



Save Time and Money with Automated Process to Extract Relevant ESI from Tape

Many organizations have stockpiles of backup tapes. Because they may be mandated to locate and retrieve specific content from these tapes for e-discovery, this poses significant potential financial and liability risks. New technology can automate culling of irrelevant tapes, indexing the remaining ones, and retrieving and archiving only the relevant content, resulting in significant savings.

Jim McGann

To safeguard their information, most organizations run a daily backup process, during which an exact copy of their files and e-mails are written to a number of tapes, which are then stored offsite. Over time, an organization will accumulate significant volumes of these backup tapes. While the bulk of the content on these tapes is no longer of interest, a fraction does have value or poses potential liability, so it is important the organization be able to locate and access that electronically stored information (ESI) when needed, such as for e-discovery.

However, backup tapes are made for

bulk protection of data, so they are not designed to make it easy to access specific content on them. Because gaining access to this ESI has been technically difficult, time consuming, and expensive, most legal teams have relied on the “undue burden” argument to avoid collecting and processing it for discovery.

But, new rules and regulations, such as the *Federal Rules of Civil Procedure* and the *California Electronic Discovery Act* (June 2009), require all ESI to be made accessible and produced to support litigation. As organizations begin to respond to these

regulations, they need to understand what their backup tapes contain. This knowledge allows them to extract relevant data into an archive and eliminate future liability posed by data that is locked away on inaccessible tapes.

Considering Remediation

It used to be expensive and difficult to understand the content of old backup tapes because first, the content had to be restored, and then it had to be analyzed to determine what to keep and what to purge. Restoring thousands of tapes was out of the question, so IT departments allowed the moun-

tain of tapes to grow taller, with no end in sight.

However, new technology has solved this problem by eliminating the need for backup restoration and applying a more intelligent approach to the process. This technology scans tapes and then searches and extracts specific files and e-mail without using the original backup software, identifying what is typically only a minute percentage of tape content that will have to be handled.

This means that in significantly less time than was required before, IT can process tapes, find what legal needs, archive it, and make it available. This efficient, cost-effective remediation process enables IT departments to recapture tape storage budgets, while providing legal departments with the necessary data.

Grasping the Backup Process

As data is written to a tape during the backup process, a header is generated, which indicates the specific data contained on the tape. The header of the tape contains metadata, which is defined by the international records management standard, *ISO 15489-1:2001 Information and Documentation – Records Management – Part 1: General*, as the data “describing context, content, and structure of records and their management through time.”

Tape headers can be quickly scanned to create a catalog, which is used to determine if further investigation of a specific tape is necessary. 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 whether tapes containing data from a specific date range are of interest based on data retention policies and litigation hold timelines.
- **Servers:** The header provides insight into what server the data

Realizing Significant Real-World Savings

A government agency implemented automated tape indexing technology for a project encompassing 20,000 backup tapes from which they wanted to extract responsive data into their records management platform and then remediate the tapes.

The initial phase of the project was to analyze a sample set of 50 tapes to gain insight into the type of header information and tape content the agency would encounter and allow it to begin formulating a processing strategy.

Phase One

The first step in this phase was scanning tape headers to generate the tape catalog. This provided information about the tapes’ date range, on which server(s) the data resided, the client type, whether it was a full or incremental backup, the tape sequence in the backup, and whether a tape was blank.

The initial set of tapes included a mix of Symantec Backup Exec and NetBackup backups, which were loaded into a 25-slot library and were cataloged in 2.75 hours. As shown below, this process excluded 36 tapes (72%) from further processing, leaving just 14 tapes for deep processing:

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

The agency generated reports on the process and recycled for future backups the tapes that were deemed irrelevant.

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

Phase Two

The next phase of this 50-tape discovery project was indexing the content on the remaining 14 tapes, approximately 2.5 TB. *Deep indexing* requires a full scan of the tape and automatically extracts the detailed metadata and textual content of the files and e-mail into a searchable index.

The scanning process is throttled by the speed of the tape reader. In this case, the majority of the tapes were LTO-2 format, which scans at 40 MB/second and can store 200 GB of data. Therefore, the deep indexing of all unstructured files and e-mail on these tapes took less than one day. The 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 duplicative. The agency then executed query sets derived by its 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 safekeeping.

This extracted content represented less than 5% of the data that was indexed from the 14 tapes. The automated tape processing technology allowed this content to be identified and extracted in two days.

If this methodology is extrapolated out to the original stockpile of 20,000 tapes, the cost and time savings are significant.

resided on before it was backed up to the tape. In some cases, IT may have configured the backup programs to back up one server at a time and should have a directory of the servers and information about what data or applications reside on each. This information can help determine whether responsive user data or, perhaps, just irrelevant system files, resided on specific servers when they were backed up.

- **Client type:** When processing data from specific applications, the backup software retains information indicating the type of client. For example, if an Exchange e-mail server is backed up, the client type “Exchange” could be noted in the tape header.
- **Full vs. incremental backups:** The header of a tape also indicates if it contains a full or incremental backup set. Daily organization 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 organizational data. When compared to performing an incremental backup, performing a full backup is a much longer process and generates a significantly larger volume of tapes. Having a full backup every week negates the need to perform e-discovery on the incremental backups.
- **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 ordering a series of tapes for detailed processing. For example, tape one in a series of three tapes needs to be processed before tape two and then tape three to ensure the complete sequence of files and e-mails are captured.
- **Blank tapes:** A tape without a header, or with a header indicating

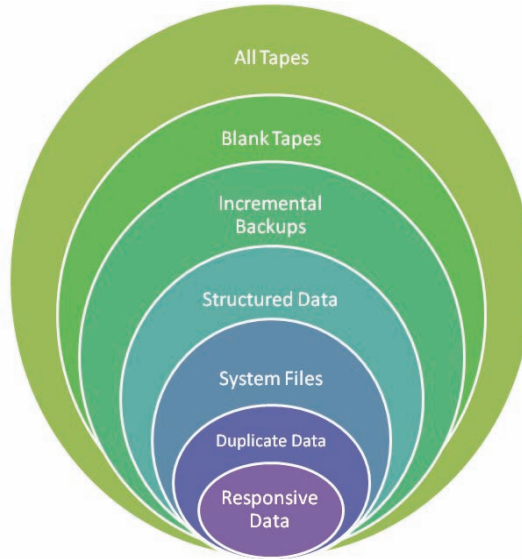


Figure 1: Analyzing Topology to Identify What Is Relevant

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. The volume of blank tapes is difficult to predict since they occur due to system failure and/or human error.

Culling Tape Content

Beyond using the tape header to identify which tapes may be irrelevant, identifying specific characteristics of contents on the tapes will allow culling large volumes of irrelevant data, reducing 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 organization. Therefore, for data backed up weekly, the majority will be exactly the same week after week. This results in large volumes of redundant files and e-mail on tapes.
- **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 e-discovery, whereas system files are insignificant. Typically, according to Index Engines, 30% of the data on tape is made up of system files that hold no value.

- **Structured vs. unstructured data:** Certain file types are more important than others during the data collection process. If only e-mail 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 historical data within the data structure. Therefore, discovery of structured data can be performed within the online version rather than within a historical copy on tape.

Understanding the backup process, types of tapes, and specific characteristics of the contents, allows for rapidly reducing historical tape volumes requiring discovery. Typically, a small percentage of historical tapes will require deep processing, and the balance containing irrelevant, redundant content can be ignored. See Figure 1 for a graphic representation of the relevant (responsive) data in relation to all data on backup tapes.

Leveraging 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 can often amount to tens of thousands of tapes. Even with 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.

It requires collaboration among corporate legal counsel, records management, and technology teams to determine how to best address this vast stockpile of potentially liable content. The corporate legal and records teams have knowledge of litigation hold parameters, target date ranges, and relevant 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.

Emphasizing Intelligent Cataloging

Reading the header(s) of a tape is known as cataloging. Automated indexing technology scans tapes from a tape library and generates the catalog, which is the first step in tape discovery. Then, the intelligent culling process can begin.

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

Figure 2 shows a detailed example of a tape catalog, including the meta-

data generated.

As discussed in the “Grasping the Backup Process” section, decisions can be made based on this catalog data. Blank tapes, those out of the relevant data range, and 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 deep indexing to obtain more detailed insight into the files’ metadata (e.g., dates, users, location) and full text content.

Once duplicates are filtered out, the index can 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 (typically a tiny percentage of the tape content) is far more efficient than restoring full tape content before culling down to the relevant files and e-mail. Once relevant data is extracted from tape and easily accessible, tapes can be eliminated, recycled, or placed back in storage.



Figure 2: Representative Catalog Report for a Backup Tape

Automating the Indexing Process

New tape indexing technology scans the tapes, using tape readers that can be loaded using an autoloader or library, and automatically generates a searchable index of the content. 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 full text of files and e-mail, as well as detailed metadata.

At the same time the data is indexed, a unique document signature is created to target duplicate files, based on actual content. Duplicate files and e-mail can be culled automatically with just the click of a button.

Perfecting Your Tape Process

Proactively processing historical tape repositories to access valuable content is now possible. Not only is automated indexing faster and more cost effective than old methods, but the cataloging process allows intelligent culling to be applied before indexing even begins.

By working together, IT, records, and legal teams can make intelligent decisions about historical organization data on tape and make the retrieval of relevant tape content a timely, cost effective, and prudent exercise. **END**

Jim McGann can be contacted at jim.mcgann@indexengines.com. See his bio on page 39.