



White Paper

The Anatomy of Tape

Dissection for Intelligent Culling, Discovery and Remediation

Author – Jim McGann, Index Engines, VP of Information Discovery

Introduction

Over time an enterprise accumulates significant volumes of backup tapes containing the historical records of their business. The bulk of the content on these tapes is no longer of interest. However, a fraction of it has value or a potential liability for the corporation. Gaining access to this data stored on offline tape archives have been technically difficult and expensive. So much so that most legal teams have relied on the undue burden argument to avoid the collection and processing of electronically stored information (ESI) from tape.

New rules and regulations, such as the Federal Rules of Civil Procedure (FRCP), focus on data that is burdensome to collect, including offline tape data, and instituting requirements to produce this content which in the past has been ignored. In fact, recent court cases have validated the issue and imposed fines and penalties for those parties who have not effectively collected evidence from old archives. See a list of representative cases in the Appendix.

As enterprises begin to focus on their historical data, it is important to understand what is contained on tape. This knowledge will allow the processing of the historical content to be streamlined. Relevant data can then be extracted and secured in an archive, thus eliminating future liability posed by data locked away on inaccessible tape. This paper will review the typical topology of backup tape. This knowledge will then be applied to a recommended tape sampling process that enables cost effective litigation support and intelligent remediation of offline tapes.

The Backup Process

Understanding the backup process, policies and schedules will facilitate intelligent decision making when dealing with large volumes of historical tapes. In general, the backup process entails making a copy of corporate data for safekeeping. During this process an exact copy of all files and email are written to tapes which are then stored offsite. Backup occurs daily within most enterprises and generates a number of tapes containing that day's representation of the data environment. To begin deciphering the information contained on tape, it is helpful to know:

- How often full data backups were scheduled
- If and how often partial or incremental backup were performed
- What servers were routinely backed up

By delving into the tape header, which can be explored through a tape cataloging process, these questions can be answered and then the analysis of tape content will become more meaningful.

Tape Header Elements

Backup tapes are not designed for easy to access specific content, but for bulk protection of enterprise data. To support the collection of tape data for legal and records management purposes it is imperative that the eDiscovery tools understand tape formats in order to streamline the process.

As data is written to tape a header is generated. The header indicates the specific data contained on the tape. Access to the header information eliminates much of the mystery and turns volumes of unknown tape cartridges into more manageable objects. The header of the tape contains metadata including the date the data was backed up, what servers contributed to the tape content, amount of content on tape, and the type of backup software used to create the tape. Tape headers can be quickly scanned to create a catalog. This catalog is then used to determine what further investigation of a specific tape is needed. As the tape headers are read during the cataloging process the following information is gathered:

Date Range: The tape header will contain information about when the tape was generated. The legal department may be able to advise that tapes containing data from a specific date range may or may not be of interest based on data retention policy and litigation hold timelines.

Servers: The header provides insight into what server the data resided on before it was backed up to tape. In some cases the IT team may have configured the backup programs to back up data one server at a time. So when the data on a server named "Saturn" is backed up the name of the server will be included in the catalog. The IT department should have a directory of the servers deployed in the organization and information about what data or applications reside on each server. With this information it can be determined what type of data resided on "Saturn" when it was backed up; responsive user data or irrelevant system files.

Client Type: When processing data from specific applications the backup software retains information indicating the type of client. For example, if an Exchange email server is being backed up, the client type "Exchange" could be noted in the tape header. Additional backup types could include: UNIX, Oracle, Windows, and more. This information will help determine the type of data that is in the backup set, and whether it warrants further processing. For example, tapes containing an Oracle database may not contain data relevant to future litigation, but those containing an Exchange email backup would be.

Full vs. Incremental Backups: The header of a tape also indicates if it contains a full or incremental backup set. Daily enterprise backups typically target only data that has changed that day. This is known as an incremental backup. A full backup usually occurs over the weekend and archives a complete copy of all corporate data. When compared to an incremental, a full backup is a much longer process, and generates a significantly larger volume of tapes.

Six incremental backups occur between each full backup. Having a full backup every week negates the need to perform eDiscovery on the incrementals. The weekly full backup will capture almost everything that has changed since the data was captured during previous week's full backup. The incrementals contain information that becomes redundant once the week's full backup is complete.

Tape Sequence: Full backups usually span many tapes. The tape header will contain information on each tape's order in the backup sequence. This is necessary information for the ordering of a series of tapes for detailed processing. For example, you will need to process tape 1 of 3 before 2 of 3 in order to ensure you capture the complete sequence of files and email.

Blank Tapes: A tape without a header or with a header indicating 0 kb of data is blank. Blank tapes can occur when a backup process fails and the administrator unknowingly sends the tapes offsite to storage. Blank tapes can also occur when the backup administrator mistakenly assumes the tape was used and contains content when in fact it is blank. The volume of blank tapes is difficult to predict since they occur due to system failure and/or human error. Index Engines experience has shown about 10% of backup tapes in storage are blank.

Tape Content

Beyond the tape header, the specific contents of the tapes will allow the culling of large volumes of data that are irrelevant. Some content can be quickly culled in order to refine the volume of data to be processed and reviewed.

Duplicate Data: The amount of data that has changed or is newly generated during a single week is insignificant when compared to the full scale of data existing within an enterprise. Therefore, since all corporate data is backed up weekly, the majority of the data on tape will be exactly the same week after week. This results in large volumes of redundant files and email on tapes. Typically 90% of the specific content on tapes is redundant. The volume of duplicate data has a large impact on traditional restoration and eDiscovery costs. Ideally duplicates should be identified as early as possible to reduce unnecessary and time consuming processing.

File Types: When a backup is executed, the data copied from the online environment to tape is a mix of user generated data and system files such as executables and help files. User generated files are interesting for records retention and eDiscovery, whereas system files are insignificant. The backup process does not separate out specific data types - it captures everything. Typically 30% of the data on tape is made up of system files that hold no value. This data can be quickly eliminated.

Determining which files are system files and which contain user data should be handled with file content analysis tools. For example, a user can "hide" critical data contained in a pdf by changing the extension to .exe or .dll. Discovery tools that use file listings to eliminate system data will quickly and unknowingly eliminate this important file. Tools that analyze the header of the file, and not simply rely on the file name and extension should be deployed during this process. Beware of vendors that offer only file listings that will be used to cull data.

Hidden/Deleted Email: The typical approach to collecting email content from tape is to pull user mailboxes. This is a flawed approach and should be used with extreme caution. A user has the ability to store email in a local file, or pst, on their desktop. Additionally, users can file individual emails anywhere on their desktop in an individual msg or eml format.

When collecting data from tape, ensure all email content is captured, EDB's, PST's, eml's, and msg's, and not just a custodian mailbox. Using individual mailboxes will ensure you are missing email

content, specifically the email files a user wants to hide. Much in the same way email can be hidden using PSTs and individual saved messages, email can also be deleted and thus not contained in a custodian mailbox. In order to perform comprehensive collection of custodian mail you need to access not only the user's mailbox, but also the email that has been deleted from their mailbox, but can still be forensically reclaimed. This includes deleted email that still exists in the mail database, in a dumpster or deleted file container. Most collection efforts ignore this email as it is difficult to obtain, however comprehensive collection will ensure this data is included.

Structured vs. Unstructured Data: Certain file types are more important than others during the data collection process. If only email is important, and unstructured user files such as spreadsheets and documents are not of interest, a significant amount of tape content can be ignored. Legal counsel may know up front what types of files are interesting. For example, an engineering or medical organization would most likely find images, such as CAD drawings or x-rays valuable, however a financial firm may not.

Discovery of backed up copies of structured databases are typically not necessary to support litigation. Structured databases, such as a patient records or transaction logs, exist online and retain the historical data within the data structure. Therefore, discovery of structured data can be performed within the online version rather than going back to a historical copy contained on tape.

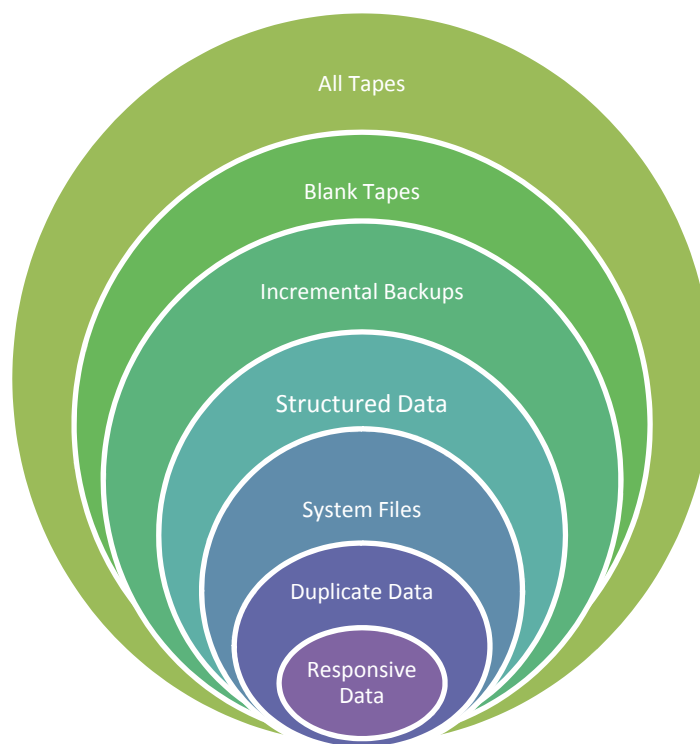


Figure 1. Analyzing Topology to Identify What is Relevant

A deeper understanding of the backup process, types of tapes, and the contents, allows for rapid reduction in historical tape volumes requiring discovery. Typically only 10 to 15% of historical tapes will require deep processing. The balance will contain irrelevant, redundant content that can be ignored once you know what to look for.

Intelligent Tape Sampling Content

When dealing with historical tape records, the objective is to migrate the relevant data to an archive or records management system as cost effectively as possible. Enterprise tape stores often amount to tens of thousands of tapes. Even with an automated collection and processing capability, full indexing of such a vast amount of tapes would be overwhelming. By first cataloging the tapes, which allows analysis of the headers, an intelligent culling approach can be applied to significantly reduce the volume of tapes that warrant full content indexing.

Corporate legal counsel and technology teams must work together to determine how to best address this vast stockpile of potentially liable content. The corporate legal team has the knowledge regarding litigation hold parameters, target date ranges, and suspect custodians. The technical team can leverage this legal insight against the intelligence gathered by cataloging the tape headers to develop a strategy that turns a massive mountain of tapes into a manageable pile. This knowledge, along with automated technology takes care of most of the work. Your effort will be managing the automated process, interpreting the results, and making decisions. The data that has been locked away on tape for years will quickly be made accessible and can then be dispatched accordingly.

Catalog for Intelligent Culling

As we have learned, each tape header contains a wealth of information. Gathering this information via cataloging can be used to cull individual tapes from further processing, or even to be remediated. A header is generated when a backup segment is written to the tape. A single tape containing content will include at least one header, or possibly dozens based on the number of times it is used for a backup. This can happen when a backup administrator creates a backup and the tape is not full. The same tape can be used again until it is full. This will result in multiple backup segments, and multiple headers, on a single tape.

Reading the header(s) of a tape is known as cataloging. The first step in tape discovery is the creation of a catalog for all tapes. The Index Engines cataloging process is fast since only a small section (the header) of the tape is read, so a typical tape takes only a few minutes to catalog. Additionally tape readers with auto loaders make this process mostly automatic. Simply load up the tapes into the library and the catalog is generated. Once the catalog is generated then the intelligent culling process can begin, which typically eliminates 85 to 90% of the tapes from further discovery.

When a catalog is generated the following information is reported for each backup segment existing on the tape:

- Date Range
- Servers backed up
- Client Types
- Full or Incremental backup
- Blank or zero data tapes

Below you can see a detailed example of a tape catalog, including all the metadata generated

The screenshot shows the 'Process Tapes' application interface. At the top, there are filters for 'Backupsets' and 'Hosts'. The 'Backupsets' list includes '192.168.17.228-WinNT', '3112387240', and '3E3A00010100096F-Arcserve tape Backup of test data for Demo paraguay_1200952396'. The 'Hosts' list includes '192.168.17.228', 'CHURCHILL', 'delaware.indexengines.com', 'paraguay', and 'ZEYA'. There are checkboxes for 'Cataloged Status' (Not Cataloged, Partially Cataloged, Fully Cataloged, Failed Catalog) and 'Indexed Status' (Not Indexed, Partially Indexed, Fully Indexed). A 'Tape Summary' shows 5 Fully Cataloged and 5 Fully Indexed tapes. Below this is a 'Tape Indexing Options' section with a checkbox to 'Exclude Selected Tapes that are Only Part of Incomplete Backupsets'. The main table has four columns: 'Select', 'Cartridge IDs', 'Volume IDs', and 'Backupsets', with a 'Tape Status' column on the right. The table lists five tapes with their respective metadata, including device names, labels, volume tags, family IDs, sequence numbers, tape formats, backupset IDs, backup times, backup types, backup formats, host types, root paths, and tape history.

| Select | Cartridge IDs | Volume IDs | Backupsets | Tape Status |
|--------------------------|---|---|--|---|
| <input type="checkbox"/> | Device: MediaChanger0 Slot Number: 1 | Label: MTF-Backups Volume Tag: NBU001 Family ID: 977113547 Sequence Number: 1 Tape Format: Microsoft Tape Format 1 | Backupset ID: Backup of test data for Demo Backup Time: Jan-29-2009 at 18:14:23 Backup Type: Full Backup Format: Microsoft Tape Format 1 Host Type: Windows Host: ZEYA | Status: Fully Cataloged, Fully Indexed Cataloged: Jan-30 at 4:25 pm Indexed: yesterday at 11:15:44 am Tape Size: 3.547GB Tape History: 59 Passes |
| <input type="checkbox"/> | Device: MediaChanger0 Slot Number: 2 | Label: TSDEMO Volume Tag: NBU002 Tape Format: Tivoli Storage Manager 5.x | Backupset ID: 192.168.17.228-WinNT Backup Time: Jan-30-2009 at 08:31:32 Backup Type: Unknown Backup Format: Tivoli Storage Manager 5.x Host Type: Windows Host: 192.168.17.228 Root Path: \\192.168.17.228\filesshare | Status: Fully Cataloged, Fully Indexed Cataloged: Jan-30 at 4:26 pm Indexed: Mar-19 at 12:36 pm Tape Size: 3.612GB Tape History: 53 Passes |
| <input type="checkbox"/> | Device: MediaChanger0 Slot Number: 3 | Label: AAC048 Label Create Time: Jan-21-2008 at 16:53:16 Volume Tag: NBU003 Tape Format: Veritas NetBackup 5 | Backupset ID: paraguay_1200952396 Backup Time: Jan-21-2008 at 16:53:16 Backup Type: Full Backup Format: Veritas NetBackup 5 Host Type: Windows Host: paraguay Copy Number: 1 | Status: Fully Cataloged, Fully Indexed Cataloged: Jan-30 at 4:26 pm Indexed: yesterday at 11:14:19 am Tape Size: 531.53MB Tape History: 51 Passes |
| <input type="checkbox"/> | Device: MediaChanger0 Slot Number: 4 | Label: delaware.indexengines.com.044 Label Create Time: Jan-30-2009 at 13:10:15 Volume Tag: NBU004 Internal ID: 00000005-ba834287-49834287-00470000-c0a8c003 Tape Format: Legato Networker 6.x or 7.x | Backupset ID: 3112387240 Backup Time: Jan-30-2009 at 13:10:48 Backup Type: Full Backup Format: Legato Networker 6.x or 7.x Host: delaware.indexengines.com Root Path: \\MSEXCHS\FIRST Storage Group | Status: Fully Cataloged, Fully Indexed Cataloged: Jan-30 at 4:26 pm Indexed: yesterday at 11:14:04 am Tape Size: 4.525GB Tape History: 97 Passes |
| <input type="checkbox"/> | Device: MediaChanger0 Slot Number: 5 | Label: 5C23-001-ARCSERVE Volume Tag: NBU005 Family ID: 23587 Sequence Number: 1 Tape Format: ARCserve | Backupset ID: 3E3A00010100096F-Arcserve tape Backup Type: Full Backup Format: ARCserve Host Type: Windows Host: CHURCHILL Root Path: C:\Arcserve_imgel | Status: Fully Cataloged, Fully Indexed Cataloged: Jan-30 at 5:55 pm Indexed: yesterday at 11:12:35 am Tape Size: 696.13MB Tape History: 61 Passes |

Figure 2. Representative Catalog Report for a Backup Tape

As discussed in the Tape Header Elements section, decisions can be made based on this catalog data. Blank tapes, those out of the relevant data range, backup sets containing irrelevant clients can all be removed from the tape population that is tagged for further discovery. The tagged tapes contain the interesting content. This subset of tapes can now be selected for indexing to obtain more detailed insight into the files' metadata (dates, users, location, etc.) and textual content.

The traditional tape discovery method requires the data to be fully restored from tape to online storage in order to make decisions about the content. Index Engines automated approach eliminates this restore process, which is known to be very expensive, time consuming and inefficient. Even though a vast amount of irrelevant tapes have been culled, the remaining tapes still contain a significant volume of irrelevant and duplicate data. Typically less than 5% of the remaining tape data will fall within the parameters that the legal and IT teams have determined to be relevant. It is counterproductive to restore and manually index 100% of tape content, of which 95% has no value. By automating the indexing process, the truly useful content can be identified and then extracted from tape in a much more efficient and cost effective approach.

Automated Approach

Index Engines Tape Engine scans the tapes, using tape readers which can be loaded using an autoloader or library, and automatically generates a searchable index. At this time, the actual data has not been restored from tape. The index contains the information that will be used for detailed culling, the raw text of files and email, as well as detailed metadata. At the same time as the data is automatically indexing, a unique document signature is created in order to target duplicate files, based on actual content. Duplicate files and email can be culled automatically with just the click of a button.

Once duplicates are filtered out, the index can then be queried to find files with specific custodians, date ranges, and keyword content. Once the relevant data is found, it can then be easily extracted from tape. Extracting only relevant data from tape is far more efficient versus the traditional requirement of restoring full tape content in order to cull down to the relevant files and email. Once relevant data is extracted from tape and easily accessible, tapes can then be eliminated, recycled, or even placed back in storage.

20,000 to 2,000 - A Real World Example

An actual Index Engines client had 20,000 tapes from which they wanted to extract responsive data into their records management platform and then recycle the tapes. The initial phase of the project occurred in the Index Engines lab, where an initial set of 50 tapes was processed. In the lab the client was trained to use the Index Engines user interface, while at the same time began their tape discovery process. It also gave them insight into the type of header information and tape content they would encounter. This allowed them to begin to formulate a processing strategy.

The first step in the process was generating the tape catalog. This initial set of tapes included a mix of Symantec Backup Exec® and NetBackup® backups. The tapes were loaded into a 25 slot library and were cataloged. The cataloging process took a total of 2.75 hours for all 50 tapes. Using this catalog the tapes were categorized as follows:

- Blank tapes: 3
- Tapes written from servers comprised of non-user data: 10
- Date range outside of discovery range: 2
- Incremental backups: 21
- Tapes requiring deep processing: 14

The analysis of the tape catalog culled 36 tapes, 72%, from further processing. Only 14 of the 50 tapes moved on for deeper processing which included indexing. The client generated reports on the process, and took the tapes that were deemed irrelevant out of storage and recycled them for future backups.

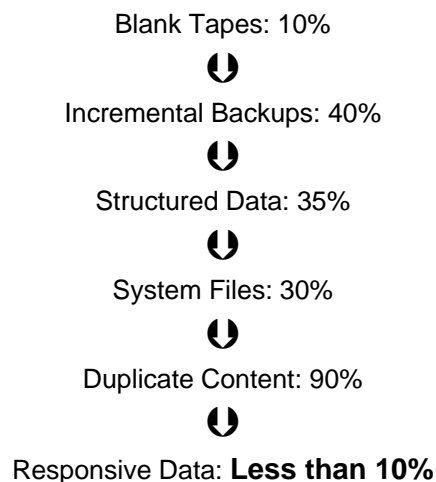
Extrapolating the 50 tape sampling process to all 20,000 tapes, allows the elimination of approximately 14,400 tapes based on a cataloging process of 1.9 weeks using four Index Engines systems and libraries with auto-loaders. This time estimate includes tape processing time, but not man hours. Since the technology is fully automated the only man hours required are minimal and are to support loading the library.

The next phase of this 50 tape discovery project was indexing of the content on the remaining 14 tapes, approximately 2.5 TB. Deep indexing automatically extracts the detailed metadata and textual content of the files and email into a searchable index. This phase requires a full scan of the tapes. The scanning process is throttled by the speed of the tape reader. In this case the majority of the tapes were LTO-2 format. This tape format scans at 40 MB/second and can store 200 GB of data. Therefore for the deep indexing of all unstructured files and email on these tapes took less than one day. Index Engines system with an autoloader processes an average of 19 LTO-2 tapes per day.

Once the deep scan was complete the resulting index was then searchable. On these 14 tapes, 91% of the content was redundant. The client executed query sets derived by their legal team, including specific custodians and content. Based on these queries a unique result set of 2.5 GB, 1% of the original data set, was extracted and migrated to an archive for safe keeping. This extracted content represents less than 5% of the data that was indexed from the 14 tapes. Intelligent sampling reduced 50 tapes down to 14 tapes, which contained less than 5% of relevant data. Index Engines technology allowed this content to be identified and extracted in two days. If this methodology is extrapolated out onto the original stockpile of 20,000 tapes, the cost and time savings are immense. The cost per tape for restoration and discovery of 20,000 tapes would approach \$2.7M. The old school approach of conducting a full restoration is 450% more costly than an investment in Index Engines technology to handle the same tape backlog.

Conclusion

Proactively processing historical tapes to access valuable content is now possible. Not only is automated indexing faster and more cost effective, but the cataloging process allows intelligent culling to be applied before indexing even begins. Typical tape topology metrics can be summarized as:



By working together, IT and Legal can make intelligent decisions about historical enterprise data on tape from the quick process of cataloging tapes. By using the knowledge specific to your corporation for the analysis of tape headers, the retrieval of relevant tape content can be made into a timely, cost effective and prudent exercise.

Methodology

Index Engines operates a lab, where clients receive training on both our technology and on how to analyze the content contained on their backup tapes. Index Engines helps teams tasked with processing large volumes of offline tape begin their discovery projects in our lab. In working with clients, Index Engines has amassed a wealth of experience that helps us to assist others in the future. Our involvement with these clients and their tape processing projects, we have compiled the metrics presented in this paper. These metrics have been developed through insight into a vast array of tape data, spanning multiple industries, many different tape formats, backup software, content types, and tape ages. The knowledge we have gained allows us to profile tape content quickly and educate others as to what to expect when they tackle their content.

Appendix

The following case law represents sanctions delivered for failing to produce stored data to support legal proceedings.

Qualcomm Inc. v. Broadcom Corp., (S.D. Cal. Jan. 7, 2008).

Plaintiff was ordered to pay \$8.5 million in attorney fees and costs.

PML North America, LLC v. ACG Enterprises of NC, Inc., (E.D. Mich. Jul. 26, 2007).

Company CEO added as a defendant to assume personal liability for electronic discovery abuses by his company.

United Medical Supply Co. v. United States, (Fed. Cl. June 27, 2007).

Finding of bad faith not prerequisite for imposing sanctions for spoliation.

Wachtel v. Health Net, Inc., (D.N.J. June 19, 2007).

Discovery abuses resulted in \$6.72 million in fees and costs.

In re September 11th Liability Insurance Coverage Cases, (S.D.N.Y. June 18, 2007).

Sanctions totaling \$1.25 million were imposed after a key document was eliminated.

Heartland Surgical Specialty Hospital, LLC v. Midwest Division, Inc., (D. Kan. Apr. 9, 2007).

Failed to provide an adequately prepared Rule 30(b)(6) witness. The CEO could not answer questions about servers and retention policies.

Exact Software North America, Inc. v. Infocon, Inc., (N.D. Ohio Dec. 5, 2006).

"Defects" in key search words not a valid excuse for failing to respond to discovery.

Optowave Co. v. Nikitin, (M.D. Fla. Nov. 7, 2006).

Sanctioned defendant knowledgeable about computers but who allowed destruction of critical email.

Quintus Corp. v. Avaya, Inc. (Bankr. D. Del. Oct. 27, 2006).

\$1.88 million judgment as a spoliation sanction

Oved & Associates Construction Services, Inc. v. Los Angeles County Metropolitan Transportation Authority, (Cal. App. Jun. 22, 2006).

Default judgments for \$5.2 million and \$978,958 against a party that destroyed the integrity of its financial data.

DaimlerChrysler Motors v. Bill Davis Racing, Inc., (E.D. Mich. Dec. 22, 2005).

Failure to suspend normal procedures for destruction resulted in sanctions even though document destruction was negligent rather than willful.