planning conference. The rule focuses on "issues relating to disclosure or discovery of electronically stored information"; the discussion is not required in cases not involving electronic discovery, and the amendment imposes no additional requirements in those cases. When the parties do anticipate disclosure or discovery of electronically stored information, discussion at the outset may avoid later difficulties or ease their resolution.

When a case involves discovery of electronically stored information, the issues to be addressed during the Rule 26(f) conference depend on the nature and extent of the contemplated discovery and of the parties' information systems. It may be important for the parties to discuss those systems, and accordingly important for counsel to become familiar with those systems before the

conference. With that information, the parties can develop a discovery plan that takes into account the capabilities of their computer systems. In appropriate cases identification of, and early discovery from, individuals with special knowledge of a party's computer systems may be helpful.

The particular issues regarding electronically stored information that deserve attention during the discovery planning stage depend on the specifics of the given case. *See Manual for Complex Litigation* (4th) §40.25(2) (listing topics for discussion in a proposed order regarding meet-and-confer sessions). For example, the parties may specify the topics for such discovery and the time period for which discovery will be sought. They may identify the various sources of such information within a party's control that should be searched for electronically stored information. They may discuss whether the information is reasonably accessible to the party that has it, including the burden or cost of retrieving and reviewing the information. *See* Rule 26(b)(2)(B). Rule 26(f)(3) explicitly directs the parties to discuss the form or forms in which electronically stored information might be produced. The parties may be able to reach agreement on the forms of production, making discovery more efficient. Rule 34(b) is amended to permit a requesting party to specify the form or forms in which it wants electronically stored information produced. If the requesting party does not specify a form, Rule 34(b) directs the responding party to state the forms it intends to use in the production. Early discussion of the forms of production may facilitate the application of Rule 34(b) by allowing the parties to determine what forms of production will meet both parties' needs. Early identification of disputes over the forms of production may help avoid the expense and delay of searches or productions using inappropriate forms.

Rule 26(f) is also amended to direct the parties to discuss any issues regarding preservation of discoverable information during their conference as they develop a discovery plan. This provision applies to all sorts of discoverable information, but can be particularly important with regard to electronically stored information. The volume and dynamic nature of electronically stored information may complicate preservation obligations. The ordinary operation of computers involves both the automatic creation and the automatic deletion or overwriting of certain information. Failure to address preservation issues early in the litigation increases uncertainty and raises a risk of disputes.

The parties' discussion should pay particular attention to the balance between the competing needs to preserve relevant evidence and to continue routine operations critical to ongoing activities. Complete or broad cessation of a party's routine computer operations could paralyze the party's activities. *Cf. Manual for Complex Litigation* (4th) §11.422 ("A blanket preservation order may be prohibitively expensive and unduly burdensome for parties dependent on computer

systems for their day-to-day operations.") The parties should take account of these considerations in their discussions, with the goal of agreeing on reasonable preservation steps.

The requirement that the parties discuss preservation does not imply that courts should routinely enter preservation orders. A preservation order entered over objections should be narrowly tailored. Ex parte preservation orders should issue only in exceptional circumstances.

Rule 26(f) is also amended to provide that the parties should discuss any issues relating to assertions of privilege or of protection as trial-preparation materials, including whether the parties can facilitate discovery by agreeing on procedures for asserting claims of privilege or protection after production and whether to ask the court to enter an order that includes any agreement the parties reach. The Committee has repeatedly been advised about the discovery difficulties that can result from efforts to guard against waiver of privilege and work-product protection. Frequently parties find it necessary to spend large amounts of time reviewing materials requested through discovery to avoid waiving privilege. These efforts are necessary because materials subject to a claim of privilege or protection are often difficult to identify. A failure to withhold even one such item may result in an argument that there has been a waiver of privilege as to all other privileged materials on that subject matter. Efforts to avoid the risk of waiver can impose substantial costs on the party producing the material and the time required for the privilege review can substantially delay access for the party seeking discovery.

These problems often become more acute when discovery of electronically stored information is sought. The volume of such data, and the informality that attends use of e-mail and some other types of electronically stored information, may make privilege determinations more difficult, and privilege review correspondingly more expensive and time consuming. Other aspects of electronically stored information pose particular difficulties for privilege review. For example, production may be sought of information automatically included in electronic files but not apparent to the creator or to readers. Computer programs may retain draft language, editorial comments, and other deleted matter (sometimes referred to as "embedded data" or "embedded edits") in an electronic file but not make them apparent to the reader. Information describing the history, tracking, or management of an electronic file (sometimes called "metadata") is usually not apparent to the reader viewing a hard copy or a screen image. Whether this information should be produced may be among the topics discussed in the Rule 26(f) conference. If it is, it may need to be reviewed to ensure that no privileged information is included, further complicating the task of privilege review.

Parties may attempt to minimize these costs and delays by agreeing to protocols that minimize the risk of waiver. They may agree that the responding party will provide certain requested materials for initial examination without waiving any privilege or protection—sometimes known

as a "quick peek." The requesting party then designates the documents it wishes to have actually produced. This designation is the Rule 34 request. The responding party then responds in the usual course, screening only those documents actually requested for formal production and asserting privilege claims as provided in Rule 26(b)(5)(A). On other occasions, parties enter agreements—sometimes called "clawback agreements"—that production without intent to waive privilege or protection should not be a waiver so long as the responding party identifies the documents mistakenly produced, and that the documents should be returned under those circumstances. Other voluntary arrangements may be appropriate depending on the circumstances of each litigation. In most circumstances, a party who receives information under such an arrangement cannot assert that production of the information waived a claim of privilege or of protection as trial-preparation material.

Although these agreements may not be appropriate for all cases, in certain cases they can facilitate prompt and economical discovery by reducing delay before the discovering party obtains access to documents, and by reducing the cost and burden of review by the producing party. A case-management or other order including such agreements may further facilitate the discovery process. Form 35 is amended to include a report to the court about any agreement regarding protections against inadvertent forfeiture or waiver of privilege or protection that the parties have reached, and Rule 16(b) is amended to recognize that the court may include such an agreement in a case- management or other order. If the parties agree to entry of such an order, their proposal should be included in the report to the court.

Rule 26(b)(5)(B) is added to establish a parallel procedure to assert privilege or protection as trial-preparation material after production, leaving the question of waiver to later determination by the court.

Changes Made After Publication and Comment. The Committee recommends a modified version of what was published. Rule 26(f)(3) was expanded to refer to the form "or forms" of production, in parallel with the like change in Rule 34. Different forms may be suitable for different sources of electronically stored information.

The published Rule 26(f)(4) proposal described the parties' views and proposals concerning whether, on their agreement, the court should enter an order protecting the right to assert privilege after production. This has been revised to refer to the parties' views and proposals concerning any issues relating to claims of privilege, including—if the parties agree on a procedure to assert such claims after production—whether to ask the court to include their agreement in an order. As with Rule 16(b)(6), this change was made to avoid any implications as to the scope of the protection that may be afforded by court adoption of the parties' agreement.

Rule 26(f)(4) also was expanded to include trial-preparation materials.

The Committee Note was revised to reflect the changes in the rule text.

The changes from the published rule are shown below. [Omitted]

Committee Notes on Rules—2015 Amendment

Rule 26(b)(1) is changed in several ways.

Information is discoverable under revised Rule 26(b)(1) if it is relevant to any party's claim or defense and is proportional to the needs of the case. The considerations that bear on proportionality are moved from present Rule 26(b)(2)(C)(iii), slightly rearranged and with one addition.

Most of what now appears in Rule 26(b)(2)(C)(iii) was first adopted in 1983. The 1983 provision was explicitly adopted as part of the scope of discovery defined by Rule 26(b)(1). Rule 26(b)(1) directed the court to limit the frequency or extent of use of discovery if it determined that "the discovery is unduly burdensome or expensive, taking into account the needs of the case, the amount in controversy, limitations on the parties' resources, and the importance of the issues at stake in the litigation." At the same time, Rule 26(g) was added. Rule 26(g) provided that signing a discovery request, response, or objection certified that the request, response, or objection was "not unreasonable or unduly burdensome or expensive, given the needs of the case, the discovery already had in the case, the amount in controversy, and the importance of the issues at stake in the litigation." The parties thus shared the responsibility to honor these limits on the scope of discovery.

The 1983 Committee Note stated that the new provisions were added "to deal with the problem of overdiscovery. The objective is to guard against redundant or disproportionate discovery by giving the court authority to reduce the amount of discovery that may be directed to matters that are otherwise proper subjects of inquiry. The new sentence is intended to encourage judges to be more aggressive in identifying and discouraging discovery overuse. The grounds mentioned in the amended rule for limiting discovery reflect the existing practice of many courts in issuing protective orders under Rule 26(c). . . . On the whole, however, district judges have been reluctant to limit the use of the discovery devices."

The clear focus of the 1983 provisions may have been softened, although inadvertently, by the amendments made in 1993. The 1993 Committee Note explained: "[F]ormer paragraph (b)(1) [was] subdivided into two paragraphs for ease of reference and to avoid renumbering of paragraphs (3) and (4)." Subdividing the paragraphs, however, was done in a way that could be

read to separate the proportionality provisions as "limitations," no longer an integral part of the (b)(1) scope provisions. That appearance was immediately offset by the next statement in the Note: "Textual changes are then made in new paragraph (2) to enable the court to keep tighter rein on the extent of discovery."

The 1993 amendments added two factors to the considerations that bear on limiting discovery: whether "the burden or expense of the proposed discovery outweighs its likely benefit," and "the importance of the proposed discovery in resolving the issues." Addressing these and other limitations added by the 1993 discovery amendments, the Committee Note stated that "[t]he revisions in Rule 26(b)(2) are intended to provide the court with broader discretion to impose additional restrictions on the scope and extent of discovery . . . ."

The relationship between Rule 26(b)(1) and (2) was further addressed by an amendment made in 2000 that added a new sentence at the end of (b)(1): "All discovery is subject to the limitations imposed by Rule 26(b)(2)(i), (ii), and (iii)[now Rule 26(b)(2)(C)]." The Committee Note recognized that "[t]hese limitations apply to discovery that is otherwise within the scope of subdivision (b)(1)." It explained that the Committee had been told repeatedly that courts were not using these limitations as originally intended. "This otherwise redundant cross-reference has been added to emphasize the need for active judicial use of subdivision (b)(2) to control excessive discovery."

The present amendment restores the proportionality factors to their original place in defining the scope of discovery. This change reinforces the Rule 26(g) obligation of the parties to consider these factors in making discovery requests, responses, or objections.

Restoring the proportionality calculation to Rule 26(b)(1) does not change the existing responsibilities of the court and the parties to consider proportionality, and the change does not place on the party seeking discovery the burden of addressing all proportionality considerations.

Nor is the change intended to permit the opposing party to refuse discovery simply by making a boilerplate objection that it is not proportional. The parties and the court have a collective responsibility to consider the proportionality of all discovery and consider it in resolving discovery disputes.

The parties may begin discovery without a full appreciation of the factors that bear on proportionality. A party requesting discovery, for example, may have little information about the burden or expense of responding. A party requested to provide discovery may have little information about the importance of the discovery in resolving the issues as understood by the requesting party. Many of these uncertainties should be addressed and reduced in the parties' Rule 26(f) conference and in scheduling and pretrial conferences with the court. But if the parties

continue to disagree, the discovery dispute could be brought before the court and the parties' responsibilities would remain as they have been since 1983. A party claiming undue burden or expense ordinarily has far better information — perhaps the only information — with respect to that part of the determination. A party claiming that a request is important to resolve the issues should be able to explain the ways in which the underlying information bears on the issues as that party understands them. The court's responsibility, using all the information provided by the parties, is to consider these and all the other factors in reaching a case-specific determination of the appropriate scope of discovery.

The direction to consider the parties' relative access to relevant information adds new text to provide explicit focus on considerations already implicit in present Rule 26(b)(2)(C)(iii). Some cases involve what often is called "information asymmetry." One party — often an individual plaintiff — may have very little discoverable information. The other party may have vast amounts of information, including information that can be readily retrieved and information that is more difficult to retrieve. In practice these circumstances often mean that the burden of responding to discovery lies heavier on the party who has more information, and properly so.

Restoring proportionality as an express component of the scope of discovery warrants repetition of parts of the 1983 and 1993 Committee Notes that must not be lost from sight. The 1983 Committee Note explained that "[t]he rule contemplates greater judicial involvement in the discovery process and thus acknowledges the reality that it cannot always operate on a self-regulating basis." The 1993 Committee Note further observed that "[t]he information explosion of recent decades has greatly increased both the potential cost of wide-ranging discovery and the potential for discovery to be used as an instrument for delay or oppression." What seemed an explosion in 1993 has been exacerbated by the advent of e-discovery. The present amendment again reflects the need for continuing and close judicial involvement in the cases that do not yield readily to the ideal of effective party management. It is expected that discovery will be effectively managed by the parties in many cases. But there will be important occasions for judicial management, both when the parties are legitimately unable to resolve important differences and when the parties fall short of effective, cooperative management on their own.

It also is important to repeat the caution that the monetary stakes are only one factor, to be balanced against other factors. The 1983 Committee Note recognized "the significance of the substantive issues, as measured in philosophic, social, or institutional terms. Thus the rule recognizes that many cases in public policy spheres, such as employment practices, free speech, and other matters, may have importance far beyond the monetary amount involved." Many other substantive areas also may involve litigation that seeks relatively small amounts of money, or no money at all, but that seeks to vindicate vitally important personal or public values.

So too, consideration of the parties' resources does not foreclose discovery requests addressed to an impecunious party, nor justify unlimited discovery requests addressed to a wealthy party. The 1983 Committee Note cautioned that "[t]he court must apply the standards in an even-handed manner that will prevent use of discovery to wage a war of attrition or as a device to coerce a party, whether financially weak or affluent."

The burden or expense of proposed discovery should be determined in a realistic way. This includes the burden or expense of producing electronically stored information. Computer-based methods of searching such information continue to develop, particularly for cases involving large volumes of electronically stored information. Courts and parties should be willing to consider the opportunities for reducing the burden or expense of discovery as reliable means of searching electronically stored information become available.

A portion of present Rule 26(b)(1) is omitted from the proposed revision. After allowing discovery of any matter relevant to any party's claim or defense, the present rule adds: "including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter." Discovery of such matters is so deeply entrenched in practice that it is no longer necessary to clutter the long text of Rule 26 with these examples. The discovery identified in these examples should still be permitted under the revised rule when relevant and proportional to the needs of the case. Framing intelligent requests for electronically stored information, for example, may require detailed information about another party's information systems and other information resources.

The amendment deletes the former provision authorizing the court, for good cause, to order discovery of any matter relevant to the subject matter involved in the action. The Committee has been informed that this language is rarely invoked. Proportional discovery relevant to any party's claim or defense suffices, given a proper understanding of what is relevant to a claim or defense. The distinction between matter relevant to a claim or defense and matter relevant to the subject matter was introduced in 2000. The 2000 Note offered three examples of information that, suitably focused, would be relevant to the parties' claims or defenses. The examples were "other incidents of the same type, or involving the same product"; "information about organizational arrangements or filing systems"; and "information that could be used to impeach a likely witness." Such discovery is not foreclosed by the amendments. Discovery that is relevant to the parties' claims or defenses may also support amendment of the pleadings to add a new claim or defense that affects the scope of discovery.

The former provision for discovery of relevant but inadmissible information that appears "reasonably calculated to lead to the discovery of admissible evidence" is also deleted. The phrase

has been used by some, incorrectly, to define the scope of discovery. As the Committee Note to the 2000 amendments observed, use of the "reasonably calculated" phrase to define the scope of discovery "might swallow any other limitation on the scope of discovery." The 2000 amendments sought to prevent such misuse by adding the word "Relevant" at the beginning of the sentence, making clear that "'relevant' means within the scope of discovery as defined in this subdivision . . . ." The "reasonably calculated" phrase has continued to create problems, however, and is removed by these amendments. It is replaced by the direct statement that "Information within this scope of discovery need not be admissible in evidence to be discoverable." Discovery of nonprivileged information not admissible in evidence remains available so long as it is otherwise within the scope of discovery.

Rule 26(b)(2)(C)(iii) is amended to reflect the transfer of the considerations that bear on proportionality to Rule 26(b)(1). The court still must limit the frequency or extent of proposed discovery, on motion or on its own, if it is outside the scope permitted by Rule 26(b)(1).

Rule 26(c)(1)(B) is amended to include an express recognition of protective orders that allocate expenses for disclosure or discovery. Authority to enter such orders is included in the present rule, and courts already exercise this authority. Explicit recognition will forestall the temptation some parties may feel to contest this authority. Recognizing the authority does not imply that cost-shifting should become a common practice. Courts and parties should continue to assume that a responding party ordinarily bears the costs of responding.

Rule 26(d)(2) is added to allow a party to deliver Rule 34 requests to another party more than 21 days after that party has been served even though the parties have not yet had a required Rule 26(f) conference. Delivery may be made by any party to the party that has been served, and by that party to any plaintiff and any other party that has been served. Delivery does not count as service; the requests are considered to be served at the first Rule 26(f) conference. Under Rule 34(b)(2)(A) the time to respond runs from service. This relaxation of the discovery moratorium is designed to facilitate focused discussion during the Rule 26(f) conference. Discussion at the conference may produce changes in the requests. The opportunity for advance scrutiny of requests delivered before the Rule 26(f) conference should not affect a decision whether to allow additional time to respond.

Rule 26(d)(3) is renumbered and amended to recognize that the parties may stipulate to case-specific sequences of discovery.

Rule 26(f)(3) is amended in parallel with Rule 16(b)(3) to add two items to the discovery plan — issues about preserving electronically stored information and court orders under Evidence Rule 502.

***

Rule 34. Producing Documents, Electronically Stored Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes

(a) In General. A party may serve on any other party a request within the scope of Rule 26(b):

(1) to produce and permit the requesting party or its representative to inspect, copy, test, or sample the following items in the responding party's possession, custody, or control:

(A) any designated documents or electronically stored information—including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations—stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form; or

(B) any designated tangible things; or

(2) to permit entry onto designated land or other property possessed or controlled by the responding party, so that the requesting party may inspect, measure, survey, photograph, test, or sample the property or any designated object or operation on it.

(b) Procedure.

(1) *Contents of the Request.* The request:

(A) must describe with reasonable particularity each item or category of items to be inspected;

(B) must specify a reasonable time, place, and manner for the inspection and for performing the related acts; and

(C) may specify the form or forms in which electronically stored information is to be produced.

(2) *Responses and Objections.*

(A) *Time to Respond.* The party to whom the request is directed must respond in writing within 30 days after being served or — if the request was delivered under Rule 26(d)(2) — within 30 days after the parties' first Rule 26(f) conference. A shorter or longer time may be stipulated to under Rule 29 or be ordered by the court.

(B) *Responding to Each Item.* For each item or category, the response must either state that inspection and related activities will be permitted as requested or state with specificity the grounds for objecting to the request, including the reasons. The responding party may state that it will produce copies of documents or of electronically stored information instead of permitting inspection. The production must then be completed no later than the time for inspection specified in the request or another reasonable time specified in the response.

(C) *Objections.* An objection must state whether any responsive materials are being withheld on the basis of that objection. An objection to part of a request must specify the part and permit inspection of the rest.

(D) *Responding to a Request for Production of Electronically Stored Information.* The response may state an objection to a requested form for producing electronically stored information. If the responding party objects to a requested form—or if no form was specified in the request—the party must state the form or forms it intends to use.

(E) *Producing the Documents or Electronically Stored Information.* Unless otherwise stipulated or ordered by the court, these procedures apply to producing documents or electronically stored information:

(i) A party must produce documents as they are kept in the usual course of business or must organize and label them to correspond to the categories in the request;

(ii) If a request does not specify a form for producing electronically stored information, a party must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms; and

(iii) A party need not produce the same electronically stored information in more than one form.

(c) Nonparties. As provided in Rule 45, a nonparty may be compelled to produce documents and tangible things or to permit an inspection.

Notes

(As amended Dec. 27, 1946, eff. Mar. 19, 1948; Mar. 30, 1970, eff. July 1, 1970; Apr. 29, 1980, eff. Aug. 1, 1980; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 30, 1991, eff. Dec. 1, 1991; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2006 Amendment

Subdivision (a). As originally adopted, Rule 34 focused on discovery of "documents" and "things." In 1970, Rule 34(a) was amended to include discovery of data compilations, anticipating that the use of computerized information would increase. Since then, the growth in electronically stored information and in the variety of systems for creating and storing such information has been dramatic. Lawyers and judges interpreted the term "documents" to include electronically stored information because it was obviously improper to allow a party to evade discovery obligations on the basis that the label had not kept pace with changes in information technology. But it has become increasingly difficult to say that all forms of electronically stored information, many dynamic in nature, fit within the traditional concept of a "document." Electronically stored information may exist in dynamic databases and other forms far different from fixed expression on paper. Rule 34(a) is amended to confirm that discovery of electronically stored information stands on equal footing with discovery of paper documents. The change clarifies that Rule 34 applies to information that is fixed in a tangible form and to information that is stored in a medium from which it can be retrieved and examined. At the same time, a Rule 34 request for production of "documents" should be understood to encompass, and the response should include, electronically stored information unless discovery in the action has clearly distinguished between electronically stored information and "documents."

Discoverable information often exists in both paper and electronic form, and the same or similar information might exist in both. The items listed in Rule 34(a) show different ways in which information may be recorded or stored. Images, for example, might be hard-copy documents or electronically stored information. The wide variety of computer systems currently in use, and the rapidity of technological change, counsel against a limiting or precise definition of electronically stored information. Rule 34(a)(1) is expansive and includes any type of information that is stored electronically. A common example often sought in discovery is electronic communications, such as e-mail. The rule covers—either as documents or as electronically stored information—information "stored in any medium," to encompass future developments in computer technology. Rule 34(a)(1) is intended to be broad enough to cover all current types of computer-based information, and flexible enough to encompass future changes and developments.

References elsewhere in the rules to "electronically stored information" should be understood to invoke this expansive approach. A companion change is made to Rule 33(d), making it explicit that parties choosing to respond to an interrogatory by permitting access to responsive records may do so by providing access to electronically stored information. More generally, the term used in Rule 34(a)(1) appears in a number of other amendments, such as those to Rules 26(a)(1), 26(b)(2), 26(b)(5)(B), 26(f), 34(b), 37(f), and 45. In each of these rules, electronically stored information has the same broad meaning it has under Rule 34(a)(1). References to "documents" appear in

discovery rules that are not amended, including Rules 30(f), 36(a), and 37(c)(2). These references should be interpreted to include electronically stored information as circumstances warrant.

The term "electronically stored information" is broad, but whether material that falls within this term should be produced, and in what form, are separate questions that must be addressed under Rules 26(b), 26(c), and 34(b).

The Rule 34(a) requirement that, if necessary, a party producing electronically stored information translate it into reasonably usable form does not address the issue of translating from one human language to another. *See In re Puerto Rico Elect. Power Auth.*, 687 F.2d 501, 504–510 (1st Cir. 1989).

Rule 34(a)(1) is also amended to make clear that parties may request an opportunity to test or sample materials sought under the rule in addition to inspecting and copying them. That opportunity may be important for both electronically stored information and hard-copy materials. The current rule is not clear that such testing or sampling is authorized; the amendment expressly permits it. As with any other form of discovery, issues of burden and intrusiveness raised by requests to test or sample can be addressed under Rules 26(b)(2) and 26(c). Inspection or testing of certain types of electronically stored information or of a responding party's electronic information system may raise issues of confidentiality or privacy. The addition of testing and sampling to Rule 34(a) with regard to documents and electronically stored information is not meant to create a routine right of direct access to a party's electronic information system, although such access might be justified in some circumstances. Courts should guard against undue intrusiveness resulting from inspecting or testing such systems.

Rule 34(a)(1) is further amended to make clear that tangible things must—like documents and land sought to be examined—be designated in the request.

Subdivision (b). Rule 34(b) provides that a party must produce documents as they are kept in the usual course of business or must organize and label them to correspond with the categories in the discovery request. The production of electronically stored information should be subject to comparable requirements to protect against deliberate or inadvertent production in ways that raise unnecessary obstacles for the requesting party. Rule 34(b) is amended to ensure similar protection for electronically stored information.

The amendment to Rule 34(b) permits the requesting party to designate the form or forms in which it wants electronically stored information produced. The form of production is more important to the exchange of electronically stored information than of hard-copy materials, although a party might specify hard copy as the requested form. Specification of the desired form

or forms may facilitate the orderly, efficient, and cost-effective discovery of electronically stored information. The rule recognizes that different forms of production may be appropriate for different types of electronically stored information. Using current technology, for example, a party might be called upon to produce word processing documents, e-mail messages, electronic spreadsheets, different image or sound files, and material from databases. Requiring that such diverse types of electronically stored information all be produced in the same form could prove impossible, and even if possible could increase the cost and burdens of producing and using the information. The rule therefore provides that the requesting party may ask for different forms of production for different types of electronically stored information.

The rule does not require that the requesting party choose a form or forms of production. The requesting party may not have a preference. In some cases, the requesting party may not know what form the producing party uses to maintain its electronically stored information, although Rule 26(f)(3) is amended to call for discussion of the form of production in the parties' prediscovery conference.

The responding party also is involved in determining the form of production. In the written response to the production request that Rule 34 requires, the responding party must state the form it intends to use for producing electronically stored information if the requesting party does not specify a form or if the responding party objects to a form that the requesting party specifies. Stating the intended form before the production occurs may permit the parties to identify and seek to resolve disputes before the expense and work of the production occurs. A party that responds to a discovery request by simply producing electronically stored information in a form of its choice, without identifying that form in advance of the production in the response required by Rule 34(b), runs a risk that the requesting party can show that the produced form is not reasonably usable and that it is entitled to production of some or all of the information in an additional form. Additional time might be required to permit a responding party to assess the appropriate form or forms of production.

If the requesting party is not satisfied with the form stated by the responding party, or if the responding party has objected to the form specified by the requesting party, the parties must meet and confer under Rule 37(a)(2)(B) in an effort to resolve the matter before the requesting party can file a motion to compel. If they cannot agree and the court resolves the dispute, the court is not limited to the forms initially chosen by the requesting party, stated by the responding party, or specified in this rule for situations in which there is no court order or party agreement.

If the form of production is not specified by party agreement or court order, the responding party must produce electronically stored information either in a form or forms in which it is ordinarily maintained or in a form or forms that are reasonably usable. Rule 34(a) requires that, if necessary,

a responding party "translate" information it produces into a "reasonably usable" form. Under some circumstances, the responding party may need to provide some reasonable amount of technical support, information on application software, or other reasonable assistance to enable the requesting party to use the information. The rule does not require a party to produce electronically stored information in the form it [sic] which it is ordinarily maintained, as long as it is produced in a reasonably usable form. But the option to produce in a reasonably usable form does not mean that a responding party is free to convert electronically stored information from the form in which it is ordinarily maintained to a different form that makes it more difficult or burdensome for the requesting party to use the information efficiently in the litigation. If the responding party ordinarily maintains the information it is producing in a way that makes it searchable by electronic means, the information should not be produced in a form that removes or significantly degrades this feature.

Some electronically stored information may be ordinarily maintained in a form that is not reasonably usable by any party. One example is "legacy" data that can be used only by superseded systems. The questions whether a producing party should be required to convert such information to a more usable form, or should be required to produce it at all, should be addressed under Rule 26(b)(2)(B).

Whether or not the requesting party specified the form of production, Rule 34(b) provides that the same electronically stored information ordinarily be produced in only one form.

Changes Made after Publication and Comment. The proposed amendment recommended for approval has been modified from the published version. The sequence of "documents or electronically stored information" is changed to emphasize that the parenthetical exemplifications apply equally to illustrate "documents" and "electronically stored information." The reference to "detection devices" is deleted as redundant with "translated" and as archaic.

The references to the form of production are changed in the rule and Committee Note to refer also to "forms." Different forms may be appropriate or necessary for different sources of information.

The published proposal allowed the requesting party to specify a form for production and recognized that the responding party could object to the requested form. This procedure is now amplified by directing that the responding party state the form or forms it intends to use for production if the request does not specify a form or if the responding party objects to the requested form.

The default forms of production to be used when the parties do not agree on a form and there is no court order are changed in part. As in the published proposal, one default form is "a form or forms in which [electronically stored information] is ordinarily maintained." The alternative default form, however, is changed from "an electronically searchable form" to "a form or forms that are reasonably usable." "[A]n electronically searchable form" proved to have several defects. Some electronically stored information cannot be searched electronically. In addition, there often are many different levels of electronic searchability—the published default would authorize production in a minimally searchable form even though more easily searched forms might be available at equal or less cost to the responding party.

The provision that absent court order a party need not produce the same electronically stored information in more than one form was moved to become a separate item for the sake of emphasis.

The Committee Note was changed to reflect these changes in rule text, and also to clarify many aspects of the published Note. In addition, the Note was expanded to add a caveat to the published amendment that establishes the rule that documents—and now electronically stored information—may be tested and sampled as well as inspected and copied. Fears were expressed that testing and sampling might imply routine direct access to a party's information system. The Note states that direct access is not a routine right, "although such access might be justified in some circumstances."

The changes in the rule text since publication are set out below. [Omitted]

Committee Notes on Rules—2015 Amendment

Several amendments are made in Rule 34, aimed at reducing the potential to impose unreasonable burdens by objections to requests to produce.

Rule 34(b)(2)(A) is amended to fit with new Rule 26(d)(2). The time to respond to a Rule 34 request delivered before the parties' Rule 26(f) conference is 30 days after the first Rule 26(f) conference.

Rule 34(b)(2)(B) is amended to require that objections to Rule 34 requests be stated with specificity. This provision adopts the language of Rule 33(b)(4), eliminating any doubt that less specific objections might be suitable under Rule 34. The specificity of the objection ties to the new provision in Rule 34(b)(2)(C) directing that an objection must state whether any responsive materials are being withheld on the basis of that objection. An objection may state that a request is overbroad, but if the objection recognizes that some part of the request is appropriate the objection should state the scope that is not overbroad. Examples would be a statement that the responding party will limit the search to documents or electronically stored information created

within a given period of time prior to the events in suit, or to specified sources. When there is such an objection, the statement of what has been withheld can properly identify as matters "withheld" anything beyond the scope of the search specified in the objection.

Rule 34(b)(2)(B) is further amended to reflect the common practice of producing copies of documents or electronically stored information rather than simply permitting inspection. The response to the request must state that copies will be produced. The production must be completed either by the time for inspection specified in the request or by another reasonable time specifically identified in the response. When it is necessary to make the production in stages the response should specify the beginning and end dates of the production.

Rule 34(b)(2)(C) is amended to provide that an objection to a Rule 34 request must state whether anything is being withheld on the basis of the objection. This amendment should end the confusion that frequently arises when a producing party states several objections and still produces information, leaving the requesting party uncertain whether any relevant and responsive information has been withheld on the basis of the objections. The producing party does not need to provide a detailed description or log of all documents withheld, but does need to alert other parties to the fact that documents have been withheld and thereby facilitate an informed discussion of the objection. An objection that states the limits that have controlled the search for responsive and relevant materials qualifies as a statement that the materials have been "withheld."

***

Rule 45. Subpoena

(a) In General.

(1) *Form and Contents.*

(A) *Requirements—In General.* Every subpoena must:

(i) state the court from which it issued;

(ii) state the title of the action and its civil-action number;

(iii) command each person to whom it is directed to do the following at a specified time and place: attend and testify; produce designated documents, electronically stored information, or tangible things in that person's possession, custody, or control; or permit the inspection of premises; and

(iv) set out the text of Rule 45(d) and (e).

(B) *Command to Attend a Deposition—Notice of the Recording Method.* A subpoena commanding attendance at a deposition must state the method for recording the testimony.

(C) *Combining or Separating a Command to Produce or to Permit Inspection; Specifying the Form for Electronically Stored Information.* A command to produce documents, electronically stored information, or tangible things or to permit the inspection of premises may be included in a subpoena commanding attendance at a deposition, hearing, or trial, or may be set out in a separate subpoena. A subpoena may specify the form or forms in which electronically stored information is to be produced.

(D) *Command to Produce; Included Obligations.* A command in a subpoena to produce documents, electronically stored information, or tangible things requires the responding person to permit inspection, copying, testing, or sampling of the materials.

(2) *Issuing Court.* A subpoena must issue from the court where the action is pending.

(3) *Issued by Whom.* The clerk must issue a subpoena, signed but otherwise in blank, to a party who requests it. That party must complete it before service. An attorney also may issue and sign a subpoena if the attorney is authorized to practice in the issuing court.

(4) *Notice to Other Parties Before Service.* If the subpoena commands the production of documents, electronically stored information, or tangible things or the inspection of premises before trial, then before it is served on the person to whom it is directed, a notice and a copy of the subpoena must be served on each party.

(b) Service.

(1) *By Whom and How; Tendering Fees.* Any person who is at least 18 years old and not a party may serve a subpoena. Serving a subpoena requires delivering a copy to the named person and, if the subpoena requires that person's attendance, tendering the fees for 1 day's attendance and the mileage allowed by law. Fees and mileage need not be tendered when the subpoena issues on behalf of the United States or any of its officers or agencies.

(2) *Service in the United States.* A subpoena may be served at any place within the United States.

(3) *Service in a Foreign Country.* 28 U.S.C. §1783 governs issuing and serving a subpoena directed to a United States national or resident who is in a foreign country.

(4) *Proof of Service.* Proving service, when necessary, requires filing with the issuing court a statement showing the date and manner of service and the names of the persons served. The statement must be certified by the server.

(c) Place of Compliance.

(1) *For a Trial, Hearing, or Deposition.* A subpoena may command a person to attend a trial, hearing, or deposition only as follows:

(A) within 100 miles of where the person resides, is employed, or regularly transacts business in person; or

(B) within the state where the person resides, is employed, or regularly transacts business in person, if the person

(i) is a party or a party's officer; or

(ii) is commanded to attend a trial and would not incur substantial expense.

(2) *For Other Discovery.* A subpoena may command:

(A) production of documents, electronically stored information, or tangible things at a place within 100 miles of where the person resides, is employed, or regularly transacts business in person; and

(B) inspection of premises at the premises to be inspected.

(d) Protecting a Person Subject to a Subpoena; Enforcement.

(1) *Avoiding Undue Burden or Expense; Sanctions.* A party or attorney responsible for issuing and serving a subpoena must take reasonable steps to avoid imposing undue burden or expense on a person subject to the subpoena. The court for the district where compliance is required must enforce this duty and impose an appropriate sanction—which may include lost earnings and reasonable attorney's fees—on a party or attorney who fails to comply.

(2) *Command to Produce Materials or Permit Inspection.*

(A) *Appearance Not Required.* A person commanded to produce documents, electronically stored information, or tangible things, or to permit the inspection of premises, need not appear in person at the place of production or inspection unless also commanded to appear for a deposition, hearing, or trial.

(B) *Objections.* A person commanded to produce documents or tangible things or to permit inspection may serve on the party or attorney designated in the subpoena a written objection to inspecting, copying, testing or sampling any or all of the materials or to inspecting the premises—or to producing electronically stored information in the form or forms requested. The objection

must be served before the earlier of the time specified for compliance or 14 days after the subpoena is served. If an objection is made, the following rules apply:

(i) At any time, on notice to the commanded person, the serving party may move the court for the district where compliance is required for an order compelling production or inspection.

(ii) These acts may be required only as directed in the order, and the order must protect a person who is neither a party nor a party's officer from significant expense resulting from compliance.

(3) *Quashing or Modifying a Subpoena.*

(A) *When Required.* On timely motion, the court for the district where compliance is required must quash or modify a subpoena that:

(i) fails to allow a reasonable time to comply;

(ii) requires a person to comply beyond the geographical limits specified in Rule 45(c);

(iii) requires disclosure of privileged or other protected matter, if no exception or waiver applies; or

(iv) subjects a person to undue burden.

(B) *When Permitted.* To protect a person subject to or affected by a subpoena, the court for the district where compliance is required may, on motion, quash or modify the subpoena if it requires:

(i) disclosing a trade secret or other confidential research, development, or commercial information; or

(ii) disclosing an unretained expert's opinion or information that does not describe specific occurrences in dispute and results from the expert's study that was not requested by a party.

(C) *Specifying Conditions as an Alternative.* In the circumstances described in Rule 45(d)(3)(B), the court may, instead of quashing or modifying a subpoena, order appearance or production under specified conditions if the serving party:

(i) shows a substantial need for the testimony or material that cannot be otherwise met without undue hardship; and

(ii) ensures that the subpoenaed person will be reasonably compensated.

(e) Duties in Responding to a Subpoena.

(1) *Producing Documents or Electronically Stored Information.* These procedures apply to producing documents or electronically stored information:

(A) *Documents.* A person responding to a subpoena to produce documents must produce them as they are kept in the ordinary course of business or must organize and label them to correspond to the categories in the demand.

(B) *Form for Producing Electronically Stored Information Not Specified.* If a subpoena does not specify a form for producing electronically stored information, the person responding must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms.

(C) *Electronically Stored Information Produced in Only One Form.* The person responding need not produce the same electronically stored information in more than one form.

(D) *Inaccessible Electronically Stored Information.* The person responding need not provide discovery of electronically stored information from sources that the person identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the person responding must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.

(2) *Claiming Privilege or Protection.*

(A) *Information Withheld.* A person withholding subpoenaed information under a claim that it is privileged or subject to protection as trial-preparation material must:

(i) expressly make the claim; and

(ii) describe the nature of the withheld documents, communications, or tangible things in a manner that, without revealing information itself privileged or protected, will enable the parties to assess the claim.

(B) *Information Produced.* If information produced in response to a subpoena is subject to a claim of privilege or of protection as trial-preparation material, the person making the claim may notify any party that received the information of the claim and the basis for it. After being notified, a party must promptly return, sequester, or destroy the specified information and any copies it has; must not use or disclose the information until the claim is resolved; must take reasonable steps to retrieve the information if the party disclosed it before being notified; and may promptly present the information under seal to the court for the district where compliance is required for

a determination of the claim. The person who produced the information must preserve the information until the claim is resolved.

(f) Transferring a Subpoena-Related Motion. When the court where compliance is required did not issue the subpoena, it may transfer a motion under this rule to the issuing court if the person subject to the subpoena consents or if the court finds exceptional circumstances. Then, if the attorney for a person subject to a subpoena is authorized to practice in the court where the motion was made, the attorney may file papers and appear on the motion as an officer of the issuing court. To enforce its order, the issuing court may transfer the order to the court where the motion was made.

(g) Contempt. The court for the district where compliance is required — and also, after a motion is transferred, the issuing court — may hold in contempt a person who, having been served, fails without adequate excuse to obey the subpoena or an order related to it.

Notes

(As amended Dec. 27, 1946, eff. Mar. 19, 1948; Dec. 29, 1948, eff. Oct. 20, 1949; Mar. 30, 1970, eff. July 1, 1970; Apr. 29, 1980, eff. Aug. 1, 1980; Apr. 29, 1985, eff. Aug. 1, 1985; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 30, 1991, eff. Dec. 1, 1991; Apr. 25, 2005, eff. Dec. 1, 2005; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 16, 2013, eff. Dec. 1, 2013.)

Committee Notes on Rules—2006 Amendment

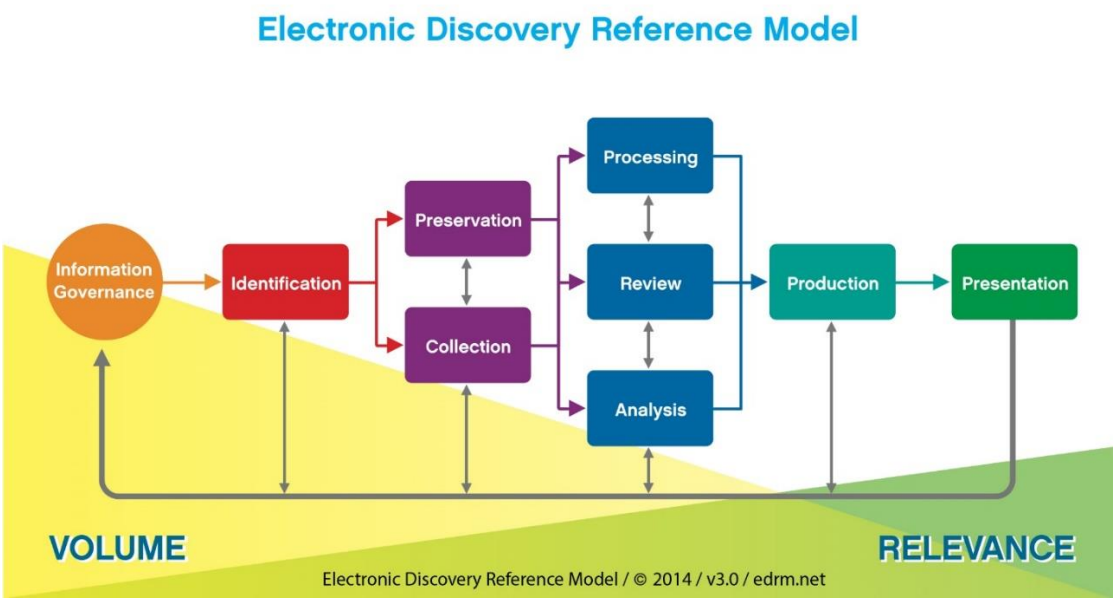
Rule 45 is amended to conform the provisions for subpoenas to changes in other discovery rules, largely related to discovery of electronically stored information. Rule 34 is amended to provide in greater detail for the production of electronically stored information. Rule 45(a)(1)(C) is amended to recognize that electronically stored information, as defined in Rule 34(a), can also be sought by subpoena. Like Rule 34(b), Rule 45(a)(1) is amended to provide that the subpoena can designate a form or forms for production of electronic data. Rule 45(c)(2) is amended, like Rule 34(b), to authorize the person served with a subpoena to object to the requested form or forms. In addition, as under Rule 34(b), Rule 45(d)(1)(B) is amended to provide that if the subpoena does not specify the form or forms for electronically stored information, the person served with the subpoena must produce electronically stored information in a form or forms in which it is usually maintained or in a form or forms that are reasonably usable. Rule 45(d)(1)(C) is added to provide that the person producing electronically stored information should not have to produce the same information in more than one form unless so ordered by the court for good cause.

As with discovery of electronically stored information from parties, complying with a subpoena for such information may impose burdens on the responding person. Rule 45(c) provides

protection against undue impositions on nonparties. For example, Rule 45(c)(1) directs that a party serving a subpoena "shall take reasonable steps to avoid imposing undue burden or expense on a person subject to the subpoena," and Rule 45(c)(2)(B) permits the person served with the subpoena to object to it and directs that an order requiring compliance "shall protect a person who is neither a party nor a party's officer from significant expense resulting from" compliance. Rule 45(d)(1)(D) is added to provide that the responding person need not provide discovery of electronically stored information from sources the party identifies as not reasonably accessible, unless the court orders such discovery for good cause, considering the limitations of Rule 26(b)(2)(C), on terms that protect a nonparty against significant expense. A parallel provision is added to Rule 26(b)(2).

Rule 45(a)(1)(B) is also amended, as is Rule 34(a), to provide that a subpoena is available to permit testing and sampling as well as inspection and copying. As in Rule 34, this change recognizes that on occasion the opportunity to perform testing or sampling may be important, both for documents and for electronically stored information. Because testing or sampling may present particular issues of burden or intrusion for the person served with the subpoena, however, the protective provisions of Rule 45(c) should be enforced with vigilance when such demands are made. Inspection or testing of certain types of electronically stored information or of a person's electronic information system may raise issues of confidentiality or privacy. The addition of sampling and testing to Rule 45(a) with regard to documents and electronically stored information is not meant to create a routine right of direct access to a person's electronic information system, although such access might be justified in some circumstances. Courts should guard against undue intrusiveness resulting from inspecting or testing such systems.

Rule 45(d)(2) is amended, as is Rule 26(b)(5), to add a procedure for assertion of privilege or of protection as trial-preparation materials after production. The receiving party may submit the information to the court for resolution of the privilege claim, as under Rule 26(b)(5)(B).



Information Governance

Getting your electronic house in order to mitigate risk & expenses should e-discovery become an issue, from initial creation of ESI (electronically stored information) through its final disposition.

Identification

Locating potential sources of ESI & determining its scope, breadth & depth.

Preservation

Ensuring that ESI is protected against inappropriate alteration or destruction.

Collection

Gathering ESI for further use in the e-discovery process (processing, review, etc.).

Processing

Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review & analysis.

Review

Evaluating ESI for relevance & privilege.

Analysis

Evaluating ESI for content & context, including key patterns, topics, people & discussion.

Production

Delivering ESI to others in appropriate forms & using appropriate delivery mechanisms.

Presentation

Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native & near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

What Every Lawyer Should Know About E-Discovery

Progress is impossible without change,
and those who cannot change their minds
cannot change anything. **--George Bernard Shaw**

We have entered a golden age of evidence, ushered in by the monumental growth of data. All who access electronically stored information (ESI) and use digital devices generate and acquire vast volumes of digital evidence. Never in human history have we had so much probative evidence, and never has that evidence been so objective and precise. Yet, lawyers are like farmers complaining of oil on their property; they bemoan electronic evidence because they haven't awoken to its value.

That's not surprising. What lawyer in practice received practical instruction in electronic evidence? Few law schools offer courses in e-discovery, and fewer teach the essential "e" that sets e-discovery apart. Continuing legal education courses shy away from the nuts and bolts of information technology needed to competently manage and marshal digital evidence. Law graduates are expected to acquire trade skills by apprenticeship; yet, experienced counsel have no e-discovery expertise to pass on. Competence in e-discovery is exceptionally rare, and there is little afoot to change that save the vain expectation that lawyers will miraculously gain competence without education or effort.

As sources of digital evidence proliferate in the cloud, on mobile devices and tablets and within the burgeoning Internet of Things, the gap between competent and incompetent counsel grows. We suffer most when standard setters decline to define competence in ways that might exclude them. Vague pronouncements of a duty to stay abreast of "relevant technology" are noble, but do not help lawyers know what they must know.[6]

So, it is heartening when the state with the second largest number of practicing lawyers in America takes a strong, clear stand on what lawyers must know about e-discovery. The State Bar of California Standing Committee on Professional Responsibility and Conduct issued an advisory opinion in which the Committee sets out the level of skill and familiarity required when, acting alone or with assistance, counsel undertakes to represent a client in a matter implicating electronic discovery.[7]

The Committee wrote:

---

[6] Rule 1.1 of the American Bar Association Model Rules of Professional Conduct provides that, "[a] lawyer shall provide competent representation to a client. Competent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation." Comment 8 to Rule 1.1 adds, "[t]o maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, *including the benefits and risks associated with relevant technology….*" Emphasis added.

[7] The State Bar of California Standing Committee on Professional Responsibility and Conduct Formal Opinion Interim No. 11-0004 (2014).

If it is likely that e-discovery will be sought, the duty of competence requires an attorney to assess his or her own e-discovery skills and resources as part of the attorney's duty to provide the client with competent representation. If an attorney lacks such skills and/or resources, the attorney must take steps to acquire sufficient learning and skill, or associate or consult with someone with appropriate expertise to assist. … Taken together generally, and under current technological standards, attorneys handling e-discovery should have the requisite level of familiarity and skill to, among other things, be able to perform (either by themselves or in association with competent co-counsel or expert consultants) the following:

1. initially assess e-discovery needs and issues, if any;
2. implement appropriate ESI preservation procedures, including the obligation to advise a client of the legal requirement to take actions to preserve evidence, like electronic information, potentially relevant to the issues raised in the litigation;
3. analyze and understand a client's ESI systems and storage;
4. identify custodians of relevant ESI;
5. perform appropriate searches;
6. collect responsive ESI in a manner that preserves the integrity of that ESI;
7. advise the client as to available options for collection and preservation of ESI;
8. engage in competent and meaningful meet and confer with opposing counsel concerning an e-discovery plan; and
9. produce responsive ESI in a recognized and appropriate manner.[8]

Thus, California lawyers face a simple mandate when it comes to e-discovery, and one that should take hold everywhere: *Learn it, get help* or *get out.* Declining the representation may be the only ethical response when the lawyer won't have sufficient time to acquire the requisite skills and the case can't sustain the cost of associating competent co-counsel or expert consultants. Most cases aren't big enough to bear the cost of two when only one is competent.

Each of the nine tasks implicate a broad range of technical and tactical skills. The interplay between technical and tactical suggests that just "asking the IT guy" some questions won't suffice. Both efficiency and effectiveness demand that, if the lawyer is to serve as decision maker and advocate, the *lawyer* needs to do more than parrot a few phrases. The lawyer needs to *understand* what the technologists are talking about.

To assess e-discovery needs and issues, a lawyer must be capable of recognizing the needs and issues that arise. This requires experience and a working knowledge of the case law and professional literature. A lawyer's first step toward competence begins with reading the leading cases and digging into the argot of information technology. When you come across an unfamiliar technical term in an opinion or article, don't elide over it. *Look it up.* Google and Wikipedia are your friends!

[8] *Id.*

Implementing appropriate ESI preservation procedures means knowing how to scope, communicate and implement a defensible legal hold. You can't be competent to scope a hold without understanding the tools and software your client uses. You can't help your client avoid data loss and spoliation if you have no idea what data is robust and tenacious and what is fragile and transitory. How do you preserve relevant data and metadata without some notion of what data and metadata exist and where it resides?

At first blush, identifying custodians of relevant ESI seems to require no special skills; but behind the scenes, a cadre of custodians administer and maintain the complex and dynamic server and database environments businesses use. You can't expect custodians no more steeped in information technology than you to preserve backup media or suspend programs purging data your client must preserve. These are tasks for IT. Competence includes the ability to pose the right questions to the right people.

Performing appropriate searches entails more than just guessing what search terms seem sensible. *Search is a science.* Search tools vary widely, and counsel must understand what these tools can and cannot do. Queries should be tested to assess precision and recall. Small oversights in search prompt big downstream costs, and small tweaks prompt big savings. How do you negotiate culling and filtering criteria if you don't understand the ways ESI can be culled and filtered?

Some ESI can be preserved in place with little cost and burden and may even be safely and reliably searched in place to save money. Other ESI requires data be collected and processed to be amenable to search. Understanding which is which is crucial to being competent to advise clients about available options.

Lawyers lacking e-discovery skills can mount a successful meet and confer on ESI issues by getting technically-astute personnel together to 'dance geek-to-geek.' But, that's can be expensive, and cautious, competent counsel will want to *understand* the risks and costs, not just trust the technologists to know what's relevant and how and when to protect privileged and sensitive data.

Competent counsel understands that there is no one form suited to production of every item of ESI and know the costs and burdens associated with alternate forms of production. Competent counsel knows that converting native electronic formats to TIFF images increases the size of the files many time and thus needlessly inflates the cost of ingestion and hosting by vendors. Competent counsel knows when it's essential to demand native forms of production to guard against data loss and preserve utility. Conversely, competent counsel knows how to make the case for TIFF production to handicap an opponent or when needed for redaction.

Clearly, there's a lot more to e-discovery than many imagine, and much of it must fall within counsel's ken. Virtually all evidence today is born digitally. It's data, and only a fraction takes

forms we've traditionally called documents. Lawyers ignored ESI for decades while information technologies changed the world. Is it any wonder that lawyers have a lot of catching up to do? Few excel at all the skills that trial work requires; but, every trial lawyer must be minimally competent in them all. Today, the most demanding of these skills is e-discovery.

Is it fair to deem lawyers incompetent, even unethical, because they don't possess skills they weren't taught in law school? It may not feel fair to lawyers trained for a vanished world of paper documents; but to the courts and clients ill-served by those old ways, it's more than just fair—it's right.

Introduction to Digital Computers, Servers and Storage

In 1774, a Swiss watchmaker named Pierre Jaquet-Droz built an ingenious mechanical doll resembling a barefoot boy. Constructed of 6,000 handcrafted parts and dubbed "L'Ecrivain" ("The Writer"), Jaquet-Droz' automaton uses quill and ink to handwrite messages in cursive, up to 40 letters long, with the content controlled by interchangeable cams. The Writer is a charming example of an early programmable computer.

The monarchs that marveled at Jaquet-Droz' little penman didn't need to understand how it worked to enjoy it. Lawyers, too, once had little need to understand the operation of their clients' information systems to conduct discovery. But as the volume of electronically stored information (ESI) has exploded and the forms and sources of ESI continue to morph and multiply, lawyers conducting electronic discovery cannot ignore the clockwork anymore. New standards of competence demand that lawyers and litigation support personnel master certain fundamentals of information technology and electronic evidence.

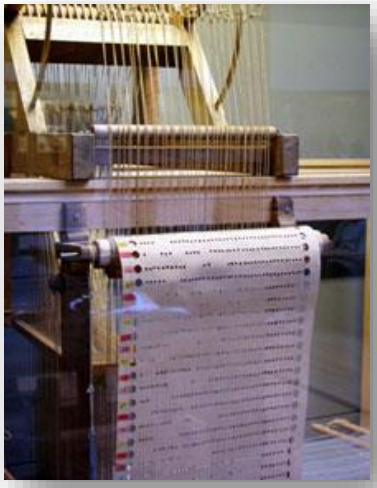
Data, Not Documents

Lawyers—particularly those who didn't grow up with computers—tend to equate data with documents when, in a digital world, documents are just one of the many forms in which electronic information exists. Documents akin to the letters, memos and reports of yore account for a dwindling share of electronically stored information relevant in discovery, and documents generated from electronic sources tend to convey just part of the information stored in the source. The decisive information in a case may exist as nothing more than a single bit of data that, in context, signals whether the fact you seek to establish is true or not. A Facebook page doesn't exist until a request sent to a database triggers the page's assembly and display. Word documents, PowerPoint presentations and Excel spreadsheets lose content and functionality when printed to screen images or paper.

With so much discoverable information bearing so little resemblance to documents, and with electronic documents carrying much more probative and useful information than a printout or screen image conveys, competence in electronic discovery demands an appreciation of data more than documents.

Introduction to Data Storage Media

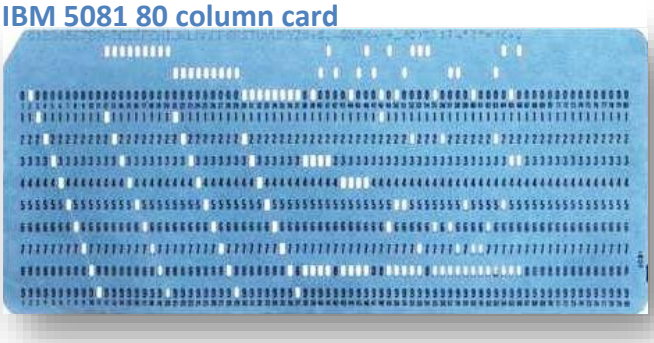
Mankind has been storing data for thousands of years, on stone, bone, clay, wood, metal, glass, skin, papyrus, paper, plastic and film. In fact, people were storing data in binary formats long before the emergence of modern digital computers. Records from 9th century Persia describe an organ playing interchangeable cylinders. Eighteenth century textile manufacturers employed perforated rolls of paper to control looms, and Swiss and German music box makers used metal drums or platters to store tunes. At the dawn of the Jazz Age, no self-respecting American family of means lacked a player piano capable (more-or-less) of reproducing the works of the world's greatest pianists.



Whether you store data as a perforation or a pin, you're storing binary data. That is, there are two data states: hole or no hole, pin or no pin. Zeroes or ones.

Punched Cards

IBM 5081 80 column card



In the 1930's, demand for *electronic* data storage led to the development of fast, practical and cost-effective binary storage media. The first of these were punched cards, initially made in a variety of sizes and formats, but ultimately standardized by IBM as the 80 columns, 12 row (7.375" by 3.25") format (right) that dominated computing well into the 1970's. [From 1975-79, the author spent many a midnight in the

basement of a computer center at Rice University typing program instructions on these unforgiving punch cards].

The 1950's saw the emergence of magnetic storage as the dominant medium for electronic data storage, and it remains so today. Although optical and solid-state storage are expected to ultimately eclipse magnetic media for local storage, magnetic storage will continue to dominate network and cloud storage well into the 2020s, if not beyond.

Magnetic Tape

The earliest popular form of magnetic data storage was magnetic tape. Compact cassette tape was the earliest data storage medium for personal computers including the pioneering Radio Shack TRS-80 and the very first IBM personal computer, the model XT.

Spinning reels of tape were a clichéd visual metaphor for computing in films and television shows from the 1950s through 1970's. Though the miles of tape on those reels now resides in cartridges and cassettes, tapes remain an enduring medium for backup and archival of electronically stored information.

The LTO-7 format tapes introduced in 2015 house 3,150 feet of half inch tape in a cartridge just four inches square and less than an inch thick; yet, each cartridge natively hold 6.0 terabytes of uncompressed data and up to 15 TB of compressed data,[9] delivered at a transfer rate of 315 megabytes per second. LTO tapes use a back-and-forth or linear serpentine recording scheme. "Linear" because it stores data in parallel tracks running the length of the tape, and "serpentine" because its path snakes back-and forth, reversing direction on each pass. Thirty-two of the LTO-7 cartridge's 3,584 tracks are read or written as the tape moves past the recording heads, so it takes 112 back-and-forth passes or "wraps" to read or write the full contents of a single LTO-7 cartridge.

That's about *67 miles* of tape passing the heads! So, it takes *hours* to read each tape.

[9] Since most data stored on backup tape is compressed, the actual volume of ESI on tape may be 2-3 times greater than the native capacity of the tape.

While tape isn't as fast as hard drives, it's proven to be more durable and less costly for long term storage; that is, so long as the data is being *stored,* not *re*stored.

LTO-7 Ultrium Tape

Sony AIT-3 Tape

SDLT-II Tape

For further information, *see* Ball, Technology Primer: Backups in Civil Discovery at http://craigball.com/Backups_in_E-Discovery_Primer_2016.pdf

Floppy Disks

8", 5.25" and 3.5" Floppy Disks

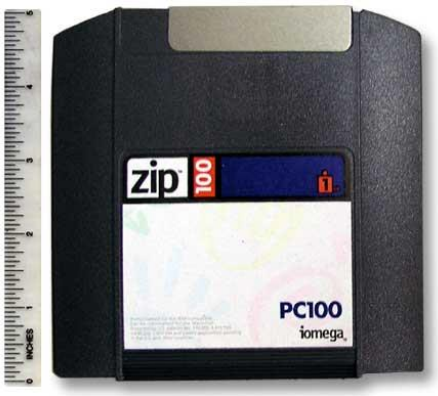
It's rare to encounter a floppy disk today, but floppy disks played a central role in software distribution and data storage for personal computing for almost thirty years. Today, the only place a computer user is likely to see a floppy disk is as the menu icon for storage on the menu bar of Microsft Office applications. All floppy disks have a spinning, flexible plastic disk coated with a magnetic oxide (e.g., rust). The disk is essentially the same composition as magnetic tape in disk form. Disks are **formatted** (either by the user or pre-formatted by the manufacturer) so as to divide the disk into various concentric rings of data called **tracks**, with tracks further subdivided into tiny arcs called **sectors**. Formatting enables systems to locate data on physical storage media much as roads and lots enable us to locate homes in a neighborhood.

Though many competing floppy disk sizes and formats have been introduced since 1971, only five formats are likely to be encountered in e-discovery. These are the 8", 5.25", 3.5 standard, 3.5 high

density and Zip formats and, of these, the 3.5HD format 1.44 megabyte capacity floppy is by far the most prevalent legacy floppy disk format.

The Zip Disk was one of several proprietary “super floppy” products that enjoyed brief success before the high capacity and low cost of recordable optical media (CD-R and DVD-R) and flash drives rendered them obsolete.

Zip Disk



8" Floppy Disk in Use

Optical Media

The most common forms of optical media for data storage are the CD, DVD and Blu-ray disks in read only, recordable or rewritable formats. Each typically exists as a 4.75” plastic disk with a metalized reflective coating and/or dye layer that can be distorted by a focused laser beam to induce pits and lands in the media. These pits and lands, in turn, interrupt a laser reflected off the surface of the disk to generate the ones and zeroes of digital data storage. The practical difference between the three prevailing forms of optical media are their native data storage capacities and the availability of drives to read them.

A **CD** (for **C**ompact **D**isk) or **CD-ROM** (for **CD R**ead **O**nly **M**edia) is read only and not recordable by the end user. It’s typically fabricated in factory to carry music or software. A **CD-R** is recordable by the end user, but once a recording session is closed, it cannot be altered in normal use. A **CD-RW** is a re-recordable format that can be erased and written to multple times. The native data storage capacity of a standard-size CD is about 700 megabytes.

A **DVD** (for **D**igital **V**ersitile **D**isk) also comes in read only, recordable (**DVD±R**) and rewritable (**DVD±RW**) iterations and the most common form of the disk has a native data storage capacity of approximately 4.7 gigabytes. So, one DVD holds the same amount of data as six and one-half CDs.

By employing the narrower wavelength of a blue laser to read and write disks, a dual layer **Blu-ray** disk can hold up to about 50 gigabytes of data, equalling the capacity of about ten and one-half DVDs. Like their predecessors, Blu-ray disks are available in recordable (BD-R) and rewritable (CD-RE) formats

Though ESI resides on a dizzying array of media and devices, by far the largest complement of same occurs within three closely-related species of computing hardware: *computers, hard drives* and *servers.* A server is essentially a computer dedicated to a specialized task or tasks, and both servers and computers routinely employ hard drives for program and data storage.

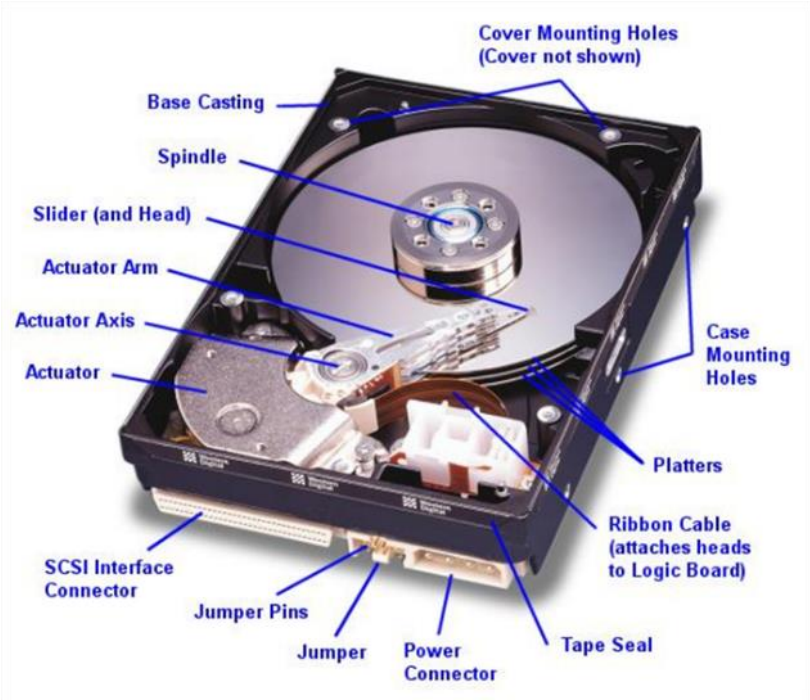
Conventional Electromagnetic Hard Drives

A hard drive is an immensely complex data storage device that's been engineered to appear deceptively simple. When you connect a hard drive to your machine, and the operating system detects the drive, assigns it a drive letter and—presto!—you've got trillions of bytes of new storage! Microprocessor chips garner the glory, but the humdrum hard drive is every bit a paragon of ingenuity and technical prowess.

A conventional personal computer hard drive is a sealed aluminum box measuring (for a desktop system) roughly 4" x 6" x 1" in height. A hard drive can be located almost anywhere within the case and is customarily secured by several screws attached to any of ten pre-threaded mounting holes along the edges and base of the case. One face of the case will be labeled to reflect the drive specifications, while a printed circuit board containing logic and controller circuits will cover the opposite face.

A conventional hard disk contains round, flat discs called **platters**, coated on both sides with a special material able to store data as magnetic patterns. Much like a record player, the platters

have a hole in the center allowing multiple platters to be stacked on a spindle for greater storage capacity.



The platters rotate at high speed—typically 5,400, 7,200 or 10,000 rotations per minute—driven by an electric motor. Data is written to and read from the platters by tiny devices called **read/write heads** mounted on the end of a pivoting extension called an **actuator arm** that functions similarly to the tone arm that carried the phonograph cartridge and needle across the face of a record. Each platter has two read/write heads, one on the top of the platter and one on the bottom. So, a conventional hard disk with three platters typically sports six surfaces and six read/write heads.

Unlike a record player, the read/write head never touches the spinning platter. Instead, when the platters spin up to operating speed, their rapid rotation causes air to flow under the read/write heads and lift them off the surface of the disk—the same principle of lift that operates on aircraft wings and enables them to fly. The head then reads the magnetic patterns on the disc while flying just .5 millionths of an inch above the surface. At this speed, if the head bounces against the surface, there is a good chance that the head will burrow into the surface of the platter, obliterating data, destroying both read/write heads and rendering the hard drive inoperable—a so-called "head crash."

The hard disk drive has been around for more than 50 years, but it was not until the 1980's that the physical size and cost of hard drives fell sufficiently for their use to be commonplace.

Introduced in 1956, the IBM 350 Disk Storage Unit pictured was the first commercial hard drive. It was 60 inches long, 68 inches high and 29 inches deep (so it could fit through a door). Called the RAMAC (for Random Access Method of Accounting and Control), it held fifty 24" magnetic disks of 50,000 sectors, each storing 100 alphanumeric characters. Thus, it held under five megabytes, or enough for about two cellphone snapshots today. It weighed a ton (literally), and users paid $3,200.00 per month to *rent* it. That's about $28,000.00 in 2017 dollars.

Today, you can buy a ten terrabyte hard drive storing *two million times* more information for a fraction of that monthly rental. That 10TB drive weighs less than two pounds and can hide behind a paperback book.

Over time, hard drives took various shapes and sizes (or "form factors" as the standard dimensions of key system components are called in geek speak). Three form factors are still in use: 3.5" (desktop drive), 2.5" (laptop drive) and 1.8" (iPod and microsystem drive, now supplanted by solid state storage).

Hard drives connect to computers by various mechanisms called "interfaces" that describe both how devices "talk" to one-another as well as the physical plugs and cabling required. The five most common hard drive interfaces in use today are:

PATA for **P**arallel **A**dvanced **T**echnology **A**ttachment (sometimes called EIDE for **E**xtended **I**ntegrated **D**rive **E**lectronics):

SATA for **S**erial **A**dvanced **T**echnology **A**ttachment

SCSI for **S**mall **C**omputer **S**ystem Interface

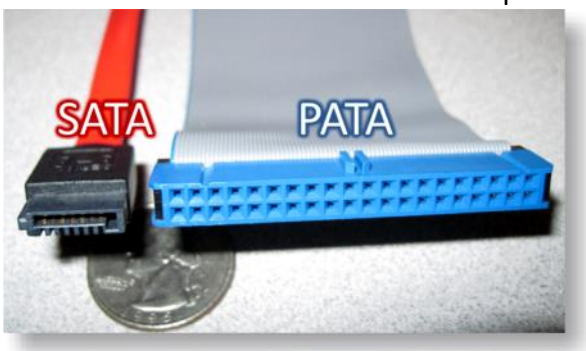
SAS for **S**erial **A**ttached **S**CSI

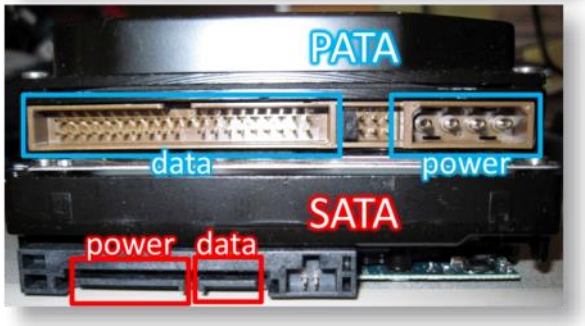
FC for **F**ibre **C**hannel

Though once dominant in personal computers, PATA drives are rarely found in machines manufactured after 2006. Today, virtually all laptop and desktop computers employ SATA drives for local storage. SCSI, SAS and FC drives tend to be seen exclusively in servers and other applications demanding high performance and reliability.

From the user's perspective, PATA, SATA, SCSI, SAS and FC drives are indistinguishable; however, from the point of view of the technician tasked to connect to and image the contents of the drive, the difference implicates different tools and connectors.

The five drive interfaces divide into two employing parallel data paths (PATA and SCSI) and three employing serial data paths (SATA, SAS and FC). Parallel ATA interfaces route data over multiple simultaneous channels necessitating 40 wires where serial ATA interfaces route data through a single, high-speed data channel requiring only 7 wires. Accordingly, SATA cabling and connectors are smaller than their PATA counterparts (see photos, right).





Fibre Channel employs optical fiber (the spelling difference is intentional) and light waves to carry data at impressive speeds. The premium hardware required by FC dictates that it will be found in enterprise computing environments, typically in conjunction with a high capacity/high demand storage device called a **SAN** (for Storage Attached Network) or a **NAS** (for Network Attached Storage).

It's easy to become confused between hard drive interfaces and external data transfer interfaces like USB or FireWire seen on external hard drives. The drive within the external hard drive housing will employ one of the interfaces described above (except FC); however, to facilitate external connection to a computer, a device called a **bridge** will convert data written to and from the hard drive to a form that can traverse a USB or FireWire connection. In some compact, low-cost external drives, manufacturers dispense with the external bridge board altogether and build the USB interface right on the hard drive's circuit board.

Flash Drives, Memory Cards, SIMs and Solid State Drives

Computer memory storage devices have no moving parts and the data resides entirely within the solid materials which compose the memory chips, hence the term, "solid state." Historically, rewritable memory was volatile (in the sense that contents disappeared when power was withdrawn) and expensive. But, beginning around 1995, a type of non-volatile memory called NAND flash became sufficiently affordable to be used for removable storage in emerging applications like digital photography. Further leaps in the capacity and dips in the cost of NAND flash led to the near-eradication of film for photography and the extinction of the floppy disk, replaced by simple, inexpensive and reusable USB storage devices called, variously, SmartMedia, Compact Flash media, SD cards, flash drives, thumb drives, pen drives and memory sticks or keys.

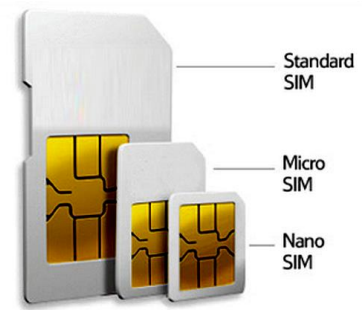
SmartMedia Card

Compact Flash Media

Secure Digital (SD) Media

USB Flash Drives

A specialized form of solid state memory seen in cell phones is the **Subscriber Identification Module** or **SIM card**. SIM cards serve both to authenticate and identify a communications device on a cellular network and to store SMS messages and phone book contacts.



SIM Cards

As the storage capacity of NAND flash has gone up and its cost has come down, the conventional electromagnetic hard drive is rapidly being replaced by **solid state drives** in standard hard drive form factors. Solid state drives are significantly faster, lighter and more energy efficient than conventional drives, but they currently cost anywhere from 10-20 times more per gigabyte than their mechanical counterparts. All signs point to the ultimate obsolescence of mechanical drives by solid state drives, and some products (notably tablets like

the iPad and Microsoft Surface or ultra-lightweight laptops like the MacBook Air) have eliminated hard drives altogether in favor of solid state storage.

Currently, solid state drives assume the size and shape of mechanical drives to facilitate compatibility with existing devices. However, the size and shape of mechanical hard drives was driven by the size and operation of the platter they contain. Because solid state storage devices have no moving parts, they can assume virtually any shape. It's likely, then, that slavish adherence to 2.5" and 3.5" rectangular form factors will diminish in favor of shapes and sizes uniquely suited to the devices that employ them.

With respect to e-discovery, the shift from electromagnetic to solid state drives is inconsequential. However, the move to solid state drives will significantly impact matters necessitating computer forensic analysis. Because the NAND memory cells that comprise solid state drives wear out rapidly with use, solid state drive controllers must constantly reposition data to insure usage is distributed across all cells. Such "wear leveling" hampers techniques that forensic examiners have long employed to recover deleted data from conventional hard drives.

RAID Arrays

Whether local to a user or in the Cloud, hard drives account for nearly all the electronically stored information attendant to e-discovery. In network server and Cloud applications, hard drives rarely work alone. That is, hard drives are ganged together to achieve greater capacity, speed and reliability in so-called **R**edundant **A**rrays of **I**ndependent **D**isks or **RAID**s. In the SAN pictured at left, the 16 hard drives housed in trays may be accessed as **J**ust a **B**unch **o**f **D**isks or **JBOD**, but it's far more likely they are working together as a RAID

RAIDs serve two ends: redundancy and performance. The redundancy aspect is obvious—two drives holding identical data safeguard against data loss due to mechanical failure of either drive—but how do multiple drives improve **performance**? The answer lies in splitting the data across more than one drive using a technique called **striping**.

A RAID improves performance by dividing data across more than one physical drive. The swath of data deposited on one drive in an array before moving to the next drive is called the "stripe." If you imagine the drives lined up alongside one-another, you can see why moving back-and-forth

the drives to store data might seem like painting a stripe across the drives. By striping data, each drive can deliver their share of the data simultaneously, increasing the amount of information handed off to the computer's microprocessor.

But, when you stripe data across drives, Information is lost if any drive in the stripe fails. You gain performance, but surrender security.

This type of RAID configuration is called a **RAID 0**. It wrings maximum performance from a storage system; but it's risky.

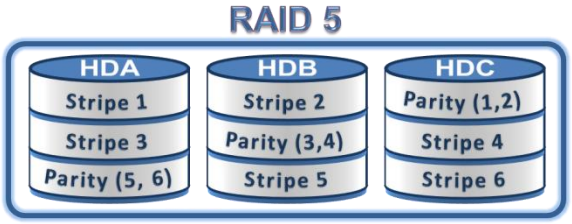
If RAID 0 is for gamblers, **RAID 1** is for the risk averse. A RAID 1 configuration duplicates everything from one drive to an identical twin, so that a failure of one drive won't lead to data loss. RAID 1 doesn't improve performance, and it requires twice the hardware to store the same information.

Other RAID configurations strive to integrate the *performance* of RAID 0 and the *protection* of RAID 1.

Thus, a "RAID 0+1" mirrors two striped drives, but demands four hard drives delivering only half their total storage capacity. Safe and fast, but not cost-efficient. The solution lies in a concept called ***parity,*** key to a range of other sequentially numbered RAID configurations. Of those other configurations, the ones most often seen are called **RAID 5 and RAID 7.**

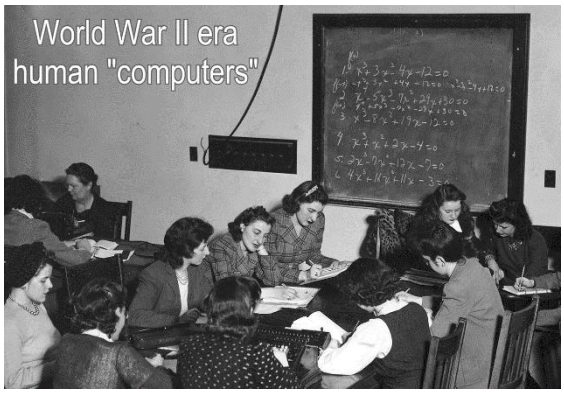
To understand parity, consider the simple equation 5 + 2 = 7. If you didn't know one of the three values in this equation, you could easily solve for the missing value, *i.e.,* presented with "5 + __ = 7," you can reliably calculate the missing value is 2. In this example, "7" is the ***parity value*** or checksum for "5" and "2."

The same process is used in RAID configurations to gain increased performance by striping data across multiple drives while using parity values to permit the calculation of any missing values lost to drive failure. In a three-drive array, any one of the drives can fail, and we can use the remaining two to recreate the third (just as we solved for 2 in the equation above).



In this illustration, data is striped across three hard drives, HDA, HDB and HDC. HDC holds the parity values for data stripe 1 on HDA and stripe 2 on HDB. It's shown as "Parity (1, 2)." The parity values for the other stripes are distributed on the other drives. Again, any one of the three drives can fail and all of the data is recoverable. This configuration is RAID 5 and, though it requires a minimum of three drives, it can be expanded to dozens or hundreds of disks.

Computers

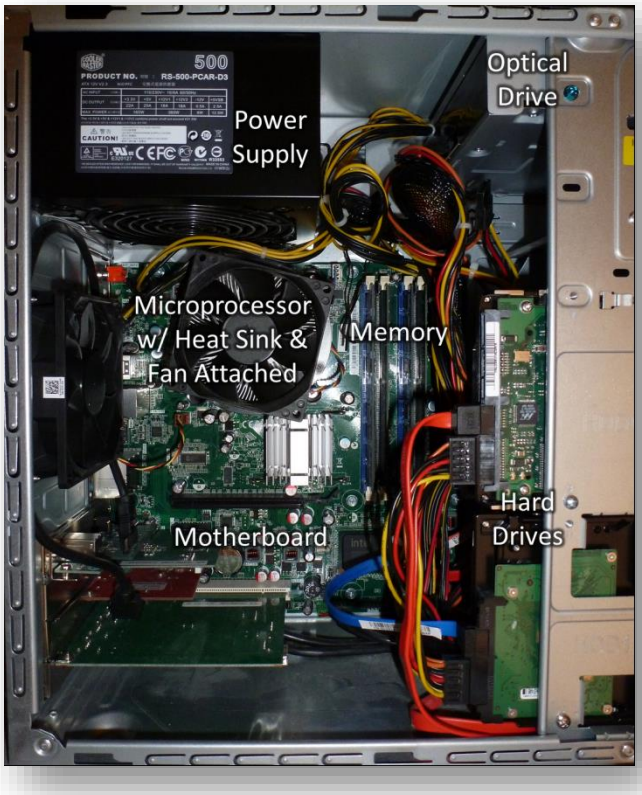


Historically, all sorts of devices—and even people—were "computers." During World War II, human computers—women for the most part—were instrumental in calculating artillery trajectories and assisting with the challenging number-crunching needed by the Manhattan Project. Today, laptop and desktop personal computers spring to mind when we hear the term "computer;" yet smart phones, tablet devices, global positioning systems, video gaming platforms, televisions and a host of other intelligent tools and toys are also computers. More precisely, the **central processing unit** (CPU) or **microprocessor** of the system is the "computer," and the various input and output devices that permit humans to interact with the processor are termed **peripherals.** The key distinction between a mere calculator and a computer is the latter's ability to be programmed and its use of **memory** and **storage**. The physical electronic and mechanical components of a computer are its **hardware,** and the instruction sets used to program a computer are its **software.** Unlike the interchangeable cams of Pierre Jaquet-Droz' mechanical doll, modern electronic computers receive their instructions in the form of digital data typically retrieved from the same electronic storage medium as the digital information upon which the computer performs its computational wizardry.

When you push the power button on your computer, you trigger an extraordinary, expedited education that takes the machine from insensible illiterate to worldly savant in a matter of seconds. The process starts with a snippet of data on a chip called the ROM BIOS storing just enough information in its Read Only Memory to grope around for the Basic Input and Output System peripherals (like the keyboard, screen and, most importantly, the hard drive). The ROM BIOS also holds the instructions needed to permit the processor to access more and more data from the hard drive in a widening gyre, "teaching" itself to be a modern, capable computer.

This rapid, self-sustaining self-education is as magical as if you lifted yourself into the air by pulling on the straps of your boots, which is truly why it's called "bootstrapping" or just "booting" a computer.

Computer hardware circa 2014 shares certain common characteristics. Within the CPU, a microprocessor chip is the computational "brains" of system and resides in a socket on the **motherboard**, a rigid surface etched with metallic patterns serving as the wiring between the components on the board. The microprocessor generates considerable heat necessitating the attachment of a heat dissipation device called a **heat sink**, often abetted by a small fan. The motherboard also serves as the attachment point for memory boards (grouped as modules or "sticks") called **RAM** for Random Access Memory. RAM serves as the working memory of the processor while it performs calculations; accordingly, the more memory present, the more information can be processed at once, enhancing overall system performance.



Other chips comprise a Graphics Processor Unit (GPU) residing on the motherboard or on a separate expansion board called a **video card** or **graphics adapter**. The GPU supports the display of information from the processor onto a monitor or projector and has its own complement of memory dedicated to superior graphics performance. Likewise, specialized chips on the motherboard or an expansion board called a **sound card** support the reproduction of audio to speakers or a headphone. Video and sound processing capabilities may even be fully integrated into the microprocessor chip.

The processor communicates with networks through an interface device called a **network adapter** which connects to the network physically, through a **LAN Port,** or wirelessly using a Wi-Fi connection.

Users convey information and instructions to computers using tactile devices like a keyboard, mouse or track pad, but may also employ voice or gestural recognition mechanisms.

Persistent storage of data is a task delegated to other peripherals: **optical drives** (CD-ROM and DVD-ROM devices), **floppy disk drives**, **solid-state media** (*i.e.,* thumb drives) and, most commonly, **hard drives**.

All the components just described require electricity, supplied by batteries in portable devices or by a **power supply** converting AC current to the lower DC voltages required by electronics.

From the standpoint of electronic discovery, it's less important to define these devices than it is to fathom the information they hold, the places it resides and the forms it takes. Parties and lawyers have been sanctioned for what was essentially their failure to inquire into and understand the roles computers, hard drives and servers play as repositories of electronic evidence. Moreover, much money spent on electronic discovery today is wasted as a consequence of parties' efforts to convert ESI to paper-like forms instead of learning to work with ESI in the forms in which it customarily resides on computers, hard drives and servers.

Servers

Servers were earlier defined as computers dedicated to a specialized task or tasks. But that definition doesn't begin to encompass the profound impact upon society of the so-called **client-server** computing model. The ability to connect local "client" applications to servers via a network, particularly to **database servers**, is central to the operation of most businesses and to all telecommunications and social networking. Google and Facebook are just enormous groupings of servers, and the Internet merely a vast, global array of shared servers.

Local, Cloud and Peer-to-Peer Servers

For e-discovery, let's divide the world of servers into three realms: Local, Cloud and Peer-to-Peer server environments.

"Local" servers employ hardware that's physically available to the party that owns or leases the servers. Local servers reside in a computer room on a business' premises or in leased equipment "lockers" accessed at a co-located data center where a lessor furnishes, *e.g.,* premises security, power and cooling. Local servers are easiest to deal with in e-discovery because physical access to the hardware supports more and faster options when it comes to preservation and collection of potentially responsive ESI.

"Cloud" servers typically reside in facilities not physically accessible to persons using the servers, and discrete computing hardware is typically not dedicated to a particular user. Instead, the Cloud computing consumer is buying services via the Internet that emulate the operation of a single machine or a room full of machines, all according to the changing needs of the Cloud consumer. Web mail is the most familiar form of Cloud computing, in a variant called SaaS (for Software as a Service). Webmail providers like Google, Yahoo and Microsoft make e-mail accounts available on their servers in massive data centers, and the data on those servers is available solely via the Internet, no user having the right to gain physical access to the machines storing their messaging.

"Peer-to-Peer" (P2P) networks exploit the fact that any computer connected to a network has the potential to serve data across the network. Accordingly, P2P networks are decentralized; that is, each computer or "node" on a P2P network acts as client and server, sharing storage space, communication bandwidth and/or processor time with other nodes. P2P networking may be employed to share a printer in the home, where the computer physically connected to the printer acts as a print server for other machines on the network. On a global scale, P2P networking is the technology behind file sharing applications like BitTorrent and Gnutella that have garnered headlines for their facilitation of illegal sharing of copyrighted content. When users install P2P applications to gain access to shared files, they simultaneously (and often unwittingly) dedicate their machine to serving up such content to a multitude of other nodes.

Virtual Servers

Though we've so far spoken of server hardware, *i.e.,* physical devices, servers may also be implemented *virtually*, through software that *emulates* the functions of a physical device. Such "hardware virtualization" allows for more efficient deployment of computing resources by enabling a single physical server to host multiple virtual servers.

Virtualization is the key enabling technology behind many Cloud services. If a company needs powerful servers to launch a new social networking site, it can raise capital and invest in the hardware, software, physical plant and personnel needed to support a data center, with the attendant risk that it will be over-provisioned or under-provisioned as demand fluctuates. Alternatively, the startup can secure the computing resources it needs by using virtual servers hosted by a Cloud service provider like Amazon, Microsoft or Rackspace. Virtualization permits computing resources to be added or retired commensurate with demand, and being pay-as-you-go, it requires little capital investment. Thus, a computing platform or infrastructure can be virtualized and leased, *i.e.,* offered as a service via the internet. Accordingly, Cloud Computing is sometimes referred to as **PaaS** *(Platform as a Service)* and **IaaS** *(Infrastructure as a Service)*. Web-based applications are **SaaS** *(Software as a Service)*.

It's helpful for attorneys to understand the role of **virtual machines** (VMs) because the ease and speed with which VMs are deployed and retired, as well as their isolation within the operating system, can pose unique risks and challenges in e-discovery, especially with respect to implementing a proper legal hold and when identifying and collecting potentially responsive ESI.

Server Applications

Computers dedicated to server roles typically run operating systems optimized for server tasks and applications specially designed to run in a server environment. In turn, servers are often

dedicated to supporting specific functions such as serving web pages (*Web Server*), retaining and delivering files from shared storage allocations (File Server), organizing voluminous data (*Database Server*), facilitating the use of shared printers (*Print Server*), running programs (*Application Server*) or handling messages (*Mail Server*). These various server applications may run physically, virtually or as a mix of the two.

Network Shares

Sooner or later, all electronic storage devices fail. Even the RAID storage arrays previously discussed do not forestall failure, but instead afford a measure of redundancy to allow for replacement of failed drives before data loss. Redundancy is the sole means by which data can be reliably protected against loss; consequently, companies routinely back up data stored on server NAS and SAN storage devices to backup media like magnetic tape or online (*i.e.,* Cloud) storage services. However, individual users often fail to back up data stored on local drives. Accordingly, enterprises allocate a "share" of network-accessible storage to individual users and "map" the allocation to the user's machine, allowing use of the share as if it were a local hard drive. When the user stores data to the mapped drive, that data is backed up along with the contents of the file server. Although **network shares** are not local to the user's computer, they are typically addressed using drive letters (*e.g.,* M: or T:) as if they were local hard drives.

Practice Tips for Computers, Hard Drives and Servers

Your first hurdle when dealing with computers, hard drives and servers in e-discovery is to identify potentially responsive sources of ESI and take appropriate steps to inventory their relevant contents, note the form and associated metadata of the potentially responsive ESI, then preserve it against spoliation. As the volume of ESI to be collected and processed bears on the expense and time required, it's useful to get a handle on data volumes, file types, metadata, replication and distribution as early in the litigation process as possible.

Start your ESI inventory by taking stock of physical computing and storage devices. For each machine or device holding potentially responsive ESI, you may wish to collect some or all of the following information:

- Manufacturer and model
- Serial number and/or service or asset tag
- Operating system
- Custodian
- Location
- Type of storage (don't miss removable media, like SD and SIM cards)

- Aggregate storage capacity (in MB, GB or TB)
- Encryption status
- Credentials (user IDs and passwords), if encrypted
- Prospects for upgrade or disposal
- If you'll preserve ESI by drive imaging, it's helpful to identify device interfaces.

For servers, further information might include:

- Purpose(s) of the server (*e.g.,* web server, file server, print server, etc.)
- Names and contact information of server administrator(s)
- Time in service and data migration history
- Whether hardware virtualization is used
- RAID implementation(s)
- Users and privileges
- Logging and log retention practices
- Backup procedures and backup media rotation and retention
- Whether the server is "mission critical" and cannot be taken offline or can be downed.

When preserving the contents of a desktop or laptop computer, it's typically unnecessary to sequester any component of the machine other than its hard drive(s) since the ROM BIOS holds little information beyond the rare forensic artifact. Before returning a chassis to service with a new hard drive, be sure to document the custodian, manufacturer, model and serial number/service tag of the redeployed chassis, retaining this information with the sequestered hard drive.

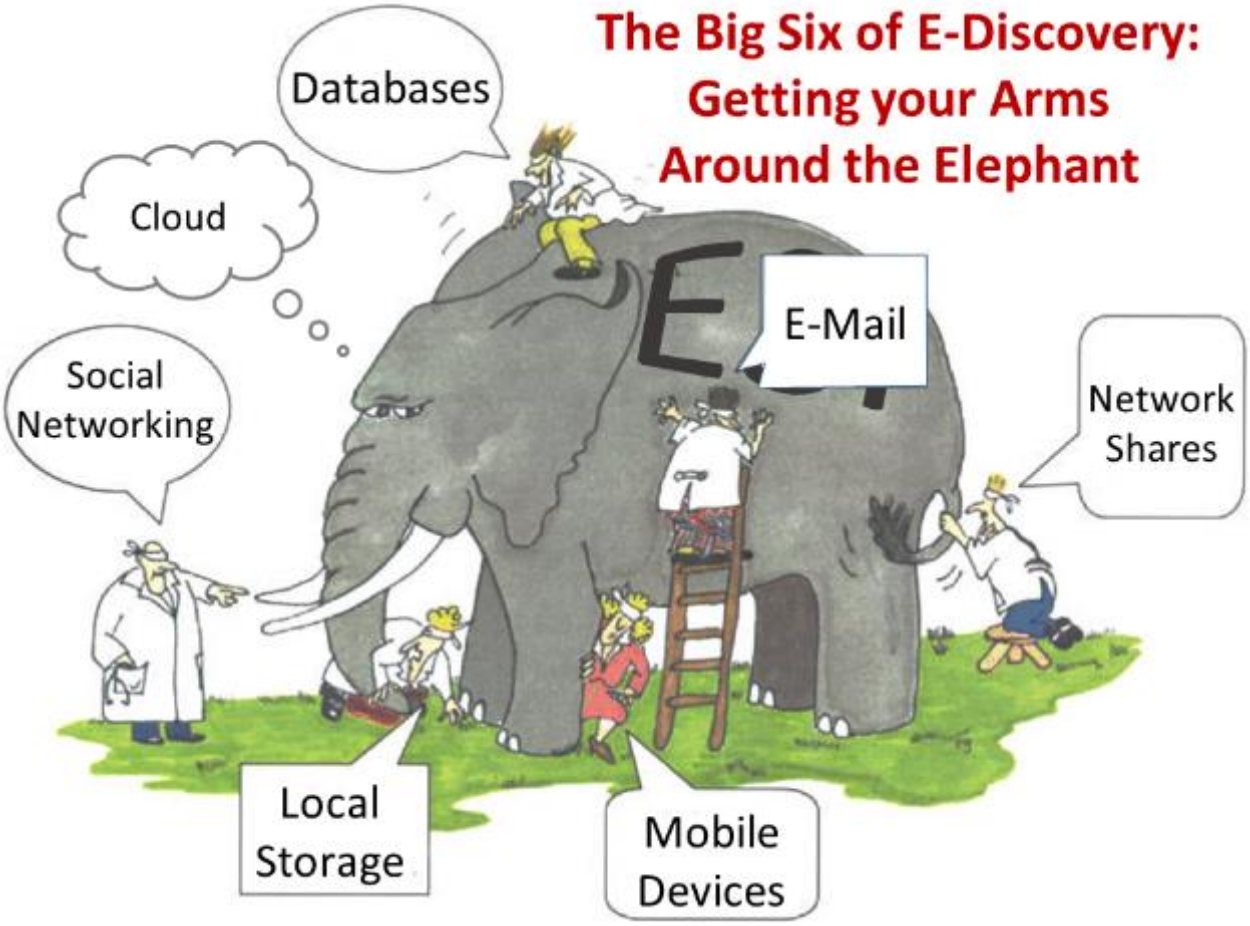
The ability to fully explore the contents of servers for potentially responsive information hinges upon the privileges extended to the user. Be sure that the person tasked to identify data for preservation or collection holds administrator-level privileges.

Above all, remember that computers, hard drives and servers are constantly changing while in service. Simply rebooting a machine alters system metadata values for large numbers of files. Accordingly, you should consider the need for evidentiary integrity before exploring the contents of a device, at least until appropriate steps are taken to guard against unwitting alteration. Note also that connecting an evidence drive to a new machine effects changes to the evidence unless suitable write blocking tools or techniques are employed.

Getting your Arms around the ESI Elephant

Many cultures and religions share the parable of the six blind men that touched an elephant. The one who grabbed the tail described the elephant as "like a snake." The blind man who grabbed the trunk said, "no, more like a tree branch," and the one with his arms around the elephant's leg said, "you're both wrong, an elephant is like a tree trunk." The man touching the ear opined that the elephant was like a large leaf, and the blind man at the tusk said, "you're all crazy. It is like a spear." None of them understood the true nature of the elephant because they failed to consider all its aspects.

In e-discovery, too, we cannot grasp the true nature of potentially responsive data until we touch many parts of the ESI elephant.



There are no forms or checklists that can take the place of understanding electronic evidence any more than a Polish phrasebook will equip you to try a case in Gdańsk. But, there are a few rules of thumb that, *applied thoughtfully*, will help you get your arms around the ESI elephant. Let's start with the Big Six and work through some geek speak as we go.

The Big Six…Plus

Without knowing anything about IT systems, you can safely assume there are at least six principal sources of digital evidence that may yield responsive ESI:

1. Key Custodians' E-Mail (Sources: server, local, archived and cloud)

Corporate computer users will have a complement of e-mail under one or more **e-mail aliases** (i.e., shorthand addresses) stored on one or more **e-mail servers**. These servers may be physical hardware managed by IT staff or **virtual machines** leased from a **cloud provider,** either running mail server software, most likely applications called **Microsoft Exchange** or **Lotus Domino.** A third potential source is a **Software as a Service (SaaS)** offering from a cloud provider, an increasingly common and important source**.** Webmail may be as simple as a single user's Gmail account or, like the Microsoft Office 365 product, a complete replication of an enterprise e-mail environment, sometimes supporting e-discovery preservation and search capabilities.

Users also tend to have a different, but overlapping complement of e-mail stored on desktops, laptops and handheld devices they've regularly used. On desktops and laptops, e-mail is found **locally** (on the user's hard drive) in **container files** with the file extensions **.pst** and **.ost** for Microsoft Outlook users or **.nsf** for Lotus Notes users. Finally, each user may be expected to have a substantial volume of **archived e-mail** spread across several on- and offline sources, including backup tapes, **journaling servers** and local archives on workstations and in network storage areas called **shares** (discussed below).

These locations are the "*where*" of e-mail, and it's crucial to promptly pin down "where" to ensure that your clients (or your opponents) don't overlook sources, especially any that may spontaneously disappear over time through **purges** (automatic deletion) or backup media **rotation** (reuse by overwriting).

Your goal here is to determine for each key custodian what they have in terms of:

- *Types of messages* (did they retain both Sent Items and Inbox contents? Have they retained messages as they were foldered by users?);
- *Temporal range of messages* (what are the earliest dates of e-mail messages, and are there significant gaps?); and
- *Volume* (numbers of messages and attachments versus total gigabyte volume—not the same thing).

Now, you're fleshing out the essential "*who, what, when, where and how*" of ESI.

2. Key Custodians' Documents and Data: Network Shares

Apart from e-mail, custodians generate most work product in the form of **productivity documents** like Microsoft Word documents, Excel spreadsheets, PowerPoint presentations and the like. These may be stored locally, *i.e.*, in a folder on the C: or D: drive of the user's computer (local

storage, see below). More often, corporate custodians store work product in an area reserved to them on a network **file server** and **mapped** to a drive letter on the user's local machine. The user sees a lettered drive indistinguishable from a local drive, except that all data resides on the server, where it can be regularly backed up. This is called the user's **network share** or **file share**.

Just as users have file shares, work groups and departments often have network storage areas that are literally "shared" among multiple users depending upon the access privileges granted to them by the network administrator. These shared areas are, at once, everyone's data and no one's data because it's common for custodians to overlook **group shares** when asked to identify their data repositories. Still, these areas must be assessed and, as potentially relevant, preserved, searched and produced. Group shares may be **hosted** on company servers or "in the cloud," which is to say, in storage space of uncertain geographic location, leased from a service provider and accessed via the Internet. Enterprises employ virtual workspaces called **deal rooms** or **work rooms** where users "meet" and collaborate in cyberspace. Deal rooms have their own storage areas and other features, including message boards and communications tools--they're like Facebook for business.

3. Mobile Devices: Phones, Tablets, IoT

Look around you in any airport, queue, elevator and waiting room or on any street corner. Chances are many of the people you see are looking at the screen of a mobile device. According to the U.S. Center for Disease Control, more than 41% of American households have no landline phone, relying on wireless service alone. For those between the ages of 25 and 29, two-thirds are wireless-only. Per an IDC report sponsored by Facebook, four out of five people start using their smartphones within 15 minutes of waking up and, for most, it's the very first thing they do, ahead of brushing their teeth or answering nature's call.

The Apple App Store supplies over 1.5 million apps accounting for over 100 billion downloads. All of them push, pull or store some data, and many of them surely contain data relevant to litigation. More people access the internet via phones than all other devices combined. Yet, in e-discovery, litigants often turn a blind eye to the content of mobile devices, sometimes rationalizing that whatever is on the phone or tablet must be replicated somewhere else. It's no; and if you're going to make such a claim, you'd best be prepared to back it up with solid metrics (such as by comparing data residing on mobile devices against data secured from other sources routinely collected and processed in e-discovery).

The bottom line is: if you're not including the data on phones and tablets, you're surely missing relevant, unique and often highly probative information.

4. Key Custodians' Documents and Data: Local Storage

Enterprises employ network shares to ensure that work product is backed up on a regular basis; but, despite a company's best efforts to shepherd custodial work product into network shares, users remain bound and determined to store data on local, physical media, including local laptop and desktop hard drives, external hard drives, thumb drives, optical disks, camera media and the like. In turn, custodians employ idiosyncratic organizational schemes or abdicate organization altogether, making their My Documents folder a huge hodgepodge of every document they've ever created or collected.

Though it's expedient to assume that no unique, potentially-responsive information resides in local storage, it's rarely a sensible or defensible assumption absent document efforts to establish that the no-local-storage policy and the local storage reality are one-and-the-same.

5. Social Networking Content

The average Facebook user visits the site 14 times daily and spends 40 minutes looking at Facebook content. That's the average; so, if you haven't visited today, some poor soul has to give Facebook 80 minutes and 28 visits. Perhaps because we believe we are sharing with "friends" or simply because nothing is private anymore, social networking content is replete with astonishingly candid photos, confessions, rants, hate speech, statements against interest and a host of other information that is evidence in the right case. Experts often blog or tweet. Spouses stray on dating and hook up sites like Tinder or Ashley Madison. Corporations receive kudos and complaints via a variety of social portals. If you aren't asking about social networking content, you're missing a lot of elephant!

6. Databases (server, local and cloud)

From Access databases on desktop machines to enterprise databases running multinational operations (think UPS or Amazon.com), databases of every stripe are embedded throughout every company. Other databases are leased or subscribed to from third-parties via the cloud (think Salesforce.com or Westlaw). Databases hold so-called **structured data**, a largely meaningless distinction when one considers that most of data stored within databases is unstructured, and much of what we deem unstructured data, like e-mail, is housed in databases. The key is recognizing that databases exist and must be interrogated to obtain the responsive information they hold.

The initial goal for e-discovery is to identify the databases and learn what they do, who uses them and what types and ranges of data they hold. Then, determine what standard reports they can generate in what formats. If standard reports aren't sufficient to meet the needs in discovery, inquire into the databases **schema** (*i.e.,* its structure) and determine what **query language** the database supports to explore how data can be extracted.

PLUS. Cloud Sources

The Big Six probably deserve to be termed the Big Seven by the escalating importance of the cloud as both a repository for replicated content and a burgeoning source of relevant and unique ESI in its own right. For now, it's Six Plus because it touches so many of the other six and because it's evolving so quickly that it's likely to ultimately differentiate into several distinct sources of unique, discoverable ESI. Whether we consider the shift of corporate applications and IT infrastructure to leased cloud environments like Amazon Web Services and Microsoft Azure or the tendency of individuals to store data in tools like Box, Dropbox, Google Drive, Microsoft OneDrive, Apple's iCloud and others, the cloud must be considered alone as adjunct to the other six sources when seeking to identify and preserve potentially responsive ESI.

The Big Six Plus don't cover the full range of ESI, but they encompass *most* potentially responsive data in *most* cases. A few more thoughts worth nailing to your forehead:

Pitfalls and Sinkholes

Few organizations preserve all legacy data (information no longer needed in day-to-day operations); however, most retain large swaths of legacy data in backups, archives and mothballed systems. Though a party isn't obliged to electronically search or produce all its potentially responsive legacy data when to do so would entail undue burden or cost, courts nonetheless tend to require parties resisting discovery to ascertain what they have and quantify and prove the burden and cost to search and produce it. This is an area where litigants often fail.

A second pitfall is that lawyers too willingly accept "it's gone" when a little wheedling and tenacity would reveal that the information exists and is not even particularly hard to access. It's an area where lawyers must be vigilant because litigation is regarded as a sinkhole by most everyone except the lawyers. Where ESI is concerned, custodians and system administrators assume too much, do too little or simply say whatever will make the lawyers go away.

Lather, Rinse and Repeat

So long as potentially responsive data is properly preserved, it's not necessary or desirable in a high-volume ESI case to seek to secure all potentially relevant data in a single e-discovery foray. It's more effective to divide and conquer. First, collect, examine and produce the most relevant and accessible ESI from what I like to call the über-key custodians; then, use that information to guide subsequent discovery requests. Research from the NIST TREC Legal Track proves that a two-tiered e-discovery effort produces markedly better results when the parties use the information gleaned from the first tier to inform their efforts through the second.

In a bygone era of e-discovery, Thomas Edison warned, "We've stumbled along for a while, trying to run a new civilization in old ways, but we've got to start to make this world over."
A century later, lawyers stumble along, trying to deal with new evidence in old ways. We've got to start to make ourselves over.

The Internet of Things Meets the Four Stages of Attorney E-Grief

I lecture about 50-70 times a year, all over the globe. Of late, my presentations start with an exploration of the Internet of Things (IoT), focused first on my own IoT-enabled life and then addressed to the proliferation of IoT data streams in all our lives. Apart from mobile phones–the apex predators of IoT–discovery from the Internet of Things remains more theoretical than real in civil litigation; and instances of IoT evidence in criminal prosecutions are still rare. That will change dramatically as lawyers come to appreciate that the disparate, detailed data streams generated by a host of mundane and intimate sensors tell a compelling human story.

With every disruptive technology, lawyers go through the Four Stages of Attorney E-Grief: **Denial, Anxiety, Rulemaking** and **Delusion**. I considered a stage called "Prattle," but that hit too close to home.

Lawyers confront disruptive technologies by pretending they don't exist, like those firms who advised clients to steer clear of cloud computing or the multitudes spending millions of client dollars to contort electronic data into printed, paginated documents. When hiding our heads in the sand fails, we worry. In the Anxiety stage, lawyers pen fretful journal articles about the fast-approaching digital Armageddon, exhorting our brethren to "prepare" without practical advice on what to…you know…*do*. We trot out a Parade of Horribles underscoring the risk and burden flowing from our rendering advice born of fear and ignorance. "Preserve it all," we counsel, without the fiduciary disclosure of, "because I don't know what 'it' is or how to figure 'it' out."

Anxiety begets Rulemaking, insuring the law marginalizes new evidence and protects parties and counsel from the consequences of business as usual. New rules buy *time.* The longer lawyers delay compulsory adaption or blunt the consequences of incompetence, the less we are compelled to change.

Rulemaking is a stopgap, of course, because we need evidence to prepare for the tragically tiny potential for trial. Even lawyers can't observe everyone spending *three hours and forty minutes a day* on average tapping on mobile screens without it dawning on us that what we need to discover in litigation may be on those phones and in the Cloud to which they connect.

So, rather than buckle down and re-train, we turn to Delusion. The delusions we cherish most are these:

- *"The big dogs don't know this e-stuff, and they're doing fine."*
- *"I don't need to understand it; I can hire someone."*
- *"I've made it this far; I can keep faking it."*

When it comes to new information tools, we delude ourselves that whatever we've been doing all along MUST be picking up the evidence we'd get from mobile, cloud and IoT. "*The same stuff's in the e-mail and on the network shares, right*?" "*Surely people half our age use e-mail just like we do?*" "*And all those apps, they're just Pokemon Go and Angry Birds.*"

Heaven forbid we seek *actual metrics* to know if our assumptions are borne out by fact. Has *your* firm or client systematically sampled custodians' phones and tablets to determine what unique information they hold that is *not* collected from other sources? No, I didn't think so.
When it comes to the Internet of Things, the bench and bar are still in Denial, and the doors to Anxiety are open; can Rulemaking be far behind?

But there's a better way. We can plan, study and *prepare* to utilize IoT evidence. *It's just data.* And when I say, "it's just data," I mean that it's manageable once you know a little bit about data and databases.
In that vein, here are some thoughts about how the IoT will change civil discovery.

What is the Internet of Things? IoT has been termed an "inextricable mixture of hardware, software, data and service." I define it as the integration and interconnection of sensors and controls in a broad range of Internet-enabled devices, some paired with living things. It's the wristband monitoring physical exertion and sleep. It's the thermostat or appliance controlled by an app. It's Internet cameras, passive transponders in shoes, biometric sensors in watches, refrigerators that track consumption and tags that track keys and wallets. It's lights you control with voice commands to Siri, Alexa or Cortana, as well as police body cams, drones and the vast array of cameras and sensors that surround autonomous vehicles. It's the milk carton that broadcasts its expiration date and the tire that tells you its tread is wearing thin. Too, it's sensors monitoring a variety of commercial, agricultural, industrial and financial processes. People, plants, animals, tools, inventories: everything that can be instrumented will be, and the data from same will feed algorithms that control processes.

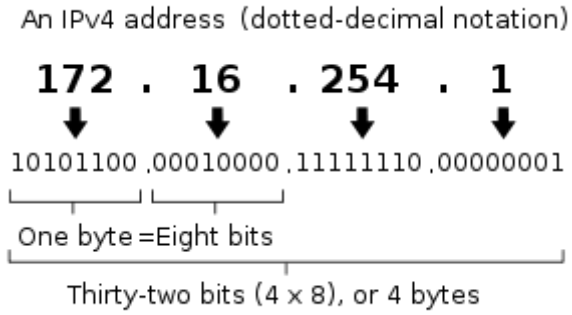
Again, I said *people* will be instrumented. and, in fact, we already are. Few of us are separated from sensor-rich mobile devices that serve as our digital Boswells. This trend will continue to

encompass real time sound- and video recording of law enforcement and all manner of persons whose conduct can prompt loss or liability. IoT will be a routine feature of wearable and implantable medical devices for birth control and pregnancy and to treat, *e.g.,*diabetes, heart disease, obesity, dementia and illness of every stripe. If it moves–a pet, a package, a pest–it will be instrumented and geolocatable.

Increasingly, "fault" will fall on flawed algorithms as much as careless or craven humans. By the same token, the IoT frees humanity from the error and drudgery of humans keeping tabs on a host of conditions, events and movements that machines can track cheaply, relentlessly and precisely.

Depending upon your point-of-view, it's wonderful or terrifying; but, the one thing it will certainly be is *probative, discoverable evidence*. **There has never been a better time to be a trial lawyer in terms of the richness, variety and accuracy of evidence to help us establish the facts.** If you are a lawyer who cares about getting the facts right, rejoice! For lawyers whose trial skills run toward fomenting fear, uncertainty and doubt, the IoT will eventually make your job harder; but, right now, you're going to have a field day attacking the integrity and admissibility of IoT data.

IP Addressing – How Do We Keep Track of So Many Things? The Internet of Things is made possible by the ability to communicate with each thing, either on demand, at intervals or in real time. This requires that everything be assigned a unique Internet Protocol or "IP" address. Early approaches to IP addressing weren't conceived with the IoT in mind; so, we almost exhausted the supply of IP addresses when we used 32-bit numbers to denote them in a system called *IPv4*. A 32-bit (four byte) binary number is customarily expressed as four decimal values separated by periods, so is sometimes called a "dotted quad."



A 32-bit number (2^{32}) is almost 4.2 billion–a big number, but nowhere near big enough to assign even one unique IP address to every person on Earth. In fact, we ran out of IPv4 addresses on February 3, 2011. Do you remember the chaos that ensued on that date?

Probably not, because the world didn't notice. Before we ran out of 32-bit IP addresses, a new system called IPv6 was put in place. The address size grew from 32 to 128 bits (16 bytes), thus providing up to 2^{128} (approximately 340 undecillion) addresses. How big is that? It's cosmic; equal to the MD5 hash address space, or let's just say that *it's way more than the number of atoms on Earth*. That should hold us for a while.

Because the number space is so large, IPv6 addresses tend to be expressed in hexadecimal values (base16), not decimal (base10) or binary (base2) notation. So, an IPv6 address will comprise eight groups of four hexadecimal numbers separated by colons and with leading zeroes suppressed*.,* **2602:304:b1b8:25e0:f86f:84a6:8d82:7b86**. *Hexadecimal notation uses the letters a-f to denote the values 10-15; so, when you see a string of characters composed of 0-9 and a-f, you're likely looking at hex.* Fear not! It's just a number written in a different notation than the decimal notation we're used to seeing.

Try This: You can find out the public IP address of your device by Googling, "*What's my IP address?*"

Communication with IoT Devices: For the IoT to work, we need more than just a way to uniquely address things, we need a way to talk with them, too. Whether by wire, radio, light or sound, IoT must push and pull data. For example, the last pair of shoes I bought came with an embedded passive **near-field communication (NFC)** tag. The tag is "passive" because it can't do anything until it receives power from a nearby electromagnetic field, enabling it to transmit its data (likely a serial number) by modulating the field. Because it needs no power to stand by, the NFC tag can function indefinitely. The tag was designed to combat counterfeit goods; but, the unique signal it supplies could as easily be exploited by a floor mat at Walmart or a multifactor security system at the airport.

Wi-Fi is another common means to communicate with IoT devices, and Wi-Fi connects many of the smart home devices in use today. Wi-Fi demands power-hungry radios, so lends itself to devices connected to AC power or batteries that can be regularly replaced or recharged,

A third conduit is **Bluetooth**, both "classic" Bluetooth, seen in wireless headsets, or the power-sipping version for data called *Bluetooth Smart* or LE (for Low Energy). Bluetooth has a limited range (~30 feet); so, Bluetooth devices typically hand off data to an intermediary device that transmits the IoT data to a database via Ethernet, W-Fi or cellular connections or, in the case of a cellphone, populates an app. The Tile locator tags in my luggage use Bluetooth LE to communicate with any nearby cellphone running the locator app. These phones then broadcast the tag's location to the company selling the tag and ultimately to the locator app on my phone. Because they sip energy, Tile trackers operate for a year using a tiny button battery.

A fourth communication mechanism like Bluetooth LE is **ZigBee**, the IoT technology seen in Phillips Hue smart light bulbs. Up to fifty bulbs speak to a gateway device called a bridge, and the bridge interfaces with the internet, enabling control of the color and brightness of the bulbs via apps on my phone or tablet.

There are other IoT communications mechanisms, including cell systems, optical interfaces like QR codes and even sound-based protocols that allow devices to speak to one-another like R2D2. Typically, IoT ecosystems employ more than one means of communication between the Internet-enabled thing and the end user (*e.g.*, IoT sensor to app via Bluetooth, app to cloud via WiFi or cellular).

Most IoT Data Won't Reside in the Thing: Apart from notable exceptions like smart phones and cars, most data generated by the Internet of Things won't be stored within the interconnected "thing" for longer than the interval required to reliably hand it off to a database in the Cloud or an app and to confirm its receipt and integrity. So, no matter how much the IoT proliferates, the burden of preserving and collecting in e-discovery won't grow commensurate with numbers of devices. The rule of thumb will be, "***preserve and collect from the database, not the device***."

Because the U.S. Internet is not uniformly fast and reliable, most IoT devices will require some local storage capability ("cache memory") that permits the device to buffer the transmission of data to the database or app that serves as repository of the data. Data may also reside on the device to save battery power and bandwidth because a device interrogated at intervals–only awakening now-and-then to send its data–uses less energy than one broadcasting all the time.

The size of the cache or buffer employed will depend on the operating environment, the criticality of the data and the cost and sophistication of the IoT device. The dimensions and form factor of

cache memory will be of little consequence as we are able to fashion capacious storage in ever-denser, low-power packages. Which is to say, storage is not only cheap, it's growing in areal density and, thus, shrinking in physical size. As storage becomes even cheaper, more capacious and compact, devices may end up storing a substantial complement of local historical data as an audit trail and failsafe. Today, we can confidently say, "preserve and collect from the database, not the device;" tomorrow, that could change.

IoT Discovery Practice Tip 1: Ascertain the nature of the 'things' connected and the data they collect; but, *target your efforts to discover relevant data to the databases and applications where the IoT data lands.*

Data or Metadata? Metadata is critical to understanding the context and reliability of IoT data. Unlike a document, spreadsheet, presentation or e-mail message, IoT data streams tend to make no sense apart from the metadata that determines where the data "fits" in an application or database. Sometimes, an IoT data stream will be little more than a bit flag serving to connote a state or status (on/off). Too, IoT streams may include a good deal of data that's not apparent in the user interface of the target application. So, counsel seeking discovery of IoT evidence needs to be conversant in database discovery (which I've addressed in other posts and articles). You will want to discover the composition of the database, its standard reporting capabilities, its data export capabilities and, when these are insufficient, its schema.

Producing parties may have very little insight into the applications they use to collect and mine IoT data. That is, they have no access to a "back end" for phone apps and most cloud implementations. Just because responding parties don't know (or refuse to learn) about the IoT systems they use doesn't mean requesting parties can't do a little homework and craft specific inquiries going to relevant information. Research the IoT device and system. Figure out what you need to know.

Crucially, be specific about the form(s) in which you seek production. Don't just demand "native with metadata" and expect that's going to prevent the other side from supplying junk. If they offer screenshots, are you expecting to run word searches against pictures? If they offer an export, will they supply the field and record data needed to make sense of the information? None of these are hard problems, but they demand a little forethought. *Hint: learn a little about SQLite, the most widely deployed Structured Query Language (SQL) database engine in the world and the dominant means by which IoT data is stored by phone and tablet apps.*

IoT Discovery Practice Tip 2: Research the IoT device and system. Figure out what you need to know, then be specific about the form(s) in which you seek production.

IoT: Curse or Blessing? One difference between the legal industry and the legal profession is that a member of the legal industry just wants to win; a legal professional wants to be on the winning side. Favorable or not, legal professionals want to uncover the true facts underlying a dispute. We want the truth. The inevitable instrumentation of people and products, plants and animals, will be problematic with respect to privacy and cybersecurity; but in litigation, *in fact-finding*, the precision, objectivity and ubiquity of IoT evidence will be a boon to the pursuit of truth.

Custodial Hold: Trust but Verify

A decade or so ago, my dear departed friend and late revered colleague, Browning Marean, presciently observed that the ability to frame and implement a legal hold would prove an essential lawyer skill. Browning understood, as many lawyers are only now coming to appreciate, that "legal hold" is more than just a communique. It's a multipronged, organic process that must be tailored to the needs of the case like a fine suit of clothes. For all the sensible emphasis on use of a repeatable process, the most successful and cost-effective legal holds demonstrate a bespoke character from the practiced hand of an awake, aware and able attorney.

Unfortunately, that deliberate, evolving character is one of the two things that people hate most about legal holds (the other being the cost). They want legal hold to be a checklist, a form letter, a one-size-fits-all tool–all of which have value, but none of which suffice, individually or collectively, to forestall the need for a capable person who understands the ESI environment and is accountable for getting the legal hold right. It's a balancing act; one maximizing the retention of relevant, material, non-duplicative information while minimizing the cost, complexity and business disruption attendant to meeting one's legal responsibilities. Achieving balance means you can't choose one or the other, you need both.

Both!

I'm talking about custodial hold. It's a very hot topic in e-discovery, and for some lawyers and companies, custodial hold is perilously synonymous with legal hold:

Q. "How do you do a legal hold in your organization?"
A. "We tell our people not to delete relevant stuff."

Custodial hold is relying upon the custodians (the creators and holders) of data to preserve it. It makes sense. They're usually the folks best informed about where the data they hold resides and what it signifies. They may be the only ones who can relate the stored information to the actions or decisions at the heart of the litigation. A custodial hold is subjective in nature because custodians choose to preserve based upon their knowledge of the data and the dispute. Absent assurance that custodians can't alter or discard potentially relevant data, you almost always need some measure of custodial hold, to the point that (Ret.) Judge Schira Schiendlin hyperbolically--and erroneously--characterized the failure to implement a written custodial hold as gross negligence *per se*.

"Okay, so a proper legal hold is a custodial hold. Check!"
"Um, sorry no, not by itself. This is where the balancing is needed."

The subjective nature of a custodial legal hold is both its strength and its weakness. It's said that three things can happen when you throw a football, and two of them are bad. The same is true for custodial hold. Custodians may do it well, but some won't bother and some will do it badly. Some won't bother because they will assume it's someone else's responsibility, or they haven't the time or any of a hundred other reasons why people fail to get something done when it's not important or beneficial to them.

Some will do it badly because they don't understand what's going on. Others will do it badly because they understand quite well what's afoot. When you make custodians think about how the information they hold relates to a dispute, you stir them to consider the consequences of others scrutinizing the information they hold. Maybe they start to worry about being blamed for the problem that gave rise to the litigation or maybe they worry about getting in trouble for something that has nothing to do with the litigation but which looms large as an item they don't want discovered. Either way, it's "their" information, and they aren't going to help it hang around if it might look bad for them, for a co-worker or for the company.

Judge Scheindlin touched upon the risk of relying solely on custodial holds in her decision in the NDLON v ICE litigation [*Nat'l Day Laborer Org. Network v. U.S. Immigration & Customs Enforcement Agency*, 10 Civ. 3488 (SAS), 2012 U.S. Dist. Lexis 97863 (S.D.N.Y. July 13, 2012)], leaving lawyers, companies and entire branches of government scratching their heads about whether they can or cannot rely upon custodial holds. "*Hrrrmph*," they sniff, "*We trust our people to do what we tell them to do*." Okay, trust, *but verif*y. It's a phrase no one who was of age when Ronald Reagan was president could ever forget, lifted from an old Russian proverb that Lenin loved, "*doveryai, no proveryai*." I much prefer the incarnation attributed to Finley Peter Dunne**:**
"Trust everyone, but cut the cards."

That means you should backstop custodial holds with objective preservation measures tending to defray the risk of reliance on custodial holds. Put another way, the limitations of custodial holds don't mean you don't use them–*you must use them in almost every case*. It means you don't use them *alone*.

Instead, design your hold around a mature recognition of human frailty. Accept that people will fail to preserve or will destroy data, and recognize that you can often easily guard against such failure by adding a measure of objective preservation to your hold strategy.

Q. Subjective custodial hold or objective systemic hold?
A. You need a measure of both.

This is where the thinking and balancing comes in. You might choose to put a hold on the e-mail and network shares of key custodians from the system/IT side before charging the custodians with preservation. That's essential when the custodians' own conduct is at issue.

Or you might quickly and quietly image the machines of persons whose circumstances present the greatest temptation to reinvent the facts or whose positions are so central to the case that their failure would carry outsize consequences.

Or you might change preservation settings at the mail server level (what used to be called Dumpster settings in older versions of Microsoft Exchange server) to hang onto double deleted messaging for key custodians. Certainly, you need to think of your client as your ally in litigation; but, you'd be a fool not to consider your client an adversary, too. Trust everyone, *but cut the cards*.

Elements of an Effective Legal Hold Notice

It's a lawyer's inclination to distill cases down to a black letter propositions and do *something*: develop a checklist, draft a form or tweak their discovery boilerplate. Modern lawyering is programmatic; necessarily so when non-hourly billing arrangements or insurance companies are involved. Thinking is a liability when carriers cap billable hours. Thus, the matter-specific instructions essential to an effective, efficient litigation hold quickly devolve into boilerplate so broad and meaningless as to serve no purpose but to enable the lawyer to say, "I told you so," if anything falls through the cracks.

How can we insure that the legal hold doesn't become just another formulaic, omnibus notice--so general as to confuse and so broad as to paralyze?

Realistically, we can't. The use of forms is too ingrained. But we can tweak our reliance on forms to avoid the *worst* abuses and produce something that better serves both lawyer and client. Accordingly, this column is not about "best practices." More like, "not *awful* practices." If you must use forms, here are some bespoke touches to consider:

Ask Why, Not Why Not: Lawyers don't eliminate risk, they manage it. Overpreservation saddles your client with a real and immediate cost that must be weighed against the potential for responsive information being lost. Your hold notice goes too far when it compels a client to "preserve everything." That's malfeasance--and the "sanction" is immediate and self-inflicted.

Get Real: It's easy to direct clients to segregate responsive matter, but the work could take them hours or days--*boring* days--even assuming they have adequate search tools and know how to use them. Some clients won't be diligent. Some will be tempted to euthanize compromising material. Naturally, you'll caution them not to deep-six evidence; but, anticipate real human behavior. Might it be safer and cheaper to shelve a complete set of their messages and lock down a copy of the user's network share?

Focus on the fragile first: You can't get in trouble for a botched legal hold if the information doesn't disappear. Fortunately, electronically stored information is tenacious, thanks to cheap, roomy hard drives and routine backup. There's little chance the company's payables or receivables will go to digital heaven. The headaches seem wedded to a handful of dumb mistakes involving e-mail and re-tasked or discarded machines. Manage these risks first.

Key custodians must receive e-mail and messaging hold notices, and IT and HR must receive machine hold notices. Is it *so* hard to put stickers on implicated devices saying, "SUBJECT TO LITIGATION HOLD: DO NOT REIMAGE OR DISCARD?" It's low tech, low cost and fairly idiot proof. Deciding whether to pull backup tapes from rotation entails a unique risk-reward assessment in every case, as does deciding whether it's safe to rely on custodians to segregate and preserve ESI. Remember: "*Trust everyone, but cut the cards.*" If there's a technology in place like journaling that serves as a backstop against sloth, sloppiness and spoliation, a supervised custodial preservation may be fine.

Forms Follow Function: Consider the IT and business units, then tailor your forms to their functions. What's the point directing a salesperson to preserve backup tapes? That's an IT function. Why ask IT to preserve material *about* a certain subject or deal? IT doesn't deal with content. *Couch preservation directives in the terms and roles each recipient understands.* Tailor your notice to each constituency instead of trying to cram it all into one monstrous directive every recipient ignores as meant for someone else.

Get Personal: Add a specific, personal instruction to each form notice--something that demonstrates you've thought about each custodian's unique role, i.e., "*Jane, you were the comptroller when these deals went through, so I trust you have electronic spreadsheets and accounting data pertaining to them, as well as checks and statements*." Personalization forces you to think about the witnesses and evidence, and personalized requests prompt diligent responses.

Don't Sound Like a Lawyer: An effective legal hold prompts action. It tells people *what* they must do, **how** to get it done and *sets a deadline*. If it's a continuing hold duty, make sure everyone understands that. Get to the point in the first paragraph. Gear your detail and language to a

bright 12-year-old. Give relevant examples of sources to be explored and material to be preserved.

Ten Elements of a "Perfect" Legal Hold Notice

1. **Timely**
2. Communicated through an **effective channel**
3. Issued by person(s) with **clout**
4. Sent to all **necessary custodians**
5. Communicates **gravity** and **accountability**
6. Supplies **context** re: claim or litigation
7. Offers clear, practical guidance re: **actions and deadlines**
8. Sensibly scopes **sources and forms**
9. Identifies mechanism and contact for **questions**
10. Incorporates **acknowledgement, follow up and refresh**

Opportunities and Obstacles: E-Discovery from Mobile Devices

Do you live two lives, one online and the other off? Millions lead lives divided between their physical presence in the real world and a deeply felt presence in virtual worlds, where they chat, post, friend, like and lurk. They are constantly checking themselves in and checking others out in cyberspace. In both worlds, they leave evidence behind. They generate evidence in the real world that comes to court as testimony, records and tangible items. Likewise, they generate vast volumes of digital evidence in cyberspace, strewn across modern electronic systems, sites, devices and applications.

Trial lawyers who know how to marshal and manage evidence from the real world are often lost when confronted with cyber evidence. Here, we take an introductory look at discovery from mobile devices.

The Blessing and Curse of ESI

Even if you don't know that data volume is growing at a compound annual rate of 42 percent, you probably sense it. This exponential growth suggests there's little point feeling overwhelmed by data volumes *today* because we are facing volumes *ten times as great in five years*, and *fifty times* as great in ten years.[10] Today is tomorrow's "good old days."

There's going to be a lot more electronic evidence; but, there's still time to *choose* how you deal with it.

A lawyer can curse electronic evidence and imagine he or she is preserving, collecting and requesting all they need without cell phones, the Cloud and all that other 'e-stuff.'

Or, the lawyer can see that electronic evidence is powerful, probative and downright amazing, and embrace it as the best thing to happen to the law since pen and ink. Never in human history have we enjoyed more or more persuasive ways to prove our cases.

Mobile Miracle

According to the U.S. Center for Disease Control, more than 41% of American households have no landline phone. They rely on wireless service alone. For those between the ages of 25 and 29, *two-thirds* are wireless-only. Per an IDC report sponsored by Facebook, four out of five people start using their smartphones within 15 minutes of waking up and for most, it's the very first thing they do, ahead of brushing their teeth or answering nature's call. For those in the lowest economic stratum, mobile phones are the principal and often sole source of Internet connectivity.

[10] Market research firm IDC predicts that digital data will grow at a compound annual growth rate of 42 percent through 2020, attributable to the proliferation of smart phones, tablets, Cloud applications, digital entertainment and the Internet of Things.

In September of 2015, Apple sold 13 million new iPhones in three days. These hold apps drawn from the more than *1.5 million* apps offered in the iOS App Store, compounding the more than *100 billion* times these apps have been downloaded and installed.

Worldwide, phones running the competing Android operating system account for three times as many activations as Apple phones. The United States Supreme Court summed it up handily: "*Today many of the more than 90% of American adults who own cell phones keep on their person a digital record of nearly every aspect of their lives*."[11]

Within this comprehensive digital record lies a cornucopia of probative evidence gathered using a variety of sensors and capabilities. The latest smart phones contain a microphone, fingerprint reader, barometer, accelerometer, compass, gyroscope, three radio systems, near field communications capability, proximity, touch, light and moisture sensors, a high-resolution still and video camera and a global positioning system.[12] As well, users contribute countless texts, email messages, social networking interactions and requests calls for web and app data.

Smart phones serve as a source of the following data:

- SIM card data
- Files
- Wi-Fi history
- Call logs
- Photographs and video
- Contacts
- Geolocation data
- E-mail
- Voicemail
- Chat
- SMS and MMS
- Application data
- Web history
- Calendar
- Bookmarks
- Task lists
- Notes
- Music and rich media

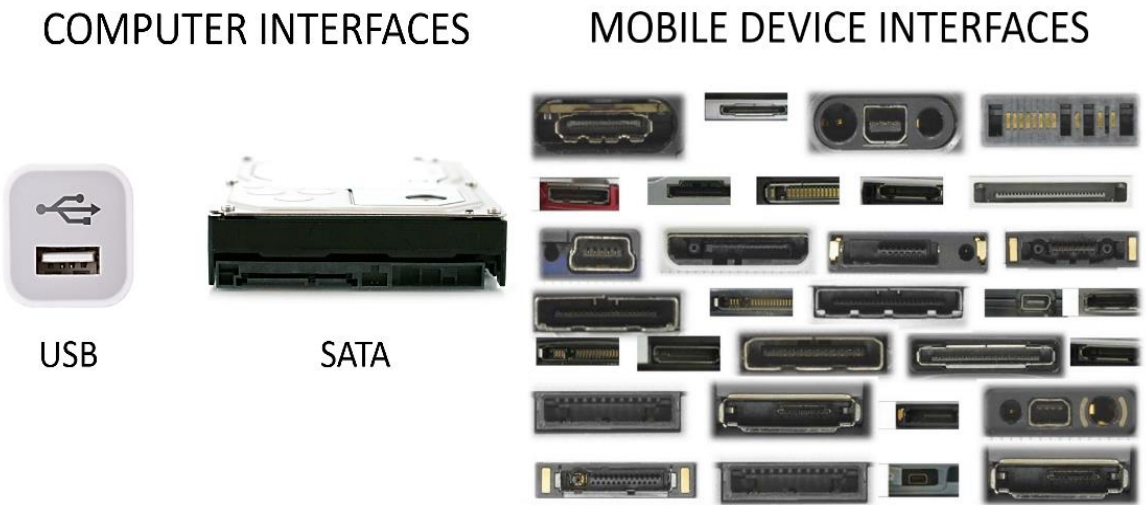
Mustering Mobile

For the last decade, lawyers have been learning to cope with electronic evidence. We *know* how to acquire the contents of hard drives. We *know* about imaging and targeted collection. We've gotten better at culling, filtering and processing PC and server data. After all, most corporate data lives within identical file and messaging systems, and even those scary databases tend to be built on just a handful of well-known platforms. Too, we've got good tools and lots of skilled personnel to call on. **Now, let's talk mobile.**

[11] *Riley v. California*, 573 U.S. ___ (2014).

[12] In support of 911 emergency services, U.S. law requires the GPS locator function when the phone is on.

Let's talk interfaces. We've been acquiring from hard drives for thirty years, using two principal interfaces: PATA and SATA. We've been grabbing data over USB for 17 years, and the USB 1, 2 and 3 interfaces all connect the same way with full backward compatibility. But phones and tablets? The plugs change almost annually (30-pin dock? Lightning? Thunderbolt?). The internal protocols change faster still: try seven generations of iOS in five years.



Let's talk operating systems. Two principal operating systems have ruled the roost in P.C. operating systems for decades: Windows and MacOS. Although the Android and iOS operating systems command huge market shares, there are still dozens of competing proprietary mobile operating systems in the world marketplace.

Let's talk encryption. There is content on phones and tablets (*e.g.,* e-mail messaging) that we cannot acquire at all because of unavoidable encryption. Apple lately claims that it has so woven encryption into its latest products that it couldn't gain access to some content on its products if it

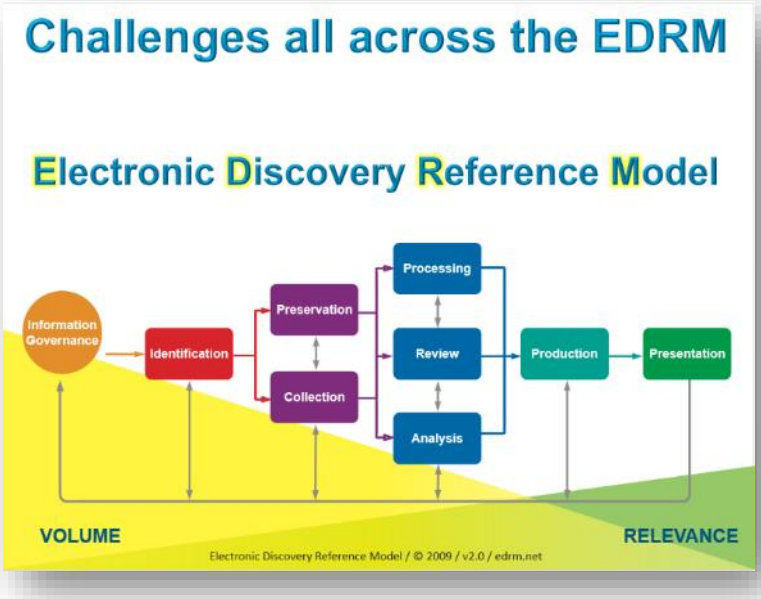
tried. The law enforcement community depends on the hacker community to come up with ways to get evidence from iPhones and iPads. What's wrong with THAT picture?

Let's talk tools. Anyone can move information off a PC. Forensic disk imaging software is free and easy to use. You can buy a write blocker suitable for forensically-sound acquisition for as little as $25.00. But, what have you got that will preserve the contents of an iPhone or iPad? Are you going to synch it with iTunes? Does iTunes grab all you're obliged to preserve? If it did (and it doesn't), what now? How are you going to get that iTunes data into an e-discovery review platform? *There's no app for that.*

Let's talk time. It takes longer to acquire a 64Gb iPhone than it does to acquire a 640Gb hard drive. A fully-loaded iPad may take 48 hours. Moreover, you can acquire several hard drives simultaneously; but, most who own tools to acquire phones and tablets can process just one at a time. It's about as non-scalable a workflow as your worst e-discovery nightmare.

Challenges All Across the EDRM

The Electronic Discovery Reference Model or EDRM is an iconic workflow schematic that depicts the end-to-end e-discovery process. It's a handy context in which to address the ways that mobile devices pose challenges in e-discovery.



Information Governance:

Businesses adopt a BYOD (Bring Your Own Device) model when they allow employees to connect their personal phones and tablets to the corporate network. Securing the ability to

access these devices for e-discovery requires employers obtaining consents in employment agreements.

Identification:
Mobile devices tend to be replaced and upgraded more frequently than laptop and desktop computers; accordingly, it's harder to maintain an up-to-date data map for mobile devices. Mobile devices also do not support remote collection software of the sort that makes it feasible to search other network-connected computer systems. Too, the variety of apps and difficulty navigating the file systems of mobile devices complicates the ability to catalog contents.

Preservation:
It's common for companies and individuals to own mobile devices, yet lack any means by which the contents of the phone or tablet can be duplicated and preserved when the need to do so arises in anticipation of litigation. Even the seemingly simple task of preserving text messages can be daunting to the user who realizes that, *e.g.*, the iPhone offers no easy means to download or print text messages.

Collection: As there are few, if any, secure ways to preserve mobile data *in situ,* preservation of mobile generally entails collection from the device, by a computer forensic expert, and tends to be harder, slower and costlier than collection from PC/server environments.

Processing: The unpacking, ingestion, indexing and volume reduction of electronically stored information on mobile devices is referred to as "Processing," and it's complicated by the fact that so many devices have their own unique operating systems. Moreover, each tends to

Geolocation

Cell phones have always been trackable by virtue of their essential communication with cell tower sites. Moreover, and by law, any phone sold in the U.S. must be capable of precise GPS-style geolocation to support 9-1-1 emergency response services. Your phone broadcasts its location all the time with a precision better than ten meters. Phones are also pinging for Internet service by polling nearby routers for open IP connections and identifying themselves and the routers. You can forget about turning off all this profligate pinging and polling. Anytime your phone can communicate by voice, text or data, you are generating and collecting geolocation data. Anytime. Every time. And when you interrupt that capability that, too, leaves a telling record.

Phones are just the tip of the iceberg. The burgeoning Internet of Things (IoT) is a cornucopia of geolocation data. My Nest thermostat knows if I'm home or away and senses my presence as I walk by. The cameras in my home store my comings and goings in the Cloud for a week at a time. When someone enters, I get a text. My cell phone controls door locks and lighting, all by conversing across the Web. I can instruct Alexa, my Amazon Echo virtual assistant to turn on and off lights, and thanks to a free service called *If This Then That* (IFTTT), I can ask iPhone's Siri to turn lights on and off by *texting* them, at the cost of leaving an indelible record of even that innocuous act. Plus, Siri is now listening *all the time* while my Phone charges, not just when I push the home button and summon her. *"Hey Siri, can you be my alibi?"*

secure data in unique, effective ways, such that encrypted data cannot be processed at all if it is not first decrypted.

Review:
Review of electronic evidence tends to occur in so-called "review platforms," including those with well-known names like Concordance and Relativity. For the most part, these (and message archival and retrieval systems) are not equipped to support ingestion and review of all the types and forms of electronic evidence that can be elicited from modern mobile devices and applications.

Analysis:
Much mobile data--particularly the shorthand messaging data that accounts for so much mobile usage—tend not to be good candidates for advanced analytics tools like Predictive Coding.

Production:
Finally, how will you produce data that's unique to a particular app in such a way that the data can be viewed by those who lack both the device and the app? Much work remains with respect to forms of production best suited to mobile data and how to preserve the integrity, completeness and utility of the data as it moves out of the proprietary phone/app environment and into the realm of more conventional e-discovery tools.

So, What Do I Do?
Though mobile is unlike anything we've faced in e-discovery and there are few affordable tools extant geared to preserving and processing mobile evidence, we are not relieved of the duty to preserve it in anticipation of litigation and produce it when discoverable.

You first hurdle will be persuading the phone's user to part with it intact. Mobile devices are unique in terms of intimacy and dependency. Unlike computers, mobile devices are constant companions, often on our person. The attachment many feel to their mobile phone cannot be overstated. It is simply inconceivable to them to part with their phones for an hour or two, let alone overnight or indefinitely. Many would be unable to contact even their spouse, children or closest friends without access to the data stored on their phones. Their mobile phone number may be the only way they can be contacted in the event of an emergency. Their phones wake them up in the morning, summon their ride to work, buy their morning bagel and serve as an essential link to almost every aspect of their social and business lives. Smart phones have become the other half of their brains.

So, when you advise a mobile user that you must take their devices away from them in order to collect information in discovery, you may be shocked at the level of resistance--even panic or duplicity--that request prompts. You need a plan and a reliable projection as to when the device will be returned. Ideally, you can furnish a substitute device that can be immediately configured to mirror the one taken without unduly altering evidence. Don't forget to obtain the credentials

required to access the device (*e.g.*, PIN code or other passwords). Further, be wary of affording users the opportunity to delete contents or wipe the device by resetting to factory settings.[13] Perhaps due to the intimate relationship users have with their devices, mobile users tend to adopt an even more proprietary and protective mien than computer users.

Four Options for Mobile Preservation

In civil cases, before you do anything with a mobile device, it's good practice to back it up using the native application (e.g., iTunes for iPhones and iPads and preserve the backup). This gives you a path back to the data and a means to provision a substitute device, if needed. Then, you have four options when it comes to preserving data on mobile devices:

1. ***Prove You Don't Have to Do It:*** If you can demonstrate that there is no information on the mobile device that won't be obtained and preserved from another more-accessible source then you may be relieved of the obligation to collect from the device. This was easier in the day when many companies employed Blackberry Enterprise Servers to redirect data to then-ubiquitous Blackberry phones. Today, it's much harder to posit that a mobile device has no unique content. But, if that's your justification to skip retention of mobile data, you should be prepared to prove that anything you'd have grabbed from the phone was obtained from another source.

It's an uphill battle to argue that a mobile device meets the definition of a "not reasonably accessible" source of discoverable data. The contents of mobile devices are readily accessible to users of the devices even if they are hard for others to access and collect.

2. ***Sequester the Device:*** From the standpoint of overall cost of preservation, it may be cheaper and easier to replace the device, put the original in airplane mode (to prevent changes to contents and remote wipes) and sequester it. Be sure to obtain and test credentials permitting access to the contents before sequestration.

3. ***Search for Software Solutions:*** Depending upon the nature of the information that must be preserved, it may be feasible to obtain applications designed to pull and preserve specific contents. For example, if you only need to preserve messaging, there are applications geared to that purpose, such as Decipher TextMessage or Ecamm PhoneView. Before using unknown software, assess what it's limitations may be in terms of the potential for altering metadata values or leaving information behind.

4. ***Get the credentials, Hire a Pro and Image It:*** Though technicians with the training and experience to forensically image phones are scarce and may be pricey, it remains the most defensible approach to preservation. Forensic examiners expert in mobile acquisition will have invested in specialized tools like Cellebrite UFED, Micro Systemation XRY, Lantern or

[13] Contents can often be erased by users entering the wrong password repeatedly, and it's not uncommon to see users making this "mistake" on the eve of being required to surrender their phones.

Oxygen Forensic Suite. Forensic imaging exploits three levels of access to the contents of mobile devices referred to as Physical, Logical and File System access. Though a physical level image is the most complete, it is also the slowest and hardest to obtain in that the device may need to be "rooted" or "jailbroken" to secure access to data stored on the physical media. Talk with the examiner about the approaches best suited to the device and matter and, again, be sure to get the user's credentials (i.e., PIN and passwords) and supply them to the examiner. Encryption schemes employed by the devices increasingly serve to frustrate use of the most complete imaging techniques. In those case, some data is simply unobtainable by any current forensic imaging methodology.

Custodian-Directed Preservation of iPhone Content: Simple. Scalable. Proportional.

Craig Ball

© 2017

This article and its appendices make the case for routine, scalable preservation of potentially-relevant iPhone and iPad data by requiring custodians back up their devices using iTunes (a free Apple program that runs on PCs and Macs), then compress and encrypt the backup for *in situ* preservation or collection.

The Need

In other settings than this seminar, most of you would likely be reading this on your iPhone or iPad. If not, it's a virtual certainty that your phone or tablet are nearby. Few of us separate from our mobile devices for more than minutes a day. On average, cell users spend four hours a day looking at that little screen. On a*verage.* If your usage is much less, someone else's is much more.

It took 30 years for e-mail to displace paper as our primary target in discovery. It's taken barely 10 for mobile data, especially texts, to unseat e-mail as the Holy Grail of probative electronic evidence. *Mobile is where evidence lives now*; yet in most cases, mobile data remains "off the table" in discovery. It's infrequently preserved, searched or produced.

> ***No one can say that mobile data isn't likely to be relevant, unique and material.*** Today, the most candid communications aren't e-mail, they're text messages. Mobile devices are our principal conduit to online information, eclipsing use of laptops and desktops. Texts and app data reside primarily and *exclusively* on mobile devices.
>
> ***No one can say that mobile data isn't reasonably accessible.*** We use phones continuously, for everything from games to gossip to geolocation. Texts are durable (the default setting on an iPhone is to keep texts "Forever"). Mobile content easily replicates as data backed up and synched to laptops, desktops and online repositories like iCloud. The mobile preservation burden pales compared to that we take for granted in the preservation of potentially-relevant ESI on servers and personal computers.
>
> ***Modest Burden.*** That's what this article is about. My goal is that you see for yourself that the preservation burden is minimal when it comes to preserving the most common and relevant mobile data. I'll go so far as to say that *the burden of preserving mobile device content, even at an enterprise scale, is less than that of preserving a comparable volume of data on laptop or desktop computers.* Too, the workflows are as *defensible and auditable as any we accept as reasonable in meeting other ESI preservation duties.*

Three Principles

The following three principles underscore the need for efficient, defensible preservation of relevant mobile content:

- When mobile data may be unique and relevant, it should be preserved in anticipation of litigation. This principle is especially compelling when the preservation burden is trivial (as by use of the backup technique described below). You can demonstrate the absence of relevant data by, *e.g.*, sampling the contents of devices; but standing alone, a policy barring the use of a device to store relevant data is *not* sufficient proof that such device has not, in fact, been used to store data. Too often, practice belies policy, particularly for messaging
- Mobile preservation should be a customary feature of a defensible litigation hold; but absent issues of spoliation, few matters warrant the added cost of mobile preservation by forensics experts or the burden and disruption of separating users from mobile devices.
- Legitimate concerns respecting personal privacy and privilege do *not* justify a failure to preserve relevant mobile data, although they *will* dictate *how* data is protected, processed, searched, reviewed and produced.

Three Provisos:
As you undertake the exemplar workflow in the exercises and ponder how you might adapt it to your needs, consider the following three provisos:

- *The method demonstrated here is but one simple, scalable and defensible method* to preserve iPhone content. It's not necessarily the only way or the optimum way.
- *Preservation isn't production.* Lawyers' abilities to search, review and produce mobile content in utile and complete forms hasn't kept pace with the obligation to do so, or on a par with other responsive sources of ESI. This article and these exercises are about routine *preservation*; they don't address downstream processes and production except insofar as ensuring that the information preserved remains readily amenable to all methods of search, review and production in e-discovery.
- *Please challenge, but don't dismiss.* The duty to preserve is real and immediate; but there's room for honest debate about what depth and exactitude of mobile preservation is warranted case to case. In weighing any method, compare it to the alternative. *If you reject a preservation method because you deem it flawed, is the alternative a superior method or nothing at all?* "None" is rarely the proper choice when it comes to mobile evidence. Preserving "most" is better than "none," but, considerations of risk may dictate that one preserve "all" over "most." In turn, considerations of proportionality may elevate "most" over "all." It's sensible to ask, "*Is the incremental cost of forensic-level preservation by experts justified by relevant and unique content? If not, might 'good' be good enough?"*

Defensibility
Ignoring mobile evidence isn't the path taken by competent, ethical attorneys. We must employ methods of preservation that aren't unduly costly or burdensome yet pose little risk that a judge will find the methods unreasonable. The essence of defensibility is the ability to show that an action was prudent per a good faith assessment of what was known, or in the exercise of diligence should have been known, when the action occurred. If mobile content required to be preserved is lost, the Court will ask: "*Was the preservation method employed reasonably calculated to guard*

against loss or corruption of potentially-relevant mobile data?" This will entail consideration of the method, its deployment and its oversight. These considerations are addressed below in Audit and Verification.

Custodian-Directed Preservation

The predominant approach to preservation in e-discovery entails use of a legal hold directive instructing custodians to act to preserve potentially-relevant ESI. This is custodian-directed preservation, and it's been justifiably criticized for its many flaws, among them that:

- It requires custodians to make judgments concerning relevance, materiality and privilege;
- It obliges custodians to complete tasks, like lexical search, without proper tools or training;
- It demands effort without affording custodians the time, resources and guidance to succeed; and
- It doesn't deter custodians who seek to destroy or change inculpatory or embarassing data.

Custodian-directed preservation is key to a defensible legal hold process; however, it's just part of a proper process and is best paired with other efforts, like IT-initiated holds, that defray its shortcomings.

So, if custodian-directed preservation is problematic, why put custodians in charge of preserving their own devices instead of handing the devices over to digital forensics experts for imaging? Isn't that inviting the fox to guard the henhouse?

The signal challenge to preserving mobile devices is persuding custodians to part with them. By empowering custodians to preserve the data themselves, custodians need never surrender custody of their devices. Accordingly, users are less threatended by the process and less inclined to fight or subvert it. Backing up an iPhone is simple and quick; and crucially, the process affords the custodian neither the need nor the practical ability to select or omit content. Compare that to tasking a custodian to collect e-mail or documents, where it's easy to overlook or deliberately omit material with little chance of detection.

The advantages of custodian-directed preservation of mobile devices by backup are:

- Custodians need not make judgments concerning relevance, materiality and privilege;
- Custodians need not run searches or require no special tools or training;
- The backup process is speedy, easy to autheticate and lets custodians retain their phone;
- It's difficult to omit content from a backup and, once created, backups are hard to alter.

Scalability and Proportionality

Scalability describes the ability of a system or process to handle a growing number of tasks or a larger volume of data. It's a crucial consideration in all phases of e-discovery, but particularly challenging when dealing with mobile data. Historically, preserving mobile data was a one-off task: seldom undertaken and typically for only a handful of devices. Preserving the contents of a single phone by engaging a digital forensics specialist to image the device was the norm, and

though costly, the obligation rarely had to scale to dozens or hundreds of far-flung devices. For one or two phones, you could do it in a day or two for, say, a thousand dollars.

Now, imagine you must preserve the texts and call data from the mobile devices of sales reps, one each in all fifty United States, the District of Columbia, Puerto Rico and Guam. Fifty-three iPhones. What are your options? Let's compare:

[My cost projections are educated guesses and—*please*--not an invitation for enterprising readers to post comments extolling their company's superior pricing.]

1. **Instruct all custodians to overnight courier their phones to your trusty forensic examiner.** In turn, the examiner will image each device and overnight each back when the work is complete.
    - Cost: Under $30,000.00 without rush or overtime fees.
    - Timing: Assuming no glitches, most users will have their phones back within about four to five business days, as few labs possess the equipment permitting them to image more than a couple of phones simultaneously. As well, 53 packages must be correctly processed, logged as evidence, re-packaged and returned to the correct custodian.
        - How many businesses can idle their national sales staff for four to five days?
        - How many reps will be willing to hand over their phones for four to five days?

2. **Send your trusty forensic examiner to 53 locations to image each phone.**
    - Cost: $50-$60,000.00 in professional time; add a comparable sum for travel costs.
    - Timing: A month or more. It's a 19-hour flight to Guam, 11 hours to Hawaii and nine to Alaska. Equipment must travel, and each custodian must part with their phone for the better part of a day.
        - Caveat: Some states license forensic examiners. It may not be legal for an unlicensed examiner to come into the jurisdiction to acquire the image.
3. **Engage 53 local, licensed (as required) examiners to image each device.**
    - Cost: $35-$50,000.00 in examiner fees, plus the professional time required to locate, vet and contract with each examiner. There will also be travel time assessed, albeit with little airfare and hotel expense.
    - Timing: Weeks, at best. Fifty-three data sets from as many senders must be correctly packaged and returned to you, and each custodian must still part with their phone.

All three options implicate proportionality concerns. All are expensive, disruptive and time-consuming. Accordingly, many litigants opt not to preserve the content of mobile devices, claiming phones don't hold relevant data in the face of compelling contrary evidence and a dearth of supportive metrics.

Let's compare the custodian-directed option:

4. **Direct and instruct 53 custodians to back up their devices, collecting the data as desired.**
    - Cost: None, insofar as discrete expenditures. Of course, discovery is never "free" because time costs money. The expense to notify the custodians and follow up on compliance is attendant to all methods, and administrative costs don't count against any. Expenses, if any, for the custodian-directed method hinge on whether you preserve backup data *in situ,* collect it via network transfer or ship it on physical media. Each method demands *some* effort of each custodian, whether that entails coordinating with an examiner to tender and retrieve a device or connecting the device to a computer for an iTunes backup. The latter is far easier and least disruptive.
    - Timing: A day or two. Sure, some custodians may be on vacation, and some may miss or ignore the request; however, such risks afflict every method. Only the custodian-directed method makes it possible to preserve the many, widespread devices in hours, not days or weeks. The custodian need only get to a computer with the device, whereas a forensic examiner must get to the device or the device must get to the examiner.

The custodian-directed method scales easily for phones and tablets. Custodians need never part with their devices, so there is no business interruption. It's speedy. It requires no special tools, cabling or software and no technical expertise. Moreover, the process poses almost no risk of loss or alteration of the relevant data and is unlikely to prompt custodians to game the process. There are no operating system compatibility issues. Remote screen-sharing handily facilitates any desired oversight and audit. In short, cost and burden are so trivial that relevance alone should be the pole star in deciding whether to preserve mobile content.

For an example of mobile backup instructions that might be directed to a custodian, look at Appendix 4. What we will ask of our clients will serve as the step-by-step of our exercises today.

Audit and Verification

Recently, my friend and fellow forensic examiner, Scott Moulton, visited New Orleans. Over beignets and café au lait in the French Quarter, I made the case for the preservation methodology described here. Scott's a brilliant examiner and hard-eyed skeptic. I wanted him to kick the tires and find flaws.

At first, Scott wouldn't take off his forensic examiner hat and don an e-discovery thinking cap. He extolled the benefits of hiring a qualified forensic examiner and the specialized forensics tools we use to dig for esoteric artifacts. "Hire me. Hire *you*!" I liked the sound of that, and Scott liked the idea of motorcycling through the lower 48 and D.C. gathering digital evidence like some two-wheeled remake of Cannonball Run Meets Revenge of the Nerds.

Still, Scott conceded that in the context of e-discovery, there really isn't much iPhone data preserved using a costly forensics tool versus preservation using iTunes. Our training and tool sets don't add much when preserving mobile data for discovery.

Once Scott warmed to the methodology for its speed and low cost, he questioned how the process could be quality checked for integrity. "What if the backup was interrupted or failed," he asked, "How would we know?"

It's a good point. Most experienced forensic examiners have found an image acquired in the field to be incomplete or unusable back in the lab. Thankfully, it's rare; but, sooner or later, it happens. *There are always gremlins*. Custodial-initiated preservation benefits from oversight and audit, if only because the risk of gremlins *feels* greater when custodians are in charge.

If iTunes successfully completes a backup, the backup event can be verified several ways:

1. In iTunes (with the device connected), by looking at the device summary for the attached device and noting the latest backups. ***Fig. 1, right top.***
2. In iTunes (with or without the device connected), under ***Edit>Preferences>Devices***. **Fig.2, right.** This lists the backed-up devices by name with time of backup. Hovering the mouse pointer over a listing will bring up further details about the device backed up (model, software version and build, serial number, phone number, IMEI and MEID). ***Fig. 3 right bottom.***
3. By confirming the date and time values for the folder containing the latest backup (stored by default in: C:\Users*user's account name*\AppData\Roaming\Apple Computer\MobileSync\Backup\). ***Fig. 4 below.***

Figure 12

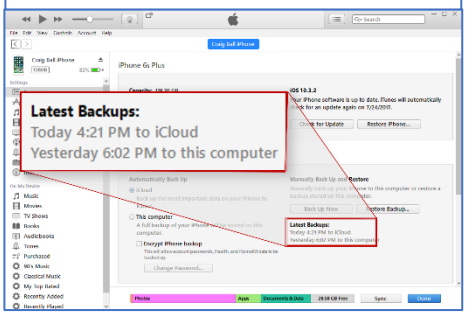


Figure 1

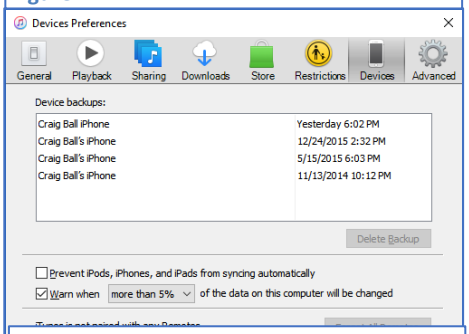


Figure 3

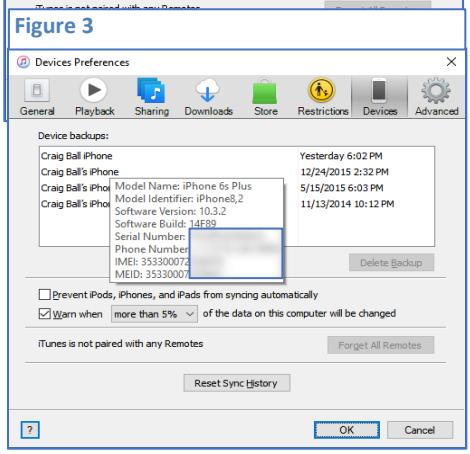
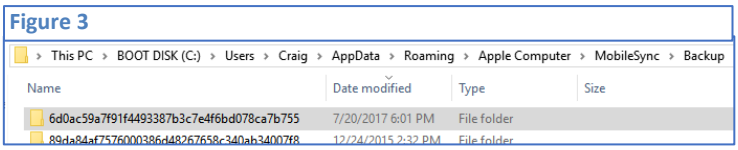


Figure 3



There are several sensible ways to verify and audit a custodian-directed preservation effort. Tailor the method to the potential for failure and the willingness of a sponsoring witness to vouch for the integrity of the process if

challenged. A proper audit trail could be as simple as the custodian supplying a screenshot (ALT-Print Screen) of the details panel for the latest backup (as seen when one hovers over backups in Devices Preferences, as described above and seen in **Fig. 3**). A second approach is the use of cryptographic hashing, and a third, the use of remote screen-sharing and -recording software to permit step-by-step oversight of the work by the sponsoring witness or designee. Also, device backup sets may be sampled and tested for accuracy and completeness. It's important to do *something* to audit and verify the effort; but proportionality suggests you needn't do *everything*.

What You Won't Get with a Backup

An iPhone backup won't preserve e-mail stored on the iPhone. This is by design. Per Apple, an unencrypted iTunes backup also won't include:

- Content from the iTunes and App Stores, or PDFs downloaded directly to iBooks
- Content synced from iTunes, like imported MP3s or CDs, videos, books, and photos
- Photos already stored in the cloud, like My Photo Stream, and iCloud Photo Library
- Touch ID settings
- Apple Pay information and settings
- Activity, Health and Keychain data

Why not use iCloud?

At some point, you *will* use iCloud for preservation; but currently, an iCloud backup is not equal to an iTunes backup. It preserves less data, and byte-for-byte, it takes more time to create than an iTunes backup. Additionally, iCloud encrypts all backups, making them a future challenge for processing and search should a user's credentials be unavailable.

Why an Unencrypted Backup?

This is a compromise. On the one hand, an encrypted iTunes backup preserves more information than an unencrypted backup. Apple won't store passwords, website history, Health data and Wi-Fi settings in an unencrypted backup. On the other hand, many tools can't process the contents of an encrypted backup, even with user credentials, and no tool can process an encrypted backup without credentials. Accordingly, we collect the data as an unencrypted backup, obviating the need for user credentials. To protect the data and add efficiency, we compress and optionally encrypt the backup set using credentials chosen for the legal hold project, not each user's credentials.

Encryption

Encryption is a crucial security tool to protect client data collected in e-discovery, but it's better to manage credentials systematically for the e-discovery project instead of according to each custodian's preference. However, because mobile devices employ layers of encryption, obtaining an unencrypted backup won't serve to unlock encrypted application data. You must obtain and preserve the user's access credentials for that data.

Many users employ the same password for multiple sources, so requiring a user to disclose credentials serves to compromise the security of sources not collected. Assuage concerns by detailing steps taken to protect users' credentials. An unlocked spreadsheet with each custodian's password(s) may be a convenience for the legal team, but it's a cybersecurity nightmare. Keep that in mind when furnishing credentials to service providers, and be sure your vendors are handling passwords securely.

Why Compress the Backup Data?
One reason we compress the data to a Zip file is to make it easier to copy to new media. Smaller data volumes move faster. However, depending upon the composition of the data backed up, the compressed Zip file may be much smaller or hardly smaller at all. My backup set compressed by just 2%. Much of the data on my iPhone consists of JPEG photos already in a compressed format, and it's hard to compress data that's already compressed as there's little 'space' to squeeze out by further compression.

So why bother compressing the backup files?

Two reasons. First, placing the preserved data in a Zip file guards against overwriting the data by a subsequent backup of the device. Second, depending upon the Zip tool employed to compress the file, the Zip process affords a means to securely encrypt the data without having to install an encryption tool. Every Windows machine can create compressed and encrypted Zip files, so will every Mac running OS X.

A New Paradigm in Mobile Device Preservation: Recently, I wrote a post (Appendix 1, next page) where I stated, "Today, if you fail to advise clients to preserve relevant and unique mobile data when under a preservation duty, you're committing malpractice." I'll go further and add that competent counsel not only tells clients what they must do but must also help clients identify practical, proportional ways to meet mobile preservation obligations. This article lays out one scalable, defensible and cost-effective way to preserve iPhone and iPad content. The purpose is to debunk claims that mobile preservation is unduly burdensome, expensive and disruptive. Practical approaches are out there for other phones and devices, too. It's our duty to insure our clients know about them and use them.

Appendix 1: Ball in Your Court, April 18, 2017

A New Paradigm in Mobile Device Preservation

Tuesday, April 18, 2017

Can anyone doubt the changes wrought by the modern "smart" cellphone? My new home sits at the corner of one-way streets in New Orleans, my porch a few feet from motorists. At my former NOLA home, my porch faced cars stopped for a street light. From my vantage points, I saw drivers looking at their phones, some so engrossed they failed to move when they could. Phones impact how traffic progresses through controlled intersections in every community. We are slow-moving zombies in cars.

Distracted driving has eclipsed speeding and drunken driving as the leading cause of motor vehicle collisions. Walking into fixed objects while texting is reportedly the most common reason young people visit emergency rooms today. Instances of "distracted walking" injury have doubled every year since 2006. Doing the math, 250 ER visits in 2006 are over half a million ER visits today, *because we walk into poles, doors and parked cars while texting*.

Look around you. CAUTION: *This will entail looking up from your phone.* How many are using their phones? At a concert, how many are experiencing it through the lens of their cell phone cameras? How many selfies? How many texts? How many apps?

Lately I've begun asking CLE attendees how many are never more than an arm's length from their phones 24/7. A majority raise their hands. These are tech-wary lawyers, and most are Boomers, not Millennials.

Smart phones have changed us. Litigants are at a turning point in meeting e-discovery duties, and lawyers ignore this sea change at peril. The "legal industry" has chosen self-deception when it comes to mobile devices. It's a lie in line with corporate bottom lines, and it once found support in the e-discovery case law and rules of procedure. But, no more.

Today, if you fail to advise clients to preserve relevant and unique mobile data when under a preservation duty, you're committing malpractice.

Yes, I used the "M" word, and not lightly.

I wouldn't have called it malpractice a few years ago. But two things have changed, and we can't hide our heads in the sand. These are paradigm shifts.

The two things are, first, the data on phones and tablets is *not* just a copy of information held elsewhere. It's unique, and often relevant, probative evidence. Second, the locking down of phone content has driven the preservation of mobile content from the esoteric realm of computer forensics to the readily accessible world of apps and backups. These developments mean that, notwithstanding the outdated rationales lawyers trot out for ignoring mobile, the time has come to accept that mobile is routinely within the scope of preservation obligations.

Too, lawyers need to stop treating mobile devices like biohazards and realize that there are easy, low-cost ways to preserve relevant mobile content *without taking phones away from users.* Because it's easy and cheap to preserve it, mobile content is accessible, and its preservation, when potentially relevant, is proportionate under the Rules.

That's a strong stand, and one some will angrily reject. I get where they're coming from. It was wonderful to be able to ignore mobile in e-discovery. Mobile was a black hole. It wasn't just that you had to hire technical experts to use expensive tools to preserve the contents of phones, it was like pulling teeth to get users to let loose of their devices for the hours or days it took to collect them. Even when they did hand them over, more than a few users claimed to have entered the wrong password too many times and "accidentally" wiped the contents of the phone. "Oops. My bad."

If that never happened to one of your clients, it may be because your client wasn't preserving phone data, indulging in the assumption that whatever they'd glean from the phone would be collected elsewhere. They deemed mobile redundant.

Lecturing about mobile and IoT in D.C. last year, an associate from a megafirm confided to me that his firm routinely advised all its litigation clients that they need not preserve the content of mobile devices because "all the relevant content would be duplicated on the servers." I asked if the firm had ever tested its advice against the relevant data to determine if there was truth in what they were telling clients. He admitted they never had, and offered that they'd never do so. The firm didn't want to know the facts because the fairy tale of "replicated elsewhere" was what the client wanted to hear.

Is it a fairy tale? I have my own views based on my own comparisons of mobile content versus other collected sources. What I see demonstrates that the claim that what's relevant on a phone is preserved elsewhere is a whopper. I am routinely finding examples of relevant data stored on mobile devices that is not found among the other sources of data routinely preserved in e-discovery. The replication fairy tale is a relic of a bygone era of Blackberry Enterprise Servers and phones with lower IQs than the brilliant devices now our constant companions and confidantes.

But, I'm not asking you (or courts) to take my word for it. *Test it yourself.*

If you're going to tell the tale, then get some metrics to make it plausible. Use sampling. Process the phones of a few key custodians and compare all the potentially relevant items collected from their mobile devices against the other sources collected for the sampled custodians. What's the differential? Is the unique evidence from the mobile device probative and material?

I've done that, and so I know replication is a fairy tale. If you want to claim it's true *for your client in your case*, how about putting some facts to work? Bear the burden of proof, or start bearing the onus of truth. When you have the facts, you'll have to let loose of the legend and preserve relevant mobile content.

That's the bad news for those who would prefer to ignore mobile. But take heart, as that will seem like great news compared to the next development. Yet, there's a silver lining. Mobile preservation is now quick, cheap and easy.

A few years ago, mobile phones shared some of the characteristics of personal computers in that they held latent data that could be recovered using specialized tools sold for princely sums by a couple of shadowy tech companies. So, the preservation of mobile devices slipped into the shadows, too. Phones and tablets were *forensic* evidence, and only forensic examiners could collect their contents.

Although users used mobile devices all day, the contents of mobile devices were dubbed "not reasonably accessible." It was too costly and burdensome to preserve a phone. Good thing, because users were holding onto their phones tighter than Willie Nelson clutches a bong. Users protested, *"the mobile phone is the only way the kids' school can reach me in an emergency, and I can't use another phone because everyone texts now, and WHO REMEMBERS PHONE NUMBERS ANYMORE?"*

So, the next altered paradigm: In e-discovery today, the forensic-level preservation of phones—the sort geared to deleted content and forensic artifacts—is a fool's errand. As the public learned from the FBI's tussle with Apple over unlocking the iPhones of the San Bernardino terrorists, modern smart phones are locked down hard. Content is encrypted and even the keys to access the encrypted content are themselves encrypted. Phone forensics isn't what it used to be. More and more, we can't get to that cornucopia of recoverable forensically-significant data.

At the same time, it's quick, easy and free for a user to generate a full, unencrypted backup of a phone without surrendering possession. The user can even place the backup in a designated location for safekeeping by counsel or IT. Will this be a "forensic image" of the contents? Strictly speaking, no. But as the phone manufacturers tighten their security, "forensic imaging" becomes less and less likely to yield up content of the sort encompassed by a routine e-discovery preservation obligation. Not every case is a job for C.S.I.—and I say that as someone who makes a living through computer forensics.

I grant that a full unencrypted backup of an iPhone isn't going to encompass all the data that might be gleaned by a pull-out-all-stops forensic preservation of the phone. But so what? As my corporate colleagues love to say, "*the standard for ESI preservation isn't perfect.*" I always agree adding, "*but it isn't lousy either.*" Preserving by backup isn't perfect; but, it isn't lousy. I've come to regard it as sufficient and proportionate. It's good *enough*, and in most cases, darn good.

I think this is important. It's a game changer for what most litigants are doing today. In a view I hope will come to be shared by all who think it through—preservation of mobile device content must become a standard component of a competent preservation effort except where the mobile content can be shown to be beyond scope. Mobile content has become so relevant and unique, and the ability to preserve it so undemanding, that the standard must be preservation.

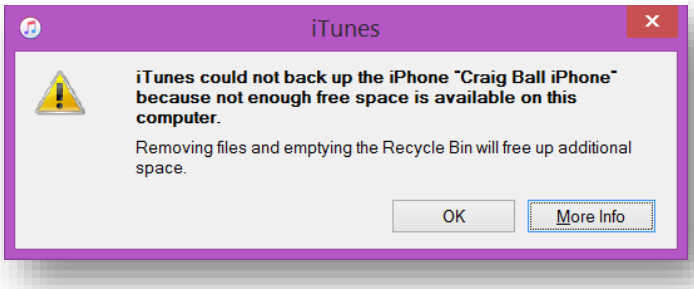
In a future post, I'll lay out the steps to make mobile preservation part of routine preservation workflows and facilitate custodial-initiated preservation of mobile device content. I'll also talk about why it's defensible, proportionate and amenable to targeted processing when it's time to move from preservation to production.

Appendix 2: Redirecting the iPhone Backup Files to External Media

Q. What if I don't have enough space on my Windows C: drive to hold the backup?

A. Smart phones have evolved to capture a *lot* of data. Ten years ago, you couldn't store more than 8GB of data on an iPhone. Today, they store up to 256GB, 32 times as much. So, an iTunes backup may fail to complete because not enough free space is available on the computer performing the backup. You may be able to resolve this by, *e.g.,* emptying the Recycle Bin; but, if you simply can't garner enough space on the boot drive where Apple stores the backup by default, you may need to "trick" your Windows machine into storing the backup on a sufficiently-sized alternate or external storage medium.

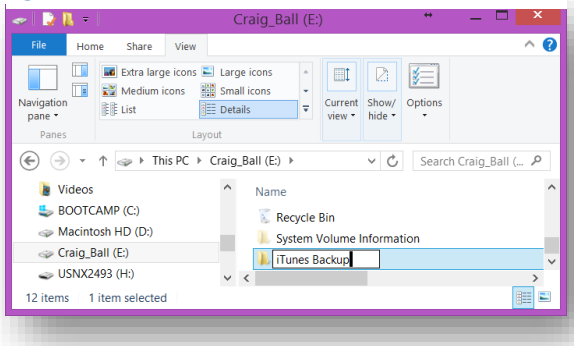
Figure 1



How to Redirect an iTunes Backup Location in Windows

Step 1. Create a new backup folder on a disk with sufficient space to create your backup (roughly. twice the capacity of your iPhone is ample). In Figure 2, I've created the new iTunes backup location on my E: drive (a 250GB thumb drive) and named it "iTunes_Backup." You can name yours anything you'd like.

Figure 2



Step 2. Rename the current iTunes backup folder
Using Windows File Explorer, navigate to your current iTunes "Backup" folder. By default, it's:
C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync
where "*your account name*" is the name of your Window's User ID on the machine.

Right click on the "Backup" folder and rename it. I called mine "Old_Backup;" but here again, call it whatever you like.

3. Redirect the Old Backup Folder Address to the New One
Here, it gets a tad tricky because you must use a Windows Command line interface. Make it easier on yourself by writing down the full paths to the old and new backup folders. *You must get both right for the redirection to work.*

The old one *should* be:

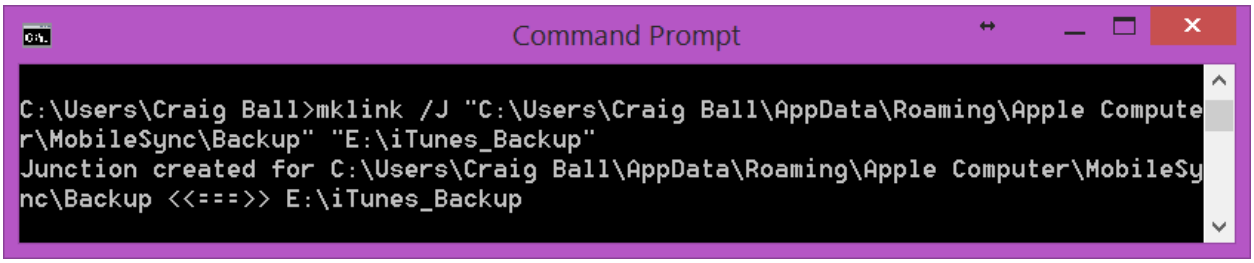
C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync\Backup

The new path is on whatever storage medium you chose, using whatever path and folder name you gave it in step 1, above (mine was "**E:\iTunes_Backup**").

Open a command prompt window by pressing the Windows key on your keyboard, then typing CMD or by pressing the Shift key on your keyboard while right clicking in an open area of any folder, then selecting "Y and selecting "*Open command window here*" from the menu.

At the command line, carefully type the following command:
mklink /J "*path to old backup location*" "*path to new backup location*"
where you substitute the old and new paths you've written down. *Be sure to enclose each path in quotation makes, as shown.*

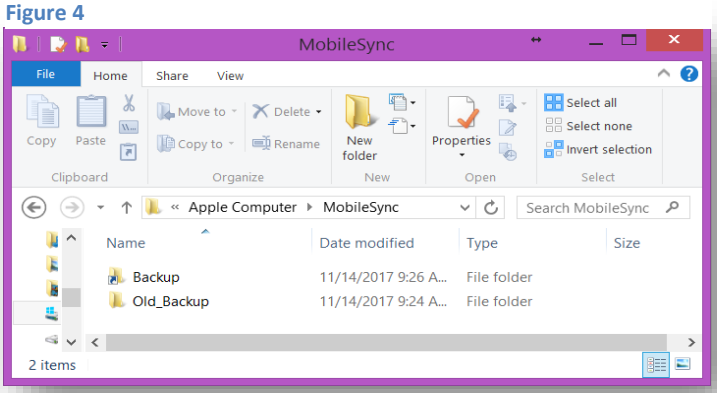
On my machine, the command and response looked like Figure 3:
Figure 3



The "junction created" refers to a Windows symbolic link, a **Directory Junction,** that will serve to redirect any actions that would have been performed on the old backup folder to be redirected to the new one.

What Note: The **mklink /J** command creates a symbolic link to the new folder from the old one. It's like creating a shortcut of D:\Backup from the original MobileSync\Backup folder. You can test the effect by double-clicking on the Backup folder in MobileSync. It will take you to the new Backup folder.

Figure 4



Now, if you look in your MobileSync folder:
(**C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync**

you will see a folder shortcut named “Backup” alongside your renamed former backup folder as mine appears in Figure 4.

4. Move your Old Backups

If desired, you can move your old iTunes backup files from your old renamed Backup folder to your new backup folder and delete them from the old location.

5. Run your iTunes Backup

Be sure the media you selected to hold the relocated backup is attached. Now, run your iTunes backup as usual and, if all is working, the backup will be created where you created the new backup folder.

Appendix 3: iPhone Backup Data Extractors

There are quite a few applications marketed as tools to extract data from iPhone backups. Though a handful are free, the ones that look promising tend to run about $40-50 single user/single computer license, a trivial sum if it obviates the need to hire a forensic examiner or technician to extract texts, call logs, photos, browsing history, contacts and the like.

My brief exploration of these applications underscored that it's hard to discern which tools work and which don't in advance of buying a license. Happily, some promise "free evaluation copies" and "money back guarantees."

iMazing: $39.99-$279.96
The only tool that worked remarkably well (which is to say, "worked *at all*") in my testing was iMazing. It allowed me to explore its capabilities and confirm that it had no trouble opening and interpreting my iPhone backup before requiring I buy a license. Licenses started at $39.99, but I paid ten dollars more for a license that runs on two machines.

iMazing seemed to have no difficulty getting the content I'd most likely need in e-discovery and, crucially, was adept at exporting the content in utile formats, including delimited CSV files for messaging and call histories.

iPhone Backup Extractor: Free
Though its interface wasn't as intuitive as iMazing's, the Lite version of iPhone Backup Extractor seemed like it might be able to extract texts to a delimited CSV format that was Excel-ready for search and review. The problem was, it hadn't done so after more than an hour and offered no way to determine if it was going to perform in due time or if it had lapsed into a coma. Neither was it a simple matter to kill the process once initiated. So, it *may* offer some functionality without buying a license, and there are basic, premium and business licenses offered for $34.95, $69.95 and $299.95 respectively.

I also tried **FonePaw, iBackup Viewer Pro, PhoneBrowse, Phone Rescue** and **Phone Trans**. All failed to perform. Though FonePaw seemed promising and worked on older iOS backups, it didn't recognize a recent iTunes backup. Likewise, purchased licenses for iBackup Viewer Pro and Phone Rescue were wasted monies in that neither could make head-nor-tails of a backup of my iPhone 6S Plus—the same backup that iMazing had no trouble parsing.

I expect that several of the tools that failed in my testing would turn out to be capable of interpreting iPhone backup files with some tweaking or when applied to other backups. I just can't prove it as I write this.

APPENDIX 4: Exemplar iPhone Backup Instruction for Custodian-Directed Backup

[[NOTE: This draft directive is offered to assist counsel in formulating language suited to the needs of the case and controlling law. It is not a form to be deployed without counsel. This example omits optional steps to encrypt the data set and transfer same to a distal repository for preservation, as such steps are frequently unnecessary to meet preservation duties].

Dear [Custodian]:

You recently acknowledged your obligation to preserve information relevant to a dispute between our company and _____. Please see the _____ hold notice for further details.

Within *48 hours of your receipt of this notice, y*ou must preserve the contents of your company-issued iPhone. If you cannot comply, please advise me at once by e-mail or phone. *Time is of the essence.*

You must make an unencrypted backup using iTunes and compress the backup folder per the instructions below. *Do not assume that you have been automatically making an unencrypted backup or preserving what's required using iCloud. You must carefully follow the procedures set out below.*

What you will need:

- Your company-issued iPhone and its USB charge/sync cable;
- Your company-issued desktop or laptop computer with the iTunes program installed. The computer must have available (unused) storage space on its boot (C:) drive exceeding *twice* the storage capacity of the iPhone. That is, if you have a 128GB capacity iPhone, use a computer with at least 256GB of unused storage space on its C: drive. You can find the capacity of the iPhone in Settings>General>About>Capacity. You can find the available storage on your computer's boot (C:) drive using File Explorer on a Windows machine or Finder on a Mac.

Time Required: One to two hours (most of it unattended "machine" time)

It will take about 10-15 minutes to follow these instructions, update iTunes, if needed, and begin the backup. The backup will complete in under 30 minutes, and you can continue to use the phone during the backup process (*but don't disconnect the charge/sync cable*). Then, it should take less than an hour to compress the data and 10 minutes or so to confirm successful compression and report on results. So long as the computer is secure and powered up throughout the process, you do not need to supervise, or leave the iPhone connected once backup completes.

Follow These Steps:

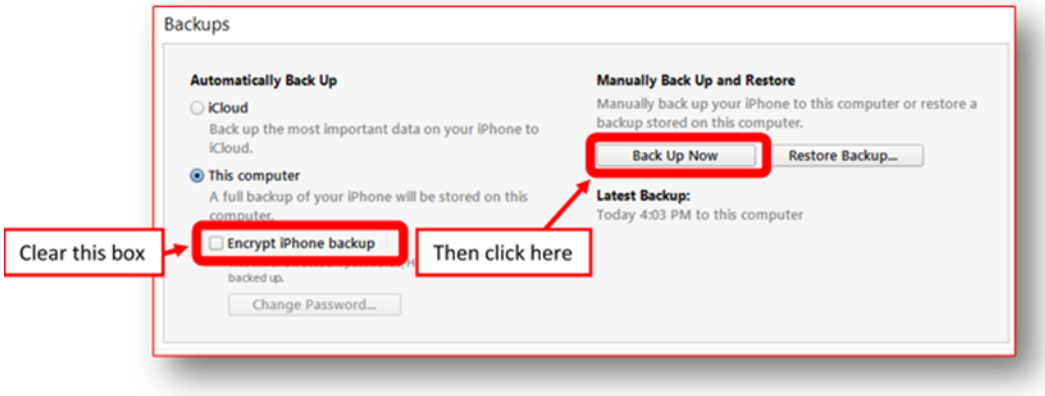
1. Open iTunes and check for updates (Help>Check for Updates). Install the latest version of iTunes if not installed.

2. Connect your iPhone to a USB 2.0 or 3.0 port on the computer using a USB charge/sync cable.

3. If a message asks for your device passcode or to Trust This Computer, follow the onscreen steps.

4. Select your iPhone when it appears in iTunes. Click Summary in the sidebar.

5. In the Summary pane, be sure to uncheck "Encrypt iPhone Backup," then click "Back Up Now." You need not otherwise modify your Backups settings.



6. Monitor the progress of the backup at the top center of the iTunes window. After the process ends, see if your backup finished successfully. If you're using iTunes for Windows, choose Edit>Preferences>Devices from the menu bar at the top of the iTunes window. If you're using iTunes for Mac, go to iTunes Preferences>Devices. You should see the name of your device with the date and time that iTunes created the backup. If you see 🔒 beside the name of your device, you need to be certain you unchecked "Encrypt iPhone Backup" and repeat the process until you do not see 🔒 beside the name of your device.
7. You can now disconnect your phone from the computer.

8. Locate the backup folder:

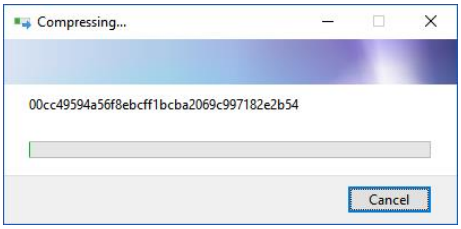
   1. **Windows:** Using File Explore, navigate to: **C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync\Backup** where "*your account name*" is the name of your Window's User ID on the machine.

   2. **Mac:** Using Finder, select Go>Go to Folder on the Finder menu and enter: **~/Library/Application Support/MobileSync/Backup/**

In both Windows and Mac, the Backup folder will contain one or more subfolders with 40-character names like *12da34bf5678900386c48267658d340eb34007f8*. **If there are**

multiple subfolders, i*dentify the subfolder that has the last modified date and time that matches the time you started this backup.*

9. **Compress the contents of the subfolder**: In Windows, right click on the subfolder just identified and select "***Send to>Compressed (zipped) folder***." A progress panel like the one at right should appear. On a Mac, right click on the subfolder and select "Compress." Do not turn off your computer or reboot. Allow the compression process to complete. It could take less than an hour to finish depending upon the type and volume of data backed up.



10. Once compression has completed, Windows users should again navigate to the backup folder (see step 8 above) to confirm the presence of a file with the same name as the subfolder you identified but with the file extension .zip. Record the name, date/time and size of the zip file. *[If you cannot see file extensions on your Windows machine, open "My Computer," click "Tools" and click "Folder Options" or click "View" and then "Options" depending on your version of Windows. In the Folder Options window, click the "View" tab. Uncheck the box that says, "Hide file extensions for known file types." This should make file extensions visible.]*

11. By reply e-mail, send the **name, date/time and size of the zip file you just created**. *Do not delete or open this file. It must be preserved without alteration until further notice.*

Your supervisor is copied here to insure you are afforded the time, oversight and support needed to comply in a timely way. Thank you for your cooperation. Call me at _____ with any questions.

Introduction to Metadata

In the old joke, a balloonist descends through the fog to get directions. "Where am I?" she calls out to a man on the ground, who answers, "You're in a yellow hot air balloon about sixty-seven feet above the ground." The frustrated balloonist replies, "Thanks for nothing, Counselor." Taken aback, the man on the ground asks, "How did you know I'm a lawyer?" "Simple," says the balloonist, "your answer was 100% accurate and totally useless."

If you ask a tech-savvy lawyer, "What's metadata?" there's a good chance you'll hear, "Metadata is data about data." Another answer that's 100% accurate and totally useless.

It's time to get past defining metadata as *data about data.*

It's time to move past "data about data" and embrace more useful ways to describe metadata—ways that enable counsel to rationally assess relevance and burden attendant to metadata. Metadata may be the most misunderstood topic in electronic discovery. Requesting parties demand discovery of "the metadata" without specifying what metadata is sought, and producing parties fail to preserve or produce metadata of genuine value and relevance.

It's Information *and* Evidence

Metadata is information that helps us use and make sense of other information. More particularly, metadata is information stored electronically that describes the characteristics, origins, usage, structure, alteration and validity of other electronic information. Many instances of metadata in many forms occur in many locations within and without digital files. Some is supplied by the user, but most metadata is generated by systems and software. Some is crucial evidence and some is merely digital clutter. Appreciating the difference--knowing what metadata exists and understanding its evidentiary significance—is a skill essential to electronic discovery.

Metadata is Evidence!

If evidence is anything that tends to prove or refute an assertion as fact, then clearly metadata is evidence. Metadata sheds light on the origins, context, authenticity, reliability and distribution of electronic evidence, as well as provides clues to human behavior. It's the electronic equivalent of DNA, ballistics and fingerprint evidence, with a comparable power to exonerate and incriminate.

In ***Williams v. Sprint/United Mgmt. Co.***, 230 F.R.D. 640 (D. Kan. 2005), the federal court ruled:

[W]hen a party is ordered to produce electronic documents as they are maintained in the ordinary course of business, the producing party should produce the electronic documents with their metadata intact, unless that party timely objects to production of metadata, the parties agree that the metadata should not be produced, or the producing party requests a protective order.

Within the realm of metadata lies discoverable evidence that litigants are obliged to preserve and produce. There's as much or more metadata extant as there is information and, like information, you don't deal with every bit of it. You *choose* wisely.

A lawyer's ability to advise a client about how to find, preserve and produce metadata, or to object to its production and discuss or forge agreements about metadata, hinges upon how well he or she understands metadata.

It's Just Ones and Zeroes

Understanding metadata and its importance in e-discovery begins with awareness that electronic data is, fundamentally, just a series of ones and zeroes. Though you've surely heard that before, you may not have considered the implications of information being expressed so severely. There are no *words*. There are no spaces or punctuation. *There is no delineation of any kind*.

How, then, do computers convert this unbroken sequence of ones and zeroes into information that makes sense to human beings? There must be some *key*, some *coherent structure* imposed to divine their meaning. But where does it come from? We can't derive it *from* the data if we can't first make sense *of* the data.

It's Encoded

Consider that written English conveys all information using fifty-two upper- and lowercase letters of the alphabet, ten numerical digits (0-9), some punctuation marks and a few formatting conventions, like spaces, lines, pages, etc. You can think of these collectively as a seventy- or eighty-signal "code." In turn, much of the same information could be communicated or stored in Morse code, where a three-signal code composed of dot, dash and pause serves as the entire "alphabet."

We've all seen movies where a tapping sound is heard and someone says, "Listen! It's Morse code!" Suddenly, the tapping is an encoded *message* because someone has furnished metadata ("It's Morse code!") *about* the data (tap, tap, pause, tap). Likewise, all those ones and zeroes on a computer only make sense when other ones and zeroes—the metadata—communicate the framework for parsing and interpreting the data stream.

All those ones and zeroes on a computer only make sense when *other* ones and zeroes—the metadata—communicate the framework for parsing and interpreting the data.

So, we <u>need</u> data *about* the data. We need information that tells us the data's encoding scheme. We need to know when information with one purpose ends and different information begins. And

we need to know the context, purpose, timeliness and origin of information for it to help us. That's ***metadata***.

The Metadata Continuum

Sometimes metadata is elemental, like the contents of a computer's master file table detailing where the sequences of one and zeroes for files begin and end. This metadata is invisible to a user without special tools called hex editors capable of peering through the walls of the Windows interface into the utilitarian plumbing of the operating system. Without file location metadata, every time a user tries to access a file or program, the operating system would have to examine every one and zero to find it. It'd be like looking for someone by knocking on every door in town!

At other times, metadata supports enhanced functionality not essential to the operation of the system. The metadata that tracks a file's name or the dates a file was created or last modified may only occasionally be probative of a claim or defense in a case, but that information *always* makes it easier to locate, sort and segregate files.

Metadata is often instrumental to the intelligibility of information, helping us use and make sense of it. "Sunny and 70 degrees" isn't a very useful forecast without metadata indicating *when* and *where* it's predicted to be the weather. Similarly, fully understanding information on a website or within a database, a collaborative environment like Microsoft's SharePoint or a social network like Facebook depends on metadata that defines its location, origin, timing and structure. It's even common for computerized information to comprise *more* metadata than data, in the same way that making sense of the two data points "sunny" and "70 degrees" requires *three* metadata points: location, date and time of day.

There's No Such Thing as "The Metadata"

As we move up the evolutionary ladder for metadata, some is recorded just in case it's needed to support a specialized task for the operating system or an application. Standard system metadata fields like "Camera Model" or "Copyright" may seem an utter backwater to a lawyer concerned with spreadsheets and word-processed documents; but, if the issue is the authenticity of a photograph or pirated music, these fields can make or break the case. ***It's all about relevance and utility.***

The point is, there's really no such thing as "the metadata" for a file or document. Instead, there's a continuum of metadata that enlightens many aspects of ESI. The metadata that matters depends upon the issues presented in the case and the task to be accomplished; consequently, the metadata preserved for litigation should reasonably reflect the issues that should be reasonably anticipated, and it must also address the file management and integrity needs attendant to identification, culling, processing, review and presentation of electronic evidence.

Up by the Bootstraps

When you push the power button on your computer, you trigger an extraordinary expedited education that takes the machine from an insensible, illiterate lump of silicon to a worldly savant in a matter of seconds. The process starts with a snippet of data on a chip called the **ROM BIOS** storing just enough information in its **R**ead **O**nly **M**emory to grope around for the **B**asic **I**nput and **O**utput **S**ystem devices like the keyboard, screen and hard drive. It also holds the metadata needed to permit the computer to begin loading ones and zeroes from storage and to make just enough sense of their meaning to allow more metadata to load from the disk, in turn enabling the computer to access more data and, in this widening gyre, "teach" itself to be a modern, capable computer.

This rapid, self-sustaining self-education is as magical as if you hoisted yourself into the air by pulling on the straps of your boots, which is truly why it's called "bootstrapping" or just "booting" a computer.

File Systems and Relative Addressing

So now that our computer's taught itself to read, it needs a library. Most of those ones and zeroes on the hard drive are files that, like books, are written, read, revised and referenced. Computers use file systems to keep track of files just as libraries once used card catalogues and the Dewey Decimal system to track books.

Imagine you own a thousand books without covers that you stored on one very long shelf. You also own a robot named Robby that can't read, but Robby can count books very accurately. How would you instruct Robby to get a book?

If you know the order in which the books are stored, you'd say, "Robby, bring me the 412th book." If it was a 24-volume set of encyclopedias, you might add: "…and the next 23 books." The books don't "know" where they're shelved. Each book's location is metadata *about* the book.

Locating something by specifying that it's so many units from a particular point is called *relative addressing*. The number of units the destination is set off from the specified point is called the *offset*. Computers use offset values to indicate the locations of files on storage devices as well as to locate information inside files.

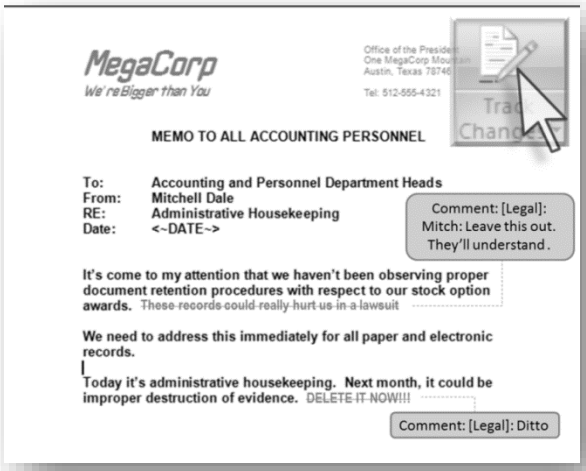
Computers use various units to store and track information, so offsets aren't always expressed in the same units. A "bit" stores a one or zero, eight bits is a "byte," (sufficient to hold a letter in the Latin alphabet), 512 bytes is often a *sector or block* (see **Appendix A**) and (typically) eight contiguous sectors or blocks is a *cluster*. The cluster is the most common unit of logical storage, and modern computers tend to store files in as many of these 4,096-byte clusters, or "data

baskets," as needed. Offset values are couched in bytes when specifying the location of information within files and as sectors when specifying the location of files on storage media.

Metadata Mix-up: Application Metadata

To the extent lawyers have heard of metadata at all, it's likely in the context of just one species of metadata called **application metadata** with the fearsome potential to inadvertently reveal confidential or privileged information embedded within electronic documents. Computer programs or "applications" store work product in files "native" to them, meaning that the data is structured and encoded to support the application. As these applications added features--like the ability to undo changes in or collaborate on a document--the native files used to store documents had to retain those changes and collaborations.

An oft-cited culprit is Microsoft Word, and a cottage industry has grown up offering utilities to strip embedded information, like comments and tracked changes, from Word documents. Because of its potential to embarrass lawyers or compromise privilege, metadata has acquired an unsavory reputation amongst the bar. But metadata is much more than simply the embedded *application* metadata that affords those who know how to find it the ability to dredge up a document's secrets.

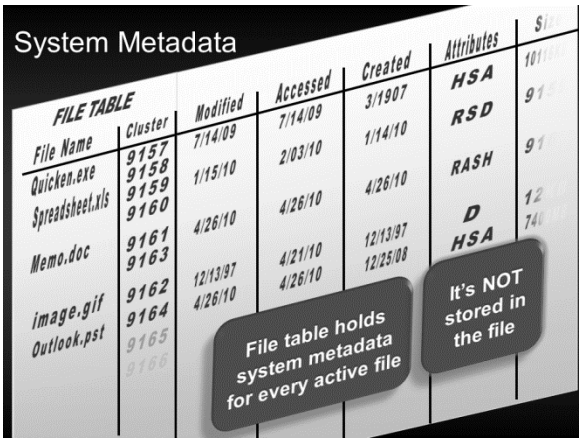


By design, **application metadata is embedded in the file it describes and moves with the file when you copy it.** However, not all metadata is embedded (for the same reason that cards in a library card catalog aren't stored between the pages of the books). You have to know where the information resides to reach it.

System Metadata

Unlike books, computer files aren't neatly bound tomes with names embossed on spines and covers. Often, files don't internally reflect the name they've been given or other information about their location,

history or ownership. The information about the file which is *not* embedded within the file it describes but is stored apart from the file is its **system metadata**. The computer's file management system uses *system* metadata to track file locations and store demographics about each file's name, size, creation, modification and usage.



System metadata is crucial to electronic discovery because so much of our ability to identify, find, sort and cull information depends on its system metadata values. For example, system metadata helps identify the custodians of files, what the file is named, when files were created or altered and the folders in which they are stored. System metadata stores much of the *who, when, where* and *how* of electronic evidence.

Every computer employs one or more databases to keep track of system metadata. In computers running the Windows operating system, the principal "card catalog" tracking system metadata is called the Master File Table or "MFT." In the predecessor DOS operating system, it was called the File Allocation Table or "FAT." The more sophisticated and secure the operating system, the greater the richness and complexity of the system metadata in the file table.

Windows Shell Items

In the Windows world, Microsoft calls any single piece of content, such as a file, folder or contact, a "Shell item." Any individual piece of metadata associated with a Shell item is called a "property" of the item. Windows tracks 284 distinct metadata properties of Shell items in 28 property categories**.** To see a list of Shell item properties on your own Windows system, right click on the column names in any folder view and select "More…." Examining a handful of these properties in four key categories reveals metadata of great potential evidentiary value existing within and without files, messages and photos:

Category	Properties	
Document	ClientID	LastAuthor
	Contributor	RevisionNumber
	DateCreated	Template
	DatePrinted	TotalEditingTime
	DateSaved	Version
	DocumentID	
Message	AttachmentContents	FromAddress
	AttachmentNames	FromName
	BccAddress	HasAttachments
	BccName	IsFwdOrReply
	CcAddress	SenderAddress

	CcName	SenderName
	ConversationID	Store
	ConversationIndex	ToAddress
	DateReceived	ToDoFlags
	DateSent	ToDoTitle
	Flags	ToName
Photo	CameraManufacturer	CameraSerialNumber
	CameraModel	DateTaken
System	ApplicationName	ItemAuthors
	Author	ItemDate
	Comment	ItemFolderNameDisplay
	Company	ItemFolderPathDisplay
	ComputerName	ItemName
	ContainedItems	OriginalFileName
	ContentType	OwnerSID
	DateAccessed	Project
	DateAcquired	Sensitivity
	DateArchived	SensitivityText
	DateCompleted	SharedWith
	DateCreated	Size
	DateImported	Status
	DateModified	Subject
	DueDate	Title
	EndDate	FileOwner
	FileAttributes	FlagStatus
	FileCount	FullText
	FileDescription	IsAttachment
	FileExtension	IsDeleted
	FileName	IsEncrypted
	IsShared	

Much More Metadata

The 284 Windows Shell item properties are by no means an exhaustive list of metadata. Software applications deploy their own complements of metadata geared to supporting features unique to each application. E-mail software, word processing applications and spreadsheet, database, web browser and presentation software collectively employ hundreds of additional fields of metadata.

For example, digital photographs can carry dozens of embedded fields of metadata called EXIF data detailing information about the date and time the photo was taken, the camera, settings, exposure, lighting, even precise geolocation data. Photos taken with cell phones having GPS capabilities contain detailed information about where the photo was taken *to a precision of about ten meters*.

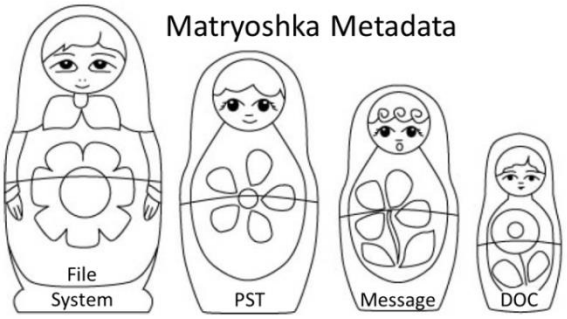
Photos taken with cell phones having GPS capabilities contain detailed information about where the photo was taken.

The popular Microsoft Outlook e-mail client application provides for more than 180 standard application metadata fields which users may select to customize their view.

But, even this broad swath of metadata is still only *part* of the probative information about information recorded by computers. Within the Master File Table and index records used by Windows to track all files, still more attributes are encoded in hexadecimal notation. In fact, an ironic aspect of Windows is that the record used to track information about a file may be larger than the file itself! Stored within the hives of the System Registry—the "Big Brother" database that tracks attributes covering almost any aspect of the system—are thousands upon thousands of attribute values called "registry keys." Other records and logs track network activity and journal virtually every action.

Matryoshka Metadata

Matryoshka are carved, cylindrical Russian dolls that nest inside one another. It's helpful to think of computer data the same way. If the evidence of interest is a Word document attached to an e-mail, the document has its usual complement of application metadata that moves with the file; but, as it nests within an e-mail message, its "system" metadata is only that which is contained within the transporting message. The transporting message, in turn, carries its own metadata concerning transit, addressing, structure, encoding and the like. The message is managed by Outlook, which maintains a rich complement of metadata about the message and about its own configuration. As configured, Outlook may store all messages and application metadata in a container file called Outlook.PST. This container file exists within a file system of a computer that stores system metadata about the container file, such as where the file is stored, under whose user account, when it was last modified, its size, name, associated application and so on.



Within this Matryoshka maelstrom of metadata, some information is readily accessible and comprehensible while other data is so Byzantine and cryptic as to cause even highly skilled computer forensic examiners to scratch their heads.

Forms of Metadata

Now that *your* head is spinning from all the types, purposes and sources of metadata, let's pile on another complexity concern: the *form* of the metadata. Metadata aren't presented the same way from field to field or application to application. For example, some of the standard metadata fields for Outlook e-mail are simply bit flags signifying "true" or "false" for, *e.g.,* "Attachment," "Do Not Auto Archive," "Read" or "Receipt Requested." Some fields reference different *units*, *e.g.*, "Size" references bytes, where "Retrieval Time" references minutes. Several fields even use the *same* value to mean *different* things, e.g., a value of "1" signifies "Completed" for "Flag Status," but denotes "Normal for "Importance," "Personal" for "Sensitivity" and "Delivered" for "Tracking Status."

The form of metadata is a key consideration when deciding how to preserve and produce the information. Not everyone would appreciate a response like, "for this message, item type 0x0029 with value type 0x000b was set to 0x00," when the question posed was whether the sender sought a read receipt. Because some metadata items are simply bit flags or numeric values and make sense only as they trigger an action or indication in the native application, preserving metadata can entail more than just telling opposing counsel, "we will grab it and give it to you." Context must be supplied.

It's not that locating and interpreting any particular item is difficult, but you have to know whether your firm, client or service provider has the tools and employs a methodology that makes it easy. That's why it's crucial to know what metadata is routinely collected and amenable to production before making commitments to opposing counsel or the court. Any e-discovery vendor you employ should be able to readily identify the system and application metadata values they routinely collect and process for production. Any still-existing metadata value can be readily collected and processed—after all, it's just data like any other; but, a few items will require specialized tools, custom programming or tweaks to established workflows.

Relevance and Utility

How much of this metadata is relevant and discoverable? Would I be any kind of lawyer if I didn't answer, "It depends?" In truth, it *does* depend upon what issues the data bears upon, its utility and the cost and burden of preservation and review.

Metadata is unlike almost any other evidence in that its import in discovery may flow from its probative value (relevance as evidence), its utility (functionally abetting the searching, sorting and interpretation of ESI) or both. If the origin, use, distribution, destruction or integrity of electronic evidence is at issue, the relevant "digital DNA" of metadata is essential, probative evidence that needs to be preserved and produced. Likewise, if the metadata materially facilitates the searching sorting and management of electronic evidence, it should be preserved and produced for its utility.[14] Put simply, *metadata is an important part of ESI and should be considered for production in every case. Too, much of what is dismissed (and suppressed) as "mere metadata" is truly substantive content, such as embedded comments between collaborators in documents, speaker notes in presentations and formulas in spreadsheets.*

Does this then mean that every computer system and data device in every case must be forensically imaged and analyzed by experts? Absolutely not! *Once we understand what metadata exists and what it signifies, a continuum of reasonableness will inform our actions*. A competent police officer making a traffic stop collects relevant information, such as, *e.g.*, the driver's name, address, vehicle license number, driver's license number and date, time and location of offense. We wouldn't expect the traffic cop to collect a bite mark impression, DNA sample or shoe print from the driver. But, make it a murder case and the calculus changes.

Addressing just the utility aspect of metadata in the context of forms of production, The Sedona Conference guideline states:

Absent party agreement or court order specifying the form or forms of production, production should be made in the form or forms in which the information is ordinarily maintained or in a reasonably usable form, **taking into account the need to produce reasonably accessible metadata that will enable the receiving party to have the same ability to access, search, and**

[14] This important duality of metadata is a point sometimes lost by those who read the rules of procedure too literally and ignore the comments to same. Federal Rules of Civil Procedure Rule 26(b) states that, "Parties may obtain discovery regarding any nonprivileged matter that is ***relevant to any party's claim or defense*** and proportional to the needs of the case..." (emphasis added). The Comments to Rules revisions made in 2015 note, "[a] portion of present Rule 26(b)(1) is omitted from the proposed revision. After allowing discovery of any matter relevant to any party's claim or defense, the present rule adds: "including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter." Discovery of such matters is so deeply entrenched in practice that it is no longer necessary to clutter the long text of Rule 26 with these examples. The discovery identified in these examples should still be permitted under the revised rule when relevant and proportional to the needs of the case. ***Framing intelligent requests for electronically stored information, for example, may require detailed information about another party's information systems and other information resources***" (emphasis added). Though the Committee could have been clearer in its wording and have helpfully used the term "metadata," the plain import is that relevance "to a party's claims or defenses" is not the sole criterion to be used when determining the scope of discovery as it bears on metadata. Metadata is discoverable for its utility as well as its relevance.

display the information as the producing party where appropriate or necessary in light of the nature of the information and the needs of the case.

The Sedona Principles Addressing Electronic Document Production, Second Edition (June, 2007), Principle 12 (emphasis added).

The crucial factors are burden and cost balanced against utility and relevance. The goal should be a level playing field between the parties in terms of their ability to see and use relevant electronic evidence, including its metadata.

So where do we draw the line? Begin by recognizing that the advent of electronic evidence hasn't changed the fundamental dynamics of discovery: *Litigants are entitled to discover relevant, non-privileged information, and relevance depends on the issues before the court.* Relevance assessments aren't static, but change as new evidence emerges and new issues arise. Metadata irrelevant at the start of a case may become decisive when, *e.g.*, allegations of data tampering or spoliation emerge. Parties must periodically re-assess the adequacy of preservation and production of metadata and act to meet changed circumstances.

Periodically re-assess the adequacy of preservation and production of metadata, and act to meet changed circumstances.

Metadata Musts

There are easily accessible, frequently valuable metadata that, like the information collected by the traffic cop, we should expect to routinely preserve. Examples of essential system metadata fields for any file produced are:

- **Custodian;**
- **Source Device;**
- **Originating Path** (file path of the file as it resided in its original environment);
- **Filename** (including extension);
- **Last Modified Date;** and
- **Last Modified Time.**

Any party producing or receiving ESI should be able to state something akin to, "This spreadsheet named *Cash Forecast.xls* came from the My Documents folder on Sarah Smith's Dell laptop and was last modified on January 16, 2016 at 2:07 PM CST."

One more metadata "must" for time and date information is the UTC time zone offset applicable to each time value (unless all times have been normalized; that is, processed to a common time zone). UTC stands for both for Temps Universel Coordonné and Coordinated Universal Time. It's

a fraction of a second off the better known Greenwich Mean Time (GMT) and identical to Zulu time in military and aviation circles. Why UTC instead of TUC or CUT? It's a diplomatic compromise, for neither French nor English speakers were willing to concede the acronym. Because time values may be expressed with reference to local time zones and variable daylight savings time rules, you need to know the UTC offset for each item.

Application metadata is, by definition, embedded within native files; so, native production of ESI obviates the need to selectively preserve or produce application metadata. It's in the native file. But when ESI is converted to other forms, the parties must assess what metadata will be lost or corrupted by conversion and identify, preserve and extract relevant or utile application metadata fields for production.

For e-mail messages, this is a straightforward process, notwithstanding the dozens of metadata values that may be introduced by e-mail client and server applications. The metadata "musts" for e-mail messages are, as available:

- **Custodian** – Owner of the mail container file or account collected;
- **To** – Addressee(s) of the message;
- **From** – The e-mail address of the person sending the message;
- **CC** – Person(s) copied on the message;
- **BCC** – Person(s) blind copied on the message;
- **Subject** – Subject line of the message;
- **Date Sent (or Received**)– Date the message was sent (or received);
- **Time Sent (or Received)** – Time the message was sent (or received);
- **Attachments** – Name(s) or other unique identifier(s) of attachments/families;
- **Mail Folder Path** – Path of the message to its folder in the originating mail account; and,
- **Message ID** – Microsoft Outlook or similar unique message thread identifiers.[15]

E-mail messages that traverse the Internet contain so-called header data detailing the routing and other information about message transit and delivery. Whether header data should be preserved and produced depends upon the reasonable anticipation that questions concerning authenticity, receipt or timing of messages will arise. A more appropriate inquiry might be, "since header data is an integral part of every message, why should any party be permitted to discard this part of the evidence absent cause shown?"

[15] In fact, few of these items are truly "metadata" in that they are integral parts of the message (*i.e.,* user-contributed content); however, message header fields like To, From, CC, BCC and Subject are so universally labeled "metadata," it's easier to accept the confusion than fight it.

The metadata essentials must further include metadata values generated by the discovery and production process itself, such as Bates numbers and ranges, hash values, production paths and names, family relationships and the like.

When ESI other than e-mail is converted to non-native forms, it can be enormously difficult to preserve, produce and present relevant or necessary application metadata in ways that don't limit its utility or intelligibility. For example, tracked changes and commentary in Microsoft Office documents may be incomprehensible without seeing them in context, i.e., superimposed on the document. By the same token, furnishing a printout or image of the document with tracked changes and comments revealed can be confusing and deprives a recipient of the ability to see the document as the user ultimately saw it. As well, it often corrupts the extraction of searchable text using optical character recognition. If native forms will not be produced, the most equitable approach may be to produce the document twice: once with tracked changes and comments hidden and once with them revealed.

For certain ESI, there is simply no viable alternative to native production with metadata intact. The classic example is a spreadsheet file. The loss of functionality and the confusion engendered by rows and columns that break and splay across multiple pages mandates native production. A like loss of functionality occurs with sound files (e.g., voice mail), video, animated presentations (i.e., PowerPoint) and databases, web content, SharePoint, social networking sites and collaborative environments where the structure and interrelationship of the information--reflected in its metadata—defines its utility and intelligibility.

The Path to Production of Metadata

The balance of this section discusses steps typically taken in shepherding a metadata production effort, including:

- Gauge spoliation risks before you begin
- Identify potential forms of metadata
- Assess relevance
- Consider authentication and admissibility
- Evaluate need and methods for preservation
- Collect metadata
- Plan for privilege and production review
- Resolve production issues

Gauge spoliation risks before you begin

German scientist Werner Heisenberg thrilled physicists and philosophy majors alike when he posited that the very act of observing alters the reality observed. Heisenberg's Uncertainty

Principal speaks to the world of subatomic particles, but it aptly describes a daunting challenge to lawyers dealing with metadata: When you open any document in Office applications without first employing specialized hardware or software, metadata often changes and prior metadata values may be lost. Altered metadata implicates not only claims of spoliation, but also severely hampers the ability to filter data chronologically. How, then, can a lawyer evaluate documents for production without reading them?

Begin by gauging the risk. Not every case is a crime scene, and few cases implicate issues of computer forensics. Those that do demand extra care be taken immediately to preserve a broad range of metadata evidence. Further, it may be no more difficult or costly to preserve data using forensically sound methods that reliably preserve all data and metadata.

For the ordinary case, a working knowledge of the most obvious risks and simple precautions are sufficient to protect the metadata most likely to be needed.

Windows systems typically track at least three date values for files, called "MAC dates" for Last Modified, Last Accessed and Created. Of these, the Last Accessed date is the most fragile, yet least helpful. Historically, last accessed dates could be altered by previewing files and running virus scans. Now, last accessed dates are only infrequently updated in Windows (after Vista and Win7/8/10).

Similarly unhelpful in e-discovery is the Created date. The created date is often presumed to be the authoring date of a document, but it more accurately reflects the date the file was "created" *within the file system of a storage medium*. So, when you copy a file to new media, you're "created" it on the new media as of the date of copying, and the created date changes accordingly. Conversely, when you use an old file as a template to create a new document, the creation date of the template stays with the new document. Created dates may or may not coincide with authorship; so, it's a mistake to assume same.

The date value of greatest utility in e-discovery is the Last Modified date. The last modified date of a file is not changed by copying, previewing or virus scans. It changes only when a file is opened and saved; however, it is not necessary that the user-facing content of a document be altered for the last accessed date to change. Other changes—including subtle, automatic changes to application metadata--may trigger an update to the last modified date when the file is re-saved by a user.

Apart from corruption, application metadata does not change unless a file is opened. So, the easiest way to preserve a file's application metadata is to keep a pristine, unused copy of the file and access only working copies. By always having a path back to a pristine copy, inadvertent loss

or corruption of metadata is harmless error. Calculating and preserving hash values for the pristine copies is a surefire way to demonstrate that application metadata hasn't changed

An approach favored by computer forensic professionals is to employ write blocking hardware or software to intercept all changes to the evidence media.

Finally, copies can be transferred to read only media (e.g., a CD-R or DVD-R), permitting examination without metadata corruption.

Identify potential forms of metadata

To preserve metadata and assess its relevance, you must know it exists. So, for each principal file type subject to discovery, assemble a list of associated metadata of potential evidentiary or functional significance. You'll likely need to work with an expert the first time or two, but once you have a current and complete list, it will serve you in future matters. You'll want to know not only what the metadata fields contain, but also their location and significance.

For unfamiliar or proprietary applications and environments, enlist help identifying metadata from the client's IT personnel. Most importantly, seek your opponent's input, too. Your job is simpler when the other side is conversant in metadata and can expressly identify fields of interest. The parties may not always agree, but at least you'll know what's in dispute.

Assess relevance

Are you going to preserve and produce dozens and dozens of metadata values for every document and e-mail in the case? Probably not, although you may find it easier to preserve all than selectively cull out just those values you deem relevant.

Metadata is like the weather reports from distant cities published in the daily newspaper. Though only occasionally relevant, we want the information available when we need it.[16]

Relevance is always subjective and is as fluid as the issues in the case. Case in point: two seemingly innocuous metadata fields common to Adobe Portable Document Format (PDF) files are "PDF Producer" and "PDF Version." These are listed as "Document Properties" under the "File" menu in any copy of Adobe Acrobat. Because various programs can link to Acrobat to create PDF files, the PDF Producer field stores information concerning the source application, while the PDF Version field tracks what release of Acrobat software was used to create the PDF document. These metadata values may seem irrelevant, but consider how that perception changes if the dispute turns on a five-year-old PDF contract claimed to have been recently forged. If the

[16] Of course, we are more likely go to the internet for weather information; but even then, we want the information available when we need it.

metadata reveals the PDF was created using a scanner introduced to market last year and the latest release of Acrobat, that metadata supports a claim of recent fabrication. In turn, if the metadata reflects use of a very old scanner and an early release of Acrobat, the evidence bolsters the claim that the document was scanned years ago. Neither is conclusive on the issue, but both are relevant evidence needing to be preserved and produced.

Assessing relevance is another area where communication with an opponent is desirable. Often, an opponent will put relevance concerns to rest by responding, "I don't need that." For every opponent who demands "all the metadata," there are many who neither know nor care about metadata.

Consider Authentication and Admissibility

Absent indicia of authenticity like signatures, handwriting and physical watermarks, how do we establish that electronic evidence is genuine or that a certain individual created an electronic document? Computers may be shared or unsecured and passwords lost or stolen. Software permits alteration of documents sans the telltale signs that expose paper forgeries. Once, we relied upon dates in correspondence to establish temporal relevance, but now documents may generate a new date each time they are opened, inserted by a word processor macro as a "convenience" to the user.

Where the origins and authenticity of evidence are in issue, preservation of original date and system user metadata is essential. When deciding what metadata to preserve or request, consider, *inter alia,* network access logs and journaling, evidence of other simultaneous user activity and version control data. For more on this, review the material on digital forensics, *supra.*

An important role of metadata is establishing a sound chain of custody for ESI. Through every step of e-discovery--collection, processing, review, and production—the metadata should facilitate a clear, verifiable path back to the source ESI, device and custodian.

In framing a preservation strategy, balance the burden of preservation against the likelihood of a future need for the metadata, but remember, if you act to preserve metadata for documents supporting your case, it's hard to defend a failure to preserve metadata for items bolstering the opposition's case. Failing to preserve metadata could deprive you of the ability to challenge the relevance or authenticity of material you produce.

Evaluate Need and Methods for Preservation

Not every item of metadata is important in every case, so what factors should drive preservation? The case law, rulings of the presiding judge and regulatory obligations are paramount concerns, along with obvious issues of authenticity and relevance; but another aspect to consider is the stability of metadata. As discussed, some essential metadata fields, like Last Modified Date, change when a file is used and saved. If you don't preserve dynamic data, you lose it. Where a preservation duty has attached, by, *e.g.*, issuance of a preservation order or operation of law, the loss of essential metadata may, at best, require costly remedial measures be undertaken or, at worst, could constitute spoliation subject to sanctions.

If you fail to preserve metadata at the earliest opportunity, you may never be able to replicate what was lost.

How, then, do you avoid spoliation occasioned by review and collection? What methods will preserve the integrity and intelligibility of metadata? Poorly executed collection efforts can corrupt metadata. When, for example, a custodian or reviewer copies responsive files to new media, prints documents or forwards e-mail, metadata is altered or lost. Consequently, metadata preservation must be addressed before a preservation protocol is implemented. Be certain to document what was done and why. Advising your opponents of the proposed protocol in sufficient time to allow them to object, seek court intervention or propose an alternate protocol helps to protect against belated claims of spoliation.

Collect Metadata

Because metadata is stored both within and without files, simply duplicating a file without capturing its system metadata may be insufficient. However, not all metadata preservation efforts demand complex and costly solutions. It's possible to tailor the method to the case in a proportional way. As feasible, record and preserve system metadata values before use or collection. This can be achieved using software that archives the basic system metadata values to a table, spreadsheet or CSV file. Then, if an examination results in a corruption of metadata, the original values can be ascertained. Even just archiving files ("zipping" them) may be a sufficient method to preserve associated metadata. In other cases, you'll need to employ tools purpose-built for e-discovery, undertake forensic imaging or use vendors specializing in electronic discovery.

Whatever the method chosen, be careful to preserve the association between the data and its metadata. For example, if the data is the audio component of a voice mail message, it may be of little use unless correlated with the metadata detailing the date and time of the call and the identity of the voice mailbox user. This is often termed, "preserving family relationships."

When copying file metadata, know the limitations of the environment and medium in which you're working. I learned this lesson the hard way many years ago while experimenting with recordable CDs to harvest files and their metadata. Each time I tried to store a file and its MAC dates (modified/accessed/created) on a CD, I found that the three different MAC dates derived from the hard drive would always emerge as three identical MAC dates when read from the CD! I was corrupting the data I sought to preserve. I learned that optical media like CD-Rs aren't formatted in the same manner as magnetic media like hard drives. Whereas the operating system formats a hard drive to store three distinct dates, CD-R media stores just one. In a sense, a CD file system has no place to store all three dates, so discards two. When the CD's contents are copied back to magnetic media, the operating system re-populates the slots for the three dates with the single date found on the optical media. Thus, using a CD in this manner served to both corrupt and misrepresent the metadata. Similarly, different operating systems and versions of applications maintain different metadata; so, test your processes for alteration, truncation or loss of metadata.

Plan for Privilege and Production Review

The notion of reviewing metadata for privilege may seem odd unless you consider that application metadata potentially contains deleted content and commentary. The industry (sub)standard has long been to simply suppress the metadata content of evidence, functionally deleting it from production. This has occurred without any apparent entitlement springing from privilege. Producing parties didn't want to review metadata so simply, *incredibly* purged it from production for their own convenience. *But, that dog don't hunt no more.* Metadata must be assessed like any other potentially-responsive ESI and produced when tied to a responsive and non-privileged information item.

When the time comes to review metadata for production and privilege, the risks of spoliation faced in harvest may re-appear during review. Ponder:

- How will you efficiently access metadata?
- Will the metadata exist in a form you can interpret?
- Will your examination alter the metadata?
- How will you flag metadata for production?
- How can you redact privileged or confidential metadata?

If a vendor or in-house discovery team has extracted the metadata to a slip-sheet in an image format like TIFF or PDF, review is as simple as reading the data. However, if review will take place in native format, some metadata fields may be inaccessible, encoded or easily corrupted unless you use tools that make the task simple. Good e-discovery tools are designed to do so. If the

review set is hosted online, be certain you understand which metadata fields are accessible and intelligible via the review tool and which are not. Don't just assume: test.

Application Metadata and Review

As noted, many lawyers deal with metadata in the time-honored way: **by pretending that it doesn't exist.** That is, they employ review methods that don't display application metadata, such as comments and tracked changes present in native Microsoft Office productivity documents. These lawyers review only what prints instead of all the information in the document. Rather than adjust their methods to the evidence, they refuse to produce ESI with its application metadata intact lest they unwittingly produce privileged or confidential content.

They defend this behavior by claiming that the burden to review application metadata for privileged or confidential content is greater than the evidentiary value of that content. To ensure that requesting parties cannot access all that metadata the producing counsel ignored, producing parties instead strip away all metadata, either by printing the documents to paper or hiring a vendor to convert the ESI to static images (i.e., TIFFs). Doing so successfully removes the metadata, but wrecks the utility and searchability of most electronic evidence.

Sometimes, counsel producing TIFF image productions will undertake to reintroduce some of the stripped metadata and searchable text as ancillary productions called **load files.** The production of document images and load files is a high-cost, low utility, error-prone approach to e-discovery; but, its biggest drawback is that it's increasingly unable to do justice to the native files it supplants. When produced as images, spreadsheets often become useless and incomprehensible. Multimedia files disappear. Any form of interactive, animated or structured information ceases to work. In general, the richer the information in the evidence, the less likely it is to survive production in TIFF.

Despite these shortcomings, lawyers cling to cumbersome TIFF productions, driving up e-discovery costs. This is troubling enough, but raises a disturbing question: *Why does any lawyer assume he or she is free to unilaterally suppress--without review or proffer of a privilege log—integral parts of discoverable evidence?* Stripping away or ignoring metadata that's an integral part of the evidence seems little different from erasing handwritten notes in medical records because you'd rather not decipher the doctor's handwriting!

In ***Williams v. Sprint/United Mgmt Co.,*** 230 F.R.D. 640 (D. Kan. 2005), concerns about privileged metadata prompted the defendant to strip out metadata from the native-format spreadsheet files it produced in discovery. The court responded by ordering production of all metadata as maintained in the ordinary course of business, save only privileged and expressly protected metadata.

The court was right to recognize that privileged information need not be produced, wisely distinguishing between surgical redaction and blanket excision. One is redaction following examination of content and a reasoned judgment that matters are privileged. The other excises data in an overbroad and haphazard fashion, grounded only on an often-unwarranted concern that the data pared away ***might*** contain privileged information. The baby goes out with the bathwater. Moreover, blanket redaction based on privilege concerns doesn't relieve a party of the obligation to log and disclose such redaction. The defendant in *Williams* not only failed to examine or log items redacted, it left it to the plaintiff to figure out that something was missing.

The requesting party is entitled to the metadata benefits that are available to the producing party.

The underlying principle is that the requesting party is entitled to the metadata benefits available to the producing party. That is, the producing party may not vandalize or hobble electronic evidence for production without adhering to the same rules attendant to redaction of privileged and confidential information from paper documents.

Resolve Production Issues

Like other forms of electronic evidence, metadata may be produced in its native and near-native formats, as a database or a delimited load file, in an image format, hosted in an online database or even furnished as a paper printout. However, metadata presents more daunting production challenges than other electronic evidence. One hurdle is that metadata is often unintelligible outside its native environment without processing and labeling. How can you tell if an encoded value describes the date of creation, modification or last access without both decoding the value *and* preserving its significance with labels?

Another issue is that metadata isn't always textual. It may consist of no more than a flag in an index entry—just a one or zero—wholly without meaning unless you know what it denotes. A third challenge to producing metadata lies in finding ways to preserve the relationship between metadata and the data it describes and, when obliged to do so, to present both the data and metadata to be electronically searchable.

When files are separated from their metadata, we lose much of the ability to sort, manage and authenticate them. Returning to the voice mail example, unless the sound component of the message (e.g., the WAV file) is paired with its metadata, a reviewer must listen to the message in real time, hoping to identify the voice and deduce the date of the call from the message. It's a Herculean task without metadata, but a task made much simpler if the producing party, *e.g.,* drops

the WAV file into an Adobe PDF file as an embedded sound file, then inserts the metadata in the image layer. Now, a reviewer can both listen to the message and search and sort by the metadata.

Sometimes, simply producing a table, spreadsheet or load file detailing originating metadata values will suffice. On other occasions, only native production will suffice to supply relevant metadata in a useful and complete way. Determining the method of metadata production best suited to the case demands planning, guidance from experts and cooperation with the other side.

Beyond Data about Data

The world's inexorable embrace of digital technologies serves to escalate the evidentiary and functional value of metadata in e-discovery. Today, virtually all information is born electronically, bound to and defined by its metadata as we are bound to and defined by our DNA. The proliferation and growing importance of metadata dictates that we move beyond unhelpful definitions like "data about data," toward a fuller appreciation of metadata's many forms and uses.

Appendix A: Just Ones and Zeros

The binary data above comprises a single hard drive sector storing a binary encoding of the text below (excerpted from *David Copperfield* by Charles Dickens):

I was born with a caul, which was advertised for sale, in the newspapers, at the low price of fifteen guineas. Whether sea-going people were short of money about that time, or were short of faith and preferred cork jackets, I don't know; all I know is, that there was but one solitary bidding, and that was from an attorney connected with the bill-broking business, who offered two pounds in cash, and the balance in sherry, but declined to be guaranteed from drowning on any higher bargain. Consequently the advertisement was withdrawn at a dead loss--for as to sherry, my poor dear mother's own sherry was in the market then--and ten years afterwards, the caul was put up in a raffle down in our part of the country, to fifty members at half-a-crown a head, the winner to spend five shillings. I was present myself, and I remember to have felt quite uncomfortable and confused, at a part of myself being disposed of in that way. The caul was won, I recollect, by an old lady with a hand-basket, who, very reluctantly, pr
[end of sector]

A 3 terabyte hard drive ($85 at your local Wal-Mart) contains more than *5.8 billion* 512 byte sectors.

Appendix B: Exemplar Native Production Protocol

The following protocol is an example of how one might designate forms and fields for production. Its language and approach should be emulated only when careful analysis suggests so doing is likely to be effective, economical and proportionate, as well as consistent with applicable law and rules of practice.

It's important to recognize that there is no omnibus complement of metadata applicable to all forms of ESI. You must identify and select the fields with particular relevance and utility for your case and applicable to the particular types and forms of ESI produced. But see "Metadata Musts," *supra*.

Note also that names assigned to the load file fields are arbitrary. How one names fields in load files is largely immaterial so long as the field name chosen is unique. In practice, when describing the date an e-mail was sent, some label the field "Sent_Date," others use "Datesent" and still others use "Date_Sent." There is no rule on this, nor need there be. What matters is that the information that will be used to populate the field be clearly and unambiguously defined and not be unduly burdensome to extract. Oddly, the e-discovery industry has not settled upon a standard naming convention for metadata fields.

NATIVE FORMAT PRODUCTION PROTOCOL

1. "Information items" as used here encompasses individual documents and records (including associated metadata), whether on paper, as discrete "files" stored electronically, optically or magnetically, or as a database, archive, or container file. The term should be read broadly to include all forms of electronically stored information (ESI), including but not limited to e-mail, messaging, word processed documents, digital presentations, social media posts, webpages, and spreadsheets.

2. Responsive ESI shall be produced in its native form; that is, in the form in which the information was created, used, and stored by the native application employed by the producing party in the ordinary course of business.

3. If it is infeasible or unduly burdensome to produce an item of responsive ESI in its native form, it may be produced in an agreed upon near-native form; that is, in a form in which the item can be imported into an application without a material loss of content, structure, or functionality as compared to the native form. Static image production formats serve as near-native alternatives only for information items that are natively static images (*i.e.,* faxes and scans).

4. Examples of agreed-upon native or near-native forms in which specific types of ESI should be produced are:

Source ESI	Native or Near-Native Form or Forms Sought
Microsoft Word documents	.DOC, .DOCX
Microsoft Excel spreadsheets	.XLS, .XLSX
Microsoft PowerPoint presentations	.PPT, .PPTX
Microsoft Access Databases	.MDB, .ACCDB
WordPerfect documents	.WPD
Adobe Acrobat documents	.PDF
Photographs	.JPG, .PDF
E-mail	.PST, .MSG, .EML [17]
Webpages	.HTML

5. Where feasible, when a party produces reports from databases that can be generated in the ordinary course of business (*i.e.,* without specialized programming skills), these shall be produced in a delimited electronic format preserving field and record structures and names. The parties will meet and confer regarding programmatic database productions, as necessary.

6. Information items that are paper documents or that require redaction shall be produced in static image formats, *e.g.,* single-page .TIF or multipage .PDF images. If an information item contains color, it shall be produced in color unless the color is merely decorative (*e.g.,* company logo or signature block).

7. Individual information items requiring redaction shall (as feasible) be redacted natively or produced in .PDF or .TIF format and redacted in a manner that does not downgrade the ability to electronically search the unredacted portions of the item. The unredacted content of each redacted document should be extracted by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Parties shall take reasonable steps to ensure that text extraction methods produce usable, accurate and complete searchable text.

8. Except as set out in this Protocol, a party need not produce identical information items in more than one form and may globally deduplicate identical items across custodians using each document's unique MD5 or other mutually agreeable hash value. The content, metadata, and utility of an information item shall all be considered in determining whether information items are identical, and items reflecting different information shall not be deemed identical. Parties

[17] Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in RFC 5322 (the Internet e-mail standard). For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished if source foldering metadata is preserved and produced. For Lotus Notes mail, furnish .NSF files or convert messages to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.

may need to negotiate alternate hashing protocols for items (like e-mail) that do not lend themselves to simple hash deduplication.

9. Production should be made using commercially reasonable electronic media of the producing party's choosing, provided that the production media chosen not impose an undue burden or expense upon a recipient.

10. Each information item produced shall be identified by naming the item to correspond to a Bates identifier according to the following protocol:

a. The first four (4) or more characters of the filename will reflect a unique alphanumeric designation identifying the party making production.
b. The next nine (9) characters will be a unique, consecutive numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length.
c. The final six (6) characters are reserved to a sequence beginning with a dash (-) followed by a four (4) or five (5) digit number reflecting pagination of the item when printed to paper or converted to an image format for use in proceedings or when attached as exhibits to pleadings.
d. By way of example, a Microsoft Word document produced by ABC Corporation in its native format might be named: ABCC000000123.docx. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ABCC000000123-00006.

11. Information items designated "Confidential" may, at the Producing Party's option:

a. Be separately produced on electronic production media or in a folder prominently labeled to comply with the requirements of paragraph ___ of the Protective Order entered in this matter; or, alternatively,
b. Each such designated information item shall have appended to the file's name (immediately following its Bates identifier) the following protective legend: ~CONFIDENTIAL-SUBJ TO PROTECTIVE ORDER IN CAUSE MDL-16-0123.

When any "Confidential" item is converted to a printed or imaged format for use in any submission or proceeding, the printout or page image shall bear the protective legend on each page in a clear and conspicuous manner, but not so as to obscure content.

12. The producing party shall furnish a delimited load file supplying the metadata field values listed below for each information item produced (to the extent the values exist and as applicable):

Field
BeginBates
EndBates
BeginAttach
EndAttach

Custodian/Source
Source File Name
Source File Path
From/Author
To
CC
BCC
Date Sent
Time Sent
Subject/Title
Last Modified Date
Last Modified Time
Document Type
Redacted Flag (yes/no)
Hidden Content/Embedded Objects Flag (yes/no)
Confidential flag (yes/no)
E-mail Message ID
E-mail Conversation Index
Parent ID
MD5 or other mutually agreeable hash value
Hash De-Duplicated Instances (by full path)

13. Each production should include a cross-reference load file that correlates the various files, images, metadata field values and searchable text produced

Deep Diving into Deduplication

In the 2008 BBC 6-part series *Stephen Fry in America,* Stephen Fry, the wry English entertainer races about all fifty U.S. states in his trademark London cab. In Boston, Fry discussed contradictions in the American character with the late Peter Gomes, a pastor and Harvard professor of divinity who Fry described as "a black, gay, Republican Baptist." Gomes observed that, "One of the many things one can say about this country is that we dislike complexity, so we will make simple solutions to everything that we possibly can, even when the complex answer is obviously the correct answer or the more intriguing answer. We want a simple 'yes' or 'no,' or a flat out 'this' or an absolutely certain 'that.'"

Gomes wasn't talking about electronic discovery, but he could have been.

For a profession that revels in convoluted codes and mind-numbing minutiae, lawyers and judges are queerly alarmed at the complexity and numerousity of ESI. They speak of ESI only in terms that underscore its burden, never extoling its benefits. They demand simple solutions without looking beyond the (often misleading) big numbers to recognize that the volume they vilify is mostly just the same stuff, replicated over and over again. It's a sad truth that much of the time and money expended on e-discovery in the U.S. is wasted on lawyers reviewing duplicates of information that could have been easily, safely and cheaply culled from the collection. Sadder still, the persons best situated to eradicate this waste are the ones most enriched by it.

The oft-overlooked end of discovery is proving a claim or defense in court. So, the great advantage of ESI is its richness and revealing character. It's *better* evidence in the sense of its more-candid content and the multitude of ways it sheds light on attitudes and actions. Another advantage of ESI is the ease with which it can be disseminated, collected, searched and deduplicated. This post is about deduplication, and why it might be attorney malpractice not to understand it well and use it routinely.

A decade or three ago, the only way to know if a document was a copy of something you'd already seen was to look at it again…and again…and again. It was slow and sloppy; but, it kept legions of lawyers employed and minted fortunes in fees for large law firms.

With the advent of electronic document generation and digital communications, users eschewed letters and memos in favor of e-mail messages and attachments. Buoyed by fast, free e-mail, paper missives morphed into dozens of abbreviated exchanges. Sending a message to three or thirty recipients was quick and cheap. No photocopies, envelopes or postage were required, and the ability to communicate without the assistance of typists, secretaries or postal carriers extended the work day.

But we didn't start doing much more *unique* work. That is, human productivity didn't burgeon, and sunsets and sunrises remained about 12 hours apart. In the main, we merely projected

smaller slices of our work into more collections. And, I suspect any productivity gained from the longer workday was quickly surrendered to the siren song of eBay or Facebook.

Yes, there *is* more stuff. *Deduplication alone is not a magic bullet*. But there is not *as much* more stuff as the e-discovery doomsayers suggest. Purged of replication and managed sensibly with capable tools, ESI volume is still quite wieldy.

And that's why I say a lot of the fear and anger aimed at information inflation is misplaced. If you have the tools and the skills to collect the relevant conversation, avail yourself of the inherent advantages of ESI and eradicate the repetition, e-discovery is just…discovery.

Some organizations imagine they've dodged the replication bullet through the use of single-instance archival storage solutions. But were they to test the true level of replication in their archives, they'd be appalled at how few items actually exist as single instances. In their messaging systems alone, I'd suggest that upwards of a third of the message volume are duplicates despite single instance features. In some collections, forty percent wouldn't surprise me.

But in e-discovery—and especially in that platinum-plated phase called "attorney review"—just how much replication is too much, considering that replication risk manifests not only as wasted time and money but also as inconsistent assessments? Effective deduplication isn't something competent counsel may regard as being optional. I'll go further*: Failing to deduplicate substantial collections of ESI before attorney review is tantamount to cheating the client*.

Just because so many firms have gotten away with it for so long doesn't make it right.

I've thought more about this of late as a consequence of a case where the producing party sought to switch review tools and couldn't figure out how to exclude the items they'd already produced from the ESI they were loading to the new tool. This was a textbook case for deduping, because no one benefits by paying lawyers to review items already reviewed and produced; no one, that is, but the producing party's counsel, who was unabashedly gung-ho to skip deduplication and jump right to review.

I pushed hard for deduplication before review. This isn't altruism; responding parties aren't keen to receive a production bloated by stuff they'd already seen. Replication wastes the recipient's time and money, too.

The source data were Outlook .PSTs from various custodians, each under 2GB in size. The form of production was single messages as .MSGs. Reportedly, the new review platform (actually a rather old concept search tool) was incapable of accepting an overlay load file that could simply tag the items already produced, so the messages already produced would have to be culled from the .PSTs before they were loaded. Screwy, to be sure; but, we take our cases as they come, right?

A somewhat obscure quirk of the .MSG message format is that when the same Outlook message is exported as an .MSG at different times, each exported message generates a different hash value because of embedded time of creation values. [A hash value is a unique digital "fingerprint" that can be calculated for any digital object to facilitate authentication, identification and deduplication]. The differing hash values make it impossible to use hashes of .MSGs for deduplication without processing (i.e., normalizing) the data to a format better suited to the task.

Here, a quick primer on deduplication might be useful.

Mechanized deduplication of ESI can be grounded on three basic approaches:

1. Hashing the ESI as a file (i.e., a defined block of data) containing the ESI using the same hash algorithm (e.g., MD5 or SHA1) and comparing the resulting hash value for each file. If they match, the files hold the same data. This tends not to work for e-mail messages exported as files because, when an e-mail message is stored as a file, messages that we regard as identical in common parlance (such as identical message bodies sent to multiple recipients) are not identical in terms of their byte content. The differences tend to reflect either variations in transmission seen in the message header data (the messages having traversed different paths to reach different recipients) or variations in time (the same message containing embedded time data when exported to single message storage formats as discussed above with respect to the .MSG format).

2. Hashing segments of the message using the same hash algorithm and comparing the hash values for each corresponding segment to determine relative identicality. With this approach, a hash value is calculated for the various parts of a message (e.g., Subject, To, From, CC, Message Body, and Attachments) and these values are compared to the hash values calculated against corresponding parts of other messages to determine if they match. This method requires exclusion of those parts of a message that are certain to differ (such as portions of message headers containing server paths and unique message IDs) and normalization of segments, so that contents of those segments are presented to the hash algorithm in a consistent way.
3. Textual comparison of segments of the message to determine if certain segments of the message match to such an extent that the messages may be deemed sufficiently "identical" to allow them to be treated as the same for purposes of review and exclusion. This is much the same approach as (2) above, but without the use of hashing as a means to compare the segments.

Arguably, a fourth approach entails a mix of these methods.

These approaches can be frustrated by working from differing forms of the "same" data because, from the standpoint of the tools which compare the information, the forms are significantly

different. Thus, if a message has been 'printed' to a TIFF image, the bytes which make up the TIFF image bear no digital resemblance to the bytes which comprise the corresponding e-mail message, any more than a photo of a rose smells or feels like the rose.

In short, changing forms of ESI changes data, and changing data changes hash values. Deduplication by hashing requires the same source data and the same algorithms be employed in a consistent way. This is easy and inexpensive to accomplish, but requires that a compatible work flow be observed to ensure that evidence is not altered in processing so as to prevent the application of simple and inexpensive mechanized deduplication.

When parties cannot deduplicate e-mail, the reasons will likely be one or more of the following:

1. They are working from different forms of the ESI;
2. They are failing to consistently exclude inherently non-identical data (like message headers and IDs) from the hash calculation;
3. They are not properly normalizing the message data (such as by ordering all addresses alphabetically without aliases);
4. They are using different hash algorithms;
5. They are not preserving the hash values throughout the process; or
6. They are changing the data.

Once I was permitted to talk to the sensible technical personnel on the other side, it was clear there were several ways to skin this cat and exclude the items already produced from further review. It would require use of a tool that could more intelligently hash the messages, and not as a monolithic data block; but, there several such tools extant. Because the PSTs were small (each under 2GB), the tool I suggested would cost the other side only $100.00 (or about ten Big Law billing minutes). I wonder how many duplicates must be excluded from review to recoup that princely sum?

Deduplication pays big dividends even in imperfect implementations. Any duplicate that can be culled is time and money saved at multiple points in the discovery process, and deduplication delivers especially big returns when accomplished before review. Deduplication is not a substitute for processes like predictive coding or enhanced search that also foster significant savings and efficiencies; but, few other processes allow users to reap rewards as easily, quickly or cheaply as effective deduplication.

Deduplication: Why Computers See Differences in Files that Look Alike

An employee of an e-discovery service provider asked me to help him explain to his boss why deduplication works well for native files but frequently fails when applied to TIFF images. The question intrigued me because it requires we dip our toes into the shallow end of cryptographic hashing and dispel a common misconception about electronic documents.

Most people regard a Word document file, a PDF or TIFF image made from the document file, a printout of the file and a scan of the printout as being essentially "the same thing." Understandably, they focus on content and pay little heed to form. But when it comes to electronically stored information, the form of the data—the structure, encoding and medium employed to store and deliver content--matters a great deal. As data, a Word document and its imaged counterpart are radically different data streams from one-another and from a digital scan of a paper printout. *Visually*, they are alike when viewed as an image or printout; but *digitally*, they bear not the slightest resemblance.

Having just addressed the challenge of deduplicating e-mail messages, let's look at the same issue with respect to word processed documents and their printed and imaged counterparts.

I'll start by talking about hashing, as a quick refresher (read on, if you just can't stand to have me explain hashing again); then, we will look at how hashing is used to deduplicate files and wrap up by examining examples of the "same" data in a variety of common formats seen in e-discovery and explore why they will and won't deduplicate. At that point, it should be clear why deduplication works well for native files but frequently fails when applied to TIFF images.

Hashing

We spend a considerable time here learning that all ESI is just a bunch of numbers. The readings and exercises about Base2 (binary), Base10 (decimal), Base16 (hexadecimal) and Base64; as well as about the difference between single-byte encoding schemes (like ASCIII) and double-byte encoding schemes (like Unicode) may seem like a wonky walk in the weeds; but the time is well spent if you make the crucial connection between numeric encoding and our ability to use math to cull, filter and cluster data. It's a necessary precursor to their gaining Proustian "new eyes" for ESI.

Because ESI is just a bunch of numbers, we can use algorithms (mathematical formulas) to distill and compare those numbers. In e-discovery (as I hope you are coming to see), one of the most used and –useful family of algorithm are those which manipulate the very long numbers that comprise the content of files (the "message") to generate a smaller, fixed length value called a "Message Digest" or "hash value." This now familiar calculation process is called "hashing," and

the most common hash algorithms in use in e-discovery are **MD5** (for Message Digest five) and **SHA-1** (for Secure Hash Algorithm one).

From the preceding exercises, we've seen that, using hash algorithms, any volume of data—from the tiniest file to the contents of entire hard drives and beyond—can be uniquely expressed as an alphanumeric sequence of fixed length. When I say, "fixed length," I mean that no matter how large or small the volume of data in the file, the hash value computed will (in the case of MD5) be distilled to a value written as 32 hexadecimal characters (0-9 and A-F). Now that you've figured out Base16, you appreciate that those 32 characters represent 340 *trillion, trillion, trillion* different possible values (2^{128} or 16^{32}).

Being one-way calculation, a hash value identifies a sequence of data but reveals nothing about the data; much as a fingerprint uniquely identifies an individual but reveals nothing about their appearance or personality.[18]

Hash algorithms are simple in their operation: a number is inputted (and here, the "number" might be the contents of a file, a group of files, i.e., all files produced to the other side, or the contents of an entire hard drive or server storage array), and a value of fixed length emerges at a speed commensurate with the volume of data being hashed.

Hashing for Deduplication

A modern hard drive holds trillions of bytes, and even a single Outlook e-mail container file typically comprises billions of bytes. Accordingly, it's easier and faster to compare 32-character/16 byte "fingerprints" of voluminous data than to compare the data itself, particularly as the comparisons must be made repeatedly when information is collected and processed in e-discovery. In practice, each file ingested and item extracted is hashed and its hash value compared to the hash values of items previously ingested and extracted to determine if the file or item has been seen before. The first file is sometimes called the "pivot file," and subsequent files with matching hashes are suppressed as duplicates, and the instances of each duplicate and certain metadata is typically noted in a deduplication or "occurrence" log.

When the data is comprised of loose files and attachments, a hash algorithm tends to be applied to the full contents of the files. Notice that I said to "*contents*." Recall that some data we associate with files is not actually stored inside the file but must be gathered from the file system of the device storing the data. Such "system metadata" is not contained within the file and, thus, is not included in the calculation when the file's content is hashed. A file's name is perhaps the best example of this. Recall that even slight differences in files cause them to generate different

[18] There's more to say on this issue; so, if you are really into this, search Google or Wikipedia for "rainbow tables."

hash values. But, since a file's name is not typically housed within the file, you can change a file's name without altering its hash value.

So, the ability of hash algorithms to deduplicate depends upon whether the numeric values that serve as building blocks for the data differ from file-to-file. Keep that firmly in mind as we consider the many forms in which the informational payload of a document may manifest.

A Word .DOCX document is constructed of a mix of text and rich media encoded in Extensible Markup Language (XML), then compressed using the ubiquitous Zip compression algorithm. It's a file designed to be read by Microsoft Word.

When you print the "same" Word document to an Adobe PDF format, it's reconstructed in a *page description language* specifically designed to work with Adobe Acrobat. It's structured, encoded and compressed in an entirely different way than the Word file and, as a different format, carries a different binary header signature, too.

When you take the printed version of the document and scan it to a Tagged Image File Format (TIFF), you've taken a picture of the document, now constructed in still another different format—one designed for TIFF viewer applications.

To the uninitiated, they are all the "same" document and might look pretty much the same printed to paper; but as ESI, their structures and encoding schemes are radically different. Moreover, even files generated in the same format may not be *digitally* identical when made at different times. For example, no two optical scans of a document will produce identical hash values because there will always be some variation in the data acquired from scan to scan. Small differences perhaps; but, any difference at all in content is going to frustrate the ability to generate matching hash values.

Opinions are cheap; testing is truth; so to illustrate this, I created a Word document of the text of Lincoln's Gettysburg Address. First, I saved it in the latest .DOCX Word format. Then, I saved a copy in the older .DOC format. Next, I saved the Word document to a .PDF format, using both the Save as PDF and Print to PDF methods. Finally, I printed and scanned the document to TIFF and PDF. Without shifting the document on the scanner, I scanned it several times at matching and differing resolutions.

I then hashed all the iterations of the "same" document and, as the table below demonstrates, none of them matched hash wise, not even the successive scans of the paper document:

FILENAME	MD5 HASH	FILE SIZE
GBA.docx	5074fbb210ed4e9e498e4908a946a871	21Kb
GBA.doc	1aacf60b523eb8cf2829208ffee58005	26Kb
GBA-Save as.pdf	c8d68e84ea573772d14dc536fbe8594e	83Kb
GBA-Word generated.pdf	2be09d776682fee46c79be8ecac03ec5	27Kb
GBA-scan1.tiff	0f5fdbbcbc96abc05b43f356c4e24818	967Kb
GBA-scan2.tiff	04c93ac7eb6716bc96bc3a396fed882a	967Kb
GBA-scan3_600BW.tiff	93e726efa56fe7f25956da6664a32957	1,060Kb
GBA-scan4_600BW.tiff	8d97df97c28414d4b61bb8b88b1db343	1,060Kb
GBA_scan5_300GS.pdf	b558eccee1bdcc5f26de53763f89aef4	2,950Kb
GBA_scan6_300GS.pdf	520be78a7ec81ebebece5a19e9c6e425	2,930Kb

Thus, file hash matching--the simplest and most defensible approach to deduplication--won't serve to deduplicate the "same" document when it takes different forms or is made optically at different times.

Now, here's where it can get confusing. If you copied any of the electronic *files* listed above, the duplicate files <u>would</u> hash match the source originals, and would handily deduplicate by hash. Consequently, multiple copies of the same electronic files will deduplicate, but that is because the files being compared have the same *digital* content. But, we must be careful to distinguish the identicality seen in multiple iterations of the same file from the pronounced differences seen when different electronic versions are generated at different times from the same content. One notable exception seen in my testing was that successively saving the same Word document to a PDF format in the same manner sometimes generated identical PDF files. It didn't occur consistently (i.e., if enough time passed, changes in metadata in the source document triggered differences prompting the calculation of different hash values); but it happened, so was worth mentioning.

Mastering E-Mail in Discovery

Introduction

Get the e-mail! It's long been the war cry in e-discovery. It's a recognition of e-mail's enduring importance and ubiquity. We go after e-mail because it accounts for the majority of business communications and because, despite years of cautions and countless headlines tied to e-mail improvidence, e-mail users still let their guards down and reveal plainspoken truths they'd never put in a memo.

If you're on the producing end of a discovery request, you not only worry about what the messages say, but also whether you and your client can find, preserve and produce all responsive items. Questions like these *should* keep you up nights:

- Will the client simply conceal damning messages, leaving counsel at the mercy of an angry judge or disciplinary board?
- Will employees seek to rewrite history by deleting "their" e-mail from company systems?
- Will the searches employed prove reliable and be directed to the right digital venues?
- Will review processes unwittingly betray privileged or confidential communications?

Meeting these challenges begins with understanding e-mail technology well enough to formulate a sound, defensible strategy. For requesting parties, it means grasping the technology well enough to assess the completeness and effectiveness of your opponent's e-discovery efforts.

Not Enough Eyeballs

Futurist Arthur C. Clarke said, "Any sufficiently advanced technology is indistinguishable from magic." E-mail, like television or refrigeration, is one of those magical technologies we use every day without really knowing how it works. "It's magic to me, your Honor," won't help you when the e-mail pulls a disappearing act. Judges expect you to pull that e-mail rabbit out of your hat.

A lawyer managing electronic discovery is obliged to do more than just tell their clients to "produce the e-mail." The lawyer must endeavor to understand the client's systems and procedures, as well as ask the right questions of the right personnel. Too, counsel must know when he or she isn't getting trustworthy answers. That's asking a lot, but virtually all business documents are born digitally and only a tiny fraction are ever printed.[19] Hundreds of billions of e-

[19] Extrapolating from a 2003 updated study compiled by faculty and students at the School of Information Management and Systems at the University of California at Berkeley. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ (visited 5/18/2013)

mails traverse the Internet *daily,* far more than telephone and postal traffic combined,[20] and the average business person sends and receives roughly 123 e-mails daily. And the e-mail volumes continue to grow even as texting and other communications channels have taken off.

Neither should we anticipate a significant decline in users' propensity to retain their e-mail. Here again, it's too easy and, at first blush, too cheap to expect users to selectively dispose of e-mail and still meet business, litigation hold and regulatory obligations. Our e-mail is so twisted up with our lives that to abandon it is to part with our personal history.

This relentless growth isn't happening in just one locale. E-mail lodges on servers, cell phones, laptops, home systems, thumb drives and in the cloud. Within the systems, applications and devices we use to store and access e-mail, most users and even most IT professionals don't know where messages lodge or exactly how long they hang around.

Test Your E.Q.

Suppose opposing counsel serves a preservation demand or secures an order compelling your client to preserve electronic messaging. Are you assured that your client can and will faithfully back up and preserve responsive data? Even if it's practicable to capture and set aside the current server e-mail stores of key custodians, are you *really* capturing all or even most of the discoverable communications? How much is falling outside your net, and how do you assess its importance?

Here are a dozen questions you should be able to confidently answer about your client's communication systems:

1. What messaging environment(s) does your client employ? Microsoft Exchange, IBM Domino, Office 365 or something else?
2. Do *all* discoverable electronic communications come in and leave via the company's e-mail server?
3. Is the e-mail system configured to support synchronization with local e-mail stores on laptops and desktops?
4. How long have the current e-mail client and server applications been used?
5. What are the message purge, retention, journaling and archival settings for each key custodian?
6. Can your client disable a specific custodian's ability to delete messages?
7. Does your client's backup or archival system capture e-mail stored on individual user's hard drives, including company-owned laptops?

[20] http://www.radicati.com/wp/wp-content/uploads/2013/04/Email-Statistics-Report-2013-2017-Executive-Summary.pdf (visited 5/26/2016)

8. Where are e-mail container files stored on laptops and desktops?
9. How should your client collect and preserve relevant web mail?
10. Do your clients' employees use home machines, personal e-mail addresses or browser-based e-mail services like Gmail for discoverable business communications?
11. Do your clients' employees use instant messaging on company computers or over company-owned networks?
12. Does your client permit employee-owned devices to access the network or e-mail system?

If you are troubled that you can't answer these questions, you should be; but know you're not alone. Despite decades of dealing with e-mail in discovery, most lawyers still can't. And if you're a lawyer, don't delude yourself that these are someone else's issues, *e.g.,* your litigation support people or IT expert. These are your issues when it comes to dealing with the other side and the court about the scope of e-discovery.

Staying Out of Trouble

Fortunately, the rules of discovery don't require you to do the impossible. All they require is diligence, reasonableness and good faith. To that end, you must be able to establish that you and your client acted swiftly, followed a sound plan, and took such action as reasonable minds would judge adequate to the task. It's also important to keep the lines of communication open with the opposing party and the court, seeking agreement with the former or the protection of the latter where fruitful. I'm fond of quoting Oliver Wendell Holmes' homily, "Even a dog knows the difference between being stumbled over and being kicked." Judges, too, have a keen ability to distinguish error from arrogance. There's no traction for sanctions when the failure to produce electronic evidence occurred despite good faith and due diligence.

…And You Could Make Spitballs with It, Too

Paper discovery enjoyed a self-limiting aspect because businesses tended to allocate paper records into files, folders and cabinets according to persons, topics, transactions or periods of time. The space occupied by paper and the high cost to create, manage and store paper records served as a constant impetus to cull and discard them, or even to avoid creating them in the first place. By contrast, the ephemeral character of electronic communications, the ease of and perceived lack of cost to create, duplicate and distribute them and the very low direct cost of data storage have facilitated a staggering and unprecedented growth in the creation and retention of electronic evidence. At 123 e-mails per day, a company employing 100,000 people could find itself storing almost 4.*5 billion* e-mails annually.

Did You Say *Billion*?

But volume is only part of the challenge. Unlike paper records, e-mail tends to be stored in massive data blobs. My e-mail comprises almost 25 gigabytes of data and contains over 100,000 messages, many with multiple attachments covering virtually every aspect of my life and many other people's lives, too. In thousands of those e-mails, the subject line bears only a passing connection to the contents as "Reply to" threads strayed further and further from the original topic. E-mails meander through disparate topics or, by absent-minded clicks of the "Forward" button, lodge in my inbox dragging with them, like toilet paper on a wet shoe, the unsolicited detritus of other people's business.

To respond to a discovery request for e-mail on a topic, I'd either need to skim/read a horrific number of messages or I'd have to naively rely on keyword search to flush out all responsive material. If the request for production implicated material I no longer kept on my current computer or web mail collections, I'd be forced to root around through a motley array of archival folders, old systems, obsolete disks, outgrown hard drives, ancient backup tapes (for which I currently have no tape reader) and unlabeled CDs. Ugh!

Net Full of Holes

I'm just one guy. What's a company to do when served with a request for "all e-mail" on a matter in litigation? Surely, I mused, someone must have found a better solution than repeating the tedious and time-consuming process of accessing individual e-mail servers at far-flung locations along with the local drives of all key players' computers?

In researching this text, I contacted colleagues in both large and small electronic discovery consulting groups, inquiring about "the better way" for enterprises, and was struck by the revelation that, if there was a better mousetrap, they hadn't discovered it either. Uniformly, we recognized such enterprise-wide efforts were gargantuan undertakings fraught with uncertainty and concluded that counsel must somehow seek to narrow the scope of the inquiry—either by data sampling, use of advanced analytics or through limiting discovery according to offices, regions, time span, business sectors or key players. Trying to capture *everything,* enterprise-wide, is trawling with a net full of holes.

New Tools

The market has responded in recent years with tools that either facilitate search of remote e-mail stores, including locally stored messages, from a central location (*i.e.,* enterprise search) or which agglomerate enterprise-wide collections of e-mail into a single, searchable repository (*i.e.,* e-mail archiving), often reducing the volume of stored data by so-called "single instance deduplication," rules-based journaling and other customizable features.

These tools, especially enterprise archival and advanced analytics termed "TAR" or "Predictive Coding," promise to make it easier, cheaper and faster to search and collect responsive e-mail, but they're costly and complex to implement. Neither established standards nor a leading product has emerged. Further, it remains to be seen whether the practical result of a serial litigant employing an e-mail archival system is that they—for all intents and purposes--end up keeping every message for every employee and becoming increasingly dependent upon fraught electronic search to cull wheat from chaff.

E-Mail Systems and Files

The "behind-the-firewall" corporate and government e-mail environment is dominated by two well-known, competitive product pairs: Microsoft Exchange Server and its Outlook e-mail client and IBM Lotus Domino server and its Lotus Notes client. A legacy environment called Novell GroupWise occupies a negligible third place, largely among government users.

Increasingly, corporate and government e-mail environment no longer live behind-the-firewall but are ensconced in the Cloud. Cloud products such as Google Apps and Microsoft Office 365 now account for an estimated 20-25% market shares, with Microsoft claiming that 4 out of 5 Fortune 500 companies use Office 365.

When one looks at personal and small office/home office business e-mail, it's rare to encounter LOCAL server-based systems. Here, the market belongs to Internet service providers (*e.g.,* the major cable and telephone companies) and web mail providers (*e.g.,* Gmail and Yahoo! Mail). Users employ a variety of e-mail client applications, including Microsoft Outlook, Apple Mail and, of course, their web browsers and webmail. This motley crew and the enterprise behemoths are united by common e-mail *protocols* that allow messages and attachments to be seamlessly handed off between applications, providers, servers and devices.

Mail Protocols

Computer network specialists are always talking about this "protocol" and that "protocol." Don't let the geek-speak get in the way. An *application protocol* or API is a bit of computer code that facilitates communication between applications, *i.e.,* your e-mail client and a network like the Internet. When you send a snail mail letter, the U.S. Postal Service's "protocol" dictates that you place the contents of your message in an envelope of certain dimensions, seal it, add a defined complement of address information and affix postage to the upper right-hand corner of the envelope adjacent to the addressee information. Only then can you transmit the letter through the Postal Service's network of post offices, delivery vehicles and postal carriers. Omit the address, the envelope or the postage—or just fail to drop it in the mail—and Grandma gets no Hallmark this year! Likewise, computer networks rely upon protocols to facilitate the transmission

of information. You invoke a protocol—*Hyper Text Transfer Protocol*—every time you type *http://* at the start of a web page address.

Incoming Mail: POP, IMAP, MAPI and HTTP E-Mail

Although Microsoft Exchange Server rules the roost in enterprise e-mail, it's by no means the most common e-mail system for the individual and small business user. If you still access your personal e-mail from your own Internet Service Provider, chances are your e-mail comes to you from your ISP's e-mail server in one of three ways: POP3, IMAP or HTTP, the last commonly called web- or browser-based e-mail. Understanding how these three protocols work—and differ—helps in identifying where e-mail can be found.

POP3 (Post Office Protocol, version 3) is the oldest and was once the most common of the three approaches and the one most familiar (by function, if not by name) to users of the Windows Mail, Outlook Express and Eudora e-mail clients. But, it's rare to see many people using POP3 e-mail today. Using POP3, you connect to a mail server, download copies of all messages and, unless you have configured your e-mail client to leave copies on the server, the e-mail is deleted on the server and now resides on the hard drive of the computer you used to pick up mail. Leaving copies of your e-mail on the server seems like a great idea as it allows you to have a backup if disaster strikes and facilitates easy access of your e-mail, again and again, from different computers. However, few ISPs afforded unlimited storage space on their servers for users' e-mail, so mailboxes quickly became "clogged" with old e-mails, and the servers started bouncing new messages. As a result, POP3 e-mail typically resides only on the local hard drive of the computer used to read the mail and on the backup system for the servers which transmitted, transported and delivered the messages. In short, POP is locally-stored e-mail that supports some server storage; but, again, this once dominant protocol is little used anymore.

IMAP (Internet Mail Access Protocol) functions in much the same fashion as most Microsoft Exchange Server installations in that, when you check your messages, your e-mail client downloads just the headers of e-mail it finds on the server and only retrieves the body of a message when you open it for reading. Else, the entire message stays in your account on the server. Unlike POP3, where e-mail is searched and organized into folders locally, IMAP e-mail is organized and searched on the server. Consequently, the server (and its backup tapes) retains not only the messages but also the way the user *structured* those messages for archival.

Since IMAP e-mail "lives" on the server, how does a user read and answer it without staying connected all the time? The answer is that IMAP e-mail clients afford users the ability to synchronize the server files with a local copy of the e-mail and folders. When an IMAP user reconnects to the server, local e-mail stores are updated (synchronized) and messages drafted

offline are transmitted. So, to summarize, IMAP is server-stored e-mail, with support for synchronized local storage.

A notable distinction between POP3 and IMAP e-mail centers on where the "authoritative" collection resides. Because each protocol allows for messages to reside both locally ("downloaded") and on the server, it's common for there to be a difference between the local and server collections. Under POP3, the *local* collection is deemed authoritative whereas in IMAP the *server* collection is authoritative. But for e-discovery, the important point is that the contents of the local and server e-mail stores can and do *differ.*

MAPI (Messaging Application Programming Interface) is the e-mail protocol at the heart of Windows and Microsoft's Exchange Server applications. Simple MAPI comes preinstalled on Windows machines to provide basic messaging services. A more sophisticated version of MAPI (Extended MAPI) is installed with Microsoft Outlook and Exchange. Like IMAP, MAPI e-mail is typically stored on the server and not necessarily on the client machine. The local machine may be configured to synchronize with the server mail stores and keep a copy of mail on the local hard drive (typically in a Personal Storage file with the extension .PST or an Offline Synchronization file with the extension .OST), but this is user- and client application-dependent. Though it's rare (especially for laptops) for there to be no local e-mail stores for a MAPI machine, it's nonetheless possible and companies have lately endeavored to do away with local e-mail storage on laptop and desktop computers. When machines are configured to bar creation of local PST and OST files, e-mail won't be found on the local hard drive except to the extent fragments may turn up through computer forensic examination.

HTTP (Hyper Text Transfer Protocol) mail, or web-based/browser-based e-mail, dispenses with the local e-mail client and handles all activities on the server, with users managing their e-mail using their Internet browser to view an interactive web page. Although most browser-based e-mail services support local POP3 or IMAP synchronization with an e-mail client, most users have no local record of their browser-based e-mail transactions except for messages they've affirmatively saved to disk or portions of e-mail web pages which happen to reside in the browser's cache (*e.g.,* Internet Explorer's Temporary Internet Files folder). Gmail and Yahoo! Mail are popular examples of browser-based e-mail services, although many ISPs (including all the national providers) offer browser-based e-mail access in addition to POP and IMAP connections.

The protocol used to carry e-mail is not especially important in electronic discovery except to the extent that it signals the most likely place where archived and orphaned e-mail can be found. Companies choose server-based e-mail systems (*e.g.,* IMAP and MAPI) for two principal reasons. First, such systems make it easier to access e-mail from different locations and machines. Second, it's easier to back up e-mail from a central location. Because IMAP and MAPI systems store e-mail

on the server, the backup system used to protect server data can yield a mother lode of server e-mail.

Depending upon the backup procedures used, access to archived e-mail can prove a costly and time-consuming task or a relatively easy one. The enormous volume of e-mail residing on backup tapes and the potentially high cost to locate and restore that e-mail makes discovery of archived e-mail from backup tapes a major bone of contention between litigants. In fact, most reported cases addressing cost-allocation in e-discovery seem to have been spawned by disputes over e-mail on server backup tapes.

Outgoing Mail: SMTP and MTA

Just as the system that brings water into your home works in conjunction with a completely different system that carries wastewater away, the protocol that delivers e-mail to you is completely different from the one that transmits your e-mail. Everything discussed in the preceding paragraphs concerned the protocols used to *retrieve* e-mail from a mail server.

Yet another system altogether, called SMTP for *Simple Mail Transfer Protocol*, takes care of outgoing e-mail. SMTP is indeed a very simple protocol and doesn't even require authentication, in much the same way as anyone can anonymously drop a letter into a mailbox. A server that uses SMTP to route e-mail over a network to its destination is called an MTA for *Message Transfer Agent*. Examples of MTAs you might hear mentioned by IT professionals include Sendmail, Exim, Qmail and Postfix. Microsoft Exchange Server is an MTA, too. In simplest terms, an MTA is the system that carries e-mail between e-mail servers and sees to it that the message gets to its destination. Each MTA reads the code of a message and determines if it is addressed to a user in its domain and, if not, passes the message on to the next MTA after adding a line of text to the message identifying the route to later recipients. If you've ever set up an e-mail client, you've probably had to type in the name of the servers handling your outgoing e-mail (perhaps *SMTP.yourISP.com*) and your incoming messages (perhaps *mail.yourISP.com* or *POP.yourISP.com*).

Anatomy of an E-Mail

Now that we've waded through the alphabet soup of protocols managing the movement of an e-mail message, let's look inside the message itself. Considering the complex systems on which it lives, an e-mail is astonishingly simple in structure. The Internet protocols governing e-mail transmission require electronic messages to adhere to rigid formatting, making individual e-mails easy to dissect and understand. The complexities and headaches associated with e-mail don't really attach until the e-mails are stored and assembled into databases and local stores.

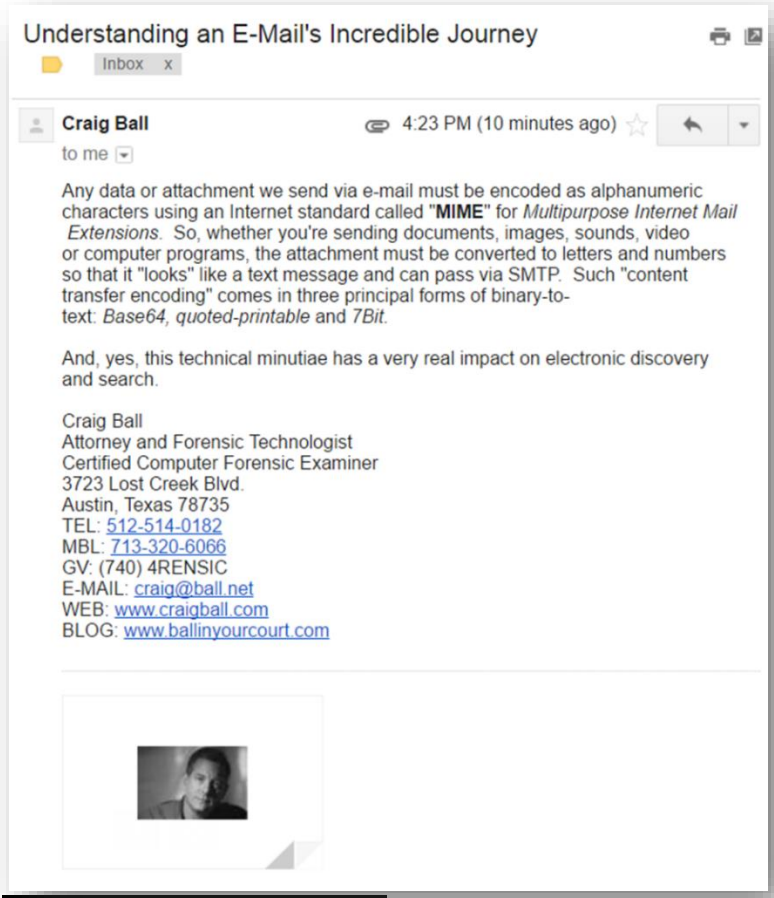
An e-mail is just a plain text file. Though e-mail can be "tricked" into carrying non-text binary data like application files (*i.e.,* a Word document) or image attachments (*e.g.,* GIF or JPEG files), this

piggybacking requires binary data be *encoded into text* for transmission. Consequently, even when transmitting files created in the densest computer code, *everything in an e-mail is plain text*.

E-Mail Autopsy: Tracing a Message's Incredible Journey

The image below left is an e-mail I sent to computerforensics@gmail.com from my alias craig@ball.net using my Gmail account craigball@gmail.com. A tiny JPG photograph was attached. A user might see the e-mail presentment at left and mistakenly assume that what they see is all of the information in the message. Far from it!

The image below right contains the source code of the same e-mail message.[21] Viewed in its "true" and complete format, it's too long to legibly appear on one page. So, let's dissect it by looking at its constituent parts: message header, message body and encoded attachment.



[21] While viewing a Gmail message, you may switch the screen display to show the source code for both the message header and body by selecting "Show original" from the message options drop-down menu. By default, Outlook makes only some encoded header content readily viewable at message Properties—the complete source code of incoming e-mail is not recorded absent a system Registry edit, which is not a casual operation!

MESSAGE HEADER

```
Delivered-To: computerforensics@gmail.com
Received: by 10.129.51.76 with SMTP id z73csp352014ywz;
        Fri, 27 May 2016 14:23:19 -0700 (PDT)
X-Received: by 10.55.116.69 with SMTP id p66mr16076053qkc.129.1464384199340;
        Fri, 27 May 2016 14:23:19 -0700 (PDT)
Return-Path: <craigball@gmail.com>
Received: from mail-qg0-f47.google.com (mail-qg0-f47.google.com. [209.85.192.47])
        by mx.google.com with ESMTPS id 137si18766698qkg.117.2016.05.27.14.23.19
        for <computerforensics@gmail.com>
        (version=TLS1_2 cipher=ECDHE-RSA-AES128-GCM-SHA256 bits=128/128);
        Fri, 27 May 2016 14:23:19 -0700 (PDT)
Received-SPF: pass (google.com: domain of craigball@gmail.com designates 209.85.192.47 as
permitted sender) client-ip=209.85.192.47;
Authentication-Results: mx.google.com;
       dkim=pass header.i=@gmail.com;
       spf=pass (google.com: domain of craigball@gmail.com designates 209.85.192.47 as
permitted sender) smtp.mailfrom=craigball@gmail.com
Received: by mail-qg0-f47.google.com with SMTP id j92so547qga.0
        for <computerforensics@gmail.com>; Fri, 27 May 2016 14:23:19 -0700 (PDT)
 ALyK8tIcImgplC/t/05YlqsL5XZlnGadMdH2YdrrgfI+NSRlzPmBen8BeYNhnLPG/ZwnCIX9uY6s9n/BteHGzA==
MIME-Version: 1.0
X-Received: by 10.140.238.66 with SMTP id j63mr15641024qhc.48.1464384198894;
 Fri, 27 May 2016 14:23:18 -0700 (PDT)
Sender: craigball@gmail.com
Received: by 10.55.209.142 with HTTP; Fri, 27 May 2016 14:23:18 -0700 (PDT)
Date: Fri, 27 May 2016 16:23:18 -0500
X-Google-Sender-Auth: SqZ326PT6-VsMBA9F0hP7iLVyro
Message-ID: <CALckR-as13NxO0q1pb6wSa14V1m-9tY8y093SkrBV8eNMvcZvQ@mail.gmail.com>
Subject: Understanding an E-Mail's Incredible Journey
From: Craig Ball <craig@ball.net>
To: Craig Ball <computerforensics@gmail.com>
Content-Type: multipart/mixed; boundary=001a1135933cfe0c350533d98387
```

In an e-mail header, each line beginning with "Received" or X-Received" represents the transfer of the message between two e-mail servers. The transfer sequence is reversed chronologically such that those closest to the top of the header were inserted after those that follow, and the topmost line reflects delivery to the recipient's e-mail server and account, in this instance, *computerforensics@gmail.com*. As the message passes through intervening hosts, each adds its own identifying information along with the date and time of transit.

The area of the header labeled **(A)** contains the parts of the message designating the sender, addressee, date, time and subject line of the message. These are the only features of the header most recipients ever see. Note that the 24-hour message time has been recast as to a 12-hour format when shown in Gmail.

In the line labeled "Date," both the date and time of transmittal are indicated. The time indicated is 16:23:18, and the "-0500" which follows denotes the time difference between the sender's local time (the system time on my computer in New Orleans, Louisiana during daylight savings time) and Coordinated Universal Time (UTC), roughly equivalent to Greenwich Mean Time. As the offset from UTC was minus five hours on May 27, 2016, we deduce that the message was sent from a

machine set to Central Daylight Time, giving some insight into the sender's location. Knowing the originating computer's time and time zone can occasionally prove useful in demonstrating fraud or fabrication.

E-mail must adhere to structural conventions. One of these is the use of a Content-Type declaration and setting of content boundaries, enabling systems to distinguish the message header region from the message body and attachment regions. The line labeled **(B)** advises that the message will be "multipart/mixed," indicating that there will be multiple constituents to the item (*i.e.,* header/message body/attachment), and that these will be encoded in different ways, hence "mixed." To prevent confusion of the boundary designator with message text, a complex sequence of characters is generated to serve as the content boundary. The first boundary, declared as "001a1135933cfe0c350533d98387," serves to separate the message header from the message body and attachment. It also signals the end of the message.

The message was created and sent using Gmail web interface; consequently, the first hop **(C)** indicates that the message was transmitted using HTTP and first received by IP (Internet Protocol) address 10.55.209.142 at 14:23:18 -0700 (PDT). Note that the server marks time in Pacific Daylight Time, suggesting it may be located on the west coast. The message is immediately handed off to another IP address 10.140.238.66 using Simple Mail Transfer Protocol, denoted by the initials SMTP. Next, we see another SMTP hand off to Google's server named "mail-qg0-f47.google.com" and so on until delivery to my account, *computerforensics@gmail.com*.

In the line labeled **(D),** the message header declares the message as being formatted in MIME (MIME-Version: 1.0).[22] Ironically, there is no other version of MIME than 1.0; consequently, trillions of e-mails have dedicated vast volumes of storage and bandwidth to this useless version declaration.

Proceeding to dissect the message body seen on the next page, at line **(E)**, we see our first boundary value (--001a1135933cfe0c350533d98387) serving to delineate the transition from header to message body. At line **(F)**, another Content-Type declaration advises that this segment of the message will be multipart/alternative (the alternatives being plain text or HTML) and a second boundary notation is declared as 001a1135933cfe0c350533d98385. Note that the first boundary ends in 387 and the second in 385. The second boundary is used at **(G)** to mark the start of the first alternative message body, declared as text/plain at line **(H)**.in plain text.

[22] MIME, which stands for Multipurpose Internet Mail Extensions, is a seminal Internet standard that supports non-US/ASCII character sets, non-text attachments (e.g., photos, video, sounds and machine code) and message bodies with multiple parts. Virtually all e-mail today is transmitted in MIME format.

We then see the second boundary value used at line **(I)** to denote the start of the second alternative message body, and the Content-Type declared to be text/html at line **(J).** The second boundary notation is then used to signal the conclusion of the multipart/alternative content.

MESSAGE BODY

```
--001a1135933cfe0c350533d98387 E
Content-Type: multipart/alternative; boundary=001a1135933cfe0c310533d98385 F

--001a1135933cfe0c310533d98385 G
Content-Type: text/plain; charset=UTF-8 H

Any data or attachment we send via e-mail must be encoded as alphanumeric
characters using an Internet standard called "*MIME*" for
*Multipurpose Internet Mail Extensions*.  So, whether you're sending
documents, images, sounds, video or computer programs, the attachment
must be converted to letters and numbers so that it "looks" like a text
message and can pass via SMTP.  Such "content transfer encoding" comes in
three principal forms of binary-to-text: *Base64,* *quoted-printable *and
* 7Bit*.

And, yes, this technical minutiae has a very real impact on electronic
discovery and search.

Craig Ball
Attorney and Forensic Technologist
Certified Computer Forensic Examiner
3723 Lost Creek Blvd.
Austin, Texas 78735
TEL: 512-514-0182
MBL: 713-320-6066
GV: (740) 4RENSIC
E-MAIL: craig@ball.net
WEB: www.craigball.com
BLOG: www.ballinyourcourt.com

--001a1135933cfe0c310533d98385 I
Content-Type: text/html; charset=UTF-8 J
Content-Transfer-Encoding: quoted-printable

<div dir=3D"ltr"><div style=3D"font-size:12.8px">Any data or attachment we =
send via e-<span class=3D"">mail</span>=C2=A0must be encoded as alphanumeri=
c characters using an=C2=A0<span class=3D"">Internet</span>=C2=A0standard c=
alled &quot;<strong>MIME</strong>&quot; for=C2=A0<em><span class=3D"">Multi=
purpose</span>=C2=A0<span class=3D"">Internet</span>=C2=A0<span class=3D"">=
Mail=C2=A0</span><span class=3D"">Extensions</span></em>.=C2=A0=C2=A0So, wh=
ether you&#39;re sending documents, images, sounds, video or=C2=A0computer =
programs, the attachment must=C2=A0be converted to letters and numbers so t=
hat it &quot;looks&quot; like a text message and can pass via SMTP.=C2=A0 S=
uch &quot;content transfer encoding&quot;=C2=A0comes in three principal for=
ms of binary-to-text:=C2=A0<em>Base64,</em>=C2=A0<em>quoted-printable=C2=A0=
</em>and<em>=C2=A07Bit</em>.</div><div style=3D"font-size:12.8px">=C2=A0</d=
iv><div style=3D"font-size:12.8px">And, yes, this technical minutiae has a =
very real impact on electronic discovery and search.</div><div><div class=
=3D"gmail_signature" data-smartmail=3D"gmail_signature"><div dir=3D"ltr"><d=
iv><div>=C2=A0</div><div>Craig Ball<br>Attorney and Forensic Technologist<b=
r>Certified Computer Forensic Examiner<br>3723 Lost Creek Blvd.<br>Austin, =
Texas 78735<br><span style=3D"font-size:12.8px">TEL: 512-514-0182</span><br=
 style=3D"font-size:12.8px"><span style=3D"font-size:12.8px">MBL: 713-320-6=
066</span></div><div>GV: (740) 4RENSIC<br>E-MAIL: <a href=3D"mailto:craig@b=
all.net" target=3D"_blank">craig@ball.net</a><br>WEB: <a href=3D"http://www=
.craigball.com" target=3D"_blank">www.craigball.com</a></div><div>BLOG: <a =
href=3D"http://www.ballinyourcourt.com" target=3D"_blank">www.ballinyourcou=
rt.com</a></div></div></div></div></div>
</div>

--001a1135933cfe0c310533d98385– K
```

I didn't draft the message in *either* plain text or HTML formats, but my e-mail service thoughtfully did both to insure that my message won't confuse recipients using (very old) e-mail software

unable to display the richer formatting supported by HTML. For these recipients, the plain text version gets the point across, albeit sans the bolding, italics, hyperlinks and other embellishments of the HTML version.

Turning to the last segment of the message, we see, at **(L)**, the transition between the message body and the attachment segments commemorated by our old friend 387, the first boundary notation.

ATTACHMENTS

```
--001a1135933cfe0c350533d98387
Content-Type: image/jpeg; name="Ball-photo_76x50 pixels_B&W.jpg"
Content-Disposition: attachment; filename="Ball-photo_76x50 pixels_B&W.jpg"
Content-Transfer-Encoding: base64
X-Attachment-Id: f_ioq8ehs40

/9j/4AAQSkZJRgABAQEAYABgAAD/2wBDAAYEBQYFBAYGBQYHBwYIChAKCgkJChQODwwQFxQYGBcU
FhYaHSUfGhsjHBYWICwgIyYnKSopGR8tMC0oMCUoKSj/2wBDAQcHBwoIChMKChMoGhYaKCgoKCgo
KCgoKCgoKCgoK
```

represented by the Latin alphabet.[23] Note the declaration in **(M)**, "**Content-Transfer-Encoding: base64**."

Accordingly, the attached JPEG photograph with the filename "Ball-photo_76x50 pixels_B&W.jpg," has been encoded from non-printable binary code into those 26 lines of apparent gibberish comprising nearly 2,000 plain text characters **(N).** It's now able to traverse the network as an e-mail, yet easily be converted back to binary data when the message reaches its destination.

Figure 3

```
/9j/4AAQSkZJRgABAQEAYABgAAD/2wBDAAYEBQYFBAYGBQYHBwYIChAKCgkJChQODwwQFxQYGBcU
FhYaHSUfGhsjHBYWICwgIyYnKSopGR8tMC0oMCUoKSj/2wBDAQcHBwoIChMKChMoGhYaKCgoKCgo
KCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgo
AhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQA
AAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3
ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWm
p6ipqrKztLW2t7i5usLDxMXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6erx8vP09fb3+Pn6/8QAHwEA
AwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAQAAQJ3AAECAxEEBSEx
BhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElK
U1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3
uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwDzOK0c
scluvrWpaWeMcn86kiT5uner8CYxQBYsYPc12umWBk0zeqn61y1iPnrstOvjBprR46DAoA5bxRrN
vodp5swLsDtSMHl2/wAK8s1bxPrusyMXupYYP4YYSUUD8Ov410vi6GTWPFKWrNiGJQox/ePJrr9P
8LW5tdkkKbVXABUUAeHtf3afeuZW/wB5if51t+Hb77U7I3Eg7Z61reL/AAxZ28rNAhQnn5Tx+VcT
YzNp+oxyIcqG2n3FAHt3gOOw51WNHGRmvbLHTUjjZYyQu7OPwFeRfDhgurxHAO4cV7jaAqjDb/F/
QUAfLsRGT9avw8jpUEOwE4Her0AzigC7YqTIOK3o0eRRFH94j8qybJcMCa6jw/bl5C5XIPFAGLp3
hlDrt1fqWmQLuQMOr4wT+nH1qxDqt2mrDT2tJH39d0RQqCOx5B/SuwOlFhmYE8Lx83XFLNJZi7Mx
CrNyqO5wMAZOPwoA8Y8e3UEbT2u4falLAr1KjPBNeVRWzXF1FECMs/UV0ep6oJ/GGty4LNcu4jdH
I2kHPX04rIOlJX1WAKQGDFmI5GKAPYPCEv2e6t3BxtYV7jpWqpcW7OpyA2POFeF+GwpB3dhxXo3h
K7K6Y6nBxK38hQB5BD1P1q/BxyxwB61mRy8sE/OpAzy27MmSynp7igDYi1a1jYKiyzEdfLTj8zxW
vceOI9MOK7SztJBfOpSJ2xhGPQ/hkvzlqfNhVh0NZmvOI7NZW+5FIrMfbetAHruiyy7Ht3kPnqOG
bkkevvWR4+1CSz0eaK4jFxGEJYoSrEYwM9upB69qsaqZEghvLN1S4jGVJ6MPSvLvHHjvUdRtZbG6
tkglDYLg9QPagDz+9tFgit5gJEEoJw/Xg4rQOS4tIGUtPH5smABzwPSuevbp5cKXJC8ZJqpk0Ae7
6DMPKBQhs+hr0DwoP+Jc+Tz5p/kK+UrHUbuxlWS0uJYmU5G1uPyr234d+O5JtAb7ba+ZOkxUujbQ
3yrzj8aAMGxP7v8ACtG24nmHbIP6UUUAWLTiacDpuHFVPEQB8OavkA/Kf6UUUAdjZOzeDLFmZixt
ojknnO0V5Z47AaCZ2GX8wDcev5OUUAedHr+NJRRQADrXpvw9UDQ5MAf69u3+ytFFAH//2Q==
```


Finally, the message transmission concludes with the first boundary notation at **(O)**.

The lesson from this is that what you see displayed in your e-mail client application isn't really the e-mail. It's an *arrangement* of selected *parts* of the message, frequently modified in some respects from the native message source that traversed the network and Internet and, as often, supplemented by metadata (like message flags, contact data and other feature-specific embellishments) unique to your software and setup. What you see handily displayed as a discrete attachment is, in reality, encoded into the message body. The time assigned to message is calculated relative to your machine's time and DST settings. Even the sender's name may be altered based upon the way your machine and contact's database is configured. What you see is not always what you get (or got).

Hashing and Deduplication

[23] A third common transfer encoding is called "quoted-printable" or "QP encoding." It facilitates transfer of non-ASCII 8-bit data as 7-bit ASCII characters using three ASCII characters (the "equals" sign followed by two hexadecimal characters: 0-9 and A-F) to stand in for a byte of data Quoted-printable is employed where the content to be encoded is predominantly ASCII text coupled with some non-ASCII items. Its principal advantage is that it allows the encoded data to remain largely intelligible to readers.

The ability to "fingerprint" data using hash algorithms makes it possible to identify identical files without the necessity of examining their content. If the hash values of two files are identical, the files are identical. As previously discussed, this file-matching ability allows hashing to be used to deduplicate collections of electronic files before review, saving money and minimizing the potential for inconsistent decisions about privilege and responsiveness for identical files.

Although hashing is a useful and versatile technology, it has a few shortcomings. Because the tiniest change in a file will alter that file's hash value, hashing is of little value in comparing files that have any differences, even if those differences have no bearing on the substance of the file. Applied to e-mail, we understand from our e-mail "autopsy" that messages contain unique identifiers, time stamps and routing data that would frustrate efforts to compare one complete message to another using hash values. Looking at the message as a whole, multiple recipients of the same message have different versions insofar as their hash values.

Consequently, deduplication of e-mail messages is accomplished by calculating hash values for selected segments of the messages and comparing those segment values. Thus, hashing e-mails for deduplication will omit the parts of the header data reflecting, *e.g.,* the message identifier and the transit data. Instead, it will hash just the data seen in, *e.g.,* the To, From, Subject and Date lines, message body and encoded attachment. If these match, the message can be said to be *practically* identical.

By hashing particular segments of messages and selectively comparing the hash values, it's possible to gauge the *relative* similarity of e-mails and perhaps eliminate the cost to review messages that are *inconsequentially* different. This concept is called "near deduplication." It works, but it's important to be aware of exactly what it's excluding and why. It's also important to advise your opponents when employing near deduplication and ascertain whether you're mechanically excluding evidence the other side deems relevant and material.

Hash deduplication of e-mail is tricky. Time values may vary, along with the apparent order of attachments. These variations, along with minor formatting discrepancies, may serve to prevent the exclusion of items defined as duplicates. When this occurs, be certain to delve into the reasons *why* apparent duplicates aren't deduplicating, as such errors may be harbingers of a broader processing problem.

Local E-Mail Storage Formats and Locations

Suppose you're faced with a discovery request for a client's e-mail and there's no budget or time to engage an e-discovery service provider or ESI expert?

Where are you going to look to find stored e-mail, and what form will it take?

"Where's the e-mail?" It's a simple question, and one answered too simply and often wrongly by, "It's on the server" or "The last 60 days of mail is on the server and the rest is purged." Certainly, much e-mail will reside on the server, but most e-mail is elsewhere; and it's never all gone in practice, notwithstanding retention policies. The true location and extent of e-mail depends on systems configuration, user habits, backup procedures and other hardware, software and behavioral factors. This is true for mom-and-pop shops, for large enterprises and for everything in-between.

Going to the server isn't the wrong answer. It's just not the entire answer. In a matter where I was tasked to review e-mails of an employee believed to have stolen proprietary information, I went first to the company's Microsoft Exchange e-mail server and gathered a lot of unenlightening e-mail. Had I stopped there, I would've missed the Hotmail traffic in the Temporary Internet Files folder and the Short Message Service (SMS) exchanges in the smartphone synchronization files. I'd have overlooked the Microsoft Outlook archive file (archive.pst) and offline synchronization file (Outlook.ost) on the employee's laptop, collectively holding thousands more e-mails, including some "smoking guns" absent from the server. These are just some of the many places e-mails without counterparts on the server may be found. Though an exhaustive search of every nook and cranny may not be required, you need to know your options in order to assess feasibility, burden and cost.

E-mail resides in some or all of the following venues, grouped according to relative accessibility:

Easily Accessible:

E-Mail Server: Online e-mail residing in active files on enterprise servers: MS Exchange e.g., (.edb, .stm, .log files), Lotus Notes (.nsf files).

File Server: E-mail saved as individual messages or in container files on a user's network file storage area ("network share").

Desktops and Laptops: E-mail stored in active files on local or external hard drives of user workstation hard drives (e.g., .pst, .ost files for Outlook and .nsf for Lotus Notes), laptops (.ost, .pst, .nsf), mobile devices, and home systems, particularly those with remote access to networks.

OLK system subfolders holding viewed attachments to Microsoft Outlook messages, including deleted messages.

Mobile devices: An estimated 65% of e-mail messages were opened using mobile phones and tablets in Q4 2015. As many of these were downloaded to a local mail app, they reside on the device and do not necessarily lose such content when the same messages are deleted from the

server. E-mail on mobile devices is readily accessible to the user, but poses daunting challenges for preservation and collection in e-discovery workflows.

Nearline e-mail: Optical "juke box" devices, backups of user e-mail folders.

Archived or journaled e-mail: e.g., HP Autonomy Zantaz Enterprise Archive Solution, EMC EmailXtender, NearPoint Mimosa, Symantec Enterprise Vault.

Accessible, but Often Overlooked:

E-mail residing on non-party servers: ISPs (IMAP, POP, HTTP servers), Gmail, Yahoo! Mail, Hotmail, etc.

E-mail forwarded and cc'd to external systems: Employee forwards e-mail to self at personal e-mail account.

E-mail threaded as text behind subsequent exchanges.

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 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. Also, e-mail on mothballed or re-tasked machines and devices.

E-mail saved to other formats: PDF, .tiff, .txt, .eml, .msg, 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 or attorneys)

Paper print outs.

Less Accessible:

Offline e-mail on server backup tapes and other media.

E-mail in forensically accessible areas of local hard drives and re-tasked/reimaged legacy machines: deleted e-mail, internet cache, unallocated clusters.

The levels of accessibility above speak to practical challenges to ease of access, not to the burden or cost of review. The burden continuum isn't a straight line. That is, it may be less burdensome or costly to turn to a small number of less accessible sources holding relevant data than to broadly search and review the contents of many accessible sources. Ironically, it typically costs much more to process and review the contents of a mail server than to undertake forensic examination of a key player's computer; yet, the former is routinely termed "reasonably accessible" and the latter not.

The issues in the case, key players, relevant time periods, agreements between the parties, applicable statutes, decisions and orders of the court determine the extent to which locations must be examined; however, the failure to diligently identify relevant e-mail carries such peril that caution should be the watchword. Isn't it wiser to invest more effort to know exactly what the client has—even if it's not reasonably accessible and will not be searched or produced—than concede at the sanctions hearing the client failed to preserve and produce evidence it didn't know it because no one looked?

Looking for E-Mail 101

Because an e-mail is just a text file, individual e-mails could be stored as discrete text files. But that's not a very efficient or speedy way to manage many messages, so you'll find that most e-mail client software doesn't do that. Instead, e-mail clients employ proprietary database files housing e-mail messages, and each of the major e-mail clients uses its own unique format for its database. Some programs encrypt the message stores. Some applications merely display e-mail housed on a remote server and do not store messages locally (or only in fragmentary way). The only way to know with certainty if e-mail is stored on a local hard drive is to look for it.

Merely checking the e-mail client's settings is insufficient because settings can be changed. Someone not storing server e-mail today might have been storing it a month ago. Additionally, users may create new identities on their systems, install different client software, migrate from other hardware or take various actions resulting in a cache of e-mail residing on their systems without their knowledge. *If they don't know it's there, they can't tell you it's not*. On local hard drives, you've simply got to know what to look for and where to look…*and then you've got to look for it.*

For many, computer use has been a decades-long adventure. One may have first dipped her toes in the online ocean using browser-based e-mail or an AOL account. Gaining computer-savvy, she may have signed up for broadband access or with a local ISP, downloading e-mail with Netscape Messenger or Microsoft Outlook Express. With growing sophistication, a job change or new technology at work, the user may have migrated to Microsoft Outlook or Lotus Notes as an e-mail

client, then shifted to a cloud service like Office 365. Each of these steps can orphan a large cache of e-mail, possibly unbeknownst to the user but still fair game for discovery. Again, you've simply got to know what to look for and where to look.

One challenge you'll face when seeking stored e-mail is that every user's storage path is different. This difference is not so much the result of a user's ability to specify the place to store e-mail—which few do, but which can make an investigator's job more difficult when it occurs—but more from the fact that operating systems are designed to support multiple users and so must assign unique identities and set aside separate storage areas for different users. Even if only one person has used a Windows computer, the operating system will be structured at the time of installation so as to make way for others. Thus, finding e-mail stores will hinge on your knowledge of the User's Account Name or Globally Unique Identifier (GUID) string assigned by the operating system. This may be as simple as the user's name or as obscure as the 128-bit hexadecimal value {721A17DA-B7DD-4191-BA79-42CF68763786}. Customarily, it's both.

Finding Outlook E-Mail

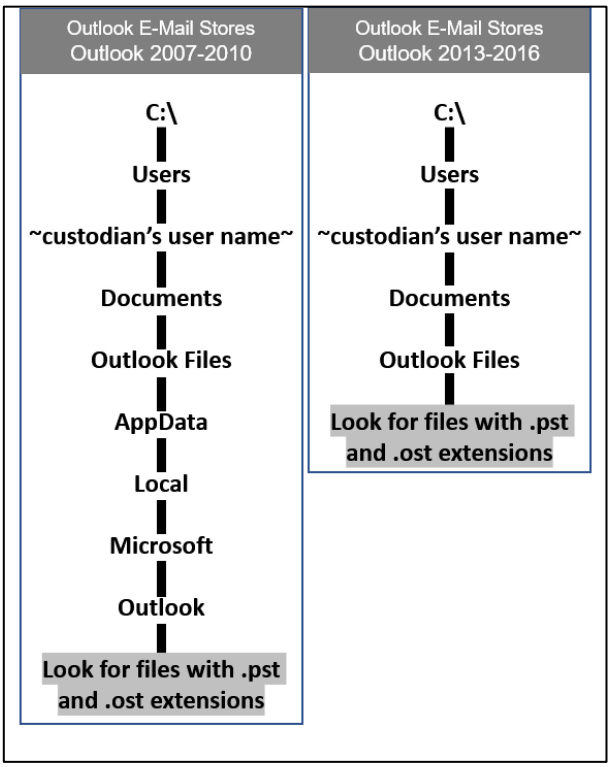
PST: Microsoft Outlook has long been the most widely used e-mail client in the business environment. Outlook encrypts and compresses messages, and all of its message data and folder structure, along with all other information managed by the program (except the user's Contact data), is stored within a single, often massive, database file with the file extension .pst.

OST: While awareness of the Outlook PST file is widespread, even many lawyers steeped in e-discovery fail to consider a user's Outlook .ost file. The OST or offline synchronization file is commonly encountered on laptops configured for Exchange Server environments. It exists for the purpose of affording access to messages when the user has no active network connection. Designed to allow work to continue on, e.g., airplane flights, local OST files often hold messages purged from the server—at least until re-synchronization. It's not unusual for an OST file to hold e-mail unavailable from any other comparably-accessible source.

Archive.pst: Another file to consider is one customarily called, "archive.pst." As its name suggests, the archive.pst file holds older messages, either stored automatically or by user-initiated action. If you've used Outlook without manually configuring its archive settings, chances are the system periodically asks whether you'd like to auto archive older items. Every other week (by default), Outlook seeks to auto archive any Outlook items older than six months (or for Deleted and Sent items older than two months). Users can customize these intervals, turn archiving off or instruct the application to permanently delete old items.

Outlook Mail Stores Paths

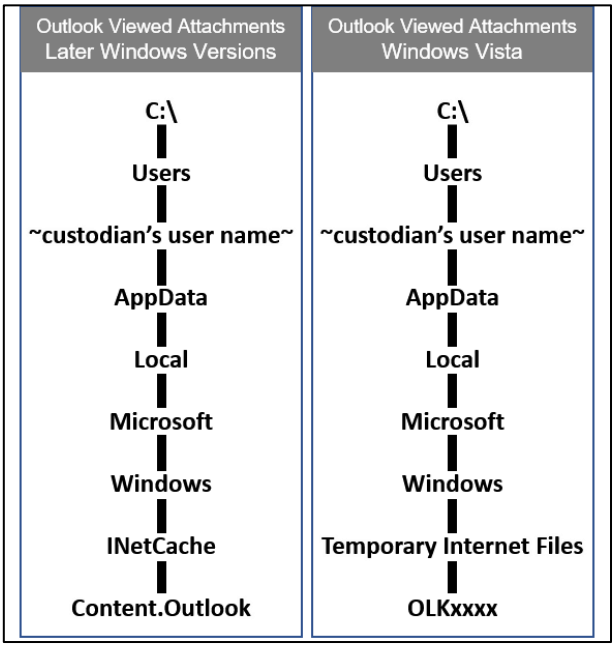
To find the Outlook message stores on Windows machines, drill down from the root directory (C:\ for most users) according to the path diagram shown for the applicable version of Outlook. The default filename of Outlook.pst/ost may vary if a user has opted to select a different designation or maintains multiple e-mail stores; however, it's rare to see users depart from the default settings. Since the location of the PST and OST files can be changed by the user, it's a good idea to do a search of all files and folders to identify any files ending with the .pst and .ost extensions.



"Temporary" OLK Folders

Note that by default, when a user opens an attachment to a message from within Outlook (as opposed to saving the attachment to disk and then opening it), Outlook stores a copy of the attachment in a "temporary" folder. But don't be misled by the word "temporary." In fact, the folder isn't going anywhere and its contents—sometimes voluminous--tend to long outlast the messages that transported the attachments. Thus, litigants should be cautious about representing that Outlook e-mail is "gone" if the e-mail's attachments are not.

The Outlook viewed attachment folder will have a varying name for every user and on every machine, but it will always begin with the letters "OLK" followed by several randomly generated numbers and uppercase letters (e.g., OLK943B, OLK7AE, OLK167, etc.). To find the OLKxxxx viewed attachments folder on machines running Windows XP/NT/2000 or Vista, drill down from the root directory according to the path diagrams on the right for the applicable operating system.[24]



[24] By default, Windows hides system folders from users, so you may have to first make them visible. This is accomplished by starting Windows Explorer, then selecting 'Folder Options' from the Tools menu in Windows XP or

Microsoft Exchange Server

Hundreds of millions of people get their work e-mail via a Microsoft product called Exchange Server. It's been sold for twenty years and its latest version is Exchange Server 2016; although, many users continue to rely on the older versions of the product.

The key fact to understand about an e-mail server is that it's a *database* holding the messages (and calendars, contacts, to-do lists, journals and other datasets) of multiple users. E-mail servers are configured to maximize performance, stability and disaster recovery, with little consideration given to compliance and discovery obligations. If anyone anticipated the role e-mail would play in virtually every aspect of business today, their prescience never influenced the design of e-mail systems. E-mail evolved largely by accident, absent the characteristics of competent records management, and only lately are tools emerging that are designed to catch up to legal and compliance duties.

The other key thing to understand about enterprise e-mail systems is that, unless you administer the system, it probably doesn't work the way you imagine. The exception to that rule is if you can distinguish between Local Continuous Replication (LCR), Clustered Continuous Replication (CCR), Single Copy Cluster (SCC) and Standby Continuous Replication (SCR). In that event, I should be reading *your* paper!

Though the preceding pages dealt with finding e-mail stores on local hard drives, in disputes involving medium- to large-sized enterprises, the e-mail server (or its cloud-based counterpart) is likely to be the initial nexus of electronic discovery efforts. The server is a productive venue in electronic discovery for many reasons, among them:

The periodic backup procedures which are a routine part of prudent server management tend to shield e-mail stores from those who, by error or guile, might delete or falsify data on local hard drives.

The ability to recover deleted mail from archival server backups may obviate the need for costly and unpredictable forensic efforts to restore deleted messages.

Data stored on a server is often less prone to tampering by virtue of the additional physical and system security measures typically dedicated to centralized computer facilities as well as the inability of the uninitiated to manipulate data in the more-complex server environment.

'Organize>Folder and Search Options' in Vista. Under the 'View' tab, scroll to 'Files and Folders' and check 'Show hidden files and folders' and uncheck 'Hide extensions for known file types' and 'Hide protected operating system files. Finally, click 'OK.'

The centralized nature of an e-mail server affords access to many users' e-mail and may lessen the need for access to workstations at multiple business locations or to laptops and home computers.

Unlike e-mail client applications, which store e-mail in varying formats and folders, e-mail stored on a server can usually be located with relative ease and adhere to common file formats.

The server is the crossroads of corporate electronic communications and the most effective chokepoint to grab the biggest "slice" of relevant information in the shortest time, for the least cost.

The latest versions of Exchange Server and the cloud tool, Office 365, feature robust e-discovery capabilities simplifying initiation and managements of legal holds and account exports.

Of course, the big advantage of focusing discovery efforts on the mail server (*i.e.,* it affords access to thousands or millions of messages) is also its biggest disadvantage (someone has to *collect and review* thousands or millions of messages). Absent a carefully-crafted and, ideally, agreed-upon plan for discovery of server e-mail, both requesting and responding parties run the risk of runaway costs, missed data and wasted time.

E-mail originating on servers is generally going to fall into two realms, being online "live" data, which is deemed reasonably accessible, and offline "archival" data, routinely deemed inaccessible based on considerations of cost and burden.[25] Absent a change in procedure, "chunks" of data routinely migrate from accessible storage to less accessible realms—on a daily, weekly or monthly basis—as selected information on the server is replicated to backup media and deleted from the server's hard drives.

The ABCs of Exchange

Because it's unlikely most readers will be personally responsible for collecting e-mail from an Exchange Server and mail server configurations can vary widely, the descriptions of system architecture here are offered only to convey a rudimentary understanding of common Exchange architecture.

Older versions of Exchange Server stored data in a Storage Group containing a Mailbox Store and a Public Folder Store, each composed of two files: an .edb file and a .stm file. Mailbox Store,

[25] Lawyers and judges intent on distilling the complexity of electronic discovery to rules of thumb are prone to pigeonhole particular ESI as "accessible' or 'inaccessible" based on the media on which it resides. In fact, ESI's storage medium is just one of several considerations that bear on the cost and burden to access, search and produce same. Increasingly, backup tapes are less troublesome to search and access while active data on servers or strewn across many "accessible" systems and devices is a growing challenge.

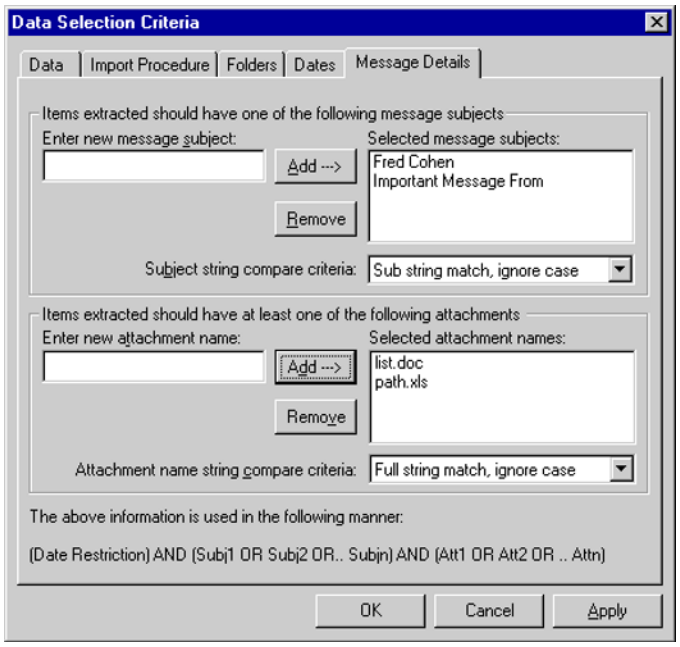
Priv1.edb, is a rich-text database file containing user's e-mail messages, text attachments and headers. Priv1.stm is a streaming file holding SMTP messages and containing multimedia data formatted as MIME data. Public Folder Store, Pub1.edb, is a rich-text database file containing messages, text attachments and headers for files stored in the Public Folder tree. Pub1.stm is a streaming file holding SMTP messages and containing multimedia data formatted as MIME data. Later versions of Exchange Server did away with STM files altogether, shifting their content into the EDB database files.

Storage Groups also contain system files and transaction logs. Transaction logs serve as a disaster recovery mechanism that helps restore an Exchange after a crash. Before data is written to an EDB file, it is first written to a transaction log. The data in the logs can thus be used to reconcile transactions after a crash.

By default, Exchange data files are located in the path X:\Program files\Exchsrvr\MDBDATA, where X: is the server's volume root. But, it's common for Exchange administrators to move the mail stores to other file paths.

Recovery Storage Groups and ExMerge

Two key things to understand about Microsoft Exchange are that, since 2003, an Exchange feature called Recovery Storage Group supports collection of e-mail from the server without any need to interrupt its operation or restore data to a separate recovery computer. The second key thing is that Exchange includes a simple utility for exporting the server-stored e-mail of individual custodians to separate PST container files. This utility, officially the Exchange Server Mailbox Merge Wizard but universally called ExMerge allows for rudimentary filtering of messages for export, including by message dates, folders, attachments and subject line content.



ExMerge also plays a crucial role in recovering e-mails "double deleted" by users if the Exchange server has been configured to support a "dumpster retention period." When a user deletes an e-mail, it's automatically relegated to a "dumpster" on the Exchange Server. The dumpster holds

the message for 30 days by default or until a full backup of your Exchange database is run, whichever comes first. The retention interval can be customized for a longer or shorter interval.

Later versions of Exchange Server and certain implementations of Exchange Online [Office 365] have done away with the dumpster feature and take an entirely different (and superior) approach to retention of double-deleted messages. As noted, these tools also offer purpose-built e-discovery preservation features that are much easier to implement and manage than earlier Exchange Server versions.

Journaling, Archiving and Transport Rules

Journaling is the practice of copying all e-mail to and from all users or particular users to one or more repositories inaccessible to most users. Journaling serves to preempt ultimate reliance on individual users for litigation preservation and regulatory compliance. Properly implemented, it should be entirely transparent to users and secured in a manner that eliminates the ability to alter the journaled collection.

Exchange Server supports three types of journaling: Message-only journaling which does not account for blind carbon copy recipients, recipients from transport forwarding rules, or recipients from distribution group expansions; Bcc journaling, which is identical to Message-only journaling except that it captures Bcc addressee data; and Envelope Journaling which captures all data about the message, including information about those who received it. Envelope journaling is the mechanism best suited to e-discovery preservation and regulatory compliance.

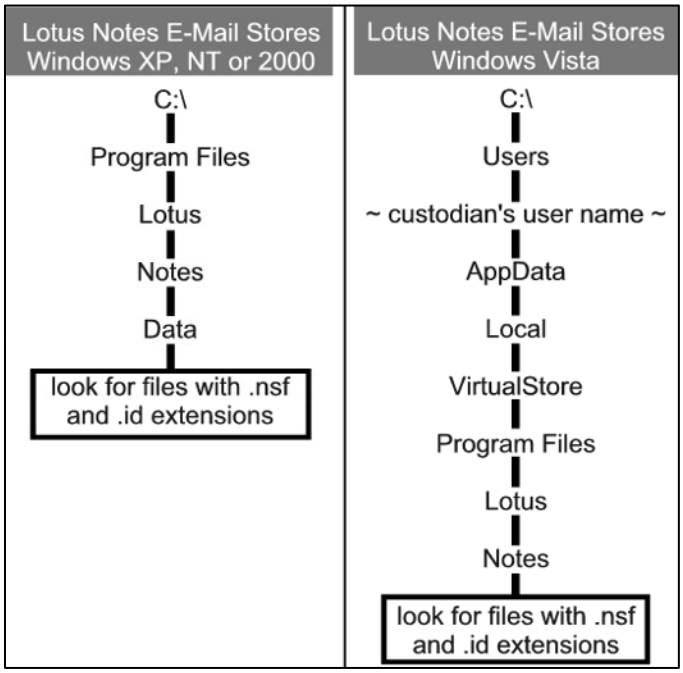
Journaling should be distinguished from e-mail archiving, which may implement only selective, rules-based retention and customarily entails removal of archived items from the server for offline or near-line storage, to minimize strain on IT resources and/or implement electronic records management. However, Exchange journaling also can implement rules-based storage, so each can conceivably be implemented to play the role of the other.

A related concept is the use of Transport Rules in Exchange, which serve, *inter alia*, to implement "Chinese Walls" between users or departments within an enterprise who are ethically or legally obligated not to share information, as well as to guard against dissemination of confidential information. In simplest terms, software called *transport rules agents* "listen" to e-mail traffic, compare the content or distribution to a set of rules (conditions, exceptions and actions) and if particular characteristics are present, intercedes to block, route, flag or alter suspect communications.

Lotus Domino Server and Notes Client

Though Microsoft's Exchange and Outlook e-mail products have a greater overall market share, IBM's Lotus Domino and Notes products hold powerful sway within the world's largest corporations, especially giant manufacturing concerns and multinationals. IBM boasts of over 300 million Notes mailboxes worldwide.

Lotus Notes can be unhelpfully described as a "cross-platform, secure, distributed document-oriented database and messaging framework and rapid application development environment." The main takeaway with Notes is that, unlike Microsoft Exchange, which is a purpose-built application designed for messaging and calendaring, Lotus Notes is more like a toolkit for *building* whatever capabilities you need to deal with documents—mail documents, calendaring documents and any other type of document used in business. Notes wasn't *designed* for e-mail—e-mail just happened to be one of the things it was tasked to do.[26] Notes is database driven and distinguished by its replication and security.



Lotus Notes is all about copies. Notes content, stored in Notes Storage facility or NSF files, are constantly being replicated (synchronized) here and there across the network. This guards against data loss and enables data access when the network is unavailable, but it also means that there can be many versions of Notes data stashed in various places within an enterprise. Thus, discoverable Notes mail may not be gone, but lurks within a laptop that hasn't connected to the network since the last business trip.

By default, local iterations of users' NSF and ID files will be found on desktops and laptops in the paths shown in the diagrams at right. It's imperative to collect the user's .id file along with the .nsf message container or you may find yourself locked out of encrypted content. It's also important to secure each custodian's Note's password. It's common for Notes to be installed in ways other than the default configuration, so

[26] Self-anointed "Technical Evangelist" Jeff Atwood described Lotus Notes this way: "It is death by a thousand tiny annoyances— the digital equivalent of being kicked in the groin upon arrival at work every day." http://www.codinghorror.com/blog/2006/02/12/ (visited 5/18/2013) In fairness, Lotus Notes has been extensively overhauled since he made that observation.

search by extension to insure that .nsf and .id files are not also found elsewhere. Also, check the files' last modified date to assess whether the date is consistent with expected last usage. If there is a notable disparity, look carefully for alternate file paths housing later replications.

Local replications play a significant role in e-discovery of Lotus Notes mail because, built on a database and geared to synchronization of data stores, deletion of an e-mail within Lotus "broadcasts" the deletion of the same message system wide. Thus, it's less common to find undeleted iterations of messages in a Lotus environment unless you resort to backup media or find a local iteration that hasn't been synchronized after deletion.

Webmail

More than 25% of the people on the planet use webmail; so any way you slice it, webmail can't be ignored in e-discovery. Webmail holding discoverable ESI presents legal, technical and practical challenges, but the literature is nearly silent about how to address them.

The first hurdle posed by webmail is the fact that it's stored "in the cloud" and off the company grid. Short of a subpoena or court order, the only legitimate way to access and search employee web mail is with the employee's cooperation, and that's not always forthcoming. Courts nonetheless expect employers to exercise control over employees and insure that relevant, non-privileged webmail isn't lost or forgotten.

One way to assess the potential relevance of webmail is to search server e-mail for webmail traffic. If a custodian's Exchange e-mail reveals that it was the custodian's practice to e-mail business documents to or from personal webmail accounts, the webmail accounts may need to be addressed in legal hold directives and vetted for responsive material.

A second hurdle stems from the difficulty in collecting responsive webmail. How do you integrate webmail content into your review and production system? Where a few pages might be "printed" to searchable Adobe Acrobat PDF formats or paper, larger volumes require a means to dovetail online content and local collections. The most common approach is to employ a POP3 or IMAP client application to download messages from the webmail account. All of the leading webmail providers support POP3 transfer, and with the user's cooperation, it's simple to configure a clean installation of any of the client applications already discussed to capture online message stores. Before proceeding, the process should be tested against accounts that don't evidence to determine what metadata values may be changed, lost or introduced by POP3 collection.

Webmail content can be fragile compared to server content. Users rarely employ a mechanism to back up webmail messages (other than the POP3 or IMAP retrieval just discussed) and webmail accounts may purge content automatically after periods of inactivity or when storage limits are

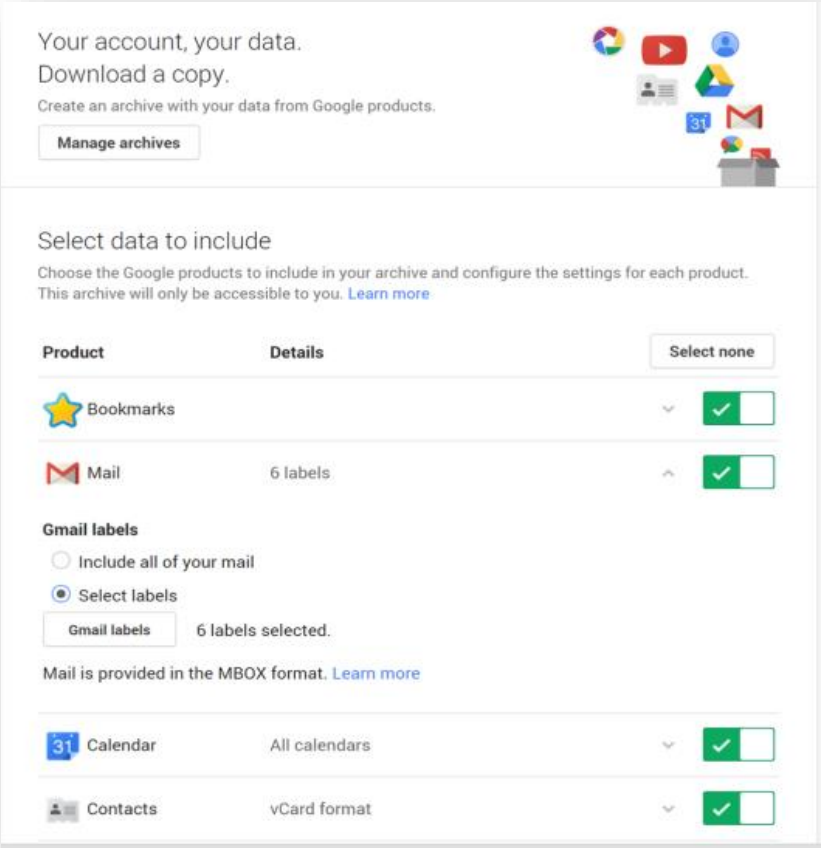
exceeded. Further, users tend to delete embarrassing or incriminating content more aggressively on webmail, perhaps because they regard webmail content as personal property or the evanescent nature of account emboldens them to believe spoliation will be harder to detect and prove.

Happily, some webmail providers—notably Google Gmail—have begun to offer effective "take out" mechanisms for user cloud content, including webmail. Google does the Gmail collection *gratis* and puts it in a standard MBOX container format that can be downloaded and sequestered. Google even incorporates custom metadata values that reflect labeling and threading. You won't see these unique metadata tags if you pull the messages into an e-mail client; but, good e-discovery software will pick them up.

MBOX might not be everyone's choice for a Gmail container file; but, it's inspired. MBOX stores the messages in their original Internet message format called RFC 2822 (now RFC 5322), a superior form for e-discovery preservation and production.

Google Data Tools

The only hard part of archiving Gmail is navigating to the right page. You get there from the Google Account Setting page by selecting "Data Tools" and looking for the "Download your Data" option on the lower right. When you click on "Create New Archive," you'll see a menu like that below where you choose whether to download all mail or just items bearing the labels you select.



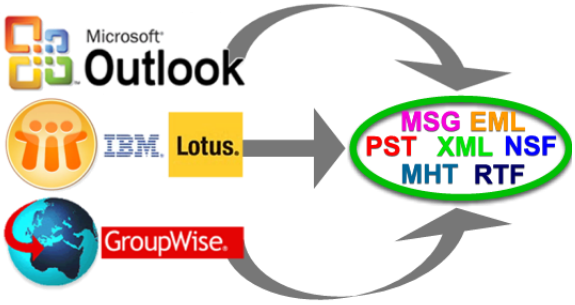
The ability to label content within Gmail and archive only messages bearing those labels means that Gmail's powerful search capabilities can be used to identify and label potentially

responsive messages, obviating the need to archive everything. It's not a workflow suited to every case; yet, it's a promising capability for keeping costs down in cases involving just a handful of custodians with Gmail.

Forms of Production

As discussed above, what users see presented onscreen as e-mail is a selective presentation of information from the header, body and attachments of the source message, determined by the capabilities and configuration of their e-mail client and engrafted with metadata supplied by that client. Meeting the obligation to produce comparable data of similar utility to the other side in discovery is no mean feat, and one that hinges on choosing suitable forms of production.

Requesting parties often demand "native production" of e-mail; but, electronic mail is rarely produced natively in the sense of supplying a duplicate of the source container file. That is, few litigants produce the entire Exchange database EDB file to the other side. Even those that produce mail in the format employed natively by the application (e.g., as a PST file) aren't likely to produce the source file but will fashion a reconstituted PST file composed of selected messages deemed responsive and non-privileged.



As applied to e-mail, "native production" instead signifies production in a form or forms that most closely approximate the contents and usability of the source. Often, this will be a form of production identical to the original (e.g., PST or NSF) or a form (like MSG or EML) that shares many of the characteristics of the source and can deliver comparable usability when paired with additional information (e.g., information about folder structures).[27] For further discussion of native forms of e-mail, see the following article, *What is Native Production of E-Mail?*

Similarly, producing parties employ imaged production and supply TIFF image files of messages, but in order to approximate the usability of the source must also create and produce accompanying load files carrying the metadata and full text of the source message keyed to its images. Collectively, the load files and image data permit recipients with compatible software (*e.g.*, Relativity, Summation, Concordance) to view and search the messages. Selection of Adobe PDF documents as the form of production allows producing parties to dispense with the load files

[27] When e-mail is produced as individual messages, the folder structure may be lost and with it, important context. Additionally, different container formats support different complements of metadata applicable to the message. For example, a PST container may carry information about whether a message was opened, flagged or linked to a calendar entry.

because much of the same data can be embedded in the PDF. PDF also has the added benefit of not requiring the purchase of review software.

Some producing parties favor imaged production formats in a mistaken belief that they are more secure than native production and out of a desire to emboss Bates numbers or other text (i.e., protective order language) to the face of each image. Imaged productions are more expensive than native or quasi-native productions, but, as they hew closest to the document review mechanisms long employed by law firms, they require little adaption. It remains to be seen if clients will continue to absorb higher costs solely to insulate their counsel from embracing more modern and efficient tools and techniques.

Other possible format choices include XML and MHT,[28] as well as Rich Text Format (RTF)--essentially plain text with improved formatting—and, for small collections, paper printouts.

There is no single, "perfect" form of production for e-mail, though the "best" format to use is the one on which the parties agree. Note also that there's likely not a single production format that lends itself to *all* forms of ESI. Instead, *hybrid productions* match the form of production to the characteristics of the data being produced. In a hybrid production, images are used where they are most utile or cost-effective and native formats are employed when they offer the best fit or value.

As a rule of thumb to maximize usability of data, hew closest to the format of the source data (i.e., PST for Outlook mail and NSF for Lotus Notes), but keep in mind that whatever form is chosen should be one that the requesting party has the tools and expertise to use.

Though there is no ideal form of production, we can be guided by certain ideals in selecting the forms to employ. Absent agreement between the parties or an order of the Court, the forms of production employed for electronic mail should be either the mail's native format or a form that will:

- Enable the complete and faithful reproduction of all information available to the sender and recipients of the message, including layout, bulleting, tabular formats, colors, italics, bolding, underlining, hyperlinks, highlighting, embedded images, emoticons and other non-textual ways we communicate and accentuate information in e-mail messages.
- Support accurate electronic searchability of the message text and header data;

[28] MHT is a shorthand reference for MHTML or MIME Hypertext markup Language. HTML is the markup language used to create web pages and rich text e-mails. MHT formats mix HTML and encoded MIME data (see prior discussion of MIME at page to represent the header, message body and attachments of an e-mail.

- Maintain the integrity of the header data (To, From, Cc, Bcc, Subject and Date/Time) as discrete fields to support sorting and searching by these data;
- Preserve family relationships between messages and attachments;
- Convey the folder structure/path of the source message;
- Include message metadata responsive to the requester's legitimate needs;
- Facilitate redaction of privileged and confidential content and, as feasible, identification and sequencing akin to Bates numbering; and
- Enable reliable date and time normalization across the messages produced.[29]

[29] E-mails carry multiple time values depending upon, e.g., whether the message was obtained from the sender or recipient. Moreover, the times seen in an e-mail may be offset per the time zone settings of the originating or receiving machine as well as for daylight savings time. When e-mail is produced as TIFF images or as text embedded in threads, these offsets may produce hopelessly confusing sequences.

What is Native Production of E-Mail?

Recently, I've weighed in on disputes where the parties were fighting over whether the e-mail production was sufficiently "native" to comply with the court's orders to produce natively. In one matter, the question was whether Gmail could be produced in a native format, and in another, the parties were at odds about what forms are native to Microsoft Exchange e-mail. In each instance, I saw two answers; the technically correct one and the helpful one.

I am a vocal proponent of native production for e-discovery. Native is complete. Native is functional. Native is inherently searchable. Native costs less. I've explored these advantages in other writings and will spare you that here. But when I speak of "native" production in the context of databases, I am using a generic catchall term to describe electronic forms with superior functionality and completeness, notwithstanding the common need in e-discovery to produce less than all of a collection of ESI.

It's a Database

When we deal with e-mail in e-discovery, we are usually dealing with database content. Microsoft Exchange, an *e-mail server application*, is a database. Microsoft Outlook, an *e-mail client application*, is a database. Gmail, a *SaaS webmail application*, is a database. Lotus Domino, Lotus Notes, Yahoo! Mail, Hotmail and Novell GroupWise—they're all *databases*. It's important to understand this at the outset because if you think of e-mail as a collection of discrete objects (like paper letters in a manila folder), you're going to have trouble understanding why defining the "native" form of production for e-mail isn't as simple as many imagine.

Native in Transit: Text per a Protocol

E-mail is one of the oldest computer networking applications. Before people were sharing printers, and long before the internet was a household word, people were sending e-mail across networks. That early e-mail was plain text, also called ASCII text or 7-bit (because you need just seven bits of data, one less than a byte, to represent each ASCII character). In those days, there were no attachments, no pictures, not even simple enhancements like **bold**, *italic* or <u>underline</u>.

Early e-mail was something of a free-for-all, implemented differently by different systems. So the fledgling internet community circulated proposals seeking a standard. They stuck with plain text in order that older messaging systems could talk to newer systems. These proposals were called

Requests for Comment or **RFC**s, and they came into widespread use as much by convention as by adoption (the internet being a largely anarchic realm). The RFCs lay out the form an e-mail should adhere to in order to be compatible with e-mail systems.

The RFCs concerning e-mail have gone through several major revisions since the first one circulated in 1973. The latest protocol revision is called RFC 5322 (2008), which made obsolete RFC 2822 (2001) and its predecessor, RFC 822 (1982). Another series of RFCs (RFC 2045-47, RFC 4288-89 and RFC 2049), collectively called Multipurpose Internet Mail Extensions or MIME, address ways to graft text enhancements, foreign language character sets and multimedia content onto plain text emails. These RFCs establish the form of the billions upon billions of e-mail messages that cross the internet.

So, if you asked me to state the native form of an e-mail *as it traversed the Internet between mail servers*, I'd likely answer, "plain text (7-bit ASCII) adhering to RFC 5322 and MIME." In my experience, this is the same as saying ".EML format;" and, it can be functionally the same as the MHT format, but only if the content of each message adheres strictly to the RFC and MIME protocols listed above. You can even change the file extension of a properly formatted message from EML to MHT and back to open the file in a browser or in a mail client like Outlook 2010. Try it. If you want to see what the native "plain text in transit" format looks like, change the extension from .EML to .TXT and open the file in Windows Notepad.

The appealing feature of producing e-mail in exactly the same format in which the message traversed the internet is that it's a form that holds the entire content of the message (header, message bodies and encoded attachments), and it's a form that's about as compatible as it gets in the e-mail universe. [30]

Unfortunately, the form of an e-mail *in transit* is often incomplete in terms of metadata it acquires upon receipt that may have probative or practical value; and the format in transit isn't native to

[30] There's even an established format for storing multiple RFC 5322 messages in a container format called mbox. The mbox format was described in 2005 in RFC 4155, and though it reflects a simple, reliable way to group e-mails in a sequence for storage, it lacks the innate ability to memorialize mail features we now take for granted, like message foldering. A common workaround is to create a single mbox file named to correspond to each folder whose contents it holds (e.g., Inbox.mbox)

the most commonly-used e-mail server and client applications, like Microsoft Exchange and Outlook. It's from these applications--*these databases*--that e-mail is collected in e-discovery.

Outlook and Exchange

Microsoft Outlook and Microsoft Exchange are database applications that talk to each other using a protocol (machine language) called MAPI, for *Messaging Application Programming Interface*. Microsoft Exchange is an e-mail server application that supports functions like contact management, calendaring, to do lists and other productivity tools. Microsoft Outlook is an e-mail client application that accesses the contents of a user's account on the Exchange Server and may synchronize such content with local (i.e., retained by the user) container files supporting offline operation. If you can read your Outlook e-mail without a network connection, you have a local storage file.

Practice Tip (and Pet Peeve): *When your client or company runs Exchange Server and someone asks what kind of e-mail system your client or company uses, please don't say "Outlook." That's like saying "iPhone" when asked what cell carrier you use. Outlook can serve as a front-end client to Microsoft Exchange, Lotus Domino and most webmail services; so saying "Outlook" just makes you appear out of your depth (assuming you are someone who's supposed to know something about the evidence in the case).*

Outlook: The native format for data stored locally by Outlook is a file or files with the extension PST or OST. Henceforth, I'm going to speak only of PSTs, but know that either variant may be seen. PSTs are container files. They hold collections of e-mail—typically stored in multiple folders—as well as content supporting other Outlook features. The native PST found locally on the hard drive of a custodian's machine will hold all of the Outlook content that the custodian can see when not connected to the e-mail server.

Because Outlook is a database application designed for managing messaging, it goes well beyond simply receiving messages and displaying their content. Outlook begins by taking messages apart and using the constituent information to populate various fields in a database. What we see as an e-mail message using Outlook is actually a report queried from a database. The native form of Outlook e-mail carries these fields and adds metadata not present in the transiting message. The

added metadata fields include such information as the name of the folder in which the e-mail resides, whether the e-mail was read or flagged and its date and time of receipt. Moreover, because Outlook is designed to “speak” directly to Exchange using their own MAPI protocol, messages between Exchange and Outlook carry MAPI metadata not present in the "generic" RFC 5322 messaging. Whether this MAPI metadata is superfluous or invaluable depends upon what questions may arise concerning the provenance and integrity of the message. Most of the time, you won’t miss it. Now and then, you’ll be lost without it.

Because Microsoft Outlook is so widely used, its PST file format is widely supported by applications designed to view, process and search e-mail. Moreover, the complex structure of a PST is so well understood that many commercial applications can parse PSTs into single message formats or assemble single messages into PSTs. Accordingly, it’s feasible to produce responsive messaging in a PST format while excluding messages that are non-responsive or privileged. It’s also feasible to construct a production PST without calendar content, contacts, to do lists and the like. You'd be hard pressed to find a better form of production for Exchange/Outlook messaging. Here, I'm defining "better" in terms of completeness and functionality, not compatibility with your ESI review tools.

MSGs: There’s little room for debate that the PST or OST container files are the native forms of data storage and interchange for a *collection* of messages (and other content) from Microsoft Outlook. But is there a native format for *individual* messages from Outlook, like the RFC 5322 format discussed above? The answer isn’t clear cut. On the one hand, if you were to drag a single message from Outlook to your Windows desktop, Outlook would create that message in its proprietary MSG format. The MSG format holds the complete content of its RFC 5322 cousin plus additional metadata; but it lacks information (like foldering data) that's contained within a PST. It’s not "native" in the sense that it’s not a format that Outlook uses day-to-day; but it’s an export format that holds more message metadata unique to Outlook. All we can say is that the MSG file is a highly compatible *near-native* format for individual Outlook messages--more complete than the transiting e-mail and less complete than the native PST. Though it’s encoded in a proprietary Microsoft format (i.e., it’s *not* plain text), the MSG format is so ubiquitous that, like PSTs, many applications support it as a standard format for moving messages between applications.

Exchange: The native format for data housed in an Exchange server is its database file, prosaically called the Exchange Database and sporting the file extension .EDB. The EDB holds the account content for everyone in the mail domain; so unless the case is the exceedingly rare one that warrants production of all the e-mail, attachments, contacts and calendars for every user, no litigant hands over their EDB.

It may be possible to create an EDB that contains only messaging from selected custodians (and excludes privileged and non-responsive content) such that you could really, truly produce in a native form. But, I've never seen it done that way, and I can't think of anything to commend it over simpler approaches.

So, if you're not going to produce in the "true" native format of EDB, the desirable alternatives left to you are properly called "near-native," meaning that they preserve the requisite content and essential functionality of the native form, but aren't the native form. If an alternate form doesn't preserve content and functionality, you can call it whatever you want. I lean toward "garbage," but to each his own.

E-mail is a species of ESI that doesn't suffer as mightily as, say, Word documents or Excel spreadsheets when produced in non-native forms. If one were meticulous in their text extraction, exacting in their metadata collection and careful in their load file construction, one could produce Exchange content in a way that's sufficiently complete and utile as to make a departure from the native less problematic—assuming, of course, that one produces the attachments in their native forms. That's a lot of "ifs," and what will emerge is sure to be incompatible with e-mail client applications and native review tools.

Litmus Test: Perhaps we have the makings of a litmus test to distinguish functional near-native forms from dysfunctional forms like TIFF images and load files: ***Can the form produced be imported into common e-mail client or server applications?***

You must admire the simplicity of such a test. If the e-mail produced is so distorted that not even e-mail programs can recognize it as e-mail, that's a fair and objective indication that the form of production has strayed too far from its native origins.

Gmail

The question whether it's feasible to produce Gmail in its native form triggered an order by U.S. Magistrate Judge Mark J. Dinsmore in a case styled, *Keaton v. Hannum,* 2013 U.S. Dist. LEXIS 60519 (S.D. Ind. Apr. 29, 2013). It's a seamy, sad suit brought *pro se* by an attorney named Keaton against both his ex-girlfriend, Christine Zook, and the cops who arrested Keaton for stalking Zook. It got my attention because the court cited a blog post I made some years ago. The Court wrote:

> Zook has argued that she cannot produce her Gmail files in a .pst format because no native format exists for Gmail (i.e., Google) email accounts. The Court finds this to be incorrect based on Exhibit 2 provided by Zook in her Opposition Brief. [Dkt. 92 at Ex. 2 (Ball, Craig: Latin: To Bring With You Under Penalty of Punishment, EDD Update (Apr. 17, 2010)).] Exhibit 2 explains that, although Gmail does not support a "Save As" feature to generate a single message format or PST, the messages can be downloaded to Outlook and saved as .eml or.msg files, or, as the author did, generate a PDF Portfolio – "a collection of multiple files in varying format that are housed in a single, viewable and searchable container." [Id.] In fact, Zook has already compiled most of her archived Gmail emails between her and Keaton in a .pst format when Victim.pst was created. It is not impossible to create a "native" file for Gmail emails.

Id. at 3.

I'm gratified when a court cites my work, and here, I'm especially pleased that the Court took an enlightened approach to "native" forms in the context of e-mail discovery. Of course, one strictly defining "native" to exclude near-native forms might be aghast at the loose lingo; but the more important takeaway from the decision is the need to strive for the most functional and complete forms when true native is out-of-reach or impractical.

Gmail is a giant database in a Google data center someplace (or in many places). I'm sure I don't know what the native file format for cloud-based Gmail might be. Mere mortals don't get to peek at the guts of Google. But, I'm also sure that it doesn't matter, because even if I *could* name the

native file format, I couldn't obtain that format, nor could I faithfully replicate its functionality locally.[31]

Since I can't get "true" native, how can I otherwise mirror the completeness and functionality of native Gmail? After all, a litigant doesn't seek native forms for grins. A litigant seeks native forms to secure the unique benefits native brings, principally functionality and completeness.

There are a range of options for preserving a substantial measure of the functionality and completeness of Gmail. One would be to produce in Gmail.

HUH?!?!

Yes, you could conceivably open a fresh Gmail account for production, populate it with responsive messages and turn over the access credentials for same to the requesting party. That's probably as close to true native as you can get (though some metadata will change), and it flawlessly mirrors the functionality of the source. Still, it's not what most people expect or want. It's certainly not a form they can pull into their favorite e-discovery review tool.

Alternatively, as the Court noted in *Keaton v. Hannum,* an IMAP[32] capture to a PST format (using Microsoft Outlook or a collection tool) is a practical alternative. The resultant PST won't look or work exactly like Gmail (i.e., messages won't thread in the same way and flagging will be different); but it will supply a large measure of the functionality and completeness of the Gmail source. Plus, it's a form that lends itself to many downstream processing options.

---

[31] It was once possible to create complete, offline replications of Gmail using a technology called Gears; however, Google discontinued support of Gears some time ago. Gears' successor, called "Gmail Offline for Chrome," limits its offline collection to just a month's worth of Gmail, making it a complete non-starter for e-discovery. Moreover, neither of these approaches employs true native forms as each was designed to support a different computing environment.

[32] IMAP (for Internet Message Access Protocol) is another way that e-mail client and server applications can talk to one another. The latest version of IMAP is described in RFC 3501. IMAP is not a form of e-mail storage; it is a means by which the structure (i.e., foldering) of webmail collections can be replicated in local mail client applications like Microsoft Outlook. Another way that mail clients communicate with mail servers is the Post Office Protocol or POP; however, POP is limited in important ways, including in its inability to collect messages stored outside a user's Inbox. Further, POP does not replicate foldering. Outlook "talks" to Exchange servers using MAPI and to other servers and webmail services using MAPI (or via POP, if MAPI is not supported).

So, What's the native form of that e-mail?

Which answer do you want; the technically correct one or the helpful one? No one is a bigger proponent of native production than I am; but I'm finding that litigants can get so caught up in the quest for native that they lose sight of what truly matters.

Where e-mail is concerned, we should be less captivated by the term "native" and more concerned with specifying the actual form or forms that are best suited to supporting what we need and want to do with the data. That means understanding the differences between the forms (e.g., what information they convey and their compatibility with review tools), not just demanding native like it's a brand name.

When I seek "native" for a Word document or an Excel spreadsheet, it's because I recognize that the entire native file—and *only* the native file—supports the level of completeness and functionality I need, a level that can't be fairly replicated in any other form. But when I seek native production of e-mail, I don't expect to receive the entire "true" native file. I understand that responsive and privileged messages must be segregated from the broader collection and that there are a variety of near native forms in which the responsive subset can be produced so as to closely mirror the completeness and functionality of the source.

When it comes to e-mail, what matters most is getting all the important information within and about the message in a fielded form that doesn't completely destroy its character as an e-mail message.

So, let's not get *too* literal about native forms when it comes to e-mail. Don't seek native to prove a point. Seek native to prove your case.

_____

Postscript: When I publish an article extolling the virtues of native production, I usually get a comment or two saying, "TIFF and load files are good enough." I can't always tell if the commentator means "good enough to fairly serve the legitimate needs of the case" or "good enough for those sleazy bastards on the other side." I suspect they mean both. Either way, it

might surprise readers to know that, when it comes to e-mail, I agree with the first assessment…with a few provisos.

First, TIFF and load file productions can be good enough for production of e-mail if no one minds paying more than necessary. It generally costs more to extract text and convert messages to images than it does to leave it in a native or near-native form. But that's only part of the extra expense. TIFF images of messages are MUCH larger files than their native or near native counterparts. With so many service providers charging for ingestion, processing, hosting and storage of ESI on a per-gigabyte basis, those bigger files continue to chew away at both side's bottom lines, month-after-month.

Second, TIFF and load file productions are good enough for those who only have tools to review TIFF and load file productions. There's no point in giving light bulbs to those without electricity. On the other hand, just because you don't pay your light bill, must I sit in the dark?

Third, because e-mails and attachments have the unique ability to be encoded entirely in plain text, a load file can carry the complete contents of a message and its contents as RFC 5322-compliant text accompanied by MAPI metadata fields. It's one of the few instances where it's possible to furnish a load file that simply and genuinely compensates for most of the shortcomings of TIFF productions. Yet, it's not done.

Finally, TIFF and load file productions are good enough for requesting parties who just don't care. A lot of requesting parties fall into that category, and they're not looking to change. They just want to get the e-mail, and they don't give a flip about cost, completeness, utility, metadata, efficiency, authentication or any of the rest. If both sides and the court are content not to care, TIFF and load files really are good enough.

Luddite Lawyer's Guide to Computer Backup Systems

Backup is the Rodney Dangerfield of the e-discovery world. It gets no respect. Or, maybe it's Milton, the sad sack with the red stapler from the movie, *Office Space*. Backup is pretty much ignored...until headquarters burns to the ground or it turns out the old tapes in the basement hold the only copy of the all-important TPS reports demanded in discovery.

Would you be surprised to learn that backup is the hottest, fastest moving area of information technology? Consider the:

- Migration of data to the "cloud" *(Minsk! Why's our data in Minsk?);*
- Explosive growth in hard drive capacities *(Four terabytes! On a desktop?);*
- Ascendency of virtual machines *(Isn't that the title of the next Terminator movie?);* and
- Increased reliance on replication *(D2D2T? That's the cute Star Wars droid, right?).*

If you don't understand how backup systems work, you can't reliably assess whether discoverable data exists or how much it will cost in terms of sweat and coin to access, search and recover that data.

The Good and Bad of Backups

Ideally, the contents of a backup system would be entirely cumulative of the active "online" data on the servers, workstations and laptops that make up a network. But because businesses entrust the power to alter and destroy data to every computer user--including those motivated to make evidence disappear—and because companies configure systems to purge electronically stored information as part of records retention programs, backup tapes may prove to be the only source of evidence beyond the reach of those who've failed to preserve evidence and who have an incentive to destroy or fabricate it. Going back as far as 1986 and Col. Oliver North's deletion of e-mail subject to subpoena in the Reagan-era Iran-Contra

Jargon Watch

Look for these key terms:

- *disaster recovery*
- *full backup*
- *differential backup*
- *incremental backup*
- *tape restoration*
- *tape rotation*
- *legacy tapes*
- *replication*
- *drive imaging*
- *bitstream*
- *backup set*
- *backup catalog*
- *tape log*
- *linear serpentine*
- *virtual tape library*
- *D2D2T*
- *RAID*
- *striping*
- *parity*
- *hash value*
- *single-instance storage*
- *non-native restoration*
- *Cloud backup*

affair, it's long been backup systems that ride to truth's rescue with "smoking gun" evidence.

Backup tapes can also be fodder for pointless fishing expeditions mounted without regard for the cost and burden of turning to backup media, or targeted prematurely in discovery, before more accessible data sources have been exhausted.

Grappling with Backup Tapes

Backup tapes are made for ***disaster recovery***, i.e., picking up the pieces of a damaged or corrupted data storage system. Some call backups "snapshots" of data, and like a photo, backup tapes capture only what's in focus. To save time and space, backups typically ignore commercial software programs that can be reinstalled in the event of disaster, so ***full backups*** typically focus on all *user created* data. ***Incremental backups*** grab just what's been created or changed since the last full or incremental backup. Together, they put Humpty-Dumpty back together again in a process called ***tape restoration*.**

Tape is cheap, durable and portable, the last important because backups need to be stored away from the systems at risk. Tape is also slow and cumbersome, downsides discounted because it's so rarely needed for restoration.

Because backup systems have but one legitimate purpose--being the retention of data required to get a business information system "back up" on its feet after disaster--a business only needs recovery data covering a brief interval. No business wants to replicate its systems as they existed six months or even six weeks before a crash. Thus, *in theory*, older tapes are supposed to be recycled by overwriting them in a practice called ***tape rotation*.**

But, as theory and practice are rarely on speaking terms, companies may keep backup tapes long past (sometimes *years* past) their usefulness for disaster recovery and often beyond the IT department's ability to access tapes created with obsolete software or hardware. These ***legacy tapes*** are business records—sometimes the last surviving copy—but are afforded little in the way of *records management*. Even businesses that overwrite tapes every two weeks replace their tape sets from time to time as faster, bigger options hit the market. The old tapes are frequently set aside and forgotten in offsite storage or a box in the corner of the computer room.

Like the DeLorean in "Back to the Future," legacy tapes allow you to travel back in time. It doesn't take 1.2 million gigawatts of electricity, just lots of cabbage.

Duplication, Replication and Backup

We save data from loss or corruption via one of three broad measures: duplication, replication and backup.

Duplication is the most familiar--protecting the contents of a file by making a copy of the file to another location. If the copy is made to another location on the same medium (e.g., another

folder on the hard drive), the risk of corruption or overwriting is reduced. If the copy is made to another medium (another hard drive), the risk of loss due to media failure is reduced. If the copy is made to a distant physical location, the risk of loss due to physical catastrophe is reduced.

You may be saying, "Wait a second. Isn't backup just a form of duplication?" To some extent, it is; and certainly, duplication is the most common "backup" method used on a personal computer. But, true enterprise backup injects other distinctive elements, the foremost being that enterprise backups are not user-initiated but occur systematically, untied to the whims and preferences of individual users.

Replication is duplication without discretion. That is, the contents of one storage medium are periodically or continuously mirrored to another storage medium. Replication may be as simple as RAID 1 mirroring of two local hard drives (where one holds exactly the same data as the other) or as elaborate as keeping a distant data operations center on standby, ready to go into service in the event of a catastrophe.

Unlike duplication and replication, backup involves (reversible) alteration of the data and logging and cataloging of content. Typically, backup entails the use of software or hardware that compresses and encrypts data. Further, backup systems are designed to support iteration, e.g., they manage the scheduling and scope of backup, track the content and timing of backup "sets" and record the allocation of backup volumes across multiple devices or media.

Major Elements of Backup Systems

Understanding backups requires an appreciation of the three major elements of a backup system: the source data, the target data ("backup set") and the catalog.

1. Source Data (Logical or Physical) Though users tend to think of the source data as a collection of files, backup may instead be drawn from the broader, logical divisions of a storage medium, called "partitions," "volumes" and "folders." ***Drive imaging,*** a specialized form of backup employed by IT specialists and computer forensic examiners, may draw from below the logical hierarchy of a drive, collecting a "bitstream" of the drive's contents reflecting the contents of the medium at the physical level. The bitstream of the medium may be stored in a single large file, but more often it's broken into manageable, like-sized "chunks" of data to facilitate more flexible storage.

2. Backup Set (Physical or Logical, Full or Changed-File) A ***backup set*** may refer to a *physical* collection of *media* housing backed up data, i.e., the collective group of magnetic tape cartridges required to hold the data, or the "set" may reference the *logical* grouping of *files* (and associated catalog) which collectively comprise the backed up data. Compare, "*those three LTO tape cartridges*" to "*the backup of the company's Microsoft Exchange Mail Server*."

Backup sets further divide between what can be termed "full backups" and "changed-file backups." As you might expect, full backups tend to copy everything present on the source (or at least "everything" that has been defined as a component of the full backup set) where changed-file backups duplicate items that have been added or altered since the last full backup.

The changed-file components further subdivide into **incremental backups**, **differential backups** and **delta block-level backups**. The first two identify changed files based on either the status of a file's archive bit or a file's created and modified date values. The essential difference is that every differential backup duplicates files added or changed since the last *full* backup, where incremental backups duplicate files added or changed since the last *incremental* backup. The delta block-level method examines the contents of a file and stores only the *differences* between the version of the file contained in the full backup and the modified version. This approach is trickier, but it permits the creation of more compact backup sets and accelerates backup and restoration.

3. Backup Catalog vs. Tape Log Unlike duplication and replication, where generally no record is kept of the files moved or their characteristics, the creation and maintenance of a catalog is a key element of enterprise backup. The ***backup catalog*** tracks*, inter alia,* the source and metadata of each file or component of the backup set as well as the location of the element within the set. The catalog delineates the quantity of target media and identifies and sequences each tape or disk required for restoration. Without a catalog setting out the logical organization of the data as

stored, it would be impossible to distinguish between files from different sources having the same names or to extract selected files without restoration of all of the backed up data.

Equally important is the catalog's role in facilitating single instance backup of identical files. Multiple computers—especially those within the same company—store many files with identical names, content and metadata. It's a waste of time and resources to backup multiple iterations of identical data, so the backup catalog makes it possible to store just a single instance of such files and employ placeholder "stubs" or pointers to track all locations to which the file should be restored.

Obviously, *lose* the catalog, and it's tough to put Humpty Dumpty back together again.

It's important to distinguish the catalog--a detailed digital record that, if printed, would run to hundreds of pages or more--from the ***tape log,*** which is typically a simple listing of backup events and dates, machines and tape identifier. *See*, *e.g.,* the sample page of a tape log attached as Appendix A.

Backup Media: Tape and Disk-to-Disk

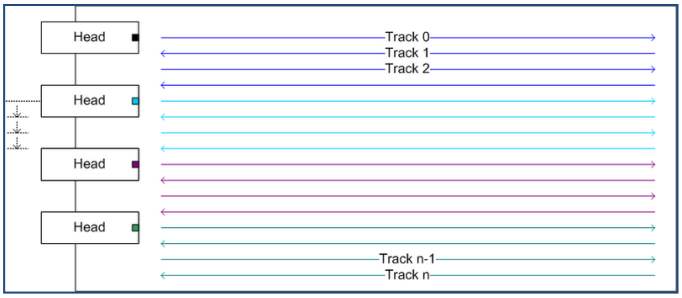
Tape Backup

Though backup tape seems almost antique, tape technology has adapted well to modern computing environments. The IBM 3420 reel-to-reel backup tapes that were a computer room staple in the 1970s and '80s employed 240 feet of half-inch tape on 10.5-inch reels. These tapes were divided into 9 tracks of data and held a then-impressive 100 megabytes of information traveling at 1.2 megabytes per second.

Today's LTO-7 tapes are housed in a 4-inch square LTO cartridge less than an inch thick and feature 3,150 feet of half-inch tape divided into 2,176 tracks holding 6 terabytes of information transferring at 300 megabytes per second.

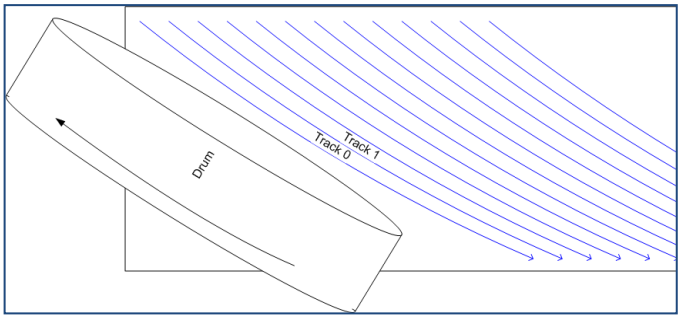
That's 240 times as many tracks, 250 times faster data transfer and *60,000 times greater* data storage capability in a far smaller package.

Mature readers may recall "auto-reverse" tape transport mechanisms, which eliminated the need to eject and turn over an audiocassette to play the other side. Many modern backup tapes use a scaled-up version of that back-and-forth or ***linear serpentine*** recording scheme. "Linear" because it stores data in parallel tracks running the length of the tape, and "serpentine" because its path snakes back-and-forth like a mountain road. Thirty-two of the LTO-7 cartridge's 3,584 tracks are read or written as the tape moves past the heads, so it takes *112 back-and-forth passes* or "wraps" to read or write the full contents of a single LTO-7 cartridge.



That's about 67 *miles* of tape passing the heads!

An alternate recording scheme employed by SAIT-2 tape systems employs a ***helical recording*** system that writes data in parallel tracks running diagonally across the tape, much like a household VCR. Despite a slower transfer rate, helical recording also achieves 800GB of storage capacity on 755 feet of 8mm tape housed in a compact cartridge like that used in handheld video cameras. Development of SAIT tape technology was abandoned in 2006 and Sony stopped selling SAIT in 2010; so, they aren't seen much beyond tape archives.



Why is Tape So Slow?

Clearly, tape is a pretty remarkable technology that's seen great leaps in speed and capacity. The latest tapes on the market can reportedly outstrip the ability of a hard drive to handle their throughput.

Still, even the best legal minds have yet to find loopholes in those pesky laws of physics.

All that serpentine shuttling back and forth over 67 miles of tape is a mechanical process. It occurs at a glacial pace relative to the speed with which computer circuits move data.

Further, backup restoration is often an incremental process. Reconstructing reliable data sets may require data from multiple tapes to be combined. Add to the mix the fact that as hard drive capacities have exploded, tape must store more and more information to keep pace. Gains in performance are offset by growth in volume.

How Long to Restore?

Several years ago, the big Atlanta tape house, eMag Solutions, LLC, weighed in on the difference between the time it *should* take to restore a backup tape considering just its capacity and data transfer rate versus the time it *really* takes considering the following factors that impact restoration:

- Tape format;
- Device interface, i.e., SCSI or fiber channel;
- Compression;
- Device firmware;
- The number of devices sharing the bus;
- The operating system driver for the tape unit;
- Data block size (large blocks fast, small blocks slow);
- File size (with millions of small files, each must be cataloged);
- Processor power and adapter card bus speed;
- Tape condition (retries eat up time);
- Data structure (e.g., big database vs. brick level mailbox accounts);
- Backup methodology (striped data? multi server?).

The following table reflects eMag's reported experience:

Drive Type	Native cartridge capacity	Drive Native Data Transfer Speed[33]	Theoretical Minimum Data Transfer Time	Typical Real World Data Transfer Time
DLT7000	35GB	3MB/sec	3.25 Hrs	6.5 Hrs
DLT8000	40GB	3MB/sec	3.7 Hrs	7.4 Hrs
LTO1	100GB	15MB/sec	1.85 Hrs	4.0 Hrs
LTO2	200GB	35MB/sec	1.6 Hrs	6.0 Hrs
SDLT 220	110GB	11MB/sec	2.8 Hrs	6.0 Hrs
SDLT 320	160GB	16MB/sec	2.8 Hrs	6.0 Hrs

The upshot is that it takes *about twice as long* to restore a tape under real world conditions than the media's stated capacity and transfer rate alone would suggest. Just to generate a catalog for a tape, the tape must be read in its entirety. Consequently, it's not feasible to deliver 3,000 tapes to a vendor on Friday and expect a catalog to be generated by Monday. The *price* to do the work has dropped dramatically, but the *time* to do the work has not.

[33] " *How Long Does it Take to Restore a Tape*," eMag blog, 7/17/2009 at http://tinyurl.com/tapetime, Some of these transfer rate values are at variance with manufacturer's stated values, but they are reported here as published by eMag.

Extrapolating from this research, we can conceive a formula to estimate the real world time to restore a set of backup tapes of consistent drive type and capacity, and considering that, employing multiple tape drives, tapes may be restored simultaneously:

$$\textbf{Real World Transfer Time (in Hours)} = \frac{\textbf{Native Cartridge Capacity (in GB)}}{\textbf{1.8 x Drive Native Transfer Speed}}$$

Applying this to a LTO-7 tape:

$$\frac{\textbf{Native Cartridge Capacity (in GB)}}{\textbf{1.8 x Transfer Speed (in MB/s)}} = \frac{\textbf{6 TB}}{\textbf{1.8 x 300}} = \frac{\textbf{6,000}}{\textbf{540}} = \textbf{11.1 hours}$$

Of course, this is merely a rule-of-thumb for a single tape. As you seek to apply it to a large-scale data restoration, be sure to factor in other real world factors impacting speed, such as the ability to simultaneously use multiple drives for restoration, the need to swap tapes and replace target drives, to clean and align drive mechanisms, the working shifts of personnel, weekend and holidays, time needed for recordkeeping, for resolving issues with balky tapes and for steps taken in support of quality assurance.

Common Tape Formats

The LTO tape format is the clear winner of the tape format wars, having eclipsed all contenders save the disk and cloud storage options that now threaten to end tape's enduring status as the leading backup medium. As noted, the recently released LTO-7 format natively holds 6.0 terabytes of data at a transfer rate of 300 megabytes per second. *These values are expected to continue to double roughly every two years through 2020.* Tape use is down, but not out—not for some time.

Too, the dusty catacombs beneath Iron Mountain still brim with all manner of legacy tape formats that will be drawn into e-discovery fights for years to come. Here are some of the more common formats seen in the last 30 years and their characteristics:

Name	Format	A/K/A	Length	Width	Capacity (GB)	Transfer Rate (MB/sec)
DLT 2000	DLT3	DLT	1200 ft	1/2"	10	1.25
DLT 2000 XT	DLT3XT	DLT	1828 ft	1/2"	15	1.25
DLT 4000	DLT 4	DLT	1828 ft	1/2"	20	1.5
DLT 7000	DLT 4	DLT	1828 ft	1/2"	35	5
DLT VS-80	DLT 4	TK-88	1828 ft	1/2"	40	3
DLT 8000	DLT 4	DLT	1828 ft	1/2"	40	6
DLT-1	DLT 4	TK-88	1828 ft	1/2"	40	3

Name	Format	A/K/A	Length	Width	Capacity (GB)	Transfer Rate (MB/sec)
DLT VS-160	DLT 4	TK-88	1828 ft	1/2”	80	8
SDLT-220	SDLT 1		1828 ft	1/2”	110	10
DLT V4	DLT 4	TK-88	1828 ft	1/2”	160	10
SDLT-320	SDLT 1		1828 ft	1/2”	160	16
SDLT 600	SDLT 2		2066 ft	1/2”	300	36
DLT-S4	DLT-S4	DLT Sage	2100 ft	1/2”	800	60
DDS-1	DDS-1	DAT	60M	4mm	1.3	.18
DDS-1	DDS-1	DAT	90M	4mm	2.0	.18
DDS-2	DDS-2	DAT	120M	4mm	4	.60
DDS-3	DDS-3	DAT	125M	4mm	12	1.1
DDS-4	DDS-4	DAT	150M	4mm	20	3
DDS-5	DAT72	DAT	170M	4mm	36	3
DDS-6	DAT160	DAT	150M	4mm	80	6.9
M1	AME	Mammoth	22M	8mm	2.5	3
M1	AME	Mammoth	125M	8mm	14	3
M1	AME	Mammoth	170M	8mm	20	3
M2	AME	Mammoth 2	75M	8mm	20	12
M2	AME	Mammoth 2	150M	8mm	40	12
M2	AME	Mammoth 2	225M	8mm	60	12
Redwood	SD3	Redwood	1200 ft	1/2”	10/25/50	11
TR-1		Travan	750 ft	8mm	.40	.25
TR-3		Travan	750 ft	8mm	1.6	.50
TR-4		Travan	740 ft	8mm	4	1.2
TR-5		Travan	740 ft	8mm	10	2.0
TR-7		Travan	750 ft	8mm	20	4.0
AIT 1	AIT		170M	8mm	25	3
AIT 1	AIT		230M	8mm	35	4
AIT 2	AIT		170M	8mm	36	6
AIT 2	AIT		230M	8mm	50	6
AIT 3	AIT		230M	8mm	100	12
AIT 4	AIT		246M	8mm	200	24
AIT 5	AIT		246M	8mm	400	24

Name	Format	A/K/A	Length	Width	Capacity (GB)	Transfer Rate (MB/sec)
Super AIT 1	AIT	SAIT-1	600M	8mm	500	30
Super AIT 2	AIT	SAIT-2	640M	8mm	800	45
3570 B	3570b	IBM Magstar MP		8mm	5	2.2
3570 C	3570c	IBM Magstar MP		8mm	5	7
3570 C	3570c XL	IBM Magstar MP		8mm	7	7
IBM3592	3592	3592	609m	1/2”	300	40
T9840A	Eagle		886 ft	1/2”	20	10
T9840B	Eagle		886 ft	1/2”	20	20
T9840C	Eagle		886 ft	1/2”	40	30
T9940A			2300 ft	1/2”	60	10
T9940B			2300 ft	1/2”	200	30
T10000	T10000	STK Titanium		1/2”	500	120
T10000B	T10000B			1/2”	1000	120
T10000C	T10000C			1/2”	5000	240
T10000D	T10000D			1/2”	8500	252
Ultrium	Ultrium	LTO 1	609M	1/2”	100	15
Ultrium	Ultrium	LTO 2	609M	1/2”	200	40
Ultrium	Ultrium	LTO 3	680M	1/2”	400	80
Ultrium	Ultrium	LTO 4	820M	1/2”	800	120
Ultrium	Ultrium	LTO 5	846M	1/2”	1,500	140
Ultrium	Ultrium	LTO 6	846M	1/2”	2,500	160
Ultrium	Ultrium	LTO 7	960M	1/2”	6,000	300

Disk-to-Disk Backup

Tapes are stable, cheap and portable—a natural media for moving data in volumes too great to transmit by wire without consuming excessive bandwidth and disrupting network traffic. But strides in deduplication and compression technologies, joined by drops in hard drive costs and leaps in hard drive capacities, have eroded the advantages of tape-based transfer and storage.

When data sets are deduplicated to unique content and further trimmed by compression, much more data resides in much less drive space. With cheaper, bigger drives flooding the market, hard drive storage capacity has grown to the point that disk backup intervals are on par with the routine rotation intervals of tape systems (e.g., 8-16 weeks), Consequently, disk-to-disk backup options once considered too expensive or disruptive are feasible.

Hard disk arrays can now hold months of disaster recovery data at a cost that competes favorably with tape. Thus, tape is ceasing to be a disaster recovery medium and is instead being used solely for long-term data storage; that is, as a place to migrate disk backups for purposes *other than* disaster recovery, *i.e.*, archival.

Of course, the demise of tape backup has been confidently predicted for years, even while the demand for tape continued to grow. But for the first time, the demand curve for tape has begun to head south.

D2D (for Disk-to-Disk) backup made its appearance wearing the sheep's clothing of tape. In order to offer a simple segue from the 50-year dominance of tape, the first disk arrays were designed to emulate tape drives so that existing software and programmed backup routines needn't change. These are ***virtual tape libraries*** or *VTL*s.

As D2D supplants tape for backup, the need remains for a stable, cheap and portable medium for long-term retention of archival data--the stuff too old to be of value for disaster recovery but comprising the digital annals of the enterprise. This need continues to be met by tape, a practice that has given rise to a new acronym: ***D2D2T*,** for Disk-to-Disk-to-Tape. By design, tape now holds the company's archives, which ensures the continued relevance of tape backup systems to e-discovery.

Essential Technologies: Compression and Deduplication

Along with big, cheap hard drives and RAID redundancy, compression and deduplication have made cost-effective disk-to-disk backup possible. But compression and deduplication are important for tape, too, and bear further mention.

Compression

The design of backup systems is driven by considerations of speed and cost. Perhaps surprisingly, the speed and expense with which an essential system can be brought back online after failure is less critical than the speed and cost of each backup. The reason for this is that (hopefully) failure is a rare occurrence whereas backup is (or should be) frequent and routine. Certainly, no one would seriously contend that restoring a failed system from a morass of magnetic tape is the fastest, cheapest way to rebuild a failed system. No, the advantage of tape is its relatively low cost per gigabyte to store data, not to restore it.

Electrons move much faster than machines. The slowest parts of any backup systems are the mechanical components: the spinning reels, moving heads and the human beings loading and unloading tape transports. One way to maximize the cost advantage and efficiency of tape is to increase the density of data that can be stored per inch of tape. The more you can store per inch,

the fewer tapes to be purchased and loaded and the fewer miles of tape to pass by the read-write heads.

Because electrons move speed-of-light faster than mechanical parts of backup systems, a lot of computing power can be devoted to restructuring data in ways that it fits more efficiently on tape or disk. For example, if a horizontal line on a page were composed of one hundred dashes, it takes up less space to describe the line as "100 dashes" or 100- than to actually type out 100 dashes. Of course, it would take some time to count the dashes, determine there were precisely 100 of them and ensure the shorthand reference "100 dashes" doesn't conflict with some other part of the text; but, these tasks can be accomplished by digital processors in infinitely less time than that required to spin a reel of tape to store the difference between the data and its shorthand reference.

This is the logic behind data compression; that is, the use of computing power to re-express information in more compact ways to achieve higher transfer rates and consume less storage space. Compression is an essential, ubiquitous technology. Without it, there would be no YouTube, Netflix, streaming music and video, DVRs, HD digital cameras, Internet radio and much else that we prize in the digital age.

And without compression, you'd need a whole lot more time, tape and money to back up a computer system.

While compression schemes for files tend to comprise a fairly small number of published protocols (e.g., Zip, LZH), compression algorithms for backup have tended to be proprietary to the backup software or hardware implementing them and to change from version-to-version. Because of this, undertaking the restoration of legacy backup tapes entails more than simply finding a compatible tape drive and determining the order and contents of the tapes. You may also need particular software to decompress the data.

Deduplication

Companies that archive backup tapes may retain years of tapes, numbering in the hundreds or thousands. Because each full backup is a snapshot of a computer system at the time it's created, there is a substantial overlap between backups. An e-mail in a user's Sent Items mailbox may be there for months or years, so every backup replicates that e-mail, and restoration of every backup adds an identical copy to the material to be reviewed. Restoration of a year of monthly backups would generate 12 copies of the same message, thereby wasting reviewers' time, increasing cost and posing a risk of inconsistent treatment of identical evidence (as occurs when one reviewer flags a message as privileged but another decides it's not). The level of duplication between ne backup to the next is often as high as 90%.

Consider, too, how many messages and attachments are dispatched to all employees or members of a product team. Across an enterprise, there's a staggering level of repetition.

Accordingly, an essential element of backup tape restoration is deduplication; that is, using computers to identify and cull identical electronically stored information before review. Deduplicating within a single custodian's mailboxes and documents is called ***vertical deduplication***, and it's a straightforward process. However, corporate backup tapes aren't geared to single users. Instead, business backup tapes hold messages and documents for multiple custodians storing identical messages and documents. Restoration of backup tapes generates duplicates within individual accounts (vertically) and across multiple users (horizontally). Deduplication of messages and documents across multiple custodians is called (not surprisingly) ***horizontal deduplication*.**

Horizontal deduplication significantly reduces the volume of information to be reviewed and minimizes the potential for inconsistent characterization of identical items; however, it can make it impossible to get an accurate picture of an individual custodian's data collection because many constituent items may be absent, eliminated after being identified as identical to another user's items.

Consequently, deduplication plays two crucial roles when backup sets are used as a data source in e-discovery. First, deduplication must be deployed to eliminate the substantial repetition from one backup iteration to the next; that is, to eliminate that 90% overlap mentioned above. Second, deduplication is useful in reducing the cost and burden of review by eliminating vertical and horizontal repetition within and across custodians.

Modern backup systems are designed to deduplicate ESI *before* it's stored; that is, to eliminate all but a single instance of recurring content, hence the name, ***single-instance storage*.** Using a method called *in-line deduplication,* a unique digital fingerprint or *hash value* is calculated for each file or data block as it's stored and that hash value is added to a list of stored files. Before being stored, each subsequent file or data block has its hash value checked against the list of stored files. If an identical file has already been stored, the duplicate is not added to the backup media but, instead, a pointer or stub to the duplicate is created. An alternate approach, called *post-process deduplication*, works in a similarly, except that all files are first stored on the backup medium, then analyzed and selectively culled to eliminate duplicates.

Data Restoration

Clearly, data in a backup set is a bit like the furniture at Ikea: It's been taken apart and packed tight for transport and storage. But, when that data is needed for e-discovery--it must be reconstituted and reassembled. It starts to take up a lot of space again. That restored data has to go *somewhere,* usually t*o* a native computing environment just like the one from which it came.

But the system where it came from may be at capacity with new data or not in service anymore. Historically, small and mid-size companies lacked the idle computing capacity to effect restoration without a significant investment in equipment and storage. Larger enterprises devote more stand-by resources to recovery for disaster recovery and may have had alternate environments ready to receive restored data, but those resources had to be at the ready in the event of emergency. It was often unacceptably risky to dedicate them, even briefly, to electronic discovery.

The burden and cost of recreating a restoration platform for backup data was a major reason why backup media came to be emblematic of ESI deemed "not reasonably accessible." But while the inaccessibility presumption endures, newer technology has largely eliminated the need to recreate a native computing environment in order to restore backup tapes. Today, when a lawyer or judge opines that "backups are not reasonably accessible, *per se*," you can be sure they haven't looked at the options in several years.

Non-Native Restoration

A key enabler of low cost access to tapes and other backup media has been the development of software tools and computing environments that support ***non-native restoration***. Non-native restoration dispenses with the need to locate copies of particular backup software or to recreate the native computing environment from which the backup was obtained. It eliminates the time, cost and aggravation associated with trying to reconstruct a sometimes decades-old system. All major vendors of tape restoration services offer non-native restoration options, and it's even possible to purchase software facilitating in-house restoration of tape backups to non-native environments.

Perhaps the most important progress has been made in the ability of vendors both to generate comprehensive indices of tape contents and extract specific files or file types from backup sets. Consequently, it's often feasible for a vendor to, e.g., acquire just certain types of documents for particular custodians without the need to restore all data in a backup. In some situations, backups are simply not that much harder or costlier to deal with in e-discovery than active data, and they're occasionally the smarter *first* resort in e-discovery.

Going to the Tape *First*?

Perhaps due to the *Zubulake*[34] opinion or the commentary to the 2006 amendments to the Federal Rules of Civil Procedure,[35] e-discovery dogma is that backup tapes are the costly, burdensome recourse of last resort for ESI.

Pity. Sometimes backup tapes are the *easiest, most cost-effective* source of ESI.

For example, if the issue in the case turns on e-mail communications between Don and Elizabeth during the last week of June of 2007, but Don's no longer employed and Elizabeth doesn't keep all

[34] Zubulake v. UBS Warburg, 217 F.R.D. 309 (S.D.N.Y. 2003

[35] Fed R. Civ. P. 26(b)(2)(B).

her messages, what are you going to do? If these were messages that should have been preserved, you could pursue a forensic examination of Elizabeth's computer (cost: $5,000-$10,000) or collect and search the server accounts and local mail stores of 50 other employees who might have been copied on the missing messages (cost: $25,000-$50,000).

Or, you could go to the backup set for the company's e-mail server from July 1 and recover just Don's or Elizabeth's mail stores (cost: $1,000-$2,500).

The conventional wisdom would be to fight any effort to go to the tapes, but the numbers show that, on the right facts, it's both faster and cheaper to do so.

Sampling

Sampling backup tapes entails selecting parts of the tape collection deemed most likely to yield responsive information and restoring and searching only those selections before deciding whether to restore more tapes. Sampling backup tapes is like drilling for oil: You identify the best prospects and drill exploratory wells. If you hit dry holes, you pack up and move on. But if a well starts producing, you keep on developing the field.

The size and distribution of the sample hinges on many variables, among them the breadth and organization of the tape collection, relevant dates, fact issues, business units and custodians, resources of the parties and the amount in controversy. Ideally, the parties can agree on a sample size or they can be encouraged to arrive at an agreement through a mediated process.

Because a single backup may span multiple tapes, and because recreation of a full backup may require the contents of one or more incremental or differential backup tapes, sampling of backup tapes should be thought of as the selection of data snapshots at intervals rather than the selection of tapes. Sensible sampling necessitates access to and an understanding of the tape catalog. Understanding the catalog likely requires explanation of both the business system hardware (*e.g., What is the SQL Server's purpose?*) and the logical arrangement of data on the source machines *(e.g., What's stored in the Exchange Data folder?*). Parties should take pains to insure that each sample is complete for a selected date or interval; that is, the number of tapes shouldn't be arbitrary but should fairly account for the totality of information captured in a single relevant backup event.

Backup and the Cloud

Nowhere is the observation that "*the Cloud changes everything*" more apt than when applied to backups. Microsoft, Amazon, Rackspace, Google and a host of other companies are making it practical and cost-effective to eschew local backups in favor of backing up data securely over the internet to leased repositories in the Cloud. The cost per gigabyte is literally pennies now and, if history is a guide, will continue to decrease to staggeringly low rates as usage explodes.

The incidence of adoption of cloud computing and storage among corporate IT departments is enormous and, assuming no high profile gaffes, will accelerate with the availability of high bandwidth network connections and as security concerns wane.

But the signal impact of the Cloud won't be as a medium for backup of corporate data but as a means to obviate *any* need for user backup. As data and corporate infrastructure migrate to the cloud, backup will cease to be a customer responsibility and will occur entirely behind-the-scenes as a perennial responsibility of the cloud provider. The cloud provider will likely fulfill that obligation via a mix of conventional backup media (e.g., tape) and redundancy across far-flung regional datacenters. But, no matter. *How the cloud provider handles its backup responsibility will be no concern of the customer so long as the system maintains uptime availability.*

Welcome to the Future

In 2009, Harvard Law professor Lawrence Lessig observed, "We are not going back to the twentieth century. In a decade, a majority of Americans will not even remember what that century was like."[36] Yet, much of what even tech-savvy lawyers understand about enterprise backup systems harkens back to a century sixteen years gone. If we do go back to the *information* of the twentieth century, it's likely to come from backup tapes.

> **We are not going back to the twentieth century. In a decade, a majority of Americans will not even remember what that century was like.**
>
> Lawrence Lessig

Backup is unlikely to play a large role in e-discovery in the twenty-first century, if only because the offline backup we knew--dedicated to disaster recovery and accreted grandfather-father-son[37]--is fast giving way to data repositories nearly as accessible as our own laptops. The distinction between inaccessible backups and accessible active data stores will soon be just a historical curiosity, like selfie sticks or Sarah Palin. Instead, we will turn our attentions to a panoply of electronic archives encompassing tape, disk and "cloud" components. The information we now pull from storage and extract tape-by-tape will simply be available to us--all the time--until someone jumps through hoops to make it go away.

Our challenge won't be in restoring information, but in making sense of it.

---

[36] Lawrence Lessig, *Against Transparency*, The New Republic, October 9, 2009.

[37] Grandfather-father-son describes the most common rotation scheme for backup media. The last daily "son" backup graduates to "father" status at the end of each week. Weekly "father" backups graduate to "grandfather" status at the end of each month. Grandfather backups are often stored offsite long past their utility for disaster recovery.

TEN PRACTICE TIPS FOR BACKUPS IN CIVIL DISCOVERY

1. **Backup ≠ Inaccessible.** Don't expect to exclude the content of backups from the scope of discovery if you haven't laid the foundation to do so. Fed. R. Civ. P. 26(b)(2)(B) requires parties identify sources deemed not reasonably accessible because of undue burden or cost. ***Be prepared to prove the cost and burden through reliable metrics and testimony.***
2. ***Determine if your client:***
   • Routinely restores backup tapes to, *e.g.*, insure the system is functioning properly or as a service to those who have mistakenly deleted files;
   • Restored the backup tapes other matters or uses them as an archive;
   • Has the system capacity and in house expertise to restore the data;
   • Has the capability to search the tapes for responsive data?
3. **Don't blindly pull tapes for preservation.** Backup tapes don't exist in a vacuum but as part of an information system. A properly managed system incorporates labeling, logging and tracking of tapes, permitting reliable judgments to be made about what's on particular tapes insofar as tying contents to business units, custodians, machines, data sets and intervals. It's costly to have to process tapes just to establish their contents. ***Always* preserve associated backup catalogues when you preserve tapes.**
4. **Be prepared to put forward a sensible sampling protocol in lieu of wholesale restoration.**
5. **Test and sample backups to determine if they hold responsive, material and unique ESI.** Judges are unlikely to force you to restore backup tapes when sensible sampling regiments demonstrate that the effort is likely to yield little of value. Backup tapes are like drilling for oil: *After a few dry holes, it's time to find a new prospect*.
6. **Be prepared to show that the relevant data on tapes is available from more accessible sources.** Sampling, testing and expert testimony help here.
7. **Know the limits of backup search capabilities.** Most backup tools have search capabilities; however, few of these are up to the task of e-discovery. Can the tool search within all common file types and compressed and container file formats?
8. **Appearances matter!** What would the Judge think if she walked through your client's tape storage area? Does it look like a dumping ground?
9. **If using a cloud-based backup system, consider bringing your e-discovery tools to the data in the Cloud instead of spending days getting the data out.**
10. **Backup tape is for disaster recovery.** If it's too stale to use to bring the systems back up, why keep it? Get rid of it!

Appendix 1: Exemplar Backup Tape Log

Tape No.	Sess. ID	Host Name	Backup Date/Time	Size in Bytes	Session Type
ABC 001	37	EX1	8/1/2007 6:15	50,675,122,176	Exchange 200x
ABC 001	38	EX1	8/1/2007 8:28	337,707,008	System state
ABC 001	39	MGT1	8/1/2007 8:29	6,214,713,344	files incremental or differential
ABC 001	40	MGT1	8/1/2007 8:45	5,576,392,704	SQL Database Backup
ABC 001	41	SQL1	8/1/2007 8:58	10,004,201,472	files incremental or differential
ABC 001	42	SQL1	8/1/2007 9:30	8,268,939,264	SQL Database Backup
ABC 001	43	SQL1	8/1/2007 9:52	272,826,368	System state
ABC 005	2	EX1	8/14/2007 18:30	51,735,363,584	Exchange 200x
ABC 005	3	EX1	8/14/2007 20:35	338,427,904	System state
ABC 005	4	MGT1	8/14/2007 20:38	6,215,368,704	files incremental or differential
ABC 005	5	MGT1	8/14/2007 20:53	5,677,776,896	SQL Database Backup
ABC 005	6	SQL1	8/14/2007 21:06	10,499,260,416	files incremental or differential
ABC 005	7	SQL1	8/14/2007 21:38	8,322,023,424	SQL Database Backup
ABC 005	8	SQL1	8/14/2007 21:57	273,022,976	System state
ABC 002	207	NT1	8/15/2007 20:19	31,051,481,088	loose files
ABC 002	18	NT1	8/16/2007 8:06	47,087,616,000	loose files
ABC 014	9	EX1	8/17/2007 6:45	52,449,443,840	Exchange 200x
ABC 014	10	EX1	8/17/2007 8:53	337,969,152	System state
ABC 014	11	MGT1	8/17/2007 8:54	6,215,368,704	files incremental or differential
ABC 014	12	MGT1	8/17/2007 9:09	5,698,748,416	SQL Database Backup
ABC 014	13	SQL1	8/17/2007 9:22	10,537,009,152	files incremental or differential
ABC 014	14	SQL1	8/17/2007 9:47	8,300,986,368	SQL Database Backup
ABC 014	15	SQL1	8/17/2007 10:08	272,629,760	System state
ABC 003	16	NT1	8/18/2007 6:15	46,850,179,072	loose files
ABC 003	17	NT1	8/18/2007 9:26	44,976,308,224	loose files
ABC 004	19	NT1	8/21/2007 6:16	46,901,690,368	loose files
ABC 004	20	NT1	8/21/2007 9:30	44,742,868,992	loose files
ABC 009	30	EX1	8/22/2007 8:52	53,680,603,136	Exchange 200x
ABC 009	31	EX1	8/22/2007 11:01	348,782,592	System state

ABC 009	32	MGT1	8/22/2007 11:03	6,215,434,240	files incremental or differential
ABC 009	33	MGT1	8/22/2007 11:18	5,715,722,240	SQL Database Backup
ABC 009	34	SQL1	8/22/2007 11:31	10,732,371,968	files incremental or differential
ABC 009	35	SQL1	8/23/2007 4:08	8,362,000,384	SQL Database Backup
ABC 009	36	SQL1	8/23/2007 4:33	272,629,760	System state
ABC 011	44	NT1	8/23/2007 6:16	46,938,193,920	loose files
ABC 011	45	NT1	8/23/2007 9:32	44,611,403,776	loose files

Databases in E-Discovery

When I set out to write this chapter on databases in electronic discovery, I went to the literature to learn prevailing thought and ensure I wasn't treading old ground. What I found surprised me.

I found there's next to no literature on the topic! What little authority exists makes brief mention of flat file, relational and enterprise databases, notes that discovery from databases is challenging and then flees to other topics.[38] A few commentators mention *In re Ford Motor Co.,*[39] the too-brief 2003 decision reversing a trial court's order allowing a plaintiff to root around in Ford's databases with nary a restraint. Although the 11th Circuit cancelled that fishing expedition, they left the door open for a party to gain access to an opponent's databases on different facts, such as where the producing party fails to meet its discovery obligations.

The constant counsel offered by any article touching on databases in e-discovery is "get help." That's good advice, but not always feasible or affordable.

Because databases run the world, we can't avoid them in e-discovery. We have to know enough about how they work to deal with them when the case budget or time constraints make hiring an expert impossible. We need to know how to identify and preserve databases, and we must learn how to gather sufficient information about them to frame and respond to discovery about databases.

Databases run the world

You can't surf the 'net, place a phone call, swipe your parking access card, use an ATM, charge a meal, buy groceries, secure a driver's license, book a flight or get admitted to an emergency room without a database making it happen.

Databases touch our lives all day, every day. Our computer operating systems and e-mail applications are databases. The spell checker in our word processor is a database. Google and Yahoo search engines are databases. Westlaw and Lexis, too. Craigslist. Amazon.com. E-Bay. Facebook. All big honkin' databases.

Yet, when it comes to e-discovery, we tend to fix our attention on documents, without appreciating that most electronic evidence exists only as a flash mob of information assembled and organized on the fly from a dozen or thousand or million discrete places. In our zeal to lay

---

[38] Happily, since I first published, others have waded in and produced more practical scholarship. Here are links to two recent, thoughtful publications on the topic:
Requests for Production of Databases: Documents v. Data, by Christine Webber and Jeff Kerr.
The Sedona Conference Database Principles Addressing the Preservation & Production of Databases & Database Information in Civil Litigation

[39] 345 F.3d 1315 (11th Cir. 2003)

hands on documents instead of data, we make discovery harder, slower and costlier. Understanding databases and acquiring the skills to peruse and use their contents gets us to the evidence better, faster and cheaper.

Databases are even changing the way we think about discovery. Historically, parties weren't obliged to *create* documents for production in discovery; instead, you produced what you had on file. Today, documents don't exist until you generate them. Tickets, bank statements, websites, price lists, phone records and register receipts are all just *ad hoc* reports generated by databases. Documents don't take tangible form until you print them out, and more and more, only the tiniest fraction of documents—one-tenth of one percent—will emerge as ink on paper, obliging litigants to be adept at both crafting queries to elicit responsive data and mastering ways to interpret and use the data stream that emerges.

Introduction to Databases

Most of us use databases with no clue how they work. Take e-mail, for example. Whether you know it or not, each e-mail message you view in Outlook or through your web browser is a report generated by a database query and built of select fields of information culled from a complex dataset. It's then presented to you in a user-friendly arrangement determined by your e-mail client's capabilities and user settings.

That an e-mail message is not a single, discrete document is confusing to some. The data segments or "fields" that make up an e-mail are formatted with such consistency from application-to-application and appear so similar when we print them out that we mistake e-mail messages for fixed documents. But each is really a customizable report from the database called your e-mail.

When you see a screen or report from a database, you experience an assemblage of information that "feels" like a document, but the data that comes together to create what you see are often drawn from different sources within the database and from different systems, locations and formats, all changing moment to moment.

Understanding databases begins with mastering some simple concepts and a little specialized terminology. Beyond that, the distinction between your e-mail database and Google's is mostly marked by differences in scale, optimization and security.

Constructing a Simple Database

If you needed a way to keep track of the cases on your docket, you'd probably begin with a simple table of columns and rows written on a legal pad. You'd start listing your clients by name. Then, you might list the names of other parties, the case number, court, judge and trial date. If you still

had room, you'd add addresses, phone numbers, settlement demands, insurance carriers, policy numbers, opposing counsel and so on.

In database parlance, you've constructed a **"table,"** and each separate information item you entered (e.g., name, address, court) is called a **"field."** The group of items you assembled for each client (probably organized in columns and arranged in a row to the right of each name) is collectively called a **"record."** Because the client's name is the field that governs the contents of each record, it would be termed the **"key field."**

Pretty soon, your table would be unwieldy and push beyond the confines of a sheet of paper. If you added a new matter or client to the table and wanted it to stay in alphabetical order by client name, you'd probably have to rewrite the list.

So, you might turn to index cards. Now, each card is a "record" and lists the information (the "fields") pertinent to each client. It's easy to add cards for new clients and re-order them by client name. Then, sometimes you'd want to order matters by trial date or court. To do that, you'd either need to extract specific data from each card to compile a report, re-sort the cards, or maintain three sets of differently ordered cards, one by name, one by trial date and a third by court.

Your cards comprise a database of three tables. They are still deemed tables even though you used a card to hold each record instead of a row. One table uses client name as its key field, another uses the trial date and the third uses the court. Each of these three sets of cards is a **"flat file database,"** distinguished by the characteristic that all the fields and records (the cards) comprise a single file (i.e., each a deck of cards) with no relationships or links between the various records and fields except the table structure (the order of the deck and the order of fields on the cards).

Of course, you'd need to keep all cards up-to-date as dates, phone numbers and addresses change. When a client has more than one matter, you'd have to write all the same client data on multiple cards and update each card, one-by-one, trying not to overlook any card. What a pain!

So, you'd automate, turning first to something like a spreadsheet. Now, you're not limited by the dimensions of a sheet of paper. When you add a new case, you can insert it anywhere and re-sort the list by name, court or trial date. You're not bound by the order in which you entered the information, and you can search electronically.

Though faster and easier to use than paper and index cards, your simple spreadsheet is still just a table in a flat file database. You must update every field that holds the same data when that data

changes (though "find and replace" functions make this more efficient and reliable), and when you want to add, change or extract information, you have to open and work with the entire table.

What you need is a system that allows a change to *one* field to update *every* field in the database with the same information, not only within a single table but across *all* tables in the database. You need a system that identifies the relationship between common fields of data, updates them when needed and, better still, uses that common relationship to bring together more related information. Think of it as adding rudimentary intelligence to a database, allowing it to "recognize" that records sharing common fields likely relate to common information. Databases that do this are called **"relational databases,"** and they account for most of the databases used in business today, ranging from simple, inexpensive tools like Microsoft Access or Intuit QuickBooks to enormously complex and costly "enterprise-level" applications marketed by Oracle and SAP.[40]

To be precise, only the tables of data are the "database," and the software used to create, maintain and interrogate those tables is called the **Database Management System** or **DBMS**. In practice, the two terms are often used interchangeably.

Relational Databases

Let's re-imagine your case management system as a relational database. You'd still have a table listing all clients organized by name. On this CLIENTS table, each client record includes name, address and case number(s). Even if a client has multiple cases in your office, there is still just a single table listing:

CLIENTS

CLT_LAST	CLT_FIRST	ST_ADD	CITY	STATE	ZIP	CASE_NO
Ballmer	Steven	3832 Hunts Point Rd.	Hunts Point	WA	98004	001, 005
Chambers	John	5608 River Way	Buena Park	CA	90621	002
Dell	Michael	3400 Toro Canyon Rd.	Austin	TX	78746	003, 007
Ellison	Lawrence	745 Mountain Home Rd.	Woodside	CA	94062	004
Gates	William	1835 73rd Ave. NE	Medina	WA	98039	001, 005
Jobs	Steven	460 Mountain Home Rd.	Woodside	CA	94062	006, 009
Palmisano	Samuel	665 Pequot Ave.	Southport	CT	06890	007

It's essential to keep track of cases and upcoming trials, so you create another table called CASES:

CASES

[40] One of the most important and widely used database applications, MySQL, is open source; so, while great fortunes have been built on relational database tools, the database world is by no means the exclusive province of commercial software vendors.

CASE_NO	TRL_DATE	MATTER	TYPE	COURT
001	2011-02-14	U.S. v. Microsoft	Antitrust	FDDC-1
002	2012-01-09	EON v Cisco	Patent	FEDTX-2
003	2011-02-15	In re: Dell	Regulatory	FWDTX-4
004	2011-05-16	SAP v. Oracle	Conspiracy	FNDCA-8
005	2012-01-09	Microsoft v. Yahoo	Breach of K	FWDWA-6
006	2010-12-06	Apple v. Adobe	Antitrust	FNDCA-8
007	2011-10-31	Dell v. Travis County	Tax	TX250
008	null	Hawkins v. McGee	Med Mal	FUSSC
009	2011-12-05	Jobs v. City of Woodside	Tax	CASMD09

You also want to stay current on where your cases will be tried and the presiding judge, so you maintain a COURTS table for all the matters on your docket:

COURTS

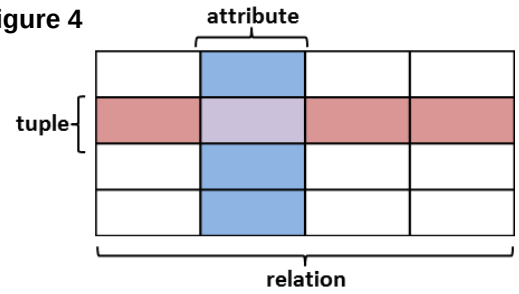
COURT	JUDGE	FED_ST	JURISDICTION
FNDCA-8	Laporte	FED	Northern District of California (SF)
FDDC-1	Kollar-Kotelly	FED	USDC District of Columbia
FWDTX-4	Sparks	FED	Western District of Texas
TX250	Dietz	STATE	250th JDS, Travis County, TX
CASMD09	Parsons	STATE	San Mateo Superior Court, CA
FEDTX-2	Ward	FED	Eastern District of Texas
FWDWA-6	Jones	FED	Western District of Washington
FUSSC	Hand	FED	United States Supreme Court

As we look at these three tables, note that each has a unique key field called the "**primary key**" for that table.[41] For the CLIENTS table, the primary key is the client's last name.[42] The primary key is the trial date for the TRIAL_DATES table and it's a unique court identifier for the COURTS table. The essential characteristic of a primary key is that it cannot repeat within the table for which it serves as primary key, and a properly-designed database will prevent a user from creating duplicate primary keys.

[41] Tables can have more than one primary key.
[42] In practice, a last name would be a poor choice for a primary key in that names tend not to be unique—certainly a law firm could expect to have multiple clients with the same surname.

Many databases simply assign a unique primary key to each table row, either a number or a non-recurring value built from elements like the first four letters of a name, first three numbers in the address, first five letters in the street name and the Zip code. For example, an assigned key for Steve Ballmer derived from data in the CLIENTS table might be BALL383HUNTS98004. The primary key is used for indexing the table to make it more efficient to search, sort, link and perform other operations on the data.

Figure 4



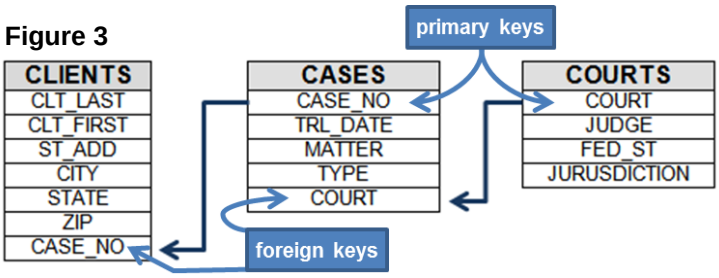
Tuples and Attributes

Now, we need to introduce some new terminology because the world of relational databases has a language all its own. Dealing with the most peculiar term first, the contents of each row in a table is called a **"tuple,"** defined as an ordered list of elements.[43] In the COURTS table above, there are seven tuples, each consisting of four elements. These elements, ordered as columns, are called **"attributes,"** and what we've called tables in the flat file world are termed **"relations"** in relational databases. Put another way, *a relation is defined as a set of tuples that have the same attributes* (See Figure 1).

Figure 5

The magic happens in a relational database when tables are **"joined"** (much like the cube in Figure 2)[44] by referencing one table from another.[45] This is done by incorporating the primary key in the table referenced as a **"foreign key"** in the referencing table. The table referenced is the **"parent table,"** and the referencing table is the **"child table"** in this joining of the two relations. In Figure 3, COURTS is the parent table to CASES with respect to the primary key field, "COURT." In the CASES table, the

Figure 3



[43] Per Wikipedia, the term "tuple" originated as an abstraction of the sequence: single, double, triple, quadruple, quintuple, sextuple, septuple, octuple...n-tuple. The unique 0-tuple is called the null tuple. A 1-tuple is called a "singleton," a 2-tuple is a "pair" and a 3-tuple is a "triple" or "triplet." The *n* can be any positive integer. For example, a complex number can be represented as a 2-tuple, a quaternion can be represented as a 4-tuple, an octonion can be represented as an octuple (mathematicians use the abbreviation "8-tuple"), and a sedenion can be represented as a 16-tuple. I include this explanation to remind readers why many of us went to law school instead of studying computer science.

[44] Although unlike the cube, a relational database is not limited to just three dimensions of attachment.

[45] The term "relation" is so confounding here, I will continue to refer to them as tables.

foreign key for the field COURT points back to the COURTS table, assuring that the most current data will populate the field. In turn, the CLIENTS table employs a foreign key relating to the CASE_NO attribute in the CASE table, again assuring that the definitive information populates the attribute in the CLIENTS table.

Remember that what you are seeking here is to ensure that you do not build a database with inconsistent data, such as conflicting client addresses. Data conflicts are avoided in relational databases by allowing the parent primary key to serves as the definitive data source. So, by pointing each child table to that definitive parent via the use of foreign keys, you promote so-called **"referential integrity"** of the database. Remember, also, that while a primary key must be unique to the parent table, it can be used as many times as desired when referenced as a foreign key. As in life, parents can have multiple children, but a child can have but one set of (biological) parents.

Field Properties and Record Structures

When you were writing case data on your index cards, you were unconstrained in terms of the information you included. You could abbreviate, write dates as words or numeric values and include as little or as much data as the space on the card and intelligibility allowed. But for databases to perform properly, the contents of fields should conform to certain constraints to insure data integrity. For example, you wouldn't want a database to accept four or ten letters in a field reserved for a Zip code. Neither should the database accept duplicate primary keys or open a case without including the name of a client. If a field is designed to store only a U.S. state, then you don't want it to accept "Zambia" or "female." You also don't want it to accept "Noo Yawk."

Accordingly, databases are built to enforce specified field property requirements. Such properties may include:

1. **Field size**: limiting the number of characters that can populate the field or permitting a variable length entry for memos;
2. **Data type**: text, currency, integer numbers, date/time, e-mail address and masks for phone numbers, Social security numbers, Zip codes, etc.;
3. **Unique fields**: Primary keys must be unique. You typically wouldn't want to assign the same case number to different matters or two Social Security numbers to the same person.
4. **Group or member lists**: Often fields may only be populated with data from a limited group of options (e.g., U.S. states, salutations, departments and account numbers);
5. **Validation rules**: To promote data integrity, you may want to limit the range of values ascribed to a field to only those that makes sense. A field for a person's age shouldn't accept negative values or (so far) values in excess of 125. A time field should not accept "25:00pm" and a date field designed for use by Americans should guard against European

date notation. Credit card numbers must conform to specific rules, as must Zip codes and phone numbers; or

6. **Required data**: The absence of certain information may destroy the utility of the record, so certain fields are made mandatory (e.g., a car rental database may require input of a valid driver's license number).

You'll appreciate why demanding production of the raw tables in a database may be an untenable approach to e-discovery when you consider how databases store information. When a database populates a table, it's stored in either **fixed length** or **variable length** fields.

Fixed-Length Field Records

Fixed length fields are established when the database is created, and it's important to appreciate that the data is stored as long sequences of data that may, to the untrained eye, simply flow together in one incomprehensible blob. A fixed length field record may begin with information setting out information concerning all of the fields in the record, such as each field's name (e.g., COURT), followed by its data type (e.g., alphanumeric), length (7 characters) and format (e.g., only values matching a specified list of courts).

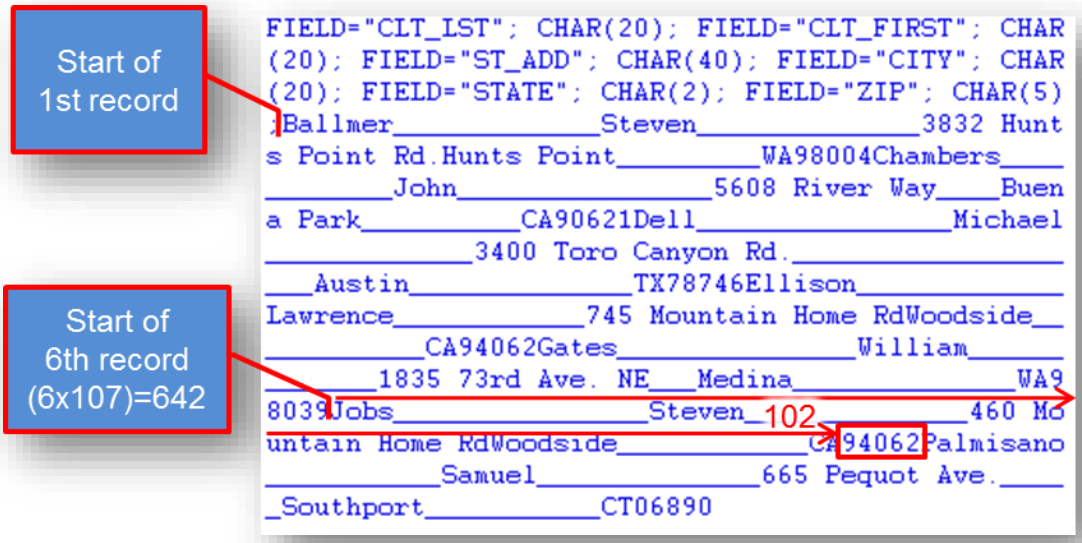
A fixed length field record for a simplified address table might look like Figure 4.

Figure 4

```
FIELD="CLT_LST"; CHAR(20); FIELD="CLT_FIRST"; CHAR
(20); FIELD="ST_ADD"; CHAR(40); FIELD="CITY"; CHAR
(20); FIELD="STATE"; CHAR(2); FIELD="ZIP"; CHAR(5)
;Ballmer_____Steven_____3832 Hunt
s Point Rd.Hunts Point_____WA98004Chambers____
_____John_____5608 River Way____Buen
a Park_____CA90621Dell_____Michael
_____3400 Toro Canyon Rd._____
___Austin_____TX78746Ellison_____
Lawrence_____745 Mountain Home RdWoodside__
_____CA94062Gates_____William_____
_____1835 73rd Ave. NE___Medina_____WA9
8039Jobs_____Steven_____460 Mo
untain Home RdWoodside_____CA94062Palmisano
_____Samuel_____665 Pequot Ave.____
_Southport_____CT06890
```

Note how the data is one continuous stream. The name, order and length of data allocated for each field is defined at the beginning of the string in all those "FIELD=" and CHAR(x) statements, such that the total length of each record is 107 characters. To find a given record in a table, the database software simply starts accessing data for that record at a distance (also called an "offset") from the start of the table equal to the number of records times the total length allocated to each record. So, as shown in **Figure 5**, the fourth record starts 428 characters from the start of the first record. In turn, each field in the record starts a fixed number of characters from the start of the record. If you wanted to extract Steve Jobs' Zip code from the exemplar table, the Jobs address record is the 6th record, so it starts 642 characters (or bytes) from the start of the first record and the Zip code field begins 102 characters from the start of the sixth record (20+20+40+20+2), or 744 bytes from the start of the first record. This sort of offset retrieval is tedious for humans, but it's a cinch for computers.

Figure 5



Variable-Length Field Records

One need only recall the anxiety over the Y2K threat to appreciate why fixed length field records can be problematic. Sometimes, the space allocated to a field proves insufficient in unanticipated ways, or you may simply need to offer the ability to expand the size of a record on-the-fly. Databases employ variable length field records whose size can change from one record to the next. Variable length fields employ **pointer fields** that seamlessly redirect data retrieval to a designated point in the memo file where the variable length field data begins (or continues). The database software then reads from the memo file until it encounters an end-of-file marker or another pointer to a memo location holding further data.

Forms, Reports and Query Language

Now that you've glimpsed the ugly guts of database tables, you can appreciate why databases employ database management software to enter, update and retrieve data. Though DBMS software serves many purposes geared to indexing, optimizing and protecting data, the most familiar role of DBMS software is as a user interface for forms and reports.

There's little difference between forms and reports except that we tend to call the interface used to input and modify data a "form" and the interface to extract data a "report." Both are simply user-friendly ways to implement commands in **"query languages."**

Query language is the term applied to the set of commands used to retrieve information from a database. The best known and most widely used of these is called **SQL** (for **Structured Query Language**, officially 'ess-cue-ell,' but most everyone calls it "sequel"). SQL is a computer language, but different from computer languages like Java or C++ that can be used to construct applications, SQL's sole purpose is the creation, management and interrogation of databases.

Though the moniker "query language" might lead anyone to believe that its raison d'être is to get data out of databases, in fact, SQL handles the heavy lifting of database creation and data insertion, too. SQL includes subset command sets for data control (DCL), data manipulation (DML) and data definition (DDL). SQL syntax is beyond the scope of this paper, but the following snippet of code will give you a sense of how SQL is used to create a table like the case management tables discussed above:

```
CREATE TABLE COURTS
        (COURT varchar(7), PRIMARY KEY,
        JUDGE varchar(18),
        FED_ST varchar(5),
        JURISDICTION varchar (40));
CREATE TABLE CASES
        (CASE_NO int IDENTITY(1,1)PRIMARY KEY,
        TRL_DATE
        MATTER varchar (60),
        TYPE varchar (40)
        COURT varchar(7));
```

In these few lines, the COURTS and CASES tables are created, named and ordered into various alphanumeric fields of varying specified lengths. Two primary keys are set and one key, CASE_NO, is implemented so as to begin with the number 1 and increment by 1 each time a new case is added to the CASES table.

Who Owns SQL?
I do, so if your firm or clients are using SQL, please have them send gobs of cash to me so I won't sue them.

In fact, nobody "owns" SQL, but several giant software companies, notably Oracle and Microsoft, have built significant products around SQL and produced their own proprietary dialects of SQL. When you hear someone mention "SQL Server," they're talking about a Microsoft product, but Microsoft doesn't own SQL; it markets a database application that's compatible with SQL.

SQL has much to commend it, being both simple and powerful; but, even the simplest computer language is too much for the average user. So, databases employ graphical user interfaces (GUIs) to put a friendly face on SQL. When you enter data into a form or run a search, you're simply triggering a series of pre-programmed SQL commands.

In e-discovery, if the standard reports supported by the database are sufficiently encompassing and precise to retrieve the information sought, great! You'll have to arrive at a suitable form of production and perhaps wrangle over scope and privilege issues; but, the path to the data is clear.

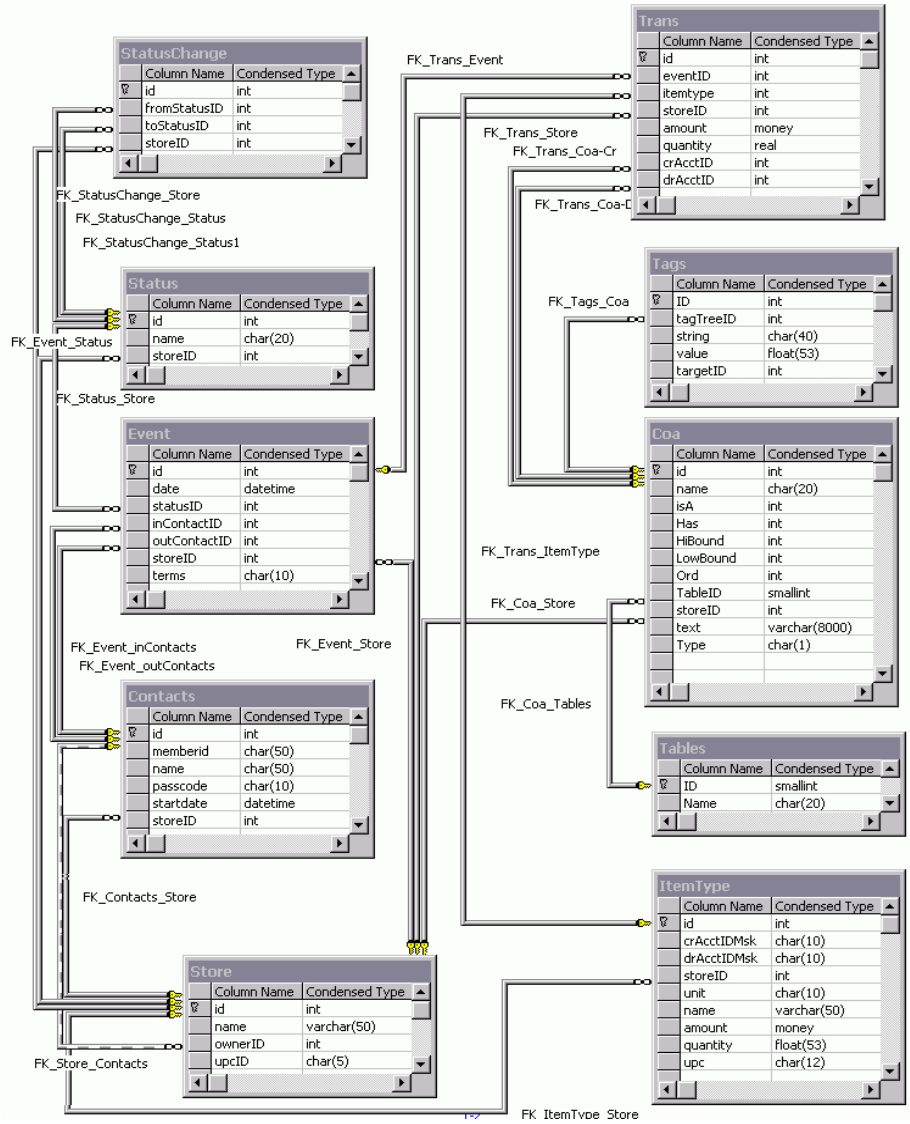
However, because most companies design their databases for operations not litigation, very often, the standard reporting capabilities won't be retrieve the types of information required in discovery. In that event, you'll need more than an SQL doctor on your team; you'll also need a good x-ray of the databases to be plumbed.

Schemas, Data Dictionaries, System Catalogs, and ERDs,
The famed database administrator, Leo Tolstoy, remarked, "Great databases are all alike, every ordinary database is ordinary in its own way." Although it's with tongue-in-cheek that I invoke Tolstoy's famous observation on happy and unhappy families, it's apt here and means that you can only assume so much about the structure of an unfamiliar database. After that, you need the manual and a map.
.
In the lingo of database land, the "map" is the database's **schema,** and it's housed in the system's **data dictionary**. It may be the system's **logical schema**, detailing how the database is designed in terms of its table structures, attributes, fields, relationships, joins and views. Or, it could be its **physical schema**, setting out the hardware and software implementation of the database on machines, storage devices and networks. As Tolstoy might have said, "A logical schema explains death; but, it won't tell you where the bodies are buried."

Information in a database is mostly gibberish without the metadata that gives it form and function. In an SQL database, the compendium of all that metadata is called the **system catalog.** In practice, the terms system catalog, schema and data dictionary seem to be used interchangeably—they are all—in essence--databases storing information about the metadata of a database. The most important lesson to derive from this discussion is that there is a map—or one can be easily generated—so get it!

Unlike that elusive Loch Ness monster of e-discovery, the "enterprise data map," the schemas of databases tend to actually exist and are usually maps; that is, graphical depictions of the database structures. **Entity-Relationship Modeling** (**ERM**) is a system and notation used to lay out the conceptual and logical schema of a relational database. The resulting diagrams (akin to flow charts) are called **Entity-Relationship Diagrams** or **ERDs (Figure 6)**.

Figure 6: ERD of Database Schema



Two Lessons from the Database Trenches

The importance of securing the schema, manuals, data dictionary and ERDs was borne out by my experience serving as Special Master for Electronically Stored Information. in a drug product liability action involving thousands of plaintiffs, I was tasked to expedite discovery from as many as 60 different enterprise databases, each more sprawling and complex than the next. The parties were at loggerheads, and serious sanctions were in the offing.

The plaintiffs insisted the databases would yield important evidence. Importantly, plaintiffs' team included support personnel technically astute enough to get deeply into the weeds with the systems. Plaintiffs were willing to narrow the scope of their database discovery to eliminate those that were unlikely to be responsive and to narrow the scope of their requests. But, to do that, they'd need to know the systems.

For each system, we faced the same questions:

i. What does the database do?
ii. What is it built on?
iii. What information does it hold?
iv. What content is relevant, responsive and privileged?
v. What forms does it take?
vi. How can it be searched effectively; using what query language?
vii. What are its reporting capabilities?
viii. What form or forms of production will be functional, searchable and cost-effective?

It took a three-step process to turn things around. First, the plaintiffs were required to do their homework, and the defense supplied the curriculum. That is, the defense was required to furnish documentation concerning the databases. First, each system had to be identified. The defense prepared a spreadsheet detailing, *inter alia*:

- Names of systems
- Applications;
- Date range of data;
- Size of database;
- User groups; and
- Available system documentation (including ERDs and data dictionaries).

This enabled plaintiffs to prioritize their demands to the most relevant systems. I directed the defendants to furnish operator's manuals, schema information and data dictionaries for the most relevant systems.

The second step was ordering that narrowly-focused meet-and-confer sessions be held between technical personnel for both sides. These were conducted by telephone, and the sole topic of each was one or more of the databases. The defense was required to make knowledgeable personnel available for the calls and plaintiffs were required to confine their questions to the nuts-and-bolts of the databases at issue.

When the telephone sessions concluded, Plaintiffs were directed to serve their revised request for production from the database. In most instances, the plaintiffs had learned enough about the databases that they were actually able to propose SQL queries to be run.

This would have been sufficient in most cases, but this case was especially contentious. The final step needed to resolve the database discovery logjam was a meeting in the nature of a mediation over which I would preside. In this proceeding, counsel and technical liaison, joined by the database specialists, would meet face-to-face over two days. We would work through each database and arrive at specific agreements concerning the scope of discovery for each system, searches run, sample sizes employed and timing and form of production. The devil is in the details, and the goal was to nail down every detail.

It took two such sessions, but in the end, disputes over databases largely ceased, the production changed hands smoothly, and the parties could refocus on the merits.

The heroes in this story are the technical personnel who collaborated to share information and find solutions when the lawyers could see only contentions. The lesson: **Get the geeks together, and then get out of their way.**

Lesson Two

In a recent case where I served as special master, the Court questioned the adequacy of defendants' search of their databases. The defendants used many databases to run their far-flung operations, ranging from legacy mainframe systems housed in national data centers to homebrew applications cobbled together using Access or Excel. But whether big or small, I found with disturbing regularity that the persons tasked to query the systems for responsive data didn't know how to use them or lacked the rights needed to access the data they were obliged to search.

The lesson: **Never assume that a DBMS query searches all of the potentially responsive records, and never assume that the operator knows what they are doing.**

Database systems employ a host of techniques to optimize performance and protect confidentiality. For example

- Older records may be routinely purged from the indices;
- Users may lack the privileges within the system to access all the potentially responsive records;
- Queries may be restricted to regions or business units;
- Tables may not be joined in the particular ways needed to gather the data sought.

Any of these may result in responsive data being missed, even by an apparently competent operator.

Establishing operator competence can be challenging, too. Ask a person tasked with running queries if they have the requisite DBMS privileges required for a comprehensive search, and they're likely to give you a dirty look and insist they do. In truth, they probably don't know. What they have are the privileges they need to do their job day-to-day; but those may not be nearly sufficient to elicit all of the responsive information the system can yield.

How do you preserve a database in e-discovery?

Talk to even tech-savvy lawyers about preserving databases, and you'll likely hear how database are gigantic and dynamic or how incomprehensibly risky and disruptive it is to mess with them. The lawyer who responds, "Don't be ridiculous. We're not preserving our databases for your lawsuit," isn't protecting her client.

Or, opposing counsel may say, "Preserve our databases? Sure, no problem. We back up the databases all the time. We'll just set aside some tapes." This agreeable fellow isn't protecting his client either. When it comes time to search the data on tape, Mr. Congeniality may learn that his client has no ability to restore the data without displacing the server currently in use, and restoration doesn't come quick or cheap.

What both of these lawyers should have said is, "Let me explain what we have and how it works. Better yet, let's get our technical advisors together. Then, we'll try to work out a way to preserve what you really need in a way you can use it. If we can't agree, I'll tell you what my client will and won't do, and you can go to the judge right away, if you think we haven't done enough."

Granted, this conversation almost never occurs for a host of reasons. Counsel may have no idea what the client has or how it works. Or the duty to preserve attaches before an opposing counsel emerges. Or counsel believes that cooperation is anathema to zealous advocacy and wants only to scorch the Earth.

In fact, it's not that daunting to subject most databases to a defensible litigation hold, if you understand how the database works and exert the time and effort required to determine what you're likely to need preserved.

Databases are dynamic by design, but not all databases change in ways that adversely impact legal hold obligations. Many databases—particularly accounting databases—are accretive in design. That is, they add new data as time goes on, but do not surrender the ability to thoroughly search data that existed in prior periods. For accretive databases, all counsel may need to do is ascertain and insure that historical data isn't going anywhere for the life of the case.

Creating snapshots of data stores or pulling a full backup set for a relevant period is a sensible backstop to other preservation efforts, as an "if all else fails" insurance policy against spoliation. If the likelihood of a lawsuit materializing is remote or if there is little chance that the tapes preserved will ultimately be subjected to restoration, preservation by only pulling tapes may prove sufficient and economical. But, if a lawsuit is certain and discovery from the database(s) is likely, the better approach is to identify ways to either duplicate and/or segregate the particular dynamic data you'll need or export it to forms that won't unduly impair searchability and utility. That is, you want to keep the essential data reasonably accessible and shield it from changes that will impair its relevance and probative value.

If the issue in litigation is temporally sensitive—e.g., wholesale drug pricing in 2010 or reduction in force decisions in 2008—you'll need to preserve the responsive data before the myriad components from which it's drawn, and the filters, queries and algorithms that govern how it's communicated, change. You'll want to retain the ability to generate the reports that should be reasonably anticipated and not lose that ability because of an alteration in some dynamic element of the reporting process.

Forms of Production

In no other corner of e-discovery are litigants quite so much as the dog that caught the car than when dealing with databases. Data from specialized and enterprise databases often don't play well with off-the-shelf applications; not surprising, considering the horsepower and high cost of the systems tasked to run these big iron applications. Still, there is always a way.

Sometimes a requesting party demands a copy of an entire database, often with insufficient consideration of what such a demand might entail were it to succeed. If the database is built in Access or on other simple platforms, it's feasible to acquire the hardware and software licenses required to duplicate the producing party's database environment sufficiently to run the application. But, if the data sets are so large as to require massive storage resources or are built on an enterprise-level DBMS like Oracle or SAP, mirroring the environment is almost out of the question. I say "almost" because the emergence of Infrastructure-as-a-Service Cloud computing

options promises to make it possible for mere mortals to acquire enterprise-level computing power for short stints

A more likely production scenario is to narrow the data set by use of filters and queries, then either export the responsive date to a format that can be analyzed in other applications (e.g., exported as extensible markup language (XML), comma separated values (CSV) or in another delimited file) or run reports (standard or custom) and ensure that the reporting takes a form that, unlike paper printouts, lends itself to electronic search.

Before negotiating a form of production, investigate the capabilities of the DBMS. The database administrator may not have had occasion to undertake a data export and so may have no clue what an application can do much beyond the confines of what it does every day. It's the rare DBMS that can't export delimited data. Next, have a proposed form of production in mind and, if possible, be prepared to instruct the DBMS administrator how to secure the reporting or export format you seek,

Remember that the resistance you experience in seeking to export to electronic formats may not come from the opposing party of the DBMS administrator. More often, an insistence on reports being produced as printouts or page images is driven by the needs of opposing counsel. In that instance, it helps to establish that the export is feasible as early as possible.

As with other forms of e-discovery, be careful not to accept production in formats you don't want because, like-it-or-not, many Court give just one bite at the production apple. If you accept it on a paper or as TIFF images for the sake of expediency, you often close the door on re-production in more useful forms.

Even if the parties can agree upon an electronic form of production, it's nevertheless a good idea to secure a test export to evaluate before undertaking a high volume export.

Closing Thoughts

When dealing with databases in e-discovery, requesting parties should avoid the trap of "You have it. I want it." Lawyers who'd never be so foolish as to demand the contents of a file room will blithely insist on production of the "database." For most, were they to succeed in such a foolish quest, they'd likely find themselves in possession of an obscure collection of inscrutable information they can't possibly use.

Things aren't much better on the producing party's side, where counsel routinely fail to explore databases in e-discovery on the theory that, if a report hasn't been printed out, it doesn't have to be created for the litigation. Even when they do acknowledge the duty to search databases, few

counsel appreciate how pervasively embedded databases are in their clients' businesses, and fewer still possess the skills needed to translate an amorphous request for production into precise, effective queries.

Each is trading on ignorance, and both do their clients a disservice.

But, these are the problems of the past and, increasingly, there's cause for cautious optimism in how lawyers and litigants approach databases in discovery. Counsel are starting to inquire into the existence and role of databases earlier in the litigation timeline and are coming to appreciate not only how pervasive databases are in modern commerce, but how inescapable it is that they take their place as important sources of discoverable ESI.

More on Databases in Discovery

I loathe the practice of law from forms, but bow to its power. Lawyers love forms; so, to get lawyers to use more efficient and precise prose in their discovery requests, we can't just harangue them to do it; we've "got to put the hay down where the goats can get it." To that end, here is some language to consider when seeking information about databases and when serving notice of the deposition of corporate designees (*e.g.,* per Rule 30(b)(6) in Federal civil practice or Rule 199(b)(1) of the Texas Rules of Civil Procedure):

For each database or system that holds potentially responsive information, we seek the following information to prepare to question the designated person(s) who, with reasonable particularity, can testify on your behalf about information known to or reasonably available to you concerning:

1. **The standard reporting capabilities of the database or system, including the nature, purpose, structure, appearance, format and electronic searchability of the information conveyed within each standard report (or template) that can be generated by the database or system or by any overlay reporting application;**
2. **The enhanced reporting capabilities of the database or system, including the nature, purpose structure, appearance, format and electronic searchability of the information conveyed within each enhanced or custom report (or template) that can be generated by the database or system or by any overlay reporting application;**
3. **The flat file and structured export capabilities of each database or system, particularly the ability to export to fielded/delimited or structured formats in a manner that faithfully reflects the content, integrity and functionality of the source data;**
4. **Other export and reporting capabilities of each database or system (including any overlay reporting application) and how they may or may not be employed to faithfully reflect the content, integrity and functionality of the source data for use in this litigation;**
5. **The structure of the database or system to the extent necessary to identify data within potentially responsive fields, records and entities, including field and table names, definitions, constraints and relationships, as well as field codes and field code/value translation or lookup tables.**
6. **The query language, syntax, capabilities and constraints of the database or system (including any overlay reporting application) as they may bear on the ability to identify, extract and export potentially responsive data from each database or system;**
7. **The user experience and interface, including datasets, functionality and options available for use by persons involved with the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;**

8. **The operational history of the database or system to the extent that it may bear on the content, integrity, accuracy, currency or completeness of potentially responsive data;**

9. **The nature, location and content of any training, user or administrator manuals or guides that address the manner in which the database or system has been administered, queried or its contents reviewed by persons involved with the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;**

10. **The nature, location and contents of any schema, schema documentation (such as an entity relationship diagram or data dictionary) or the like for any database or system that may reasonably be expected to contain information relating to the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;**

11. **The capacity and use of any database or system to log reports or exports generated by, or queries run against, the database or system where such reports, exports or queries may bear on the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT;**

12. **The identity and roles of current or former employees or contractors serving as database or system administrators for databases or systems that may reasonably be expected to contain (or have contained) information relating to the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT; and**

13. **The cost, burden, complexity, facility and ease with which the information within databases and systems holding potentially responsive data relating to the PROVIDE APPROPRIATE LANGUAGE RE THE ACTIVITIES PERTINENT TO THE MATTERS MADE THE BASIS OF THE SUIT; may be identified, preserved, searched, extracted and produced in a manner that faithfully reflects the content, integrity and functionality of the source data.**

Yes, this is the dread "discovery about discovery;" but, it's a necessary precursor to devising query and production strategies for databases. If you don't know what the database holds or the ways in which relevant and responsive data can be extracted, you are at the mercy of opponents who will give you data in unusable forms or give you nothing at all.

Remember, these are not magic words. I just made them up, and there's plenty of room for improvement. If you borrow this language, please take time to understand it, and particularly strive to know *why* you are asking for what you demand. Supplying the information requires effort that should be expended in support of a genuine and articulable need for the information. If you don't need the information or know what you plan to do with it, don't ask for it.

These few questions were geared to the feasibility of extracting data from databases so that it stays utile and complete. Enterprise databases support a raft of standardized reporting capabilities: "screens" or "reports" run to support routine business processes and decision making. An insurance carrier may call a particular report the "Claims File;" but, it is not a discrete "file" at all. It's a predefined template or report that presents a collection of data extracted from the database in a consistent way. Lots of what we think of as sites or documents are really reports from databases. Your Facebook page? It's a report. Your e-mail from Microsoft Outlook? Also a report.

In addition to supplying a range of standard reports, enterprise databases can be queried using enhanced reporting capabilities ("custom reports") and using overlay reporting tools–commercial software "sold separately" and able to interrogate the database in order to produce specialized reporting or support data analytics. A simple example is presentation software that generates handsome charts and graphics based on data in the database. The presentation software didn't come with the database. It's something they bought (or built) to "bolt on" for enhanced/overlay reporting.

Although databases are queried using a "query language," users needn't dirty their hands with query languages because queries are often executed "under the hood" by the use of those aforementioned standardized screens, reports and templates. Think of these as pre-programmed, pushbutton queries. There is usually more (and often much more) that can be gleaned from a database than what the standardized reports supply, and some of this goes to the integrity of the data itself. In that case, understanding the query language is key to fashioning a query that extracts what you need to know, both *within* the data and *about* the data.

As importantly as learning what the database can produce is understanding what the database does or does not display to end users. These are the *user experience* (UX) and *user interface* (UI). Screen shots may be worth a thousand words when it comes to understanding what the user saw or what the user might have done to pursue further intelligence.

Enterprise and commercial databases tend to be big and expensive. Accordingly, most are well documented in manuals designed for administrators and end users. When a producing party objects that running a query is burdensome, the manuals may make clear that what you seek is no big deal to obtain.

One feature that sets databases apart from many others forms of ESI is the critical importance of the fielding of data. **Preserving the fielded character of data is essential to preserving its utility and searchability** "Fielding data" means that information is stored in locations dedicated to holding just that information. Fielding data serves to separate and identify information so you can search, sort and cull using just that information. It's a capability we take for granted in databases but that is often crippled or eradicated when data is produced in e-discovery. **Be sure that you**

consider the form of production, and insure that the fielded character of the data produced will not be lost, whether supplied as a standard report or as a delimited export.

Fielding data isn't new. We did it back when data was stored as paper documents. Take a typical law firm letter: the letterhead identifies the firm, the date below the letterhead is understood to be the date sent. A *Re:* line follows, denoting matter or subject, then the addressee, salutation, etc. The recipient is understood to be named at the start of the letter and the sender at the bottom. These conventions governing where to place information are vital to our ability to understand and organize conventional correspondence.

Similarly, all the common productivity file types encountered in e-discovery (Microsoft Office formats, PDF and e-mail) employ fielding to abet utility and functionality. Native "documents" are natively fielded; that is, a file's content is structured to ensure that pieces of information reside in defined locations within the file. This structure is understood and exploited by the native application and by tools designed to avail themselves of the file architecture.

We act inconsistently, inefficiently and irrationally when we deal with fielded information in e-discovery. In contrast to just a few years ago, only the most Neanderthal counsel now challenges the need to produce the native fielding of spreadsheet data. Accordingly, production of spreadsheets in native forms has evolved to become routine and (largely) uncontentious. To get to this point, workflows were modified, Bates numbering procedures were tweaked, and despite dire predictions, none of it made the sky fall. We can and must do the same with PowerPoint presentations and Word documents.

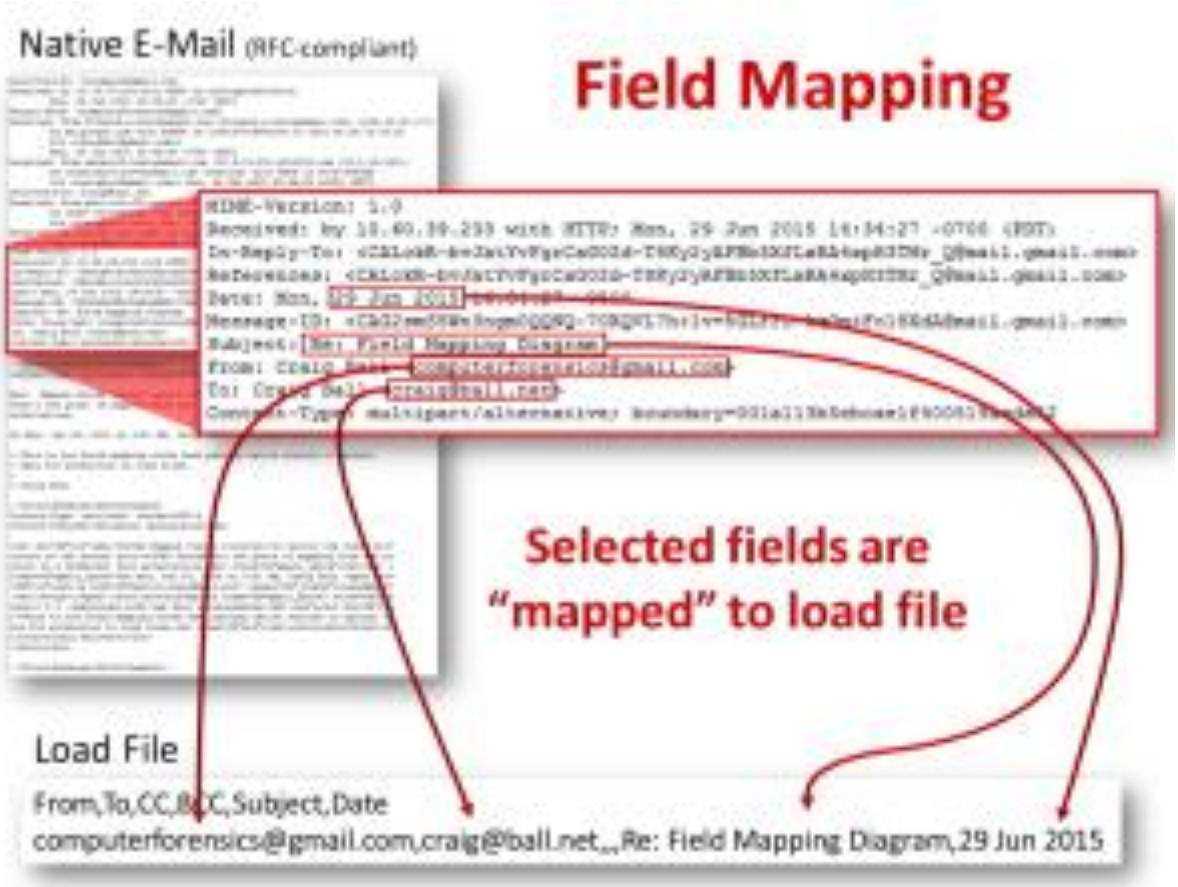
"*What's vice today may be virtue tomorrow*," wrote novelist (and jurist) Henry Fielding.

Now, take e-mail. <u>All</u> e-mail is natively fielded data, and the architecture of e-mail messages is established by published standards called RFCs—structural conventions that e-mail applications and systems must embrace to ensure that messages can traverse any server. The RFCs define placement and labeling of the sender, recipients, subject, date, attachments, routing, message body and other components of every e-mail that transits the Internet.

But when we produce e-mail in discovery, the "accepted" practice is to deconstruct each message and produce it in a cruder fielded format that's incompatible with the RFCs and unrecognizable to any e-mail tool or system. Too, the production is almost always incomplete compared to the native content.

The deconstruction of fielded data is accomplished by a process called **Field Mapping**. The contents of particular fields within the native source are extracted and inserted into a matrix that may assign the same name to the field as accorded by the native application or rename it to something else altogether. Thus, the source data is "mapped' to a new name and location. At all events, the mapped fields never mirror the field structure of the source file.

Ever? No, *never.*



The jumbled fielding doesn't entirely destroy the ability to search within fields or cull and sort by fielded content; but, it requires lawyers to rent or buy tools that can re-assemble and read the restructured data in order to search, sort and review the content. And again, information in the original is often omitted, not because it's privileged or sensitive, but because…well, um, er, *we just do it that way, dammit*!

But the information that's omitted, surely that's useless metadata, right?

Interestingly, no. In fact, the omitted information significantly aids our ability to make sense of the production, such as the fielded data that allows messages to be organized into conversational threads (*e.g.*, In-Reply-To, References and Message-ID fields) and the fielded data that enables messages to be correctly ordered across time zones and daylight savings time (*e.g.,* UTC offsets).

"Why do producing parties get to recast and omit this useful information," you ask? The industry responds: "*These are not the droids you're looking for." "Hey, is that Elvis?" "No Sedona for you!"*

The real answer is that counsel, and especially requesting counsel, are asleep at the wheel. Producing parties have been getting away with this nonsense, unchallenged, for so long,

they've come to view it as a birthright. But, reform is coming, at the glacial pace for which we lawyers are justly reviled, I mean *revered*.

E-discovery standards have indeed evolved to acknowledge that e-mail must be supplied with some fielding preserved; but, there is no sound reason to produce e-mail with shuffled or omitted fields. It doesn't cost more to be faithful to the native or near-native architecture or be complete in supplying fielded content; in fact, producing parties pay *more* to degrade the production, and what emerges costs more to review.

Perhaps the hardest thing for lawyers and judges to appreciate is the importance fielding plays in culling, sorting and search.

- It's efficient to be able to cull and sort files *only* by certain dates.
- It's efficient to be able to search *only* within e-mail recipients.
- It's efficient to be able to distinguish Speaker Notes within a PowerPoint or filter by the Author field in a Word document.

Preserving the fielded character of data makes that possible. Preserving the fielded data *and* the native file architecture allows use of a broad array of tools against the data, where restructuring fielded data limits its use to only a handful of pricey tools that understand peculiar and proprietary production formats.

It's not enough for producing parties to respond, "*But, you can reassemble the kit of data we produce to make it work somewhat like the original evidence*." In truth, you often can't, and you shouldn't have to try.

It ties back to the Typewriter Generation mentality that keeps us thinking about "documents" and seeking to define everything we seek as a "document." Most information sought in discovery today is not a purposeful precursor to something that will be printed. Most modern evidence is data, *fielded* data. Modern productivity files aren't blobs of text, they're ingenious little *databases*. *Powerful. Rich. Databases.* Their native content and architecture are key to their utility and efficient searchability in discovery. Get the fielding right, and functionality follows.

Seeking discovery from databases is a key capability in modern litigation, and it's not easy for the technically challenged (although it's probably a whole lot easier than your opponent claims). Getting the proper data in usable forms demands careful thought, tenacity and more-than-a-little homework. Still, anyone can do it, alone with a modicum of effort, or aided by a little expert assistance.

Search is a Science
The Streetlight Effect in e-Discovery

In the wee hours, a beat cop sees a drunken lawyer crawling around under a streetlight searching for something. The cop asks, "What's this now?" The lawyer looks up and says, "I've lost my keys." They both search for a while, until the cop asks, "Are you sure you lost them here?" "No, I lost them in the park," the tipsy lawyer explains, "but the light's better over here."

I told that groaner in court, trying to explain why opposing counsel's insistence that we blindly supply keywords to be run against the e-mail archive of a Fortune 50 insurance company wasn't a reasonable or cost-effective approach e-discovery. The "Streetlight Effect," described by David H. Freedman in his 2010 book *Wrong,* is a species of observational bias where people tend to look for things in the easiest ways. It neatly describes how lawyers approach electronic discovery. We look for responsive ESI only where and how it's easiest, with little consideration of whether our approaches are calculated to find it.

Easy is wonderful when it works; but looking where it's easy *when failure is assured* is something no sober-minded counsel should accept and no sensible judge should allow.

Consider *The Myth of the Enterprise Search.* Counsel within and without companies and lawyers on both sides of the docket believe that companies can run keyword searches against their myriad siloes of data: mail systems, archives, local drives, network shares, portable devices, removable media and databases. They imagine that finding responsive ESI hinges on the ability to incant magic keywords like Harry Potter. *Documentum Relevantus!*

Though data repositories may share common networks, they rarely share common search capabilities or syntax. Repositories that offer keyword search may not support Boolean constructs (queries using "AND," "OR" and "NOT"), proximity searches (Word1 near Word2), stemming (finding "adjuster," "adjusting," "adjusted" and "adjustable") or fielded searches (restricted to just addressees, subjects, dates or message bodies). Searching databases entails specialized query languages or user privileges. Moreover, different tools extract text and index such extractions in quite different ways, with the upshot being that a document found on one system will not be found on another using the same query.

But the Streetlight Effect is nowhere more insidious than when litigants use keyword searches against archives, e-mail collections and other sources of indexed ESI.

That Fortune 50 company—call it All City Indemnity—collected a gargantuan volume of e-mail messages and attachments in a process called "message journaling." Journaling copies every message traversing the system into an archive where the messages are indexed for search. Keyword searches only look at the index, not the messages or attachments; so, if you don't find it in the index, you won't find it at all.

All City gets sued every day. When a request for production arrives, they run keyword searches against their massive mail archive using a tool we'll call *Truthiness.* Hundreds of big companies use *Truthiness* or software just like it, and blithely expect their systems will find all documents containing the keywords.

They're wrong…or in denial.

If requesting parties don't force opponents like All City to face facts, All City and its ilk will keep pretending their tools work better than they do, and requesting parties will keep getting incomplete productions. To force the epiphany, consider the following interrogatory.

Interrogatory: For each electronic system or index that will be searched to respond to discovery, please state:

1. The rules employed by the system to tokenize data so as to make it searchable;
2. The stop words used when documents, communications or ESI were added to the system or index;
3. The number and nature of documents or communications in the system or index which are not searchable because of the system or index being unable to extract their full text or metadata; and
4. Any limitation in the system or index, or in the search syntax to be employed, tending to limit or impair the effectiveness of keyword, Boolean or proximity search in identifying documents or communications that a reasonable person would understand to be responsive to the search.

A court will permit "discovery about discovery" like this when a party demonstrates why an inadequate index is a genuine problem. So, let's explore the rationale behind each inquiry:

Tokenization Rules - When machines search collections of documents for keywords, they rarely search the documents for matches; instead, they consult an index of words extracted from the documents. Machines cannot read, so the characters in the documents are identified as

"words" because their appearance meets certain rules in a process called "tokenization." Tokenization rules aren't uniform across systems or software. Many indices simply don't index short words (*e.g.,* acronyms). None index single letters or numbers.

Tokenization rules also govern such things as the handling of punctuated terms (as in a compound word like "wind-driven"), case (will a search for "roof" also find "Roof?"), diacriticals (will a search for Rene also find René?) and numbers (will a search for "Clause 4.3" work?). Most people simply *assume* these searches will work. Yet, in many search tools and archives, they don't work as expected, or don't work at all, unless steps are taken to ensure that they will work.

Stop Words – Some common "stop words" or "noise words" are simply excluded from an index when it's compiled. Searches for stop words fail because the words never appear in the index. Stop words aren't always trivial omissions. For example, "all" and "city" were stop words; so, a search for "All City" will fail to turn up documents containing the company's own name! Words like side, down, part, problem, necessary, general, goods, needing, opening, possible, well, years and state are examples of common stop words. Computer systems typically employ dozens or hundreds of stop words when they compile indices.

Because users aren't warned that searches containing stop words fail, they mistakenly assume that there are no responsive documents when there may be thousands. A search for "All City" would miss millions of documents at All City Indemnity (though it's folly to search a company's files for the company's name).

Non-searchable Documents - A great many documents are not amenable to text search without special handling. Common examples of non-searchable documents are faxes and scans, as well as TIFF images and some Adobe PDF documents. While no system will be flawless in this regard, it's important to determine *how much* of a collection isn't text searchable, *what's* not searchable and whether the portions of the collection that aren't searchable are of *importance* to the case. If All City's adjusters attached scanned receipts and bids to e-mail messages, the attachments aren't keyword searchable absent optical character recognition (OCR).

Other documents may be inherently text searchable but not made a part of the index because they're password protected (*i.e.,* encrypted) or otherwise encoded or compressed in ways that frustrate indexing of their contents. Important documents are often password protected.

Other Limitations - If a party or counsel knows that the systems or searches used in e-discovery will fail to perform as expected, they should be obliged to affirmatively disclose such shortcomings. If a party or counsel is uncertain whether systems or searches work as expected, they should be obliged to find out by*, e.g.,* running tests to be reasonably certain.

No system is perfect, and perfect isn't the e-discovery standard. Often, we must adapt to the limitations of systems or software. But you have to know what a system can't do before you can find ways to work around its limitations or set expectations consistent with actual capabilities, not magical thinking and unfounded expectations.

Surefire Steps to Splendid Search

Hear that rumble? It's the bench's mounting frustration with the senseless, slipshod way lawyers approach keyword search.

It started with Federal Magistrate Judge John Facciola's observation that keyword search entails a complicated interplay of sciences beyond a lawyer's ken. He said lawyers selecting search terms without expert guidance were truly going "where angels fear to tread."

Federal Magistrate (now District) Judge Paul Grimm called for "careful planning by persons qualified to design effective search methodology" and testing search methods for quality assurance. He added that, "the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented."

Most recently, Federal Magistrate Judge Andrew Peck issued a "wake up call to the Bar," excoriating counsel for proposing *thousands* of artless search terms.

Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI. Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of 'false positives.' It is time that the Bar—even those lawyers who did not come of age in the computer era—understand this.

No Help

Despite the insights of Facciola, Grimm and Peck, lawyers still don't know what to <u>do</u> when it comes to effective, defensible keyword search. Attorneys aren't *trained* to craft keyword searches of ESI or implement quality control testing for same. And their experience using Westlaw, Lexis or Google serves only to inspire false confidence in search prowess.

Even saying "hire an expert" is scant guidance. Who's an expert in ESI search for your case? A linguistics professor or litigation support vendor? Perhaps the misbegotten offspring of William Safire and Sergey Brin?

The most admired figure in e-discovery search today—*the Sultan of Search*—is Jason R. Baron at the National Archives and Records Administration, and Jason would be the first to admit he has no training in search. The persons most qualified to design effective search in e-discovery earned

their stripes by spending thousands of hours running searches in real cases--making mistakes, starting over and tweaking the results to balance efficiency and accuracy.

The Step-by-Step of Smart Search

So, until the courts connect the dots or better guidance emerges, here's my step-by-step guide to craftsman like keyword search. I promise these ten steps will help you fashion more effective, efficient and defensible queries.

1. Start with the Request for Production
2. Seek Input from Key Players
3. Look at what You've Got and the Tools you'll Use
4. Communicate and Collaborate
5. Incorporate Misspellings, Variants and Synonyms
6. Filter and Deduplicate First
7. Test, Test, Test!
8. Review the hits
9. Tweak the Queries and Retest
10. Check the Discards

1. Start with the Request for Production

Your pursuit of ESI should begin at the first anticipation of litigation in support of the obligation to identify and preserve potentially relevant data. Starting on receipt of a request for production (RFP) is starting late. Still, it's against the backdrop of the RFP that your production efforts will be judged, so the RFP warrants careful analysis to transform its often expansive and bewildering demands to a coherent search protocol.

The structure and wording of most RFPs are relics from a bygone time when information was stored on paper. You'll first need to hack through the haze, getting beyond the "any and all" and "touching or concerning" legalese. Try to rephrase the demands in everyday English to get closer to the terms most likely to appear in the ESI. Add terms of art from the RFP to your list of keyword candidates. Have several persons do the same, insuring you include multiple interpretations of the requests and obtain keywords from varying points of view.

If a request isn't clear or is hopelessly overbroad, push back promptly. Request a clarification, move for protection or specially except if your Rules permit same. Don't assume you can trot out some boilerplate objections and ignore the request. If you can't make sense of it, or implement it in a reasonable way, tell the other side how you'll interpret the demand and approach the search

for responsive material. Wherever possible, you want to be able to say, "We told you what we were doing, and you didn't object."

2. Seek Input from Key Players

Judge Peck was particularly exercised by the parties' failure to elicit search assistance from the custodians of the data being searched. Custodians are THE subject matter experts on their own data. Proceeding without their input is foolish. Ask key players, "If you were looking for responsive information, how would you go about searching for it? What terms or names would likely appear in the messages we seek? What kinds of attachments? What distribution lists would have been used? What intervals and events are most significant or triggered discussion?" Invite custodians to show you examples of responsive items, and carefully observe how they go about conducting their search and what they offer. You may see them take steps they neglect to describe or discover a strain of responsive ESI you didn't know existed.

Emerging empirical evidence underscores the value of key player input. At the latest TREC Legal Track challenge, higher precision and recall seemed to closely correlate with the amount of time devoted to questioning persons who understood the documents and why they were relevant. The need to do so seems obvious, but lawyers routinely dive into search before dipping a toe into the pool of subject matter experts.

3. Look at what You've Got and the Tools You'll Use

Analyze the pertinent documentary and e-mail evidence you have. Unique phrases will turn up threads. Look for words and short phrases that tend to distinguish the communication as being about the topic at issue. What content, context, sender or recipients would prompt you to file the message or attachment in a responsive folder had it occurred in a paper document?

Knowing what you've got also means understanding the forms of ESI you must search. Textual content stored in TIFF images or facsimiles demands a different search technique than that used for e-mail container files or word processed documents.

You can't implement a sound search if you don't know the capabilities and limitations of your search tool. Don't rely on what a vendor tells you their tool can do, test it against actual data and evidence. Does it find the responsive data you already know to be there? If not, why not?

Any search tool must be able to handle the most common productivity formats, e.g., .doc, docx, .ppt, .pptx, .xls. .xlsx, and .pdf, thoroughly process the contents of common container files, e.g., .pst, .ost, .zip, and recurse through nested content and e-mail attachments.

As importantly, search tools need to clearly identify any "exceptional" files unable to be searched, such as non-standard file types or encrypted ESI. If you've done a good job collecting and preserving ESI, you should have a sense of the file types comprising the ESI under scrutiny. Be sure that you or your service providers analyze the complement of file types and flags any that can't be searched. Unless you make it clear that certain files types won't be searched, the natural assumption will be that you thoroughly searched all types of ESI.

4. Communicate and Collaborate

Engaging in genuine, good faith collaboration is the most important step you can take to insure successful, defensible search. Cooperation with the other side is not a sign of weakness, and courts expect to see it in e-discovery. Treat cooperation as an opportunity to show competence and readiness, as well as to assess your opponent's mettle. What do you gain from wasting time and money on searches the other side didn't seek and can easily discredit? Won't you benefit from knowing if they have a clear sense of what they seek and how to find it?

Tell the other side the tools and terms you're considering and seek their input. They may balk or throw out hundreds of absurd suggestions, but there's a good chance they'll highlight something you overlooked, and that's one less do over or ground for sanctions. Don't position cooperation as a trap nor blindly commit to run all search terms proposed. "We'll run your terms if you agree to accept our protocol as sufficient" isn't fair and won't foster restraint. Instead, ask for targeted suggestions, and test them on representative data. Then, make expedited production of responsive data from the sample to let everyone see what's working and what's not.

Importantly, frame your approach to accommodate at least two rounds of keyword search and review, affording the other side a reasonable opportunity to review the first production before proposing additional searches. When an opponent knows they'll get a second dip at the well, they don't have to make Draconian demands.

5. Incorporate Misspellings, Variants and Synonyms

Did you know Google got its name because its founders couldn't spell googol? Whether due to typos, transposition, IM-speak, misuse of homophones or ignorance, electronically stored information fairly crawls with misspellings that complicate keyword search. Merely searching for "management" will miss "managment" and "mangement."

To address this, you must either include common variants and errors in your list of keywords or employ a search tool that supports fuzzy searching. The former tends to be more efficient because fuzzy searching (also called *approximate string matching*) mechanically varies letters, often producing an unacceptably high level of false hits.

How do you convert keywords to their most common misspellings and variants? A linguist could help or you can turn to the web. Until a tool emerges that lists common variants and predicts the likelihood of false hits, try a site like http://www.dumbtionary.com that checks keywords against over 10,000 common misspellings and consult Wikipedia's list of more than 4,000 common misspellings (Wikipedia shortcut: **WP:LCM**).

To identify synonyms, pretend you are playing the board game Taboo. Searches for "car" or" automobile" will miss documents about someone's "wheels" or "ride." Consult the thesaurus for likely alternatives for critical keywords, but don't go hog wild with Dr. Roget's list. Question key players about internal use of alternate terms, abbreviations or slang.

6. Filter and Deduplicate First

Always filter out irrelevant file types and locations before initiating search. Music and images are unlikely to hold responsive text, yet they'll generate vast numbers of false hits because their content is stored as alphanumeric characters. The same issue arises when search tools fail to decode e-mail attachments before search. Here again, you have to know *how* your search tool handles encoded, embedded, multibyte and compressed content.

Filtering irrelevant file types can be accomplished various ways, including culling by binary signatures, file extensions, paths, dates or sizes and by de-NISTing for known hash values. The National Institute of Standards and Technology maintains a registry of hash values for commercial software and operating system files that can be used to reliably exclude known, benign files from e-discovery collections prior to search. http://www.nsrl.nist.gov.

The exponential growth in the volume of ESI doesn't represent a leap in productivity so much as an explosion in duplication and distribution. Much of the data we encounter are the *same* documents, messages and attachments replicated across multiple backup intervals, devices and custodians. Accordingly, the efficiency of search is greatly aided—and the cost greatly reduced—by *deduplicating* repetitious content *before* indexing data for search or running keywords. Employ a method of deduplication that tracks the origins of suppressed iterations so that repopulation can be accomplished on a per custodian basis.

Applied sparingly and with care, you may even be able to use keywords to exclude irrelevant ESI. For example, the presence of keywords "Cialis" or "baby shower" in an e-mail may reliably signal the message isn't responsive; but *testing and sampling must be used to validate such exclusionary searches*.

7. Test, Test, Test!

The single most important step you can take to assess keywords is to test search terms against representative data from the universe of machines and data under scrutiny. No matter how well you think you know the data or have refined your searches, testing will open your eyes to the unforeseen and likely save a lot of wasted time and money.

The nature and sample size of representative data will vary with each case. The goal in selection isn't to reflect the average employee's collection but to fairly mirror the collections of employees likely to hold responsive evidence. Don't select a custodian in marketing if the key players are in engineering.

Often, the optimum custodial choices will be obvious, especially when their roles made them a nexus for relevant communications. Custodians prone to retention of ESI are better candidates than those priding themselves on empty inboxes. The goal is to flush out problems *before* deploying searches across broader collections, so opting for uncomplicated samples lessens the value.

It's amazing how many false hits turn up in application help files and system logs; so early on, I like to test for noisy keywords by running searches against data having nothing whatsoever to do with the case or the parties (e.g., the contents of a new computer). Being able to show a large number of hits in wholly irrelevant collections is compelling justification for limiting or eliminating unsuitable keywords.

Similarly, test search terms against data samples collected from employees or business units having nothing to do with the subject events to determine whether search terms are too generic.

8. Review the Hits

My practice when testing keywords is to generate spreadsheet-style views letting me preview search hits in context, that is, flanked by 20 to 30 words on each side of the hit. It's efficient and illuminating to scan a column of hits, pinpoint searches gone awry and select particular documents for further scrutiny. Not all search tools support this ability, so check with your service provider to see what options they offer.

Armed with the results of your test runs, determine whether the keywords employed are hitting on a reasonably high incidence of potentially responsive documents. If not, what usages are throwing the search off? What file types are appearing on exceptions lists as unsearchable due to, e.g., obscure encoding, password protection or encryption?

As responsive documents are identified, review them for additional keywords, acronyms and misspellings. Are terms that should be finding known responsive documents failing to achieve

hits? Are there any consistent features in the documents with noise hits that would allow them to be excluded by modifying the query?

Effective search is an *iterative* process, and success depends on new insight from each pass. So expect to spend considerable time assessing the results of your sample search. It's time wisely invested.

9. Tweak the Queries and Retest

As you review the sample searches, look for ways you can tweak the queries to achieve better precision without adversely affecting recall. Do keyword pairs tend to cluster in responsive documents such that using a Boolean *and* connector will reduce noise hits? Can you approximate the precise context you seek by controlling for proximity between terms?

If very short (e.g., three letter) acronyms or words are generating too many noise hits, you may improve performance by controlling for case (e.g., all caps) or searching for discrete occurrences (i.e., the term is flanked only by spaces or punctuation).

10. Check the Discards

Keyword search must be judged both by what it *finds* and what it *misses*. That's the "quality assurance" courts demand. A defensible search protocol includes limited examination of the items <u>not</u> generating hits to assess whether relevant documents are being passed over.

Examination of the discards will be more exacting for your representative sample searches as you seek to refine and gain confidence in your queries. Thereafter, random sampling should suffice.

No court has proposed a benchmark or rule-of-thumb for random sampling, but there's more science to sampling than simply checking every hundredth document. If your budget doesn't allow for expert statistical advice, and you can't reach a consensus with the other side, be prepared to articulate why your sampling method was chosen and why it strikes a fair balance between quality assurance and economy. The sampling method you employ needn't be foolproof, but it must be rational.

Remember that the purpose of sampling the discards is to promptly *identify and resolve* ineffective searches. If quality assurance examinations reveal that responsive documents are turning up in the discards, those failures must receive prompt attention.

Search Tips

Defensible search strategies are well-documented. Record your efforts in composing, testing and tweaking search terms and the reasons for your choices along the way. Spreadsheets are handy for tracking the evolution of your queries as you add, cut, test and modify them.

Effective searches are tailored to the data under scrutiny. For example, it's silly to run a custodian's name or e-mail address against his or her own e-mail, but sensible for other collections. It's often smart to *tier* your ESI and employ keywords suited to each tier or, when feasible, to limit searches to just those file types or segments of documents (i.e., message body and subject) likely to be responsive. This requires understanding what you're searching and how it's structured.

When searching e-mail for recipients, it's almost always better to search by e-mail address than by name. In a company with dozens of Bob Browns, each must have a unique e-mail address. Be sure to check whether users employ e-mail aliasing (assigning idiosyncratic "nicknames" to addressees) or distribution lists, as these can thwart search by e-mail address or name.

Search is a Science…

…but one lawyers *can* master. I guarantee these steps will wring more quality and trim the fat from text retrieval. *It's worth the trouble*, because the lowest cost e-discovery effort is the one done right from the start.

Forms that Function

This article discusses how to request and produce electronically stored information (ESI) in *forms that function*—that is, in more utile and complete forms of production that preserve the integrity, efficiency and functionality of digital evidence. It explains the advantages of securing production in native and near-native forms, and supplies exemplar language crafted to convey forms of production and metadata values sought.

BACKGROUND

Historically, the law little concerned itself with "forms" of production because there were few alternatives to paper. Then, evidence became digital: documents, pictures, sounds, text messages, e-mail, spreadsheets, presentations, databases and more were created, communicated and recorded as a sequence of "ones" and "zeroes." Flat forms of information acquired new dimension and depth, described and supplemented by **metadata,** *i.e.,* data *about* data supporting the ability to find, use and trust digital information.

Digital photographs hold EXIF data revealing where they were taken and by what camera, spreadsheets carry formulae supporting complex calculations and Word documents store editorial histories and are laced with conversations between collaborators. Presentations feature animated text and rich media, including sound, video and dynamic connections to other data. Databases don't "store" documents as much as assemble them on demand. Even conversations—once the most ethereal of interactions—now linger as text messages and data packets traversing the internet and cellular networks.

Today, the forms in which information is supplied determine if it is intelligible, functional and complete.

FORMS OF PRODUCTION IN THE FEDERAL RULES

The Federal Rules of Civil Procedure further the goals that lawyers understand the forms of ESI in their cases and resolve forms disputes before requests for production are served. Unresolved forms disputes should be brought to court quickly.

Rule 26(f)(3)(C) requires the parties to submit a discovery plan to the Court prior to the first pretrial conference. The plan must address "any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced."

Rule 34(b)(1)(C) permits requesting parties to "specify the form or forms in which electronically stored information is to be produced," yet it's common for requests for production to be wholly silent on forms of production, despite pages of detailed definitions and instructions.

Practice Tip: Requesting parties should supply a clear and practical written specification of forms sought *before* the initial Rule 26(f) conference, affording opponents the opportunity to assess the feasibility, cost and burden of producing in specified forms. Even parties who do not know the forms in which an opponent's data natively resides can anticipate the *most common* forms of, *e.g.*, e-mail, word processed documents, presentations and spreadsheets.

The Federal Rules lay out **FIVE STEPS** to seeking and objecting to forms of production:

1. Before the first pretrial conference, parties must hash out issues related to "the form or forms in which [ESI] should be produced. FRCP 26(f)(3)(C)

2. Requesting party specifies the form or forms of production for each type of ESI sought: paper, native, near-native, imaged formats or a mix of same. FRCP 34(b)(1)(C)

3. If the responding party will supply the specified forms, the parties proceed with production. If not, the responding party must object and designate the forms in which it intends to make production. If the requesting party fails to specify forms sought, responding party must state the form or forms it intends to produce. FRCP 34(b)(2)(D)

The Notes to Rule 34(b) add: "A party that responds to a discovery request by simply producing electronically stored information in a form of its choice, without identifying that form in advance of the production . . . runs a risk that the requesting party can show that the produced form is not reasonably usable and that it is entitled to production of some or all of the information in an additional form."

4. If requesting party won't accept the forms the producing party designates, requesting party must confer with the producing party in an effort to resolve the dispute. FRCP 37(a)(1)

5. If the parties can't agree, requesting party files a motion to compel, and the Court selects the forms to be produced.

Practice Tip: Even when producing parties use native and near-native forms when reviewing for responsiveness and privilege, the final step before production is often to downgrade the evidence to images before production. Accordingly, requesting parties shouldn't wait until the response date to ascertain if an opponent refuses to furnish the forms sought. Press for a commitment; and if not forthcoming, move to compel ahead of the response date. Don't wait to hear the Court ask, "Why didn't you raise this earlier?"

WHAT ARE THE OPTIONS FOR FORMS OF PRODUCTION?

It's rarely necessary or advisable to employ a single form of production for all ESI produced in discovery; instead, tailor forms to the data. Options for forms of production include:

- Paper [where the source is paper and the volume small]
- Page Images [best for items requiring redaction and scanned paper records]
- Native [spreadsheets, electronic presentations and word-processed documents]
- Near-native [e-mail and database content]
- Hosted production

Paper

Converting searchable electronic data to paper is usually a step backward, but paper remains a reasonable choice where the items to be produced are paper documents, few in number and electronic searchability isn't required.

Page Images

Parties produce digital "pictures" of documents, e-mails and other electronic records, typically furnished in Adobe's Portable Document Format (PDF) or as Tagged Image File Format (TIFF) images. Converting ESI to TIFF images strips its electronic searchability and metadata. Accordingly, TIFF image productions are accompanied by load files holding searchable text and selected metadata. Searchable text is obtained by extraction from an electronic source or for scanned paper documents, by use of optical character recognition (OCR). Load files are composed of *delimited text*, *i.e.,* values following a predetermined sequence and separated by characters like commas, tabs or quotation marks. The organization of load files must be negotiated, and is often pegged to review software like CT Summation, LexisNexis Concordance or kCura Relativity.

> ***Pros:*** Imaged formats are ideal for production of scanned paper records, microfilm and microfiche, especially when OCR serves to add electronic searchability.
>
> ***Cons:*** Imaged production breaks down when ESI holds embedded information (*e.g.,* collaborative content like comments or formulae in spreadsheets) or non-printable information (*e.g.,* voice mail, video or animation and structured data). Imaged productions may also serve to degrade evidence when the information is fielded (*e.g.,* structured data and messaging) or functional (*e.g.,* animations in presentations, table relationships in structured data or threads in e-mail).

Native Production

Parties produce the actual data files containing responsive information, *e.g.,* Word documents in their native .DOC or .DOCX formats, Excel spreadsheets as .XLS and .XLSX files and PowerPoint

presentations in native .PPT and .PPTX. Native production is cheaper and better in competent hands using tools purpose-built for native review.

> ***Pros:*** The immediate benefits to the producing party are speed and economy—little or nothing must be spent on image conversion, text extraction or OCR.
>
> The benefits to the requesting party are substantial. Using native review tools or applications like those used to create the data (Careful here!—see *Cons* below), requesting parties see the evidence as it appeared to the producing party. Embedded commentary and metadata aren't stripped away, deduplication is facilitated, e-mail messages can be threaded into conversations, time zone irregularities are normalized and costs are reduced and utility enhanced every step of the way.
>
> ***Cons:*** Applications needed to view rare and obscure data formats may be prohibitively expensive (*e.g.,* specialized engineering applications or enterprise database software). If native applications are (unwisely) tasked to review, *e.g.,* Microsoft Word for reviewing Word documents, copies must be used to avoid altering evidence.

Near-Native Production

Some ESI cannot be feasibly or prudently tendered in true native formats. *Near*-native forms preserve the essential utility, content and searchability of native forms but are not, strictly speaking, native forms. Examples:

- **Enterprise e-mail** - When messages are exported from a corporate Exchange mail database to a container format, the container isn't native to the mail server; but it replicates the pertinent content and essential functionality of the source.
- **Databases** - Exports from databases are often produced in delimited formats not native to the database, yet supporting the ability to interpret the data in ways faithful to the source.
- **Social networking content** - Content from social networking sites like Facebook won't replicate the precise way the content is stored in the cloud, so near-native forms seek to replicate its essential utility, completeness and searchability.

Hosted Production

Hosted production is more a delivery medium than a discrete form of production. Hosted production resides on a secure website. Requesting parties access data using their web browser, searching, viewing, annotating and downloading data.

MORE ON LOAD FILES

TIFF images cannot carry the text, but PDF images can. Think pants with pockets versus skirts without pockets. When you use TIFF images for production, text has to go *somewhere* and, since TIFFs have no "pockets," the text goes into a purse called a "load file."

Load files first appeared in discovery in the 1980s to add electronic searchability to scanned paper documents and are called load files because they're used to load data to ("populate") databases called review platforms.

Different review platforms used different load file formats to order and separate information according to guidelines called "load file specifications." Load files employ characters called delimiters to field (separate) the various information items in the load file.

Load File Structure

Imagine creating a table to keep track of documents. You might use the first two columns of your table to number the first and last page of each document. The next column holds the document's name and then each succeeding column carries information about the document. To tell one column from the next, you'd draw lines to delineate the rows and columns, like so:

The lines serve as delimiters—literally delineating one field of data from the next. Vertical and

BEGDOC	ENDDOC	FILENAME	MODDATE	AUTHOR	DOCTYPE
0000001	0000004	Contract	01/12/2013	J. Smith	docx
0000005	0000005	Memo	02/03/2013	R. Jones	docx
0000006	0000073	Taxes_2013	04/14/2013	H. Block	xlsx
0000074	0000089	Policy	5/25/2013	A. Dobey	pdf

horizontal lines are excellent visual delimiters for humans, but computers work well with characters like commas or tabs. So, if the tabular data were a load file, it might be delimited as:

```
BEGDOC,ENDDOC,FILENAME,MODDATE,AUTHOR,DOCTYPE
0000001,0000004,Contract,01/12/2013,J. Smith,docx
0000005,0000005,Memo,02/03/2013,R. Jones,docx
0000006,0000073,Taxes_2013,04/14/2013,H. Block,xlsx
0000074,0000089,Policy,05/25/2013,A. Dobey,pdf
```

Each comma replaces a column divider, each line signifies another row and the first or "header" row is used to define the data that follows and the manner in which it's delimited.

Load files that use commas to separate values are called "comma separated value" or CSV files. More commonly, load files adhere to formats compatible with the Concordance and Summation review tools.

Concordance load files use the file extension .DAT and the þ (thorn, ALT-0254) and ¶ (pilcrow, ALT-0182) characters as delimiters*:*

Concordance Load File

```
þBEGDOCþ¶þenddocþ¶þfilenameþ¶þMODDATEþ¶þAUTHORþ¶þDOCTYPEþ
þ0000001þ¶þ0000004þ¶þContractþ¶þ01/12/2013þ¶þJ. Smithþ¶þdocxþ
þ0000005þ¶þ0000005þ¶þMemoþ¶þ02/03/2013þ¶þR. Jonesþ¶þdocxþ
þ0000006þ¶þ0000073þ¶þTaxes_2013þ¶þ04/14/2013þ¶þH. Blockþ¶þxlsxþ
þ0000074þ¶þ0000089þ¶þPolicyþ¶þ05/25/2013þ¶þA. Dobeyþ¶þpdfþ
```

Summation load files use the file extension .DII, and separate each record like so:

Summation Load File

```
; Record 1
@T 0000001
@DOCID 0000001
@MEDIA eDoc
@C ENDDOC 0000004
@C PGCOUNT 4
@C AUTHOR J. Smith
@DATESAVED 01/12/2013
@EDOC \NATIVE\Contract.docx
; Record 2
@T 0000005
@DOCID 0000005
@MEDIA eDoc
@C ENDDOC 0000005
@C PGCOUNT 1
@C AUTHOR R. Jones
@DATESAVED 02/03/2013
@EDOC \NATIVE\Memo.docx
@C AUTHOR A. Dobey
@DATESAVED 05/25/2013
@EDOC \NATIVE\Policy.pdf
```

Two more load files:

Opticon load files (file extension .OPT) are used in conjunction with Concordance load files to pair Bates numbered pages with corresponding page images and to define the **unitization** of each document; that is, where they begin and end. Document are unitized *physically*, as when constituent pages are joined by clips, staples or bindings, or *logically*, where constituent pages belong together even if not physically unitized (as when documents are bulk scanned or transmittals reference enclosures). Logical unitization is also a means to track family relationships between container files and contents and e-mail messages and attachments.

Opticon load files employ a simple seven-field, comma-delimited structure:

1. Page identifier,
2. Volume label (optional),
3. Path to page image,
4. New document marker,
5. Box identifier (optional),
6. Folder identifier (optional),
7. Page count (optional).

Opticon Load File

```
0000001_0001,,TIFF\001\0000001_0001.tif,Y,,,4
0000002_0002,,TIFF\001\0000002_0002.tif,,,,
0000003_0003,,TIFF\001\0000003_0003.tif,,,,
0000004_0004,,TIFF\001\0000004_0004.tif,,,,
0000005_0001,,TIFF\001\0000005_0001.tif,Y,,,1
0000006_0001,,TIFF\001\0000006_0001.tif,Y,,,68
0000007_0002,,TIFF\001\0000007_0002.tif,,,,
0000008_0003,,TIFF\001\0000008_0003.tif,,,,
0000009_0004,,TIFF\001\0000009_0004.tif,,,,
0000010_0005,,TIFF\001\0000010_0005.tif,,,,
```

Overlay load files are used to update or correct existing database content by replacing data in fields in the order in which the records occur. Thus, it's crucial that the order of data within the overlay file match the order of data replaced. Data must be sorted in the same way, and the overlay must not add or omit fields.

Making the Case against Imaged Production

Parties don't print their e-mail before reading it or emboss a document's name on every page. Parties communicate and collaborate using tracked changes and embedded comments. Parties use native forms because they are the most utile, complete and efficient forms in which to store and access data.

Lawyers come along and convert native forms to images, Bates label each page and purge tracked changes and embedded comments without disclosing the destruction.

Converting a client's ESI from its native state as kept "in its ordinary course of business" to TIFF images **injects needless expense in at least half a dozen ways**:

1. You pay to convert native forms to TIFF images and emboss Bates numbers;
2. You pay to generate load files;
3. You must produce multiple copies of documents (like spreadsheets) that are virtually incapable of production as images;
4. TIFF images and load files are <u>much</u> "fatter" files than their native counterparts (*i.e.,* bloated *5-40 times* as large), so you pay more for vendors to ingest and host them;
5. It's difficult to reliably de-duplicate documents once converted to images; and
6. You must reproduce everything when opponents recognize that imaged productions fall short of native productions.

REBUTTING THE CASE AGAINST NATIVE

When producing parties insists on converting ESI to TIFF despite a timely request for native production, they often rely on Federal Rules of Civil Procedure 34(b)(2)(E)(ii), which obliges parties

to produce ESI in "the form or forms in which it is ordinarily maintained or in a reasonably usable form or forms." This reliance is misplaced because "[i]t is only if the requesting party declines to specify a form that the producing party is offered a choice between producing in the form 'in which it is ordinary maintained'—native format—or 'in a reasonably useful form or forms.' Fed. R. Civ. P. 34(b)(2)(E)(i)-(ii)"). *The Anderson Living Trust v. WPX Energy Production,* LLC, No. CIV 12-0040 JB/LFG. (D. New Mexico March 6, 2014

Producing parties usually assert **FOUR JUSTIFICATIONS** for refusing to produce ESI in native and near-native forms. None withstand scrutiny:

1. *You can't Bates label native files.* Making the transition to modern forms of production requires acceptance of three propositions:

- Printouts and images of ESI are not "the same" as ESI;
- Most items produced in discovery aren't used in proceedings; and
- Names of electronic files can be simply changed without altering contents of files.

Native documents carry more information than their imaged counterparts, and are inherently functional, searchable and complete. Moreover, native documents are described by more and different metadata—information invaluable in identifying, sorting and authenticating evidence.

Though you can't *emboss* Bates-style identifiers on discrete pages of a native file until printed or imaged, many native forms (*e.g.,* spreadsheets, social networking content, video, and sound files) don't lend themselves to paged formats and would not be Bates labeled. When Bates-style identifiers are needed on pages for use in proceedings, simply require that file identifiers and page numbers be embossed on images or printouts. In practice, that impacts only a small subset of production.

> ***Practice tip:*** It's simple and cheap to replace, prepend, or append an incrementing Bates-style identifier to a filename. One free file renaming tool is Bulk Rename Utility, available at www.bulkrenameutility.co.uk. You can even include a protective legend like "Subject to Protective Order." **Renaming a file does not alter its content, hash value or last modified date.**

2. *Opponents will alter evidence*. Evidence tampering is not a new fear or a hazard unique to e-discovery. Page images, being black and white pictures of text, are simple to manipulate (and Adobe Acrobat has long allowed extensive revision of PDF files).

Though any form of production is prey to unscrupulous opponents, native productions support quick, reliable ways to prevent and detect alteration. Producing native files on read-only media like CDs or DVDs) guards against inadvertent alteration. Alterations are easily detected by comparing hash values (digital fingerprints) of suspect files to the files produced.

Counsel savvy enough to seek native production should be savvy enough to refrain from evidence handling practices prone to alter the evidence.

3. *Native production requires broader review*. Native forms routinely hold user-generated content (*e.g.,* collaborative comments in Word documents, animated "off-screen" and layered text in presentations and formulae in spreadsheets) that is rarely visible on page images or intelligible on extracted text. Imaged productions often obliterate such matter *without review and without disclosure, objection or logging.* Review is only "broader" because this user-contributed content has long been furtively and indefensibly stripped away.

4. *Redacting native files changes them*. Change is the sole purpose of redaction. The form of production for items requiring redaction should be the form or forms best suited to efficient removal of privileged or protected content without rendering the remaining content wholly unusable.

Some native file formats support redaction brilliantly; others do not. In the final analysis, the volume of items redacted tends to be insignificant. Accordingly, the form selected for redaction shouldn't dictate the broader forms of production when, overall, native forms have decided advantages for items not requiring.

> ***Practice Tip:*** Don't let the redaction tail wag the production dog. If an opponent wants to redact in .tiff or PDF, *let them*, but only for the redacted items and only when they restore searchability after redaction.

UPDATING YOUR REQUESTS FOR PRODUCTION

The first step in getting the information you seek in the forms you desire is to ask for it, applying the rules and eschewing dated boilerplate. Clear, specific requests are the hardest to evade and the easiest to enforce. See **Appendix: Exemplar Production Protocol, *infra.***

Most digital evidence—including e-mail—exists as data within databases. So, stop thinking about discovery as the quest for "documents" and start focusing on what you really seek: *information in utile and complete forms.*

The definition of "document" must give way to an alternate term like "information" or "information items." Instead of the usual thesaurus-like litany of types of information, consider:

"Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations, spreadsheets and database content.

Next, **cut junk prose** like "including, but not limited to" and "any and all." They don't add clarity. If you must incorporate examples of responsive items in a request, just say "including" and add an instruction that says, "Examples of responsive items set out in any request should not be construed to limit the scope of the request." If drafting a request without "any and all" makes you quake, add the instruction, "Requests for production should be read so as to encompass any and all items responsive to the request."

Before you serve discovery, **check your definitions** to be sure you've defined only terms you've used and used terms only in ways consistent with your definitions.

Specify the forms you seek

The most common error seen in requests for production is the failure to specify the forms sought for ESI production. Worse, requests often contain legacy boilerplate specifying forms the requesting party *doesn't* want.

Every request for production should specify forms of production sensibly and precisely. Don't assume that "native format" is clear or sufficient; instead, specify the formats sought for common file types, *e.g.:*

Information that exists in electronic form should be produced in native or near-native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored in the ordinary course. The table below supplies examples of the native or near-native forms in which specific types of electronically stored information (ESI) should be produced.	
Source ESI	**Native or Near-Native Form or Forms Sought**
Microsoft Word documents	.DOC, .DOCX
Microsoft Excel Spreadsheets	.XLS, .XLSX

Microsoft PowerPoint Presentations	.PPT, .PPTX
Microsoft Access Databases	.MDB, .ACCDB
WordPerfect documents	.WPD
Adobe Acrobat Documents	.PDF
Images	.JPG, .JPEG, .PNG
E-mail	Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in the internet e-mail standard, RFC 5322. For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished with folder data. For Lotus Notes mail, furnish .NSF files or convert to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.
Databases	Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schema and query language of the database along with a detailed description of its export capabilities so as to facilitate crafting a query to extract and export responsive data.
Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with unredacted OCR text furnished and logical unitization and family relationships preserved.	

Practice Tip: In settling upon a form of production for e-mail, use this inquiry as a litmus test to distinguish "native" forms from less functional forms: ***Can the form produced be imported into common e-mail client or server applications?*** If the form of the e-mail is so degraded that e-mail programs cannot recognize it as e-mail, that's a strong indication the form of production has strayed too far from functional.

Specify the Load File Format

Every electronic file has a complement of descriptive information called *system metadata* residing in the file table of the system or device storing the file. Different file types have different metadata. Every e-mail message has *"fields"* of information in the message "*header*" that support better searching, sorting and organization of messages. This may be data probative in its own right or simply advantageous to managing and authenticating electronic evidence. Either way, you want to be certain to request it sensibly and precisely. Simply demanding "the metadata" reveals you don't fully understand what you're seeking.

Develop a comprehensive production protocol tailored to the case and serve same with discovery. Always specifically request the metadata and header fields you seek, *e.g.:*

Produce delimited load file(s) supplying relevant system metadata field values for each information item by Bates number. Typical field values supplied include:

a. **Source file name** (original name of the item or file when collected from the source custodian or system);
b. **Source file path** (fully qualified file path from the root of the location from which the item was collected);
c. **Last modified date and time** (last modified date and time of the item);
d. **UTC Offset** (The UTC/GMT offset of the item's modified date and time*, e.g.,* -500).
e. **Custodian or source** (unique identifier for the original custodian or source);
f. **Document type**;
g. **Production File Path** (file path to the item from the root of the production media);
h. **MD5 hash** (MD5 hash value of the item as produced);
i. **Redacted flag** (indication whether the content or metadata of the item has been altered after its collection from the source custodian or system);
j. **Embedded Content Flag** (indication that the item contains embedded or hidden comments, content or tracked changes); and
k. **Deduplicated instances** (by full path)**.**

The following additional fields shall accompany production of e-mail messages:

l. **To** (e-mail address(es) of intended recipient(s) of the message);
m. **From** (e-mail address of the person sending the message);
n. **CC** (e-mail address(es) of person(s) copied on the message);
o. **BCC** (e-mail address(es) of person(s)blind copied on the message);
p. **Subject** (subject line of the message);
q. **Date Received** (date the message was received);
r. **Time Received** (time the message was received);
s. **Attachments** (beginning Bates numbers of attachments);
t. **Mail Folder Path** (path of the message from the root of the mail folder);and
u. **Message ID** (unique message identifier).

Hybrid productions mixing mix imaged and native formats also require that paths to images and extracted text be furnished, as well as **logical unitization data** serving as the electronic equivalent of paper clips and staples.

De-duplication and Redaction

You may wish to specify whether the production should or should not be de-duplicated, *e.g.:*

> Documents should be vertically de-duplicated by custodian using each document's hash value. Near-deduplication should not be employed so as to suppress different versions of a document, notations, comments, tracked changes or application metadata.

Because redaction tends to impact just a small part of most productions, it's important that it not co-opt the forms of production.

> Information items that require redaction shall be produced in static image formats*, e.g.,* single page .tiff or multipage PDF images with logical unitization preserved. The unredacted content of each document should be extracted by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Redactions should not be accomplished in a manner that serves to downgrade the ability to electronically search the unredacted portions of the item.

A TIFF-OCR redaction method works reasonably well for text documents, but often fails when applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can't spellcheck numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.

Specify the medium of production

A well-crafted request should address the *medium* of ESI production; that is the mechanism used to convey the electronic production to the requesting party. If you're receiving 100GB of data, you don't want it tendered on 143 CDs.

> Production of ESI should be made using appropriate electronic media of the producing party's choosing that does not impose an undue burden or expense upon a recipient.

Conclusion

It's time to take a hard look at the language of the definitions and instructions accompanying requests for production. Most are boilerplate borrowed from someone who borrowed it from someone who drafted it in 1947. It's hand-me-down verbiage long past retirement age; so, retire it and craft modern requests for a modern digital world.

We will never be less digital than we are today. Isn't it time we demand modern evidence and obtain it in the forms in which it serves us best? We must move forms of production upstream, from depleted images and load files to functional native and near native forms retaining the content and structure that supports migration into any form. *Utile* forms. *Complete* forms. *Forms that function.*

Exemplar Production Protocol

This Appendix is an example of a *production protocol*, sometimes called a *data delivery standard*. Geared to civil litigation and seeking the lowest cost approach to production of ESI, it seeks native production of common file types and relieves parties of the burden convert ESI to imaged formats except when needed for redaction. This exemplar protocol specifies near-native alternatives for production of native forms when near-native forms are preferable. For an example of a U.S. Government data delivery standard, see:
http://www.sec.gov/divisions/enforce/datadeliverystandards.pdf

Appendix: Exemplar Production Protocol

1. "Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations, spreadsheets and database content.

2. Information that exists in electronic form should be produced in native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored in the ordinary course.

3. If it is infeasible to produce an item of responsive ESI in its native form, it may be produced in an agreed-upon near-native form; that is, in a form in which the item can be imported into the native application without a material loss of content, structure or functionality as compared to the native form. Static image production formats serve as near-native alternatives only for information items that are natively static images (*i.e.*, photographs and scans of hard-copy documents).

4. The table below supplies examples of agreed-upon native or near-native forms in which specific types of ESI should be produced:

Source ESI	Native or Near-Native Form or Forms Sought
Microsoft Word documents	.DOC, .DOCX
Microsoft Excel Spreadsheets	.XLS, .XLSX
Microsoft PowerPoint Presentations	.PPT, .PPTX
Microsoft Access Databases	.MDB, .ACCDB
WordPerfect documents	.WPD
Adobe Acrobat Documents	.PDF
Photographs	.JPG, .PDF
E-mail	Messages should be produced in a form or forms that readily support import into standard e-mail

	client programs; that is, the form of production should adhere to the conventions set out in the internet e-mail standard, RFC 5322. For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished with folder data. For Lotus Notes mail, furnish .NSF files or convert to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.
Databases	Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schema and query language of the database along with a detailed description of its export capabilities so as to facilitate crafting a query to extract and export responsive data.
Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with OCR text of unredacted content furnished and logical unitization and family relationships preserved.	

5. Absent a showing of need, a party shall produce responsive information reports contained in databases through the use of standard reports; that is, reports that can be generated in the ordinary course of business and without specialized programming efforts beyond those necessary to generate standard reports. All such reports shall be produced in a delimited electronic format preserving field and record structures and names. The parties will meet and confer regarding programmatic database productions as necessary.

6. Information items that are paper documents or that require redaction shall be produced in static image formats scanned at 300 dpi e.g., single-page Group IV.TIFF or multipage PDF images. If an information item employs color to convey information (versus purely decorative use), the producing party shall not produce the item in a form that does not display color. The full content of each document will be extracted directly from the native source where feasible

or, where infeasible, by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Redactions shall be logged along with other information items withheld on claims of privilege.

7. Parties shall take reasonable steps to ensure that text extraction methods produce usable, accurate and complete searchable text.

8. Individual information items requiring redaction shall (as feasible) be redacted natively, produced in .PDF format and redacted using the Adobe Acrobat redaction feature or redacted and produced in another reasonable manner that does not serve to downgrade the ability to electronically search the unredacted portions of the item. Bates identifiers should be endorsed on the lower right corner of all images of redacted items so as not to obscure content.

9. Upon a showing of need, a producing party shall make a reasonable effort to locate and produce the native counterpart(s) of any .PDF or .TIF document produced. The parties agree to meet and confer regarding production of any such documents. This provision shall not serve to require a producing party to reveal redacted content.

10. Except as set out in this Protocol, a party need not produce identical information items in more than one form. The content, metadata and utility of an information item shall all be considered in determining whether information items are identical, and items reflecting different information shall not be deemed identical.

11. Production of ESI should be made using appropriate electronic media of the producing party's choosing that does not impose an undue burden or expense upon a recipient. Label all media with the case number, production date, Bates range and disk number (1 of X, if applicable). Organize productions by custodian, unless otherwise instructed. All productions should be encrypted for transmission to the receiving party. The producing party shall, contemporaneously with production, separately supply decryption credentials and passwords to the receiving party for all items produced in an encrypted or password-protected form.

12. Each information item produced shall be identified by naming the item to correspond to a Bates identifier according to the following protocol:

    i. The first four (4) characters of the filename will reflect a unique alphanumeric designation identifying the party making production;

    ii. The next six (6) characters will be a designation reserved to the discretionary use of the party making production for, e.g., denoting the case or matter. This value shall be padded with leading zeroes as needed to preserve its length;

iii. The next nine (9) characters will be a unique, consecutive numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length;

iv. The final six (6) characters are reserved to a sequence consistently beginning with a dash (-) or underscore (_) followed by a five-digit number reflecting pagination of the item when printed to paper or converted to an image format for use in proceedings or when attached as exhibits to pleadings.

v. By way of example, a Microsoft Word document produced by Acme in its native format might be named: ACMESAMPLE000000123.docx. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACMESAMPLE000000123_00006. Bates identifiers should be endorsed on the lower right corner of all printed pages, but not to obscure content.

vi. This format of the Bates identifier must remain consistent across all productions. The number of digits in the numeric portion and characters in the alphanumeric portion of the identifier should not change in subsequent productions, nor should spaces, hyphens, or other separators be added or deleted except as set out above.

13. Information items designated Confidential may, at the Producing Party's option:

a. Be separately produced on electronic production media prominently labeled to comply with the requirements of the **[DATE]** Protective Order entered in this matter; or, alternatively,

b. Each such designated information item shall have appended to the file's name (immediately following its Bates identifier) the following protective legend:
~CONFIDENTIAL-SUBJ_TO_PROTECTIVE_ORDER

When any item so designated is converted to a printed or imaged format for use in any submission or proceeding, the printout or page image shall bear the protective legend on each page in a clear and conspicuous manner, but not so as to obscure content.

14. Producing party shall furnish a delimited load file supplying the metadata field values listed below for each information item produced (to the extent the values exist and as applicable):

Field Name	**Sample Data**	**Description**
BegBates	ACMESAMPLE000000001	First Bates identifier of item
EndBates	ACMESAMPLE000000123	Last Bates identifier of item

AttRange	ACMESAMPLE000000124 - ACMESAMPLE000000130	Bates identifier of the first page of the parent document to the Bates identifier of the last page of the last attachment "child" document
BegAttach	ACMESAMPLE000000124	First Bates identifier of attachment range
EndAttach	ACMESAMPLE000000130	Last Bates identifier of attachment range
Parent_Bates	ACMESAMPLE000000001	First Bates identifier of parent document/e-mail message. ***This Parent_Bates field should be populated in each record representing an attachment "child" document. ***
Child_Bates	ACMESAMPLE000000004; ACMESAMPLE000000012; ACMESAMPLE000000027	First Bates identifier of "child" attachment(s); may be more than one Bates number listed depending on number of attachments. ***The Child_Bates field should be populated in each record representing a "parent" document. ***
Custodian	Houston, Sam	E-mail: mailbox where the email resided. Native: Individual from whom the document originated
Path	E-mail: \Deleted Items\Battles\ SanJac.msg Native: Z:\TravisWB\Alamo.docx	E-mail: Original location of e-mail including original file name. Native: Path where native file document was stored including original file name.
From	E-Mail: Davy@Crockett.net Native: D. Crockett	E-mail: Sender Native: Author(s) of document **semi-colons separate multiple entries **
To	Genl. A.L. de Santa Anna [mailto: sa@sa.mx]	Recipient(s) **semi-colons separate multiple entries **
CC	Jim.Bowie@bigknife.com	Carbon copy recipient(s) **semi-colons separate multiple entries **
BCC	AustinSF@state.tx.gov	Blind carbon copy recipient(s) **semi-colons separate multiple entries **
Date Sent	03/18/2015	E-mail: Date the email was sent
Time Sent	11:45 AM	E-mail: Time the message was sent
Subject/Title	Remember the Alamo!	E-mail: Subject line of the message
IntMsgID	<A1315BC17ABD4774BF779CB3E3E62B9B @gmail.com>	E-mail: For e-mail in Microsoft Outlook/Exchange, the "Unique Message ID" field; For e-mail in Lotus Notes, the UNID field. Native: empty.
Date_Mod	02/23/2015	E-mail: empty. Native: Last Modified Date
Time_Mod	01:42 PM	E-mail: empty Native: Last Modified Time
File_Type	XLSX	E-mail: empty Native: file type
Redacted	Y	Denotes that item has been redacted as containing privileged content (yes/no).
File_Size	1,836	Size of native file document/email in KB.
HiddenCnt	N	Denotes presence of hidden Content/Embedded Objects in item(s) (Y/N)

Confidential	Y	Denotes that item has been designated as confidential pursuant to protective order (Y/N).
MD5_Hash	eb71a966dcdddb929c1055ff2f1ccd5b	MD5 Hash value of the item.
DeDuped	E-mail: \Inbox\SanJac.msg Native: Z:\CrockettD\Alamo.docx	Full path of deduped instances. **semi-colons separate multiple entries **

15. Each production should include a cross-reference load file that correlates the various files, images, metadata field values and searchable text produced.

16. Parties shall respond to each request for production by listing the Bates identifiers/ranges of responsive documents produced, and where an information item responsive to these discovery requests has been withheld or redacted on a claim that it is privileged, the producing party shall furnish a privilege log.

Exercise: Forms of Production and Cost

GOALS: The goals of this exercise are for the reader to:

1. Convert evidence to PDF and TIFF with text; and

2. Assess impact of alternate forms of production in terms of impact on cost of ingestion and hosting.

OUTLINE: You will convert a Microsoft Word document to PDF, TIFF and text formats, compare file sizes and calculate the projected cost of ingestion and monthly hosting for alternate forms of production when the cost of services is assessed on a per-gigabyte pricing model.

Producing parties frequently seek to convert native file formats used by and collected from custodian into static image formats like PDF or more commonly, TIFF images plus load files holding extracted text or text generated through use of optical character recognition. Proponents of static image productions assert claims of superior document security and point to the ability to emboss page numbers and other identifiers on page images. Too, page images can be viewed using any browser application, affording users ready accessibility to some content, albeit sacrificing other content and utility.

Often overlooked in the debate over forms of production is the impact on ingestion, processing, storage and export costs engendered by use of static image formats. Most e-discovery service providers charge to ingest, process, host (store) and export electronically stored information on a per-gigabyte basis. As a result, when items produced occupy more space (measured in bytes), they cost the recipient more to use. This exercise invites students to consider what, if any, increase in cost may flow from the use of static imaged formats as forms of production.

The Myth of Page Equivalency

It's comforting to quantify electronically stored information as some number of pieces of paper or bankers' boxes. Paper and lawyers are old friends. But you can't reliably equate a volume of data with a number of pages unless you know the composition of the data. Even then, it's a leap of faith.

If you troll the Internet for page equivalency claims, you'll be astounded by how widely they vary, though each is offered with utter certitude. A gigabyte of data is variously equated to an absurd 500 million typewritten pages, a naively accepted 500,000 pages, the popularly cited 75,000 pages and a laggardly 15,000 pages. The other striking aspect of page equivalency claims is that they're blithely accepted by lawyers and judges who wouldn't concede the sky is blue without a supporting string citation.

In testimony before the committee drafting the federal e-discovery rules, Exxon Mobil representatives twice asserted that one gigabyte yields 500,000 typewritten pages. The National Conference of Commissioners on Uniform State Laws proposes to include that value in its "Uniform Rules Relating to Discovery of Electronically Stored Information." The Conference of Chief Justices cites the same equivalency in its "Guidelines for State Trial Courts Regarding Discovery of Electronically-Stored Information." Scholarly articles and reported decisions pass around the 500,000 pages per gigabyte value like a bad cold. Yet, 500,000 pages per gigabyte isn't right. It's not even particularly close to right.

Years ago, Kenneth Withers, Deputy Executive Director of The Sedona Conference and then e-discovery guru for the Federal Judicial Center, wrote a section of the fourth edition of "The Manual on Complex Litigation" that equated a terabyte of data to 500 billion typewritten pages. It was supposed to say million, not billion. Eventually, the typo was noticed and corrected; but, the echoes of that innocent thousand-fold mistake still reverberate today. Anointed by the prestige of the manual, the 500-billion-page equivalency was embraced as gospel. Even when the value was "corrected" to 500 million pages per terabyte—equal to 500,000 pages per gigabyte—we're still talking about equivalency with all the credibility of an Elvis sighting.

So, how many pages are there in a gigabyte? It's the answer lawyers love: *"It depends."*

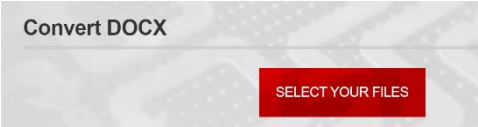
Page equivalency is a myth. One must always look at individual file types and quantities to gauge page equivalency, and there is no reliable rule of thumb geared to how many files of each type a typical user stores. It varies by industry, by user and even by the life span of the media and the evolution of particular applications. A reliable page equivalency must be expressed with reference to both the quantity and form of the data, *e.g.,* "*a gigabyte of single page TIF images of 8-1/2-inch x 11- inch documents scanned at 300 dots per inch equals approximately 18,000 pages*."

Exercise A: Convert Word Document to Imaged Formats

For this exercise, you will download an exemplar Word document and use free, online tools to convert the file to PDF, TIFF and plain text formats.

Step 1: Download the File. Download the file http://www.craigball.com/Always_and_Never.docx and save it to your Desktop or some other location where you can easily find it for this exercise. Should your system not permit download of Word files, you can download the file as a compressed .Zip file from here. Be sure to extract the .DOCX form of the file to your Desktop before proceeding. *You must undertake the conversion exercise using the .DOCX form of the file.*

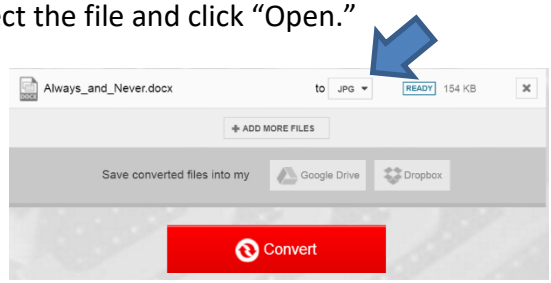
Step 2. Convert the .DOCX file to a PDF. Though there are many ways to convert a Word document to a PDF format, including by using Word itself to Save As a PDF or Print to PDF, we will use an online file converter here for consistency and simplicity.



Using your browser, go to https://convertio.co/convert-docx/ and click on the red SELECT YOUR FILES button.

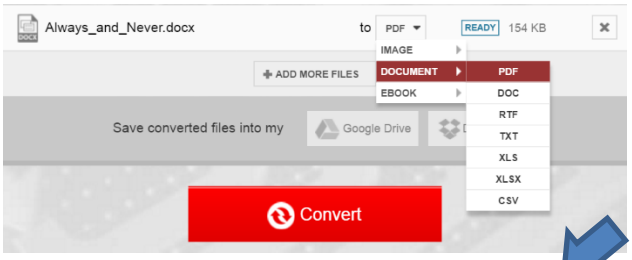
From the Select Files to Convert screen, select "Choose from Computer" then navigate to the file just downloaded called **Always_and_Never.docx.** Select the file and click "Open."

You should see the following screen:



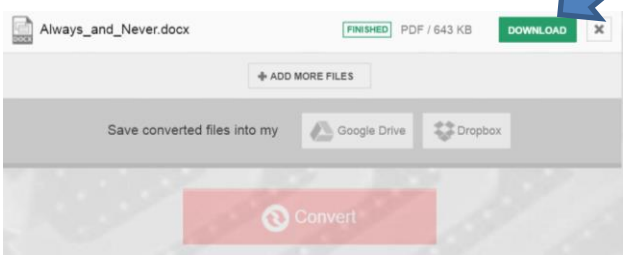
Note the pulldown menu where you may select the format for conversion (JPG in the figure at right) and select the down arrow to view options.

Select DOCUMENT and PDF from the menu and submenu (see figure at right).

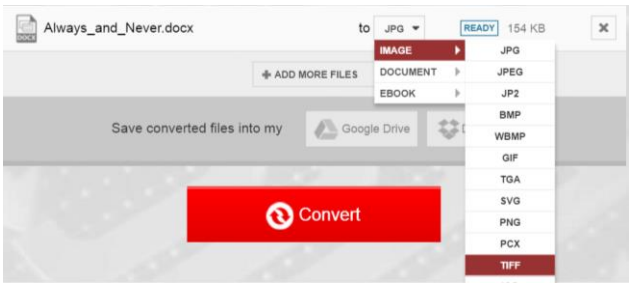


Click the red Convert button.

In the next screen, click the green DOWNLOAD button and save the Always_and_Never.PDF file to the same location where you saved the .DOCX file.

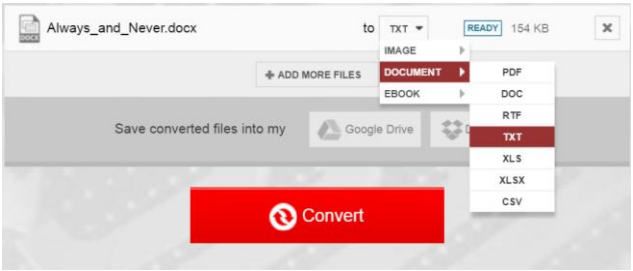


Step 3: **Convert the .DOCX file to TIFF images.** Follow the same steps as above, but this time select IMAGE>TIFF using the drop down menu (see image below) before clicking the red "CONVERT" button.



Click the green DOWNLOAD button again and save the file Always_and_Never.tiff to the same location where you placed the .DOCX and .PDF files.

Step 4: Convert the .DOCX file to plain text. Follow the same steps, but now select DOCUMENT>TXT from the drop down menu (see image below) before clicking the red "CONVERT" button.



Click the green DOWNLOAD button again and save the file Always_and_Never.txt to the same location where you placed the .DOCX and .PDF files.

Step 5: Record the file sizes. Navigate to the location where you downloaded the files and record their file sizes in the blanks below***. Be sure to note if the size value is expressed in units of bytes, kilobytes, megabytes or gigabytes.***

Always_and_Never.DOCX: 18.3 KB

Always_and_Never.PDF: 627 KB

Always_and_Never.TIFF: _____ MB

Always_and_Never.TXT: _____ KB

Exercise B: Calculate the Cost Difference Flowing from Alternate Forms of Production

There may be many variables that go into computing the cost of vendor services for e-discovery, and the charges for ingestion, processing, hosting and export are just parts of a more complicated puzzle. The purpose of this exercise is to gauge the difference that forms of production may make as a component of overall cost.

Problem: ***You are a requesting party in a federal case, and you have made a timely, compliant and unambiguous written request for production of responsive information in native and near-***

native forms. You have expressly requested that Microsoft Word documents be produced in their native .DOC or .DOCX formats. Your opponent instead produces Word documents to you as multiple .TIFF image files accompanied by a load file containing the extracted text from each document. When you object, your opponent counters that "this is what they always do" and that "TIFF plus load file is reasonably usable, so the Rules gave them the right to substitute TIFFs for natives."

Assume that your opponent has produced 1,000 different Word documents which (for ease in making the calculation) are all exactly the same size as the native and converted file sizes for the file Always_and_Never.DOCX. Assume that none of the documents are privileged or required redaction. None are hash-matching duplicates of any other items produced.

You've contracted with an e-discovery service provider to load and host the documents produced so you can review and tag the documents for use in the case. The service provider charges by the gigabyte to ingest, process and host the data month-to-month. This is the applicable fee schedule:

To Ingest and Process Data Supplied:

0 to 300 GB: $75.00 per GB

301 GB to 1 TB: $55.00 per GB

Greater than 1 TB: $40.00 per GB

Monthly Hosting Fee:

0 to 300 GB: $23.00 per GB

301 GB to 1 TB: $20.00 per GB

Greater than 1 TB: $17.00 per GB

Any fraction of a gigabyte will be rounded up to a full gigabyte when calculating charges

You intend to approach the Court to compel your opponent to produce the documents in the form you designated, and in addition to raising issues of utility, completeness and integrity, you want to determine whether the form produced to you will prove more expensive to ingest, process and host for the one-year period you expect to have the data online.

Question: If you accept the production in TIFF and load file, approximately how much more will it cost you over twelve months versus the same production in native forms?

How to Solve this Problem:

Step 1: Normalize the file sizes. Because the prices are quoted in gigabytes, you will want to express all data volumes in gigabytes, rather than as kilobytes or megabytes.

Remember: *A kilobyte is one thousand bytes. A megabyte is one thousand kilobytes. A gigabyte is one thousand megabytes and a terabyte is one thousand gigabytes.*

Step 2: Calculate the cost of Native Production using normalized values:

Native Production: One thousand files, each 18.3KB in zize, is 18,300KB or 18.3MB. Because the service provider's minimum charge is one gigabyte. The cost to ingest and host for one year would be:

Ingest and Process (1GB at $75.00/GB) + Hosting (1GB at $23.00/GB/month x 12 months) = $351.00

Step 3: Calculate the cost of TIFF andText Load File Production using normalized values:

TIFF Plus Production: One thousand files, each *(X) MB in size, is (X) GB,* where (X) is the size of the file Always_and_Never.TIFF*. We must also add the extracted text in the load file, which will be one thousand times (Y) where (Y) is the size of the the* Always_and_Never.TXT file. Any fraction of a gigabyte should be rounded up to the next whole gigabyte. Consequently, the value (Z) is the sum of X plus Y rounded up to the next whole gigabyte.

Ingest and Process (ZGB at *$75.00/GB) + Hosting* (Z GB *at $23.00/GB/month x 12 months) = $_____*

Examplar calculation using hypothetical values:

For example, if Always_and_Never.TIFF was 19MB in size and Always_and_Never.TXT file was 57KB in size, the calculation would be:

X = 1,000 (files) times 19MB = 19GB

Y = 1,000 (text extractions) times 57KB = 57MB = .057GB

Z = (19GB + .057GB) = 20GB (rounded up)

These values would make the calculation of the cost to ingest, process and host the TIFF Plus production:

Ingest and Process (20GB at *$75.00/GB) + Hosting* (20 GB *at $23.00/GB/month x 12 months) = $7,020.00*

The cost difference would be ($7,020.00 less $351.00) = $6,669.00.

Step 4: Calculate the difference using the <u>actual</u> file sizes obtained by your conversion of the file Always_and_Never.DOCX to TIFF and TXT.

What is the actual difference in cost comparing the native production to the TIFF plus TXT load file production?

Enter the actual difference here: $ _____

Preparing for Meet and Confer

Federal Rule of Civil Procedure 26(f) requires parties to confer about preserving discoverable information and to develop a proposed discovery plan addressing discovery of electronically stored information and the form or forms in which it should be produced. This conference[46], and the overall exchange of information about electronic discovery, is called "meet and confer." [47]

Meet and confer is more a process than an event. Lay the foundation for a productive process by communicating your expectations. Send a letter to opposing counsel a week or two prior to each conference identifying the issues you expect to cover and sharing the questions you plan to ask.

E-discovery duties are reciprocal. At meet and confer, be prepared to answer many of the same questions you'll pose. And while the focus will be on large data stores of ESI, don't forget that even if your client has little electronic evidence, you must nonetheless act to preserve and produce it.

If you want client, technical or vendor representatives in attendance, say so. If you're bringing a technical or vendor representative, tell them. Give a heads up on forms of production you'll seek or are prepared to offer. Study up on any load file specification you want used and keywords to search, if only to let the other side know you've done your homework. True, your requests may be ignored or even ridiculed, but it's not an empty exercise. A cardinal rule for electronic discovery, indeed for any discovery, is to tell your opponent what you seek or possess, plainly and clearly. They may show up empty-handed, but not because you failed to set the agenda.

The early, extensive attention to electronic evidence may nonplus lawyers accustomed to the pace of paper discovery. Electronic records are ubiquitous. They're more dynamic and perishable than their paper counterparts, require special tools and techniques to locate and process and implicate

---

[46] The Fed. R. Civ. P. 26(f) conference must occur "as soon as practicable and in any event at least 21 days before a scheduling conference is held or a scheduling order is due under Rule 16(b)…."

[47] *Hopson v. Mayor of Baltimore*, 232 F.R.D. 228, 245 (D. Md. 2006) details some of counsel's duties under Fed, R. Civ. P. 26(f):"[C]ounsel have a duty to take the initiative in meeting and conferring to plan for appropriate discovery of electronically stored information at the commencement of any case in which electronic records will be sought….At a minimum, they should discuss: the type of information technology systems in use and the persons most knowledgeable in their operation; preservation of electronically stored information that may be relevant to the litigation; the scope of the electronic records sought (i.e. e-mail, voice mail, archived data, back-up or disaster recovery data, laptops, personal computers, PDA's, deleted data) the format in which production will occur (will records be produced in "native" or searchable format, or image only; is metadata sought); whether the requesting party seeks to conduct any testing or sampling of the producing party's IT system; the burdens and expenses that the producing party will face based on the Rule 26(b)(2) factors, and how they may be reduced (i.e. limiting the time period for which discovery is sought, limiting the amount of hours the producing party must spend searching, compiling and reviewing electronic records, using sampling to search, rather than searching all records, shifting to the producing party some of the production costs); the amount of pre-production privilege review that is reasonable for the producing party to undertake, and measures to preserve post-production assertion of privilege within a reasonable time; and any protective orders or confidentiality orders that should be in place regarding who may have access to information that is produced."

daunting volumes and multifarious formats. These differences necessitate immediate action and unfamiliar costs. Courts judge harshly those who shirk their electronic evidence obligations.

Questions for Meet and Confer

The following exemplar questions illustrate the types and varieties of matters discussed at meet and confer. They're neither exhaustive nor unique to any type of case, but are offered merely as talking points to stimulate discussion.

1. **What's the case about?**

Relevance remains the polestar for discovery, no matter what form the evidence takes. The scope of preservation and production should reflect both claims *and* defenses. Pleadings only convey so much. Be sure the other side understands your theory of the case and the issues you believe should guide their retention and search.

2. **Who are the key players?**

Cases are still about *people* and what they did or didn't say or do. Though there may be shared repositories and databases to discover, begin your quest for ESI by identifying the *people* whose conduct is at issue. These *key players* are *custodians* of ESI, so determine what devices and applications they use and target their relevant documents, application data and electronic communications. Too, determine whether assistants or secretaries served as proxies for key players in handling e-mail or other ESI.

Like so much in e-discovery, identification of key players should be a collaborative process, with the parties sharing the information needed for informed choices.

3. **What events and intervals are relevant?**

The sheer volume of ESI necessitates seeking sensible ways to isolate relevant information. Because the creation, modification, and access dates of electronic documents tend to be tracked, focusing on time periods and particular events helps identify relevant ESI, but only if you understand what the dates signify and when you can or can't rely on them. The Created Date of a document doesn't necessarily equate to when it was written. Neither does "accessed" always mean "used." For ESI, the "last modified" date tends to be the most reliable.

4. **When do preservation duties begin and end?**

The parties should seek common ground concerning when the preservation duty attached and whether there is a preservation duty going forward. The preservation obligation generally begins

with an expectation of litigation, but the facts and issues dictate if there is a going forward obligation to preserve throughout the course of the litigation. Sometimes, events like plant explosions or corporate implosions define the endpoint for preservation, whereas a continuing tort or loss may require periodic preservation for months or years after the suit is filed. Even when a defendant's preservation duty is fixed, a claimant's ongoing damages may necessitate ongoing preservation.

5. What data are at greatest risk of alteration or destruction?

ESI is both tenacious and fragile. It's hard to obliterate but easy to corrupt. Once lost or corrupted, ESI can be very costly or impossible to reconstruct. Focus first on fragile data, like storage media slated for reuse or messaging subject to automatic deletion, and insure its preservation. Address backup tape rotation intervals, disposal of legacy systems (e.g., obsolete systems headed for the junk heap), and re-tasking of machines associated with new and departing employees or replacement of aging hardware.

6. What steps have been or will be taken to preserve ESI?

Sadly, there are dinosaurs extant who believe all they have to reveal about ESI preservation is, "We're doing what the law and the Rules require." But that's a risky tack, courting spoliation liability by denying you an opportunity to address problems before irreparable loss. More enlightened litigants see that reasonable disclosures serve to insulate them from sanctions for preservation errors.

7. What nonparties hold information that must be preserved?

ESI may reside with former employees, attorneys, agents, accountants, outside directors, Internet service providers, contractors, Cloud service providers, family members and other nonparties. Some of these non-parties may retain copies of information discarded by a party. Absent a reminder, litigants may focus on their own data stores and fail to take steps to preserve and produce data held by others over whom they have rights of direction or control.

8. What data require forensically sound preservation?

"Forensically sound" preservation of electronic media preserves, in a reliable and authenticable manner, an exact copy of all active and residual data, including remnants of deleted data residing in unallocated clusters and slack space. When there are issues of data loss, destruction, alteration or theft, or when a computer is an instrumentality of loss or injury, computer forensics and attendant specialized preservation techniques may be required. Though skilled forensic *examination* can be expensive, forensically-sound *preservation* can cost less than $500 per

system. So talk about the need for such efforts, and if your opponent won't undertake them, consider whether you should force forensic preservation, even if you must bear the cost.

9. What metadata are relevant, and how will it be preserved, extracted and produced?

Metadata is evidence, typically stored electronically, that describes the characteristics, origins, usage and validity of other electronic evidence. There are all kinds of metadata found in various places in different forms. Some is supplied by the user, and some is created by the system. Some is crucial evidence, and some is just digital clutter. You will never face the question of whether a file has metadata—all active files do. Instead, the issues are what *kinds* of metadata exist, *where* it resides and whether it's potentially *relevant* such that it must be preserved and produced. Understanding the difference—knowing what metadata exists and what evidentiary significance it holds--is an essential skill for attorneys dealing with electronic discovery.

The most important distinction is between *application metadata* and *system metadata.* The former is used by an application like Microsoft Word to embed tracked changes and commentary. Unless redacted, this data accompanies native production (that is, production in the form in which a file was created, used and stored by its associated application); but for imaged production, you'll need to ensure that application metadata is made visible before imaging or furnished in a useful form via a separate container called a "load file."

System metadata is information like a file's name, size, location, and modification date that a computer's file system uses to track and deploy stored data. Unlike application metadata, computers store system metadata outside the file. It's information essential to searching and sorting voluminous data and therefore it should be routinely preserved and produced.

Try to get your opponent to agree on the metadata fields to be preserved and produced, and be sure your opponent understands the ways in which improper examination and collection methods corrupt metadata values. Also discuss how the parties will approach the redaction of metadata holding privileged content.

10. What are the parties' data retention policies and practices?

A retention policy might fairly be called a destruction plan, and there's always a gap—sometimes a chasm—between an ESI retention policy and reality. The more onerous the policy, the greater ingenuity employees bring to its evasion to hang on to their e-mail and documents. Consequently, you can't trust a statement that ESI doesn't exist simply because a policy says it *should* be gone.

Telling examples are e-mail and backup tapes. When a corporate e-mail system imposes an onerous purge policy, employees find ways to store messages on, e.g., local hard drives, thumb

drives and personal accounts. Gone from the e-mail server rarely means gone for good. Moreover, even companies that are diligent about rotating their backup tapes and that regularly overwrite old contents with new may retain complete sets of backup tapes at regular intervals. They also fail to discard obsolete tape formats when they adopt newer formats.

To meet their discovery obligations, the defendant may need to modify or suspend certain data retention practices. Discuss what they are doing and whether they will, as needed, agree to pull tapes from rotation or modify purge settings.

11. Are there legacy systems to be addressed?

Computers and servers tend to stick around even if they've fallen off the organization's radar. That old laptop in someone's drawer can serve as a time tunnel back to evidence thought long gone. You should discuss whether potentially relevant legacy systems exist and how they will be identified and processed. Likewise, you may need to address what happens when a key custodian departs. Will the system be re-assigned, and if so, what steps will be taken to preserve potentially relevant ESI?

12. What are the current and prior e-mail applications?

E-mail systems are Grand Central Station for ESI. Understanding the current e-mail system and other systems used in the relevant past is key to understanding where evidence resides and how it can be identified and preserved. On-premise corporate e-mail systems tend to split between the predominant Microsoft Exchange Server software tied to the Microsoft Outlook e-mail client on user's machines and the less-encountered Lotus' Domino mail server accessed by the Lotus Notes e-mail client application. Increasingly, companies dispense with maintaining physical systems altogether and deploy their e-mail systems online, "in the cloud." Many companies now use Microsoft Office 365 and its virtualized version of the Exchange Server. A changeover from an old system to a new system, or even from an old e-mail client to a new one, can result in a large volume of "orphaned" e-mail on media that would not otherwise be ripe for search.

13. Are personal e-mail accounts and computer systems involved?

Those who work from home, out on the road or from abroad may use personal e-mail accounts for business or store relevant ESI on their home or laptop machines or other portable devices. Parties should address the potential for relevant ESI to reside on personal and portable machines and devices and agree upon steps to be taken to preserve and produce that data.

14. What electronic formats are common and in what anticipated volumes?

Making the right choices about how to preserve, search, produce and review ESI depends upon the forms and volume of data. Producing a Word document as a TIFF image may be acceptable where producing a native voice mail format as a TIFF is inconceivable. It's difficult to designate suitable forms for production of ESI when you don't know its native forms. Moreover, the tool you'll employ to review millions of e-mails is likely much different than the tool you'll use for thousands. If your opponent has no idea how much data they have or the forms it takes, encourage or compel them to use sampling of representative custodians to perform a "data biopsy" and gain insight into their collection.

15. How will we handle social networking, instant messaging and other challenging ESI?

Producing parties routinely ignore short-lived electronic evidence like social networking posts and instant messaging by acting too late to preserve it or deciding that the retention burden outweighs any benefit. *When it's relevant,* will the other side archive texts, voice mail messages, social networking content, mobile device application content or a host of other potentially relevant ESI that's often overlooked?

16. What relevant databases exist and how will their contents be discovered?

From R&D to HR and from finance to the factory floor, businesses run on databases. When they hold relevant evidence, you'll need to know the platform (e.g., SQL, Oracle, SAP) and how the data's structured (its "schema") before proposing sensible ways to preserve and produce it. Options include generating standard reports, running agreed queries, exporting relevant data to standard delimited formats or even (in the very rare case) mirroring the entire contents to a functional environment.

Database discovery is challenging and contentious, so know what you need and articulate why and how you need it. Be prepared to propose reasonable solutions that won't unduly disrupt operations.

17. Will paper documents be scanned, with what resolution, OCR and metadata?

Paper is still with us and ideally joins the deluge of ESI in ways that make it electronically searchable. Though parties are not obliged to convert paper to electronic forms, they commonly do so by scanning, coding and use of Optical Character Recognition (OCR). You'll want to insure that paper documents are scanned so as to be legible and suited to OCR and are accompanied by information about their source (custodian, location, container, etc.) and logical unitization (i.e., foldering and stapled and clipped groupings).

18. Are there privilege issues unique to ESI?

Discussing privilege at meet and confer entails more than just agreeing to return items that slip through the net via so-called "clawback agreements" or a Federal Rules of Evidence Rule 502 agreement or order. It's important to surface practices that overreach. If the other side uses keywords to sidetrack potentially privileged ESI, are search terms absurdly overbroad? Simply because a document has the word "law" or "legal" in it or was copied to someone in the legal department doesn't justify its languishing in privilege purgatory. When automated mechanisms replace professional judgment concerning the privileged character of ESI, those mechanisms must be closely scrutinized and challenged when flawed.

Asserting privilege is a *privilege* that should be narrowly construed to protect either genuinely confidential communications exchanged for the purpose of seeking or receiving legal counsel or the thinking and strategy of counsel. Moreover, even documents with privileged content may contain non-privileged material that should be parsed and produced. All the messages in a long thread aren't necessarily privileged because a lawyer got copied on the last one.[48]

Electronic evidence presents unique privilege issues for litigants, in part because of the potential for application metadata (like documents comments and other collaboration features) to serve as communication tools. Comments and Tracked Changes aren't fundamentally different from e-mails discussing suggested amendments to documents, yet the former tend not to be reviewed or produced by defendants. Instead, some parties will, e.g., convert Word documents to TIFF images, suppressing the embedded communications as if they never occurred so as to avoid having to review them for privilege. If these communications exist and may be relevant, you must work to insure this evidence is not ignored.

19. What search techniques will be used to identify responsive or privileged ESI?

Transparency of process is vitally important with respect to the mechanisms of automated search and filtering employed to identify or exclude information, yet opponents may resist sharing these details, characterizing it as work product. The terms and techniques facilitating an attorney's assessment of a case are protected, but search and filtering mechanisms that effectively eliminate the exercise of attorney judgment by excluding data as irrelevant should be disclosed so that they may be tested and, if flawed, challenged. Likewise, if the producing party uses mechanized search to segregate data as privileged, the requesting party should be made privy to same in case it is inappropriately exclusive, though here, redaction may be appropriate to shield searches tending to reveal privileged information. Finally, use of advanced analytic techniques like predictive

[48] See, e.g., *Muro v. Target Corporation*, 243 F.R.D. 301 (N.D. Ill. June 7, 2007) and *In re Vioxx Products Liability Litigation,* 501 F. Supp. 789 (E.D. La. Sept. 4, 2007)

coding should be thoroughly explored to insure that the processes employed are well-understood and, as feasible, the sampling and thresholds are mutually acceptable.

20. If keyword searching is contemplated, can the parties agree on keywords?

If you've been to Las Vegas, you know Keno, that game where you pick the numbers, and if enough of your picks light up on the board, you win. Keyword searching ESI is like that. The other side has you pick keywords and then goes off somewhere to run them. Later, they tell you they looked through the matches and, sorry, you didn't win. As a consolation prize, you may get the home game: a zillion jumbled images of non-searchable nonsense.

Perhaps because it performs so well in the regimented setting of online legal research, lawyers and judges invest too much confidence in keyword search. It's a seductively simple proposition: pick the words most likely to uniquely appear in responsive documents and then review for relevance and privilege just those documents containing the key words. Thanks to, *e.g.,* misspellings, acronyms, synonyms, IM-speak, noise words, OCR errors, indexing issues and the peculiar industry lexicons, keyword search performs far below most lawyers' expectations, finding perhaps 20% of responsive material on first pass.[49]

Warts and all, keyword search remains the most common method employed to tackle large volumes of ESI, and a method still enjoying considerable favor with courts.

Never allow your opponent to position keyword search as a single shot in the dark. You must be afforded the opportunity to use information gleaned from the first effort or subsequent efforts to narrow and target succeeding searches. The earliest searches are best used to acquaint both sides with the argot of the case. What shorthand references and acronyms did they use? Were products searched by their trade or technical names?"

Collaborating on search terms is optimum, but a requesting party must be wary of an opponent who, despite enjoying superior access to and understanding of its own business data, abdicates its obligation to identify responsive information. Beware of an invitation to "give us your search terms" if the plan is to review only documents "hit" by your terms and ignore the rest. Also, insure that terms are tested on representative samples of ESI to insure that search tools and queries are performing as expected. Be especially wary of stop word exclusions and documents whose textual content was not extracted and indexed.

[49] See, *e.g.*, The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (2007) (describing the famous Blair and Maron study, which demonstrated the significant gap between the assumptions of lawyers that they would find 75% of the total universe of relevant documents, versus the reality that they had in fact found only 20% of the total relevant documents in a 40,000 document collection).

21. How will deduplication be handled, and will data be re-populated for production?

ESI, especially e-mail, is characterized by enormous repetition. A message may appear in the mail boxes of thousands of custodians or be replicated dozens or hundreds of times through periodic backup. Deduplication is the process by which identical items are reduced to a single instance for purposes of review. Deduplication can be *vertical*, meaning the elimination of duplicates within a single custodian's collection, or *horizontal*, where identical items of multiple custodians are reduced to single instances.

Depending upon the review platform you employ, if production will be made on a custodial basis (person-by-person), it may be desirable to request re-population of content deduplicated horizontally so each custodian's collection is complete. This will re-inject duplicates; however, each custodian's collection will be complete, witness-by-witness.

22. What forms of production are offered or sought?

Notably, the 2006 Federal Rules amendments gave requesting parties the right to designate the form or forms in which ESI is to be produced. A responding party may object to producing the designated form or forms, but if the parties don't subsequently agree and the court doesn't order the use of particular forms, the responding party must produce ESI as it is ordinarily maintained or in a form that is reasonably usable. Moreover, responding parties may not simply dump undesignated forms on the requesting party, but must disclose the other forms before making production so as to afford the requesting party the opportunity to ask the court to compel production in the designated form or forms.[50]

Options for forms of production include native file format, near-native forms (e.g., individual e-mail messages in MSG or EML formats), imaged production (PDF or, more commonly, TIFF images accompanied by load files containing searchable text and metadata) and even paper printouts for very small collections. It is not necessary—and rarely advisable—to employ a single form of production for all items; instead, tailor the form to the data in a *hybrid* production. TIFF and load files may suffice for simple textual content like e-mail without attachments or word processed documents, but native forms are best for spreadsheets, documents with pertinent application metadata (comments and tracked changes) and social media content. Native forms are essential for rich media, like animated PowerPoint presentations or audio and video files. Quasi-native forms are well-suited to e-mail and database exports.

A requesting party uncertain of what he needs plays into the other side's hands. You must be able to articulate both what you seek *and the form in which you seek it*. The native forms of ESI dictate

[50] Fed. R. Civ. P. 34(b)

the optimum forms for its production, but rarely is there just one option. The alternatives entail tradeoffs, typically sacrificing utility or searchability of electronic information to make it function more like paper documents. Before asking for anything, know how you'll house, review and use it. That means "know your review platform."[51] That is, know the needs and capabilities of the applications or tools you'll employ to index, sort, search and access electronic evidence.

Finally, don't let your opponent confuse the medium of production with the form of production. Telling you that the data is coming on a thumb drive tells you nothing about what data you're getting.

23. How will you handle redaction of privileged or confidential content?

Defendants often seek to redact ESI in the way they once redacted paper documents: by blacking out text. To make that possible, ESI are converted to non-searchable TIFF images in a process that destroys electronic searchability. So after redaction, electronic searchability must be restored by using OCR to extract text from the TIFF image.

A TIFF-OCR redaction method works reasonably well for text documents, but it fails miserably applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can't spell check numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet

[51] If a question about your review platform gives you that deer-in-headlights look, you're probably not ready for meet and confer. Even if you're determined to look at every page of every item they produce, you'll still need a system to view, search and manage electronic information. If you wait until the data start rolling in to pick your platform, you're likely to get ESI in forms you can't use, meaning you'll have to expend time and money to convert them. Knowing your intended platform allows you to designate proper load file formats and determine if you can handle native production.

Choosing the right review platform for your practice requires understanding your work flow, your people, the way you'll search ESI and the forms in which the ESI will be produced. *You should not use native applications to review native production in e-discovery.* Instead, a platform geared to review of ESI in native formats--one able to open the various types of data received without corrupting its content or metadata--should be employed. ESI can be like Russian nesting dolls in that a compressed backup file (.BKF) may hold an encrypted Outlook e-mail container (.PST) that houses a message transmitting a compressed archive (.ZIP) attachment containing an Adobe portable document (.PDF). Clearly, a review platform needs to be able to access the textual content of compressed and proprietary formats and drill down or "recurse" through all the nested levels.

There are many review platforms on the market, including the familiar Concordance and Summation applications, Internet-accessible hosted review environments like Relativity or iConect, and proprietary platforms marketed by e-discovery service providers touting more bells and whistles than a Mardi Gras parade.

Review platforms can be cost-prohibitive for some practitioners. If you don't currently have one in-house, your case may warrant hiring a vendor offering a hosted platform suited to the ESI. When tight budgets make even that infeasible, employ whatever productivity tools you can cobble together on a shoestring. You may have to forego the richer content of native production in favor of paper-like forms such as Tagged Image File Format (TIFF) images because you can view them in a web browser.

to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.

For common productivity applications like Adobe Acrobat and Microsoft Office, it's increasingly feasible and cost-effective to redact natively so as to preserve the integrity and searchability of evidence; consequently, where it's important to preserve the integrity and searchability of redacted documents, you should determine what redaction methods are contemplated and seek to agree upon methods best suited to the task. At all events, redaction tends to implicate a relatively small population of information items in a production; so, don't let the preferred method of redaction adversely impact the form or forms of production employed for items not requiring redaction. That is, *don't let the redaction tail wag the production dog.*

24. Will load files accompany document images, and how will they be populated?

Converting ESI to TIFF images strips the evidence of its electronic searchability and metadata. Accordingly, load files accompany TIFF image productions to hold searchable text and selected metadata. Load files are constructed of delimited text, meaning that values in each row of data follow a rigid sequence and are separated by characters like commas, tabs or quotation marks. Using load files entails negotiating their organization or specifying the content and the use of a structure geared to review software such as Summation, Concordance, Ringtail or Relativity.

25. How will the parties approach file naming and Bates numbering?

It's common for file names to change to facilitate unique identification when ESI is processed for review and production. Assigned names may reflect, *e.g.*, unique values derived from a data fingerprinting process called hashing or contain sequential control numbers tied to a project management database. Native productions don't lend themselves to conventional paged formats, so aren't suited to embossed Bates numbering on a page-by-page basis; however, this is no impediment to native production in that Bates numbers can serve as filenames for native files, with page numbers embossed on the items only when converted to paged formats for use in proceedings.

26. What ESI will be claimed as not reasonably accessible, and on what bases?

Pursuant to Rule 26(b)(2)(B) of the Federal Rules of Civil Procedure, a litigant must show good cause to discover ESI that is "not reasonably accessible," but the burden of proving a claim of inaccessibility lies with the party resisting discovery. So, it's important that your opponent identify

the ESI it claims is not reasonably accessible and furnish sufficient information about that claim to enable you to gauge its merit.

The meet and confer is an opportune time to resolve inaccessibility claims without court intervention—to work out sampling protocols, cost sharing and filtering strategies—or when agreements can't be reached, at least secure commitments that the disputed data will be preserved long enough to permit the court to resolve issues.

27. Can costs be minimized by shared providers, neutral experts or special masters?

Significant savings may flow from sharing costs of e-discovery service providers and online repositories, or by eliminating dueling experts in favor of a single neutral expert for thorny e-discovery issues or computer forensics. Additionally, referral of issues to a well-qualified ESI Special Master can afford the parties speedier resolution and more deliberate assessment of technical issues than a busy docket allows.

Endgame: Transparency of Process and Cooperation

Courts and commentators uniformly cite the necessity for transparency and cooperation in electronic discovery, but old habits die hard. Too many treat meet and confer as a perfunctory exercise, reluctant to offer a peek behind the curtain. Some are paying dearly for their intransigence, sanctioned for obstructive conduct or condemned to spend obscene sums chasing data that might never have been sought had there been communication and candor. Others are paying attention and have begun to understand that candor and cooperation in e-discovery isn't a sign of weakness, but a hallmark of professionalism.

The outsize cost and complexity of e-discovery will diminish as electronic records management improves and ESI procedures become standardized, but the meet and confer process is likely to endure and grow within federal and state procedure. Accordingly, learning to navigate meet and confer—to consistently ask the right questions and be ready with the right answers—is an essential advocacy skill.

About the Author

CRAIG BALL
ESI Special Master and Attorney
Computer Forensic Examiner
Author and Educator

3251 Laurel St.
New Orleans, LA 70115

Tel: 713-320-6066
E-mail: craig@ball.net
Web: www.craigball.com
Blog: ballinyourcourt.com

Craig Ball is a board-certified Texas trial lawyer, certified computer forensic examiner, law professor and electronic evidence expert He's dedicated his career to teaching the bench and bar about forensic technology and trial tactics. After decades trying lawsuits, Craig limits his practice to service as a court-appointed special master and consultant in computer forensics and e-discovery. A prolific contributor to educational programs worldwide--having delivered nearly 2,000 presentations and papers--Craig's articles on forensic technology and electronic discovery frequently appear in the national media. For nine years, he wrote the award-winning column on computer forensics and e-discovery for American Lawyer Media called "Ball in your Court." Craig Ball has served as the Special Master or testifying expert on computer forensics and electronic discovery in some of the most challenging, front page cases in the U.S. (e.g., Enron, Madoff, In re: Seroquel, etc.).

EDUCATION
Rice University (B.A., 1979, triple major); University of Texas (J.D., with honors, 1982); Oregon State University (Computer Forensics certification, 2003); EnCase Intermediate Reporting and Analysis Course (Guidance Software 2004); WinHex Forensics Certification Course (X-Ways Software Technology 2005); Certified Data Recovery Specialist (Forensic Strategy Services 2009); Nuix Certified E-Discovery Specialist (2014); numerous other classes on computer forensics and electronic discovery.

SELECTED PROFESSIONAL ACTIVITIES
Law Offices of Craig D. Ball, P.C.; Licensed in Texas since 1982.
Board Certified in Personal Injury Trial Law by the Texas Board of Legal Specialization 1988-2018
Certified Computer Forensic Examiner, Oregon State University and NTI
Certified Computer Examiner (CCE), International Society of Forensic Computer Examiners
Certified Data Recovery Specialist
Certified E-Discovery Specialist (Nuix)
Faculty, University of Texas School of Law, Adjunct Professor teaching Electronic Discovery & Digital Evidence
Faculty and Founder, Georgetown University Law Center, E-Discovery Training Academy
Admitted to practice U.S. Court of Appeals, Fifth Circuit; U.S.D.C., Southern, Northern and Western Districts of Texas.
Board Member, Georgetown University Law Center Advanced E-Discovery Institute and E-Discovery Academy
Board Member, International Society of Forensic Computer Examiners *(agency certifying computer forensic examiners)*
Member, Sedona Conference WG1 on Electronic Document Retention and Production
Member, Maryland Committee on Federal E-Discovery Guidelines, 2014- (civil and criminal committees)
Special Master, Electronic Discovery, numerous federal and state tribunals
Instructor in Computer Forensics and Electronic Discovery, United States Department of Justice
Lecturer/Author on Electronic Discovery for Federal Judicial Center and Texas Office of the Attorney General
Instructor, HTCIA Annual 2010, 2011 Cybercrime Summit, 2006, 2007; SANS Instructor 2009, PFIC 2010, CEIC 2011, 2012
Special Prosecutor, Texas Commission for Lawyer Discipline, 1995-96

Council Member, Computer and Technology Section of the State Bar of Texas, 2003-date; Chair 2015-2016
Chairman: Technology Advisory Committee, State Bar of Texas, 2000-02
President, Houston Trial Lawyers Association (2000-01); President, Houston Trial Lawyers Foundation (2001-02)
Director, Texas Trial Lawyers Association (1995-2003); Chairman, Technology Task Force (1995-97)
Member, High Technology Crime Investigation Association and International Information Systems Forensics Assn.
Member, Texas State Bar College
Member, Continuing Legal Education Comm., 2000-04, Civil Pattern Jury Charge Comm., 1983-94, State Bar of Texas
Life Fellow, Texas and Houston Bar Foundations
Adjunct Professor, South Texas College of Law, 1983-88

Recipient of Lifetime Achievement Awards from the State Bar of Texas Computer and Technology Section (2006) and the Association of Certified E-Discovery Specialists (2016); LTN Consultant of the Year, 2009

Selected Publications available at www.craigball.com