

Discovery and Databases: Understanding the Basics of Structured Data in Litigation (Part I)

By Lauren Allen and Angela Reeves, IE Discovery

During the past decade, the amount of electronic data created by a typical organization has grown at a staggering rate. Most organizations have responded to this trend by managing critical information—for example, payroll, employee, and contract data—in structured environments like databases and spreadsheets. Electronic discovery, however, often focuses mainly on email and electronic documents, despite the fact that structured data sources frequently contain critical evidence.

The reason structured data tends to get neglected during discovery is fairly straightforward: Gathering and analyzing data from structured sources is a complex process that can quickly overwhelm any attorney, including those who are familiar with the principles of electronic data management. The good news is that, given a little guidance, even attorneys with limited technical expertise can conduct a comprehensive and defensible discovery.

What is structured data?

Structured data is a broad concept that encompasses a variety of information environments—from very basic spreadsheets to enterprise-level relational database management systems (RDBMS). While there is a world of difference between a Microsoft Excel spreadsheet and an Oracle database in terms of both complexity and functionality, these systems do share a common premise: At their core, both are formally defined schemas of rows and columns that are designed to promote efficient data storage and retrieval.

How is structured data important to discovery?

Each individual data point in a structured environment holds little meaning by itself; but when pieced together the information often exposes a larger story. To illustrate, consider a fictional company that is sued for age discrimination. Examining the data points of just a few employees—for example, date of birth, date of hire, job history, and performance reviews—probably won't reveal much about the company's motives. Extracting and analyzing data on new all hires and promotions across a specified date range and a variety of job openings, however, may reveal patterns of hiring and promotion that can be used as evidence to resolve the case.

How can I create a comprehensive and accurate structured data collection?

Getting your arms around structured data can be overwhelming, especially in a large organization with a complex network and data architecture. By following these four tips, you can at least feel confident you've covered the basics.

1. Start by reviewing data retention policies and database documentation. If your organization lacks thorough documentation, learn what you can from existing sources and work on improving documentation processes going forward.
2. Conduct extensive interviews. Not surprisingly, your best resource for locating structured data is the people who create and maintain that data. Interview the workers who regularly interact with your organization's data to uncover sources you might otherwise overlook.
3. Get IT involved in the process. The IT workers who maintain your network and data architecture can probably alert you to backend systems that even your in-the-trenches staff doesn't know about. In fact, it's possible they already have a network or data map that details the architecture of existing hardware and data sources-if not, you should consider working with them to create one. Also, since databases are frequently updated, ask your IT staff to regularly archive every database to document the contents as they currently stand-also known as taking a "snapshot" of the database. This effort to preserve data is especially important if you anticipate upcoming litigation.
4. Secure agreement from opposing counsel on the date ranges and types of information required for your specific suit. Then, build your queries based on the constraints of the case, so you extract only the data you need. If the amount of data requiring analysis isn't particularly dynamic, you can export it into a spreadsheet program like Excel. If your structured data collection is large, however, you may need to use a database program with more functionality like Microsoft Access.

What are the primary issues to consider when collecting and producing structured data?

Even the most whole-hearted attempts to compile an exhaustive structured data collection will likely face some challenges. One of the most common-and arguably the most frustrating-is dealing with missing or corrupted data. If you encounter this dilemma, your first course of action should be examining the original source to make sure the data wasn't damaged when it was extracted from the database. If you find that the data is

corrupted at its source, all you can do is piece it together as best you can and have supporting information available to defend the choices you make.

Another challenge you may face is collecting data from outdated systems, often called "legacy" systems. The challenge of legacy systems is the lack of intellectual capital available to support them. Finding experts who are well versed enough to first get the data out of the legacy system and then clearly describe their extraction methods to the court can be both expensive and time consuming, but is necessary to constructing a defensible collection.

A third challenge of structured data sources is that they are often very large, containing millions-and even hundreds of millions-of rows of data. Databases this large are expensive to collect and produce, and they also require greater expertise when it comes time to analyze and clean up data.

A final challenge is that structured data sources frequently contain sensitive data. As is true with documents, the court may be willing to make exceptions on producing competitive or proprietary data, but only if you can clearly explain and defend your need. In some cases, opposing counsel will insist on hiring a neutral third party to analyze sensitive data, thus mitigating potential conflicts of interest while still ensuring the collection includes all relevant data.

What should I do with the structured data I collect?

Once your initial collection is complete, consider engaging an experienced data analyst to determine whether you've collected all the information you were targeting. For example, imagine you are litigating a medical claims case in which a hospital is suing an insurance provider. While you believe you've collected comprehensive data on all the claims and payments in question, further scrutiny by a data analyst might uncover payments that don't specify the individual claims they covered. So, perhaps those claims were paid and perhaps not-you'll have to collect more information before you can be certain. Cases involving complex data often require counsel to spend time on back-and-forth analysis before enough facts are available to answer the questions involved in the case. The time spent on detailed analysis is well justified, however, as it helps shape the most accurate production possible.

On the topic of production, the meet and confer is generally the best time to negotiate format. Excel and Access are both common choices, but sometimes it makes sense to export the raw data into a comma-delimited file instead. In addition to production format,

also consider negotiating a clawback agreement in case sensitive or proprietary data is inadvertently produced.

Whatever format you choose, make sure to conduct a final review of the collection prior to production. This last appraisal should include both experts who understand the native systems from which the data originated and experts who can analyze the data in the context of the individual case. As with the initial back-and-forth analysis, the resources you spend on final review will be well justified by the comprehensiveness of your collection.

Lauren Allen is a Program Manager with IE Discovery, a leading provider of Discovery Management services. She has been employed at IE Discovery since 2001 and Lauren is a licensed attorney with the Commonwealth of Virginia and a certified Project Management Professional. She can be reached at lallen@iediscovery.com.

Angela Reeves is the Manager of the Standards Department for Information Services at IE Discovery, She is responsible for the standardization and oversight of automated processes across all clients. She can be reached at ARees@iediscovery.com.

Originally published in FindLaw's Legal Technology Center, <http://technology.findlaw.com>.

© 2009 IE Discovery