

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

By Belinda Longoria, IE Discovery

Part I of our series on structured data discussed how it is an important part of electronic discovery. We learned how to create a comprehensive and accurate structured data collection; the primary issues to consider when collecting and producing structured data; and what to do with the collected data. In the second part of this series, we will delve deeper into how to make the most of your structured data.

What do I do if my structured data isn't in an electronic format?

Keep in mind that structured data is not always contained in an electronic format-it may exist on printed forms such as W-2 documents, payroll forms, etc. Even though printed data isn't in a readily usable format, you still have options for capturing and utilizing this data.

Manually entering printed data into a database during document review is a common approach. If you decide to go this route, you must take time to understand the data included on the forms so you can implement validation processes that maintain the quality of data you manually capture.

Alternatively, depending on the quality of your documents and the consistency of the printed data, you may be able to use Optical Character Recognition (OCR) software to automatically capture form data and enter it into a structured electronic format. Utilizing an OCR service can significantly reduce the amount of time spent on document review, while also providing the ability to search and analyze form-based data.

How do I analyze my structured data collection?

After you collect your data and examine it for gaps, what's next? If your data originates from multiple sources, your next step is to validate its comprehensiveness and accuracy. An experienced data analyst can help you determine whether your collection includes duplicate, missing, or improperly formatted data by analyzing a subset of the data to identify patterns and exceptions. Once inconsistencies are identified, the analyst will normalize the data until it is as consistent as possible throughout the collection. Also, missing or inconsistent data can sometimes be fixed during document review.

Once your data is as consistent and complete as possible, you can decide which data points are most appropriate for building queries. Consider, for example, an insurance case that requires identifying both individual claimants and the total number of claimants. Individual claimants can be identified by any one of several of data points-member ID, name, home address, date of birth, Social Security number, or phone number. Theoretically, you can query any of these values to determine the total number of claimants. But, just because a data point can be used to identify claimants doesn't mean that it's the best possible choice-missing, incomplete, or changed data may skew your query results.

So, the first question to address in this scenario is which data points are likely to be the most accurate and complete. In other words, are any of the data points required fields that should contain data for every single claimant? Another issue to consider is whether any of the data points are changeable-for example, a home address will change when a claimant moves, possibly rendering that value outdated. Also, does the organization consistently use the same format for assigned values (like the member ID), and have any company mergers or process changes occurred that may have caused the format of assigned values to change? Static values like birth dates and Social Security numbers are often good values on which to base your query, but remember that they can contain typos or inconsistent formatting..

Once you are confident in the comprehensiveness and accuracy of your data, you're ready to combine the distinct data sources into an integrated collection. An experienced data analyst can review your data sources and recommend the most efficient and reliable way of linking them together. You also must determine whether documents should be associated with the data you've captured. For instance, if you have a claim for a certain member, you may want to associate the initial claim form, the member eligibility form, and the check showing payment so that you can easily retrieve these documents when reviewing the source data. A data analyst can provide recommendations on how to best link documents, and also assist you with matching documents to source data.

In addition to relying on a data analyst, you may want to consider using business intelligence tools-for example, an analytical engine. An analytical engine takes your summary-level data and slices and dices it based on your needs. For example, continuing the insurance claim example from above, you could enter monthly payment totals, claim amounts, and allowable or negotiated claim amounts by provider into the analytical engine, which will then aggregate the data to show annual totals, compare individual providers, and even compare one provider's data to another. It could even examine trends

to determine which providers had the highest average of claims in excess of negotiated amounts.

The primary benefit of automated analytical services is that they enable quick examination of data from a variety of different perspectives. The key to using these packages is to determine what kinds of information and analysis you will need to see at a summary level, so you can determine how to most efficiently load your source data.

How do I defend the analysis of my data collection?

As an attorney, you might have some trepidation about defending electronic data. After all, by the time your collection is complete, the data has likely gone through several rounds of processing: first, you had to collect the data from disparate sources, then you had to normalize the varying formats and, finally, you had to integrate it into one unified collection.

Fortunately, the key to defending your analysis is simple- document every step you take with your data, and implement change control processes that validate what you did and allow you to repeat the process as necessary.

Thoroughly documenting data preservation begins with an appropriate chain of custody that tracks the initial amount of data and values collected, including inventory-level information on the files received, data sources, etc. Then, as data is normalized and sources are unified, the documentation should explain in non-technical terms what changes were made and why. Every change you make should relate to your ultimate goal of using the data to defend the case.

Although creating this level of documentation may seem burdensome at first glance, it's really nothing more than an aggregation of your data presented in the context of how individual decisions were made and work was performed. The end result is a map of what data came in, how that data was processed and why, and what the result of the processing was, so the data input can be reliably compared to the final output.

The ultimate test of effective data analysis is repeatability. In other words, if you can repeat your process and get the same results, then you know you have a good understanding of the data and you'll be able to defend your analysis as needed. That's why experienced analysts take measures to track and document analytic processes with appropriate change control measures-effective change control enables you to explain exactly what processes were performed on what data, whether the processes changed at

any point and why, what processes were used for each step and in which order, and how the processes can be recreated to produce the same results.

How should I present/produce relevant data?

When deciding how to present relevant data, remember that you have two audiences: your internal organization and opposing counsel. While your internal organization needs data presented in a way that facilitates case review and management, production to external parties carries different considerations. Ultimately, you must determine which data points are relevant for each audience. You also must consider what functions your internal team requires, such as whether they will need to perform intensive data searches and, if so, what data and time periods they might need to search.

The most basic way to present data is in a data file. Perhaps the most commonly used data file format is Excel, but you also have the option of using plain text files, database files such as Access, or an integrated data storage system.

Some benefits of Excel are that it allows you to easily format data, and it provides sorting, filtering, and basic searching capabilities. A drawback of Excel is that it limits the amount of data that can be included in an individual file; the data limit will vary based on the version of Excel that you use. Excel also provides hyperlink functionality, in case you need to link to associated documents.

Access offers the same benefits as Excel, with some additional search and filtering capabilities and larger space constraints. The drawback of Access is that it can be challenging for non-technical people to use, depending on how the database is structured.

If you plan to export your data to a proprietary data system, a text or comma-delimited file is often your best choice. These file types don't provide embedded formatting, search, or filtering capabilities, however, the data is highly portable. Also, text files don't restrict the amount of data that can be included in an individual file.

A final possibility is to utilize an integrated system that displays custom data along with associated documents. These systems usually provide an efficient way to organize data and documents. They also generally allow you to search for data and documents via custom fields, so you can quickly and efficiently create custom reports to use during litigation. For example, in an integrated system, you could run a custom search for provider payments greater than \$200 from the period of 2/2/1989 - 1/16/2000 and quickly receive that data-along with all the associated claims, member eligibility forms, and

checks-in a simple, organized format. The system may even allow you to enter your own comments and analysis at the data or document level.

When producing data to an external party, you will want to consider a solution that is both simple and cost effective. Often, counsel will agree to produce data in a basic data file-such as Excel or a text file-but you may choose instead to convert data to image files, such as PDF or TIFF, so you can be sure that hidden data, formulas, and other confidential information is not included. Remember, if you are producing data files or databases, you may need to provide data definitions that describe the structure of your files. All of these decisions should be addressed in the meet and confer.

How should I set privileges on my structured data?

During the collection process, you determined what information you needed to collect to support your position, thus beginning the process of excluding potentially privileged information from your collection. For example, you may have determined that because your case only impacts people living in the state of Arizona, you can safely exclude information from all other states. Once you made that decision, you had more detailed decisions to make-will your collection include only those people who currently reside in the state, those who work in the state, anyone who does business in the state, or all of the above?

In the end, your collection should only include the data you need for the case at hand. During the meet and confer, you should have agreed on what fields to produce to opposing counsel and how to address protected information that may be covered by legislation such as HIPPA.

Before you actually produce your structured data to opposing counsel, your final step should be to search for privileged information within the data you are turning over. Just as document searches can be performed to find privileged terms within documents, data queries and searches can be conducted on structured data to identify privileged terms within data. An efficient strategy is to look for data in specific formats, such as standard formats for Social Security numbers and dates. If you do find privileged data, you must determine what you will do with it based on the specifics of your case. Perhaps you'll leave the data blank or replace it with text that indicates a redaction, such as XXX-XX-XXXX for a Social Security number.

Summary

In short, the basic principles for creating an effective and defensible discovery of structured data are the same as those you should follow for documents:

Know what information your internal team requires and what data you must share.

Interview your internal staff to discover all of your potential data sources.

Implement change control practices that ensure you can repeat your processes.

Thoroughly document chain of custody, inventory, and any processes you perform on your structured data.

Organize your data in a format that's usable for both your internal and external audiences.

Determine what data is relevant and what is privileged so that you produce only what is essential.

Belinda Longoria serves as the Data Analysis Supervisor at IE Discovery. She leads the Data Consulting and Data Solutions teams in providing custom programming and data analytic solutions to support litigation clients including Fortune 1000 companies and government agencies. She can be reached at blongoria@iediscovery.com.

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

© 2009 IE Discovery