



## **Smart Discovery**

Michael Daniec, President, National Data  
Conversion

All rights reserved 2005, 2006

**January 11, 2006**

This document was prepared by National Data Conversion. It may contain confidential or proprietary information. Any distribution or copying of the contents of this document, in whole or in part, requires the express written permission of National Data Conversion. All product or brand names are trademarks or registered trademarks of their respective owners.

## **The First Mile is the Most Important Mile**

The legal industry faces unprecedented challenges in attempting to lower the costs of litigations. The discovery component, one of the more prominent and fastest rising components of litigation is receiving much scrutiny for any possible costs savings.

As the legal industry attempts to address the rising costs of litigation and specifically the discovery part, everyone is searching for the most effective way to respond

The explosion of electronic documents in business and the corresponding and increasing mass of eDiscovery associated with litigation have placed the spotlight, within the legal industry, on finding the most effective means of reducing costs associated with the discovery phase of litigation.

Although the review phase is receiving a lot of attention because of the high cost on a per document (or any measurable quantity) of the human element associated with review, the smart bet is that the first mile, the collection and initial processing, is the foremost place to look for cost savings.

The logic behind this is simple. Any reduction of discoverable material in the initial phase of the collection process has a percolating and multiplier effect on costs savings associated with any subsequent work. In other words, the earlier and the more you can reduce the amount of material that needs to undergo any subsequent steps, the better off you are, from a cost, time and scope of work perspective.

Viewed from another perspective, the costs per any measurable quantity of data (such as a Gigabyte) to be processed escalate the further down stream you go in the processing cycle. It costs less to process a GB of data in the very first steps of the process ie the collection and processing than any subsequent processing steps such as review.

Therefore, you can achieve a multiplier effect on any feasible costs savings the more effectively you can narrow down the amount of data to be processed after the first step.

## **What can be done**

In practice, there are many means to achieve this type of costs savings. Some of these steps are objective enough that they should withstand judicial scrutiny without requiring opposing counsel approval. Other steps should probably involve engaging the other side to ensure a common understanding & acceptance of the available methodologies.

The most obvious way of limiting the universe of data under consideration is to narrow down the scope of the initial discovery request by crafting an appropriate discovery motion. Since this is not the focus of this article, we will not address the techniques associated with this. Hopefully, the new proposed FRCP rules will stimulate a productive dialog between the opposing sides, as suggested in the meet and confer guidelines. Our focus is what can be done after the discovery motion has been submitted and the associated work needs to be undertaken.

In a simplified view, ediscovery typically consists of culling data from either live systems or backup tapes. Live systems require specific forensic collection techniques and involve various issues such as avoiding disruption of live systems, ensuring no spoliation occurs etc. I am not addressing the various other potential collection sources such as PDA's, printer memory, fax memory, voice mail tapes etc etc. since these tend not to represent the mainstream or the bulk of most discovery projects and tend not to represent major cost components,

In practice, in the largest civil litigations, backup tapes typically provide the most substantive and therefore costly source of discovery. Dealing effectively with backup tapes, then, becomes a critical component of cost management in the discovery cycle. Our focus will be how to deal with tapes in the most effective manner to realize the greatest potential cost savings.

## **How to go about it**

Here are some simple underlying elements that position you to most effectively respond:

- understand the universe of data sources you need to address and how they are represented on backup tapes. This involves understanding the basic overall IT architecture and the backup regimes that were implemented

- based on above understanding pinpoint where the associated backups are. Make sure you are comprehensive, as the most embarrassing, potentially damaging and, strangely enough, most common, situation in litigation comes up when you need to modify your representations to the court because you just discovered another batch of tapes that you hadn't realized existed. Don't leave out sources such a third party records storage company or the tapes stored next to someone's desk.
- pinpoint details as early as possible. Tapes are tapes, are tapes etc. but in fact not all tapes are created equal and details matter, in fact, this business is all about details. You should have specific tape types, quantities, backup software used, and as much definition for expected content as may be available.

### **Now that you have the tapes**

Once you have collected the tapes of interest, work with a knowledgeable vendor to design the most effective strategy for processing the tapes.

Here are the most common and most effective strategies to achieve the greatest impact:

- design the tape processing around an understanding of the IT architecture, the backup regimen and the operating dynamics of the firm such as the number and location of email servers, email server purging schedules etc. For instance, you may not need to process daily backups if you have confidence in the integrity and completeness of the data on weekly or even monthly backups etc.
- sample to confirm assumptions and eliminate variables. Given a sufficiently large opportunity your vendor should be happy to engage in this either free of charge or at very limited cost, based on setting the stage for a larger project
- limit scope by techniques outlined below

### **Effective scope limitation**

Here are the various scope limitation techniques that are most effective

On the side of the culling techniques least likely to be questioned by the opposition:

- catalog tapes prior to any other processing. This may reveal important scope limiting opportunities and save substantial costs associated with tape restoration and further work.
- system and program file elimination (using commonly accepted file signatures developed by the NIST (National Institute of Standards and Technology), or similar techniques, or on a grosser scale by the definition of the backup source (this way you can potentially eliminate
- de-duplication of exact copies. As long as you de-duplicate exact copies, no issue should arise with the other side (except to be able to document that those were exact copies, ie the methodology behind ascertaining "exact"). Further efficiencies can be accomplished down stream, in the review process, for instance, in utilizing near-duplicate identification technologies. Some current solution providers even suggest a two tiered review process, the initial step utilizing near-deduplication methodologies to cull down the review set to documents requiring closer review.

Scope limiters more likely to require agreement from the other side:

- Sample the backup sets. Choose representative samples to ensure full coverage of the data universe, for instance, if a pool of tapes represents a data range of backups make sure you choose a set from the beginning of the date range, the middle and the end. Sampling is a recognized technique for assessing whether a pool of data may contain relevant items.
- limit by email server, or some other IT architecture component that may be identifiable on the backup tapes
- limit by division, geographic location or some other objective means that fits the matter specifics
- limit by date, or multiple date ranges, if appropriate
- limit by custodian if you can establish a mutually acceptable list of key and supporting players and have the other side agree to limit

discovery to those, you can have tremendous impact on the scope of discovery work

- keyword searches (this whole area is evolving quickly from a technical perspective and requires its own discussion regarding the available search methodologies including concept searches etc.)
- limit by type of documents you are looking for ie loose documents, emails, databases or other types of electronic documents/data

A successful implementation of any combination of the above approaches can have a tremendous impact on the scope of work and costs of discovery. To give an example, a client recently requested the processing of emails for a set of some 12 custodians from a set of 1,400 tapes containing GroupWise emails. After some initial consultations, we realized that, based on the email server architecture of the source environment and the associated backup regimen the client followed, only a small subset of the tapes could possibly contain the emails of interest.

We suggested that we first catalog the full pool of tapes to establish their exact contents, and on the basis of that, we would only process the tapes that actually contained the appropriate content. After cataloging, we were able to limit processing to only less than a hundred of the tapes!

The associated cost savings were tremendous, on the order of several hundred thousand dollars, on the tape processing side alone. These types of cost savings are not unusual when approaching such work in a smart way. As always, all such strategies need to reflect effectively the matter specifics including issues in dispute, key players and such.