Deloitte.

Discovery insights

What government agencies should consider when collecting data

An interview with
Patrick McColloch, Managing Director, Deloitte Discovery,
Deloitte Transactions and Business Analytics LLP

It can be a challenge for government agencies to collect electronic data in response to their litigation and investigation matters. The volume of electronic data, as well, as the number of custodians and the far-reaching locations of the custodians can present a challenge during the collection process. Here are five questions regarding the collection process that may help you be more efficient in your collection efforts.

Questions	Pat's take
Why is it important for government agencies to stay flexible during the identification and collection process?	The identification of custodians and potential data sources and data stores is an iterative process. Based on the subject of the complaint or request, you are likely to immediately identify primary sources (e.g. contracts cases would lead you to a Contracting Officer; employment cases lead you to the Human Resources Office). However, during the subsequent interviews with the initial custodians you should always ask open ended questions such as "Who else might have information related to this? What other systems or records group should I search?" The answers to these questions should be fully documented in your interview notes and you should follow through with new potential leads. An iterative processcombined with open ended interview questions will lead to many more data sources and data stores. You should be flexible enough to head down every reasonable trail that presents itself. This becomes particularly important when it comes to protracted dates. Cases tend to drag out over long periods of time, or in addition, can have a start date that goes back some time. Within the government, organizational changes occur over time and the role and responsibilities of departments change. Similarly, individuals change jobs and may even leave an organization through retirement or job rotations. If you are dealing with a matter that dates back 15 years, you are likely going to have to work back through several different custodians and possibly even different organizational structures. You must be prepared to deal with the issue of organizational and employee changes over time. One tool to assist with this challenge is a datamap or a spreadsheet that documents the geographic and chronological scope of the matter, the types of data housed at each location and the custodians in charge of that data over time.
What are some of the strategies to analyze your data for problems and pitfalls once you have collected it?	After you have collected your initial wave of information, it is important to start analyzing the data to help identify collection gaps or additional sources that need to be considered. For example, after emails and other Electronically Stored Information (ESI) have been loaded to a review platform, you can utilize simple analytic features to: · **Review the timeframe of the information you have collected.** Do you have emails for 2009, 2011 and 2012? What happened to the 2010 emails, why are they missing? · **View communication threads.** For example, an individual might be involved in a lot of the email discussions, but she was never identified as a potential custodian. Based upon the volume of traffic and the contents of some of the messages, you might have identified a new source that needs to be investigated. As noted in Question 1, people rotate throughout government positions over time so you should pay attention to custodians that might not have been initially targeted. · **Identify nicknames or other references.** Projects and cases are often given shorter names or other common reference points. Government agencies and their contractors are always using new acronyms and references for initiatives, systems and contracts. Text analytics can help identify these patterns that can help you with additional discovery searches.

Questions	Pat's take
I was expecting less than 5 GB of data, but it has exploded. Now what?	This challenge is clearly not unique to the government but can be exacerbated by some of the issues identified above. When you are faced with a much larger amount of data than you anticipated, there are several techniques you can consider to help manage the volume. A few of these include: First, consider refining the keywords used during the collection process. Additional culling can be implemented after your initial collection to reduce the volume. Run search term hit reports to understand which terms brought back disproportionately larger amounts of data than expected and look at samples of documents to determine if narrower search terms can be applied to bring back only what is relevant. For instance, your initial search term might be "monthly report" but that brings back both hiring/training reports and security activity reports. If you are only interested in the training reports, modify your search term to include training within the text so that it is more likely to only bring back those reports. Alternatively, during this investigation you may find large groups of irrelevant information and can develop search terms that will isolate the irrelevant documents so that you can exclude them from additional processing and analysis. Second, consider using technology assisted review to expand the capacity and throughput of your human review team to review larger amounts of documents with less human and chronological time. Technology assisted review techniques and tools will take the decisions made by humans on relevance and privilege on a small sample, or seed set, of documents and apply those decisions to documents that the technology determines have the same characteristics of the documents in the seed set. This technology allows your review team to only have to look at a much smaller set of documents and then rely on the technology to systematically apply those decisions to the rest of the collection. Third, sometimes the volume of relevant data is just much larger than anticipated prior to your collection process and even if you employ some of the technical solutions discussed it will still take you much longer than planned to complete your production. Negotiating a schedule of rolling productions with the receiving party is a good way to satisfy both sides. With a rolling production you agree to produce data in smaller production sets on a set schedule instead of one large production at the end of discovery. The receiving party gets the benefit of receiving data sooner and the producing party is able to extend out the final discovery deadline. It can be even more effective if the parties agree on the priority and order of the types of data produced, allowing the receiving party to ask for what they most need to work with and the producing party to hand over the "low hanging fruit". This type of communication and cooperation can help avoid discovery disputes.
What is the difference between a forensic image and a forensic collection?	At times individuals may confuse making a "forensically sound collection" with the act of performing a full forensic image of media. In fact, there is a huge difference in terms of process and more importantly the resulting volume of information. Creating a full forensic image of a media (such as a PC Hard drive) is a technical approach, employing specialized hardware and software to make a "bit by bit" image of the original media. While this is often the simplest approach from a collection perspective, this is analogous to checking out the entire library when in fact all you were interested in were the books dealing with contract law. The resulting information is potentially exponentially more than was required and will require expense and effort later to eliminate. Conversely, a forensically sound collection methodology simply means that all files are properly collected and the corresponding metadata (dates, file paths, host) have been saved and not altered during the process. Similar to the forensic image process, specialized software can be used to assist in this process, but the resulting universe is much more targeted based upon the direction and input of the collection analyst. It should be noted, however, that there may be times when a full forensic image is the best approach. Examples include: (1) Time constraints such as investigations or business interactions that prevent a targeted collection and necessitate a "just image everything" approach, (2) instances where the full collection scope is not final and there are concerns that it will not be possible to go back and conduct additional collections, or (3) there is concern that files may be purposefully or inadvertently altered or deleted.

Questions	Pat's take
What is structured data, and how do you identify it and deal with it?	Each government agency has designed databases and systems that are unique to their mission. During the collection process, you will likely run into applications and databases used by custodians to perform their jobs. Although these systems may not contain documents that we traditionally search for during Discovery requests they can, and likely do, contain information that is relevant and potentially responsive to a request. This type of information is classified as structured data. Structured data is information stored in discrete pieces and is categorized by type and organized into groups. A simple way to think of structured data is information that is stored in columns (fields) and rows (records). Spreadsheets and databases are examples of structured data. Structured data systems are used to house information and facilitate many business functions and can be built in a wide variety of tools for an even wider variety of functions. Some examples include systems to track Human Resources, contracts and grants, public reports and regulatory filings, medical records and insurance claims. These systems are typically accessed through an enterprise system over a shared network or intranet but may also be caches of spreadsheets or databases on a custodian's machine. During custodian interviews, be on the lookout for any indication that a system like this is in use. Once you have identified a structured data system that contains relevant information, a plan of action specific to that system must be designed and implemented in order to collect information in a forensically sound manner and in a format that will be feasible for review and production. Tools and techniques specific to the handling of structured data must be employed in addition, or in place of, traditional document collection and processing tools. You will need someone familiar with the data and how it is input, stored and used within the system as well as someone who has the technical ability to work with the database system to extract relevant information into a useable format. A structured data collection project can be complicated as these systems can contain vast amounts of complex information. However, as the information is stored in a rigid, consistent manner with the right technical ability and data insight you can use this information to not only meet discovery obligations but also to dig into the data and understand much more about your case.

My take:

Arguably, data collection is one of the most important phases of the discovery process. Therefore, becoming educated about different collection protocols and tools and staying flexible during the collection phase is extremely beneficial. As with other phases of discovery, employ a team approach to share the tasking, document all key decisions and assumptions, and rely upon others such as IT, Records Management and Subject Matter Experts to point you in the right direction.

Contact

For more information, please contact:

Patrick McColloch
Managing Director, Deloitte Discovery
Deloitte Transactions and Business Analytics LLP
+1 703 236 3050
pmccolloch@deloitte.com