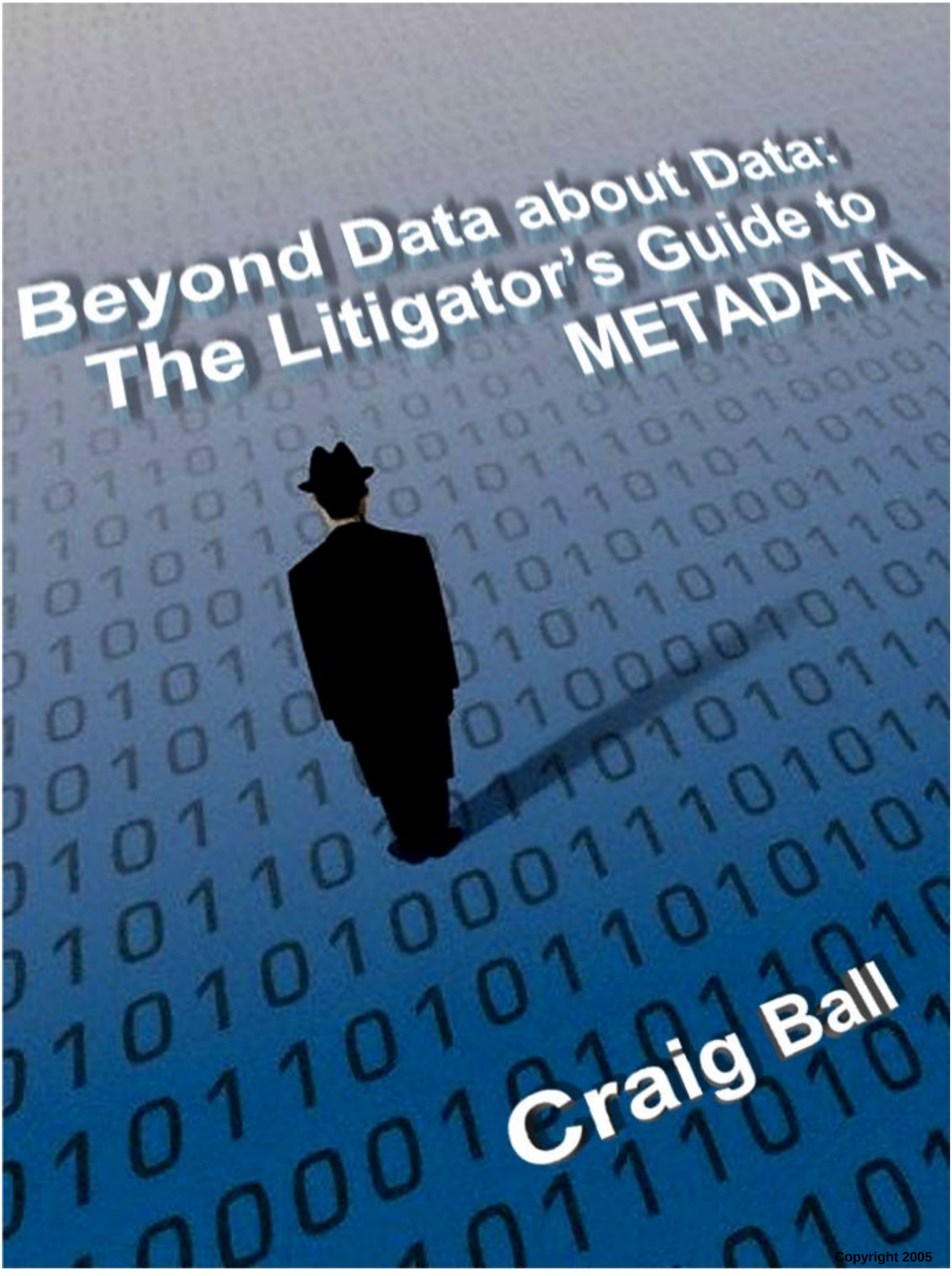

Beyond Data about Data: The Litigator's Guide to Metadata

By Craig Ball

In the old joke, a balloonist descends to ask directions, calling out, "Where am I?" A man on the ground yells back, "You're in a hot air balloon about a hundred feet above the ground." When the frustrated balloonist replies, "Thanks for nothing, Counselor," the man on the ground says, "Hey, how did you know I'm a lawyer?" "Simple," says the balloonist, "your answer was 100% accurate and totally useless."

It's time to get beyond defining metadata as "data about data."

Ask an electronic evidence expert, "What's metadata?" and there's a good chance you'll hear, "Metadata is data about data"--another answer that's 100% accurate, and totally useless!

Perhaps it's more helpful to say that, "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 created by the system. Some is crucial evidence and some just digital clutter. Understanding the difference--knowing what metadata exists and what evidentiary significance it holds--is an essential skill for attorneys dealing with electronic discovery.

It's time to move beyond defining metadata as "data about data" and establish labels and classifications that better describe and distinguish metadata in ways that allow lawyers to better assess relevance and accessibility.

Why Should You Care About Metadata?

In ***Williams v. Sprint/United Mgmt Co.***, 2005 WL 2401626 (D. Kan. Sept. 29, 2005), the Federal court ruled that "when 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."

Our ability to advise a client about how to find, preserve and produce metadata, or to object to production, discuss or forge agreements about metadata, hinges upon how well we understand metadata.

Metadata is discoverable evidence that our clients are obliged to preserve and produce. Metadata sheds light on the origins, context, authenticity, reliability and distribution of electronic evidence, as well as providing clues to human behavior. It's the electronic equivalent of DNA, ballistics and fingerprint evidence, with a comparable power to exonerate and incriminate.

Finally, we need to care about metadata because it's some of the most fragile electronic evidence around. It's a short trip from mishandled metadata to spoliation sanctions.

Misunderstood Metadata

To the extent lawyers have heard of metadata at all, it's likely in the context of its potential to reveal confidential or privileged information hidden within electronic documents. The oft-cited culprit is Microsoft Word, and a cottage industry has grown up offering utilities to strip embedded information from Office applications. Because of the potential to embarrass lawyers—or worse—metadata has acquired an unsavory reputation amongst the bar. But metadata is much more than simply the ***application* metadata** that affords those who know how to find it to dredge up a document's secrets. That's just one species of metadata.

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. Contrast application metadata with ***system* metadata**, which is *not* embedded within the file it describes but stored externally and used by the computer's file system to track file locations and store demographics about each file's name, size, creation, modification and usage. Having both embedded application metadata and external system metadata is advantageous because, when metadata is stored both within and without a file, *discrepancies* between the metadata can expose data tampering.

Like all data, embedded application metadata is just a sequence of ones and zeroes and, in that respect, no less "accessible" than any other data. Accessibility is a measure of an application's ability to convert those ones and zeroes into intelligible information. A programmer configures applications to display selected information—but not necessarily *all* information—by default. Information not displayed by default may be accessible by reconfiguring the program's default settings (such as when a user sets a spreadsheet program to display formulae instead of calculated values). Viewing other embedded data may require drilling down through application menus, such as when a user explores file properties for a Microsoft Office document. These properties are at hand and comprehensible, but tend not to lend themselves to easy printing. None of this is surreptitious data—it's there if the user elects to review it. In fact, despite the common practice to call metadata "hidden," the only application metadata to warrant that description is the information the program employs internally to track, replicate or manage its actions. This data is, indeed, not readily accessible to the user via the program's menus and user-configurable settings, instead requiring specialized computer forensic tools and expertise to extract and interpret.

Every active file stored on a computer has at least one corresponding external block of system metadata—every one, no exceptions. Files may also have multiple associated metadata blocks as well as embedded metadata fields. 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.

Every active file stored on a computer has at least one corresponding external block of system metadata—every one, *no exceptions*.

Is Metadata Just the Unprintable Information?

Some commentators mistakenly characterize metadata as the part of a file that "doesn't print out." While that flawed definition might have some merit if all files were Microsoft Word

documents, it's far afield as applied it to other applications and formats. Consider a spreadsheet. The user enters formulae in various cells to produce calculated values. The information keyed in by the user is certainly not metadata. It's the data. Yet, the formulae typically do not "print out." Or consider voicemail, which exists as both the recorded digitized sound of the message and the textual or encoded information describing the time and date of the call, mailbox identifier, etc. Though the sound doesn't print out, it's the *data*, whereas the dates and times may print but are the *metadata*.

The ability to print metadata varies within applications. It's often a simple task to display metadata onscreen—such as by a review of a document's "properties"—and not too difficult to print out. Though printable, it's metadata. As graphical user interfaces proliferate and applications become multimedia, computer data stray farther and farther from printable information. Sound and animation don't print at all; consequently, the animated PowerPoint, MPEG training video, even the Flash web page, fail "the data is what prints out" test. Moreover, data is also increasingly three-dimensional. Excel spreadsheets and database files, as well as hyperlinked documents, structure data not just across the X- and Y-axes of the printed page but "into" the page as well, layering "invisible" or unprintable linked values and sites, formulae, pivot tables and sounds on a Z-axis beneath onscreen information.

Metadata from the Ground Up

The best way to understand metadata from the ground up is to start with the fundamental building block of computerized information: the binary digit or "bit." Perhaps you already know that all computer data exists as a series of ones and zeroes, but have you stopped to consider what that really *means*? What is the consequence of constructing *everything* stored on digital devices from just two signals, on and off? 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, line feeds, 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 a *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 a 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 a framework for parsing and interpreting the data stream.

Sometimes metadata is elemental, like the contents of a computer's master file table detailing where the sequences of one and zeroes for particular files begin and end. This is metadata altogether 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, each time a user sought to access a file or program, the operating system

would be either wholly unable to find it or required to examine every stored byte. 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 the date a file is created, last accessed or last modified might be expendable, but makes it much easier to manage important functions like system back ups. Likewise, metadata indicating who has access privileges to particular files is unimportant to the user, but a network administrator would be hard-pressed to run a secure network without it.

As we move up the evolutionary ladder for system metadata, some metadata 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.***

Metadata is like the weather reports from distant cities which run in the daily paper. Though only occasionally relevant, you want the information available when you need it. Likewise, metadata preservation should be considered in every case involving digital evidence, even when you're unsure you'll need it.

Much More Metadata

Modern operating systems record a ream of data detailing the creation, use and status of files as well as the use and configuration of associated applications. Windows users see a few these characteristics tracked in the Details view of a folder. By default, only a file's name, size, type and date modified are displayed; however, right click on the column titles in Windows XP and another thirty-four-odd metadata fields can be displayed, including creation date, author and comments. But even this broad swath of metadata is just *part* of the information about the file recorded by the operating system.

Within the Master File Table and index records used by Windows XP 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. Within this 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.

An ironic aspect of Windows is that the record used to track information about a file may be larger than the file itself!

Relevance

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. If

the origin, use, distribution, destruction or integrity of electronic evidence is at issue, the "digital DNA" of metadata is essential evidence that needs to be preserved and produced.

Does this then mean that every computer system and data device in every case must be forensically imaged and analyzed by experts? Certainly 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 is expected to collect certain 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 officer to collect a bite mark impression, cheek swab or shoe print from the driver; but make the matter a murder investigation, and the investigator is far more interested in a DNA sample than a driver's license number. The crucial factor isn't burden. It's *relevance*, assessed by those with the knowledge and experience to recognize and gauge relevance

There are easily accessible, frequently valuable metadata that, like the information collected by the traffic cop, we should expect to preserve routinely. Examples of these might be originating path and filename, and origination MAC (Modified-Accessed-Created) dates for each file. For e-mail, the obligatory metadata might include complete header data, not just the To/From/Date/Subject items culled from the header by the e-mail program. Proper evidence handling entails a sound chain-of-custody, even in civil matters. Metadata functions as the tag attached to evidence in a police property room. The preservation of a file's external system metadata, in particular its name, system origins and dates of creation, last access and modification, is as fundamental to meeting chain-of-custody obligations as Bates numbering or the elements of the business records exception, perhaps more important because metadata is so fluid. Fail to preserve metadata at the earliest opportunity and you may never be able to replicate what was lost.

Fail to preserve metadata at the earliest opportunity and you may never be able to replicate what was lost.

So where do we draw the relevance 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 determination of relevance hinges on the issues before the court. Relevance assessments aren't static, but can change as evidence emerges and issues arise. Metadata irrelevant at the start of a case may be decisive when allegations of data tampering or spoliation enter the fray. A producing party must periodically re-assess the adequacy of preservation and production and act to meet changed circumstances.

Periodically re-assess the adequacy of preservation and production and act to meet changed circumstances.

The Path to Production of Metadata

The balance of this paper discusses steps typically taken in shepherding a metadata production effort. Don't look for a recitation of established best practices—those rules are very much in flux—or expect a comprehensive checklist of "Do's" and "Don'ts." Instead, the goal is to introduce challenges unique to discovery of metadata and explore the "Why" behind them. These steps include:

- Gauge spoliation risks before you begin: *Don't peek!*
- 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: *Don't peek!*

German scientist Werner Heisenberg thrilled physicists and philosophy majors alike when he posited that the very act of observing alters the reality being 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 Windows without first employing specialized hardware or software to intercept changes, metadata is altered and prior metadata values are 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?

One solution is to preserve original metadata values *before* examination. This can be achieved using software that archives the source metadata to a table or spreadsheet. Then, if an examination results in a corruption of metadata, the original values can be established. Another approach is to conduct the examination using only a forensically qualified duplicate of the data. The techniques used to create a forensically qualified image guard against alteration of the original evidence, which remains available and unaltered should original metadata values be needed. A third approach is to use write blocking hardware or software to intercept all changes to the evidence media. Finally--and most commonly—an electronic discovery vendor can harvest and preserve all metadata on read-only media (e.g., a CD-R or DVD-R) or in a hosted environment, permitting examination without metadata corruption.

Identify potential forms of metadata

To preserve metadata and assess its relevance, you have to know it exists. So, for each category of data subject to discovery, assemble a list of associated metadata. 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 field contains, but also its location and its significance.

There are at least eighty accessible application and system metadata fields tracked for a Microsoft Word

The numbers may surprise you. There are at least ***eighty*** easily accessible application and system metadata fields tracked for each Microsoft Word, PowerPoint and Excel document, *excluding* tracked changes, comments and Registry data (though a few are redundant and the majority of them rarely used). 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*. When the other side is conversant in metadata and can expressly identify fields of interest, your job is made simpler. You may not 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.

Relevance is always subjective and is as fluid as the issues in the case. For example, 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 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.

Consider Authentication and Admissibility

Absent indicia of authenticity like signature, handwriting and physical watermarks, how do we establish that electronic evidence is genuine or tie an individual to the creation of 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 reflect a new date each time they are opened, inserted by a word processor macro as a "convenience" to the user. Without metadata, it's often impossible to establish authenticity or establish relevance. 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, evidence of other simultaneous user activity and version control data.

Without metadata, it's often impossible to establish authenticity or establish relevance.

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 particular metadata. As discussed, some metadata fields, like Last Access Date, change the instant the file is opened for review, copied or even checked for viruses. If you don't preserve these fragile fields, you lose the ability to go back to the source

data and extract metadata when needed. Where a preservation duty has attached, by, e.g., issuance of a preservation order or operation of law, the loss of metadata may constitute spoliation subject to sanction.

Often, collection activities required by litigation hold notices themselves serve to corrupt metadata.

How, then, do you avoid spoliation occasioned by review and collection? What methods will preserve the integrity and intelligibility of metadata? Often, collection activities required by litigation hold notices themselves serve to 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 make objection, 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. For example, if only a handful of files are implicated, the simplest and most expedient way to preserve and produce metadata might be to simply record the relevant metadata values by hand. As the number of files increase, you might create a file listing or spreadsheet detailing the original metadata. Even just archiving the files ("zipping") may be a sufficient method to preserve associated metadata. In other cases, forensic imaging or use of vendors specializing in electronic discovery is warranted.

Whatever method is 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.

When copying file metadata, know the limitations of the environment and medium in which you're working. I learned this lesson the hard way several years ago while experimenting with recordable CDs as a means 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! What I l found out was a CD-R isn't formatted in the same manner as magnetic media. Whereas the operating system formats a hard drive to store three distinct MAC dates, CD-R media stores only one date. In a sense, a CD hasn't the "slots" to store all three dates. When the CD media is 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 serves to both corrupt and misrepresent the metadata. Similarly, different operating systems maintain different metadata fields, so, e.g., moving data from a Windows XP environment to a Windows 98 environment results in truncation or loss of metadata.

Plan for Privilege and Production Review

The notion of reviewing metadata for privileged communications may seem odd unless you consider that application metadata potentially contains deleted content and commentary. When the time comes to review metadata for production and privilege, the risks of spoliation faced in harvest may re-appear during review. Consider:

- 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 particular 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. 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.

In ***Williams v. Sprint/United Mgmt Co.***, 2005 WL 2401626 (D.Kan. Sept. 29, 2005), concerns about privileged metadata prompted the defendant to strip out *all* 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, but offered no guidance as to how one might *effect* selective redaction of application metadata.

The court was right to recognize that privileged information need not be produced, but properly distinguished between surgical redaction and blanket excision. One is redaction following examination of content and a reasoned judgment that particular 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 available to the producing party.

None of this obliges a party to use applications that generate substantial metadata or refrain from steps taken to minimize metadata in the creation of electronically stored information. 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 file format, as a database load file, exported to a compatible delimited dataset, in an image format, hosted in an online database or even 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 producing metadata lies in finding ways to preserve the relationship between metadata and the data it describes and, when obliged to do so, present both the data and metadata so as to be electronically searchable.

When a file becomes separated from its metadata, the recipient loses much of the ability to sort and search files. Returning to the voice mail example, unless the sound component of the message (e.g., the WAV file) is paired with the 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 had the producing party dropped the WAV file into an Adobe PDF file as an embedded sound file then inserted 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 or spreadsheet detailing originating metadata values will suffice. On other occasions, only native production or forensically sound imaging will suffice to carry forward relevant metadata. 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 evidentiary value of metadata will only increase as the world moves inexorably to digitization and electronic communications. Already, some 95% of all information is born electronically, the data bound to and defined by its metadata as we are by our DNA. Metadata grows ever more vital in discovery; dictating that we move beyond unhelpful definitions like "data about data," toward an effective vocabulary to describe metadata in its many forms, and toward sensible standards governing its preservation, relevance and production. We also must foster improved electronic records management practices and systems, all the while encouraging those who design and build operating systems and software to offer products better suited, not just to litigation—at best, a tertiary concern for them—but to the goals of data verifiability, portability, integrity, stability and eradication shared by businesses and litigators alike.

Already, some 95% of all information is born electronically, the data bound to and defined by its metadata as we are by our DNA.

Craig Ball is a board certified Texas trial lawyer and computer forensics expert who serves as a court-appointed special master and consultant in computer forensics and electronic data discovery. He can be contacted as craig@ball.net, by telephone at 936-582-5040 or via the website www.craigball.com.