Fordham Law Review

Volume 73 | Issue 1 Article 2

2004

Technical Aspects of Document Production and E-Discovery

Joan E. Feldman

George J. Socha, Jr.

Kenneth J. Withers

Technical Aspects of Document Production and E-Discovery

Cover Page Footnote

Founder, Computer Forensics, Inc. *Founder, Socha Consulting, LLC *Senior Judicial Education Attorney, Federal Judicial Center.

PANEL DISCUSSIONS

JUDICIAL CONFERENCE ADVISORY COMMITTEE ON THE FEDERAL RULES OF CIVIL PROCEDURE

CONFERENCE ON ELECTRONIC DISCOVERY*

PANEL ONE: TECHNICAL ASPECTS OF DOCUMENT PRODUCTION AND E-DISCOVERY

PANELISTS

*Joan E. Feldman***

*George J. Socha, Jr.****

Kenneth J. Withers†

KENNETH J. WITHERS: My mission in the next ten minutes or so is to spell out the differences between conventional discovery of paper documents and the emerging world of electronic discovery, discovery of information that is created, stored, or best manipulated and viewed using computers or computer media.

There are differences in degree and there are differences in kind. But first, and probably most important, is a difference in degree that dwarfs all other: the sheer volume of information.

Statistics from the University of California claimed that ninety-two percent of all information being created in the world today is created and stored in digital form on magnetic media—that is, on computers and disks and tapes.[1] George Socha at the end is going to go into a

---

* This Conference was held on February 20-21, 2004, at Fordham University School of Law. The text of the Conference transcripts has been lightly edited. The term "e-discovery" will be used throughout as a shorthand for the discovery of data that is used or stored on, and retrievable from, a computer or other electronic source or platform.

** Founder, Computer Forensics, Inc.

*** Founder, Socha Consulting, LLC.

† Senior Judicial Education Attorney, Federal Judicial Center.

little more detail on what that statistic really means. I simply want to demonstrate a few of the ways that this has occurred.

The fundamental difference between the way people create and communicate information on paper and on computers is that computer data is not tied to any artifact, like a piece of paper or a clay tablet. Computer data is digital, it's a sequence of zeroes and ones, positives and negatives, ons and offs, a stream of energy. When it is transmitted, there is no transmission of a physical object, like a piece of paper, but of energy, which takes patterns from one medium and places them on another, like a computer hard drive or a disk. No physical object is moved.

This replication results in the buildup of massive volumes of data, mostly redundant but often containing subtle changes made by people or automated systems along the way. That is why one printed document that may surface in conventional discovery, if it is for instance a word processed document or the result of some other automated system, may represent hundreds of copies or versions to be found on computers and on network servers and on disks and on tapes.

The fact that data can be sent to the next cubicle or around the world, to one person or to a million people, with the same click of a mouse creates a buildup of data entirely unlike anything that we have seen in human history. But computers have created whole new categories of data that do not have easy comparisons in the paper world.

The first one that I want to mention briefly is metadata. Metadata is a made-up Greek word. Roughly translated, it is information about information. It is essential for the functioning of a computer. It is contained within each computer file. It tells the computer such things as the file's creation date, the location of where it was created, how often it has been edited and on what other computers, and the date and time it was last viewed or altered. Metadata is usually generated automatically, although it can be designed and manipulated by humans.

It is not difficult to view. But computer files themselves may contain data which was neither printed on paper nor viewed on the screen. This is an example of the same word processing document showing the editorial changes that were made, what we call embedded edits. When one looks at the data on a computer hard drive not through the lens of the operating system, which arranges it much like physical documents in a file cabinet, but through the lens of computer forensic software, we see a totally different world. We can see

1. Peter Lyman & Hal R. Varian, Executive Summary, How Much Information?, 2003, *at* http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm (last visited July 19, 2004).

documents that have been supposedly deleted. References to that file, that document, have been removed from the visible operating system, but the data is still present and still intact on the hard drive.

Because of the almost magical nature of digital data, to be transmitted into any medium we have many more places in which data relevant to discovery or an investigation can be found. And also because of the magical ability of digital data to transform itself in the process of attaching to these different media, we have any number of formats in which the data can be found, as though we have to conduct discovery simultaneously in a number of countries and in a number of languages.

While the volume of discovery increases on a macro level with the number of places, the number of formats, and the sheer numbers of documents that need to be looked at, it also increases on a micro level as each electronic file becomes in essence a little database unto itself.

The typical word processing or e-mail file or other electronic file contains, of course, the visible data, the things that one can see if the file is printed out or is shown on a screen. But below that there is another strata: the metadata that we have seen, the formatting commands, the formulas used to create the spreadsheets, and the hidden and embedded edits that may be contained within that file.

And below that there is yet another strata: the bedrock on which that file rests, which is the hard drive or the medium itself, which may contain residual data from past files; it may contain what one of our speakers here, Dan Regard, in the past has called "digital packing peanuts," which is data that is used to fill out the sector on a hard drive. These are the bedrock elements underneath any particular file.

If we are simply looking at paper or the electronic equivalents of paper, what we call .pdf or .tiff images, all we see is the visible file. If we look at data in its native format, in the way that it is kept in the normal course of business and manipulated and used, then we see the second strata, the metadata, the formatting, the formulas, etc. If we take the step of going to on-site inspection of the computer media itself—the computers, the disks, the tapes—or we take what the forensic scientists call a bitstream image or a bit-by-bit copy of the data, then we have the ability to look at the residual data.

Each way we view the computer file reveals a different layer, and this may go to the question of relevance. At what stage does this become irrelevant?

But documents themselves as tangible objects are actually disappearing. Today most commercial, governmental, and even personal communications and information are not reduced to immutable physical objects, like paper. When we conduct discovery, we are actually querying databases to generate selected data which we then arrange and present in a particular way. We are no longer looking at existing objects, like paper and file cabinets.

So the primary focus must be on the relevance of the questions being asked and the efficacy of the process being used to obtain the answers, not on the nature of the physical documents involved.

JOAN FELDMAN: As was pointed out, twenty-five years ago I started out in the paper cut brigade reviewing what was considered to be huge volumes of material, courtesy of another big technological change, the photocopy machine.

We'll fast-forward to twelve years ago, when I began trying to explain to people that a deleted file could be restored and that there were such things as embedded data and so on.

I believe that today we are in the middle of the next revolution in electronic discovery, and it concerns the overwhelming volume of material that we are facing. There has been a lot of focus on this issue, and for good reason.

I like to term it "the tsunami effect." I also know that George when he follows me today will be talking about the fact that this current problem is actually only going to grow and continue to grow.

I would like to talk to you today about what people are doing in the real world to deal with electronic discovery. To that end, again, I want to encourage you to have the mindset that there is an enormous amount of material out there, that it is often difficult to identify where the real value is going to be in going through that material. That is not an idle subject because there is so much information out there and there is so little in a way that actually turns out to be truly responsive.

There has been a big push in applying technology to solve this problem. There have been some amazing developments in tools for sifting through the electronic documents, for acquiring it more easily, for doing text searching and concept searching. All of these are most helpful because we are trying to deal with a tsunami.

I would just like to tell you that in many ways it is like having a snorkel when you are out in the ocean, that it is a good tool, but we are dealing with a huge volume of material and it is growing.

Let me put this in context for you in a real world example. We were recently called upon to be a mediator in a case involving a large Fortune 100 company. The special magistrate was at a stalemate with both parties. At issue was a huge volume of material. A well-respected large litigation support company and a well-respected large law firm had assisted the Fortune 100 company in identifying the documents to preserve and produce. They came up with a total volume of 42,000 backup tapes—that's another issue—and they identified twenty hard drives. For the judges sitting here today, I think that you might be familiar with hearing this type of number brought before you.

What was at issue? Plaintiffs dug their heels in and insisted that 42,000 backup tapes be restored and reviewed. Defendant producing

the documents said, "It's expensive, it's a fishing expedition, it's not going to yield anything."

Between the two of them, they probably spent over $75,000 just on motion practice, and the lucky judge got to hear the debates about what a tape was, what was on the hard drives, and embedded data. At the end of the day, nothing really had been produced, nothing really had been reviewed.

"Break the deadlock," that was our charge. As expert witnesses often do, we gently guided the court, and although our mandate was to actually see if it was really worth the money to look at 42,000 backup tapes, we suggested something else. We suggested that they begin focusing on where the evidence might actually be. "Well, we have 42,000 backup tapes and we have twenty hard drives. How much is it going to cost?"

We gently suggested that they needed to focus on where the evidence might actually be. We applied some techniques that the attorneys sitting here today are familiar with. We questioned witnesses. We went from the point of departure as to where the evidence that they were looking for might actually be stored, in whose hands; what was the evidence—it was a trade secret theft case; who worked on the documents that might actually have something to do with that; and who worked on the product at question.

In one and one-half days of interviews with the systems people and with some of the key witnesses who had actually created some of the documents that we felt would be at issue, we actually located another server that had not been disclosed that had been set up by the engineers, as they often do. We like to refer to engineers as our "rogue" folks because they often set up their own systems. They had established their own system, including their own e-mail server. It's like Mount Everest—you know, it's there.

We located this fact that there had been a server. By the way, in the course of the discovery in a six-month period, they had disabled and mothballed the server. Actually they destroyed the server because they wanted to use it for some other application. They had not been informed by the attorneys or didn't pay attention to it. But, as engineers will often do, they were also packrats and they had created two backup tapes for that server. That is actually where the evidence was. It had been previously unidentified.

What about the 42,000 backup tapes, the subject of much fevered debate about cost and cost sharing? Were they impenetrable? Was it this monolithic dataset that was going to cost at a conservative estimate $4 to $5 million? Through questioning of the IT staff, we were able to find the Rosetta Stone that helped us begin to piece apart that monolithic dataset to identify particular tapes that might actually contain evidence.

Through a closer look at some of those tapes as we began this process, we were able to narrow the 42,000 set to thirty-seven tapes—thirty-seven backup tapes, previously undisclosed data—not as a result of some technological marvel or breakthrough in text-searching technology. Despite the fixation with blue screens as solutions for electronic discovery issues, I would just suggest to you that a good background in technology, an understanding of how enterprises use their computers, and the same principles that guide experienced litigation attorneys and jurists in their decision making process, in terms of finding and refining and looking for responsive information, is critical here.

There is a dynamic tension in my field these days because there is such an emphasis on the ability to process massive amounts of data. That is fine. I am not saying that there thirty-seven backup tapes was a lot of data; it was good to have those tools—but the volume is increasing and we do not necessarily see a corresponding interest in just understanding some of these basic fundamentals.

So that is one issue that we are dealing with and, conversely, all of you are dealing with.

There are a few ways to begin chipping away at these issues:

- You must start at the preservation phase because you are going to have to make some decisions about what needs to be preserved; and if you do not and you are continuing to overwrite tapes or reuse or format hard drives, you are going to destroy critical information. So you have to start there.

- You have to learn how to distinguish what kind of data you are looking for. Are you looking for Word documents? Are you looking for e-mail? Are you looking for database types? These questions need to be answered early on.

- Data elements. Ken did a masterful job of explaining things like metadata and embedded data. These are data elements that you may be concerned with. Or you may not; the parties may make a decision that they are not, they don't care, they just want what is on the face of the documents or that compilation. That's fine, but those decisions have to be made.

- Common terminology needs to be developed between the parties. We suggest the adoption of a glossary of terms and that they agree to it, so that you do not have this shifting target as you move through as to what is a database, what is a relational database, what is a file. You need some basic terminology that you agree upon.

- And you also have to make some decisions even at the earliest stages as to how you are going to produce that information. Mention was made of a .tiff image versus a native file. There is a big distinction there because .tiff images do not contain embedded information; they do not contain the original metadata. When you

are producing those documents you need to have some idea of what it is you are going to be producing to each other and, unfortunately for everybody, you have to make those decisions early on.

MR. SOCHA: Next is the question of more data. The volume of data is expanding rapidly.

Here is a little bit more detail from the 2003 study done by the University of California at Berkeley.[2] That followed up on a 2001 study, I believe, where the authors made at that time what they considered to be outrageous projections as to the growth in the volume of material there in electronic form. In the executive summary to the 2003 report, the authors said that they had no idea, and what they thought was outrageous didn't even come close to what appears to have happened.[3]

And, importantly for this discussion, is the row on magnetic [in the cited data].[4] Now, they are talking here about 4 million terabytes of data. That is a volume that I think none of us can even begin to conceive of. There is nothing like that in paper out there. So we have got this enormous volume of information that we potentially need to deal with. If we keep trying to buy hog farms instead of just ham sandwiches, we are going to be in a lot of trouble.

There are also more types of data and in more places than I think many people really recognize:

- Of course we've got e-mail. There is a lot of discussion about that. That is what captures people's attention. That's the easy picking, though; that's the low-hanging fruit. E-mail is almost like paper in many ways. You can pull it up on the screen, most people now are used to dealing with it, and you can read what is right there.

- Instant messages, though, it is predicted will be equal in volume to e-mail within a couple of years, perhaps sooner. That is a much more difficult medium to deal with for electronic discovery purposes.

- Text messages, such as the ones sent back and forth by cell phone users, are rapidly growing in use.

- Relational databases, while they have been around for a while, have not for the most part been the subject of discovery requests. I think lawyers have avoided going there because relational databases are simply too esoteric, too complicated, too confusing for most people who have not had to deal with them in some other aspect. If I were to bring up on the screen—and I will not do this to you—the plan for a relatively simple relational database, what you would see would be lots of boxes over the screen. It's like the anthropology

2. *See id.*
3. *Id.*
4. *Id.*

> class I had as a freshman in college. The professor put up boxes on different subjects and then started drawing lines to each other. By the time he was done, there were about thirty boxes on the wall and lines from everything to everything.
>
> If you look at that on the screen as a user would, you will see something that is coherent and makes sense, provided they built the relational database properly. If you try to go in without knowing what that database is and without the benefit of that user's experience and expertise, you might find yourself just with gobbledygook. But that is where a huge amount of data is stored these days.
>
> • XML datasets. The word processing document we see today is nothing like what you think it is. It may look like something that just gets printed out on a sheet of paper. There is not just metadata there, though. It may be broken up into all sorts of constituent parts that are not even part of that file but elsewhere. So the information is all over the place.
>
> • Digital photos.

And then, with expansion also comes better processes and tools, some of the stuff Joan was talking about. People are learning how to do this better and how to move forward with it. Well, with expansion then follows routinization. We get used to this stuff. It becomes part of what we are doing. Bigger projects can be done than ever before.

In 1996, I handled what was probably one of the largest tape cases that year, with 461 backup tapes. I had a hard time finding any vendor who could handle that work. The largest backup tape case I know of from last year involved 10,000 tapes. Now, most vendors cannot handle that, but there are some who can.

In the late 1980s we were talking about kilobytes; in 1996, 10 gigabytes was huge; and now 10 terabytes is not unusual. We have to figure out how to deal with that volume because it is only going to get larger.

But with routinization also the impossible becomes possible. We discover that we can do things now that we simply could not handle a few years ago. There are vendors out there offering services that were unimaginable six or seven years ago. The data that was essentially inaccessible not long ago is routinely available now, and that is only going to continue to change and be so moving forward.

And then, finally, with this routinization our expectation level goes up. Because we can do this much, we want to be able to do this much. As we demand more, the people who are providing the services and capabilities turn around and, so far continuously, have been able to offer more to us, which then takes us right back around to expansion.

So looking into the future as best one can at this point, there is an enormous growth in the volume of information we have to deal with, a

growth so far beyond what we are capable of doing that we cannot even really begin to imagine how we are going to be handling this information in a few years, except to know that most of the issues we are dealing with right now are going to at a very technical and detailed level be yesterday's news, at best.