



The Data Accessibility Gamble: What is Your Enterprise Wagering?

Maximize Accessibility and Minimize the Risk Stored
in Your Enterprise Data

This document does not provide legal or other professional advice and should not be relied upon as anything other than a starting point for research and information.

Copyright © 2010 Kroll Ontrack Inc. All Rights Reserved.

No part of this publication may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into a language or computer language, in any form by any means, electronic, mechanical, optical, chemical, manual or otherwise, without the express written consent of Kroll Ontrack Inc.

4	Manage Stored Data Efficiently: Maximize Accessibility, Minimize Risk
4	Information Life Cycle Management
6	The Data Accessibility Gamble
7	Four Tips To Mitigating The Data Accessibility Gamble Tip One: Define The Project Tip Two: Analyze The Data Tip Three: Manage and Refine the Data Tip Four: Review Data Conversion or Manipulation Needs
10	The Result: An Efficient, Cost-Effective Solution
10	A Real Life Example
11	Data Accessibility Solved
11	About Kroll Ontrack

Manage Stored Data Efficiently: Maximize Accessibility, Minimize Risk

Against the backdrop of an increasingly complex technical infrastructure, a mountain of data, and flat or decreasing budgets, cost-effectively and defensibly managing the life cycle of information continues to be a challenge for all organizations. When managed efficiently, a well-defined life cycle incorporates the needs of the organization and adheres to regulatory compliance, security, and litigation readiness requirements. It mitigates risk, protects information assets,

and ensures continued data accessibility. When managed poorly, all data is kept and stored “just in case,” bloating IT budgets unnecessarily and increasing e-discovery risks and costs when an investigation, litigation, or merger & acquisition occurs. This whitepaper will discuss best practices in managing the corporate information life cycle and offer tips to mitigate risks associated with storing and converting critical data.

Information Life Cycle Management

The Storage Networking Industry Association (www.SNIA.org) defines Information Life Cycle Management (ILM) as:

- The policies, processes, practices, services, and tools used to align the business value of information with the most appropriate and cost-effective infrastructure from the time the information is created through its final disposition.
- Information is aligned with business requirements through management policies and service levels associated with applications, metadata, and data.

Interestingly, the word storage is not included in SNIA’s definition. While the definition is relevant to other technical disciplines such as information assurance, security, enterprise architecture, etc., it is also designed to get organizations thinking about information management strategies, tactics, and methods.

More companies are using disk for on-site backup however, off-site tape capacity is expected to increase.

There are many fundamental questions that are integral to the concept of corporate information management. What content is the organization currently storing? Is it located on-site or at a storage vendor and if so, on which tape? What data is truly necessary for business continuity or legal purposes versus duplicate or irrelevant data that should be disposed of? Some information never expires and may already be in storage, such as proprietary drawings, prototypes, or formulas. If this, and other types of data, must be retained for longer and longer time periods, what is the plan to ensure it remains accessible as current technology becomes obsolete and the operational costs to maintain legacy systems purely for restoration purposes is no longer practical?

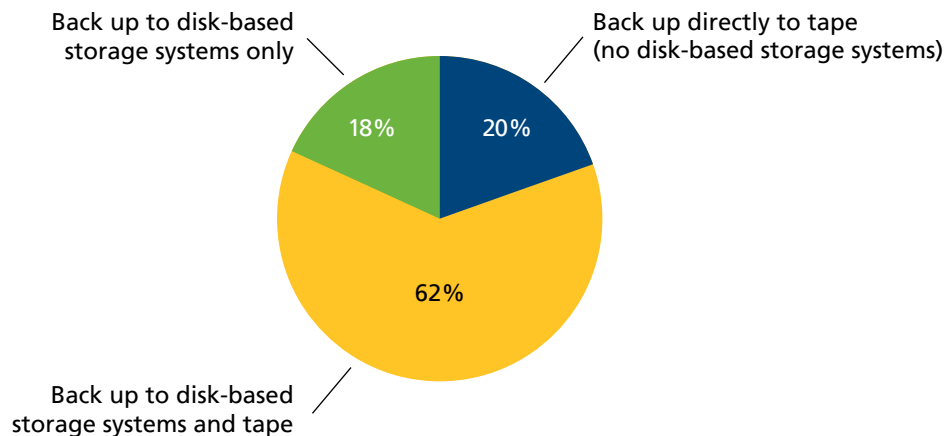
If an organization doesn't address these questions via a well-defined ILM process, the default answer for both stored data and legacy systems automatically becomes "keep it." As a practical matter, information about managing types of content is rarely tracked and communicated via an existing Service Level Agreement (SLA) and the unnecessary expense and risk multiplies.

According to recent Enterprise Strategy Group (ESG) research, "82% of organizations still use tape to support all or a portion of onsite backup processes (see Figure 1)." The report confirms "there will be a shift as more companies use disk for on-site backup processes, but off-site tape capacity is expected to increase. This is largely due to primary data

growth, representing more data that needs to be protected.

Despite predictions of its death, tape's demise as a protector of newly generated information is far from imminent. It still remains the most predominant repository for historical information—specifically vital business records, some of which could be a liability if requested as part of a regulatory or legal matter."

**Which of the following best describes your organization's onsite data backup process?
(Percent of respondents, N=441)**



Source: Enterprise Strategy Group, 2010

Figure 1: Tape is part of most organization's onsite backup processes

The Data Accessibility Gamble

Organizations routinely backup and store information, thinking their processes are strong and the data is sound. However, a variety of issues can hinder data retrieval, some of which are never discovered until the organization is in a purely reactive crisis mode and scrambling for alternatives. Regardless of the technical issue, an organization has a duty to preserve relevant data upon reasonable notice of litigation. Moreover, it is clear that managing data in a manner that

makes it difficult, if not impossible, to recover will not excuse a corporation from that duty. Some organizations are gambling that their legacy data will not be an issue and will be accessible and usable if and when they need it. The table below highlights some of the most common threats to data accessibility.

Tape remains the most predominant repository for historical information—specifically vital business records, some of which could be a liability if requested as part of a regulatory or legal matter.”

Threats to Accessibility

Backup Software Failure	Backup software is set up correctly and the process is kicked off. However, the actual backup data itself is never verified.
Storage Media Failures	Tape drive failure; corrupt or inaccessible tapes; the information written to tape can't be read (logical errors in the data). There is a significant difference between data from the last backup versus data from the point of failure.
Human Error	Errors such as accidentally re-initializing a tape or forgetting to enable the append option before starting a backup are common.
Volume of Data and "Findability"	Sheer volume of data and the ability to find specific content within the corporate memory. How do you know if data is lost or missing? For example, when companies merge, the operational, accounting, and client data of both companies needs to continue to be available. The various backup landscapes have to be harmonized (e.g., proprietary backup systems in Windows environments.)
Aging Systems & Obsolescence	Need to maintain legacy data; converting old static systems to another format or newer technology. Auditors may also request submission of old data records – such as, in the case of one bank, the submission of 17,000 sets of entries from the 1980s. The tapes were available, but the software and drives were no longer serviceable.
Disaster	Fire, water damage, mud, extraordinary cold, heat or other natural catastrophes are often the reasons for tapes becoming contaminated, damaged and no longer legible using the standard means.
Forensically Unsound Methods	The data may be “readable” by a human but moving data incorrectly can modify the file or system metadata relied upon for compliance, investigative, and e-discovery purposes.

Four Tips To Mitigating The Data Accessibility Gamble

Within the larger context of information life cycle management, organizations are looking to data management experts to help them manage stored data more efficiently and reduce the load on IT personnel and infrastructure. As part of your solution, consider the following four tips:

Tip One: Define The Project

The success of a project involving the manipulation of stored data depends on the ability of those tasked with the work to identify and understand the project scope and challenges so they may plan accordingly. For example:

- What does the data landscape look like? Have all data storage systems and media been identified?
- Is there experience in delivering solutions across disparate systems?
- What is the motivation and available budget for the project?
- Are there legal or regulatory requirement deadlines?

Recording the type of media and its condition is just as important as clarifying the suitable target medium. Even with apparently devastating damage (such as through water and fire) there is usually some sort of recovery possible offering the opportunity to arrange the company's long-term backups better at the same time. In this scenario, it is important to work quickly, before media ultimately become unusable because of adhesion or

corrosion. What data protection requirements exist? For example, if data carries cannot leave the company, then conversion must be done on-site. Or perhaps obsolete servers need to be rebuilt in such a way that the previous access rights can be reconstructed as well.

Defining the project, the scope, and identifying the required technical and personnel resources is a step that cannot be skipped or completed half-heartedly.

Tip Two: Analyze The Data

An organization must identify the contents of the media in order to make informed decisions later about data retention, destruction, or suitability for compliance or litigation readiness. Depending on the business needs, scanning, cataloging, or indexing the media can help an organization narrow their focus to the relevant media. However, enterprise backup

software is designed for managing large quantities of data, not for identifying and accessing specific content (see Figure 2). It is complex and requires a relational database to manage backup parameters, sessions, schedules, errors, and other statistics.

While the backup software tracks what it is backing up systemically, getting details about the actual contents of the backup may be elusive. A classic example is a business acquisition. All of the new company's backups become the asset of the parent company and it is likely that both companies employed different backup software. A few years after the acquisition, a lawsuit occurs and during the discovery process, all of the company's long-term data stores need to be reviewed and extracted by legal counsel. If this requirement is not met, all of the backup tapes can be subpoenaed.

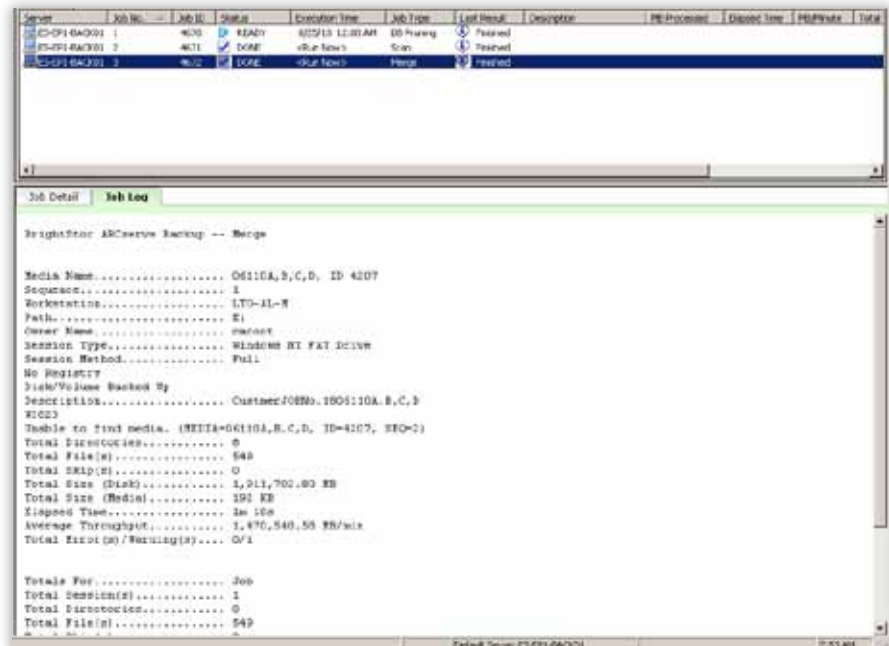


Figure 2 : Backup software records metadata about the backup rather than the specific media contents.

Four Tips To Mitigating The Data Accessibility Gamble

Without the original backup software and or the specific tape machine equipment that recorded that data, content identification will be the biggest hurdle and one of the larger costs of the project.

“Cataloging” and “indexing” has different meanings among long-term backup media software vendors. A long-term backup’s catalog usually refers to the backup sessions on a media set. Some backup vendors save this identification metadata on the tape itself. However, there are a growing number of backup software vendors that put the Media ID, Backup ID, or Session ID on the media which references back to the software’s relational database. Additionally, backup

sessions that are linear in scope and recording on the media are becoming rare. In an effort to maintain backup IOPs (Input/Output Operations Per Second) and system performance, many backup platforms employ distributed recording and data slicing. In this case, multiple data streams or processes are executed concurrently. In order for the media hardware to keep up with multiple backups, the software will store one session with a specific MB or GB size and then switch to another current backup stream. The only differentiator between what’s actually written to the media is the Media ID, Backup ID, or Session ID —the relational database stores the rest of the associated metadata.

When accessing the backup media via the command prompt, the storage administrator may be faced with ambiguous information (see Figure 3). Different meanings and terminology can cloud an assessment of media or a backup set. The command-line output could be interpreted as “Indexing the backup media.” As shown in Figure 3, all of the mounted backup media is displayed and the status of that media is listed. However, it does not display the contents to the storage administrator or the project manager of a media consolidation project.

Data management organizations that provide data accessibility services can identify tapes with sessions and then provide

Determining what data is stored on media can be a challenge when:

- Updating the backup software
- The database used to manage all of the media and session information is purged or deleted
- Expectations for metadata collection may not meet the needs of the project. For example, the enterprise database may have identified the backup sets as “media sets” - a vague reference to what was actually being backed up.

```
Up since: Wed Aug 18 16:24:46 2010 Version: Networker 7.5.1.Build.360 Eval
Saves: 1 session(s), 588 MB total Recovers: 0 sessions(s)
Device type volume
/backup adv_file (none)
/backup2 adv_file (none)
/dev/nst0 (J) sdt600 80084553 mounted sds1t600 tape 80084553
/dev/nst1 (J) sdt600 80084653 mounted sds1t600 tape 80084653
/dev/nst2 (J) sdt600 80084753 mounted sds1t600 tape 80084753
/dev/nst3 (J) sdt600 80084853 mounted sds1t600 tape 80084853
/dev/nst4 (J) sdt600 80084453 mounted sds1t600 tape 80084453
/dev/nst5 (J) sdt600 80084353 mounted sds1t600 tape 80084353
d=fawn:backup adv_file (none)
fawn:/dev/nst0(J) sdt600 90084353 writing, done (full)
fawn:/dev/nst1(J) sdt600 90084453 writing, done
fawn:/dev/nst2(J) sdt600 90084553 mounted sds1t600 tape 80084553
fawn:/dev/nst3(J) sdt600 90084653 mounted sds1t600 tape 80084653

Sessions

Messages
Wed 04:45:51 PM fawn:/root saving to pool 'Default' (90084353)
Wed 04:46:05 PM media warning: rd=fawn:/dev/nst0 writing: No space left on device,
Wed 04:46:05 PM media notice: sdt600 tape 90084353 on rd=fawn:/dev/nst0 is full
Wed 04:46:05 PM media notice: sdt600 tape 90084353 used 510 MB of 40 GB capacity
Wed 04:46:06 PM media info: verification of volume "90084353", valid 4110649123 suc
Wed 04:46:07 PM write completion notice: Writing to volume 90084353 completed
Wed 04:46:09 PM fawn:root saving to pool 'Default' (90084453) 508 MB
Wed 04:46:11 PM 588 MB are saved to pool 'Default' (90084453) of fawn:/root
Wed 04:45:43 PM write completion notice: Writing to volume 90084453 completed

Pending:
```

Figure 3 : Command-line output from the backup software's storage engine

Four Tips To Mitigating The Data Accessibility Gamble

reporting one level deeper—the exact contents contained within the long-term backup. Indexing can be done by reading the tape directly without requiring the original backup software that created the copy initially. By working outside of the backup software's layers of complexity and metadata database reliance, data management organizations can provide a complete list of the extracted and compiled files. This level of detailed analysis ensures that a data consolidation project remains within budget.

Tip Three: Manage and Refine The Data

Organizations regularly complete incremental (daily/weekly) and full (month-end) backups. Although this is an industry “best practice”, the result creates multiple copies of the same data. Based on the previous analysis and knowledge of an organization’s backup procedures, the relevant data set can be culled further and assuming there is no active legal hold on the data, the duplicate data can be deleted. If the data must be retained, backups can be consolidated by restoring them to higher capacity tapes. Additionally, irrelevant system files can also be deleted – a process known as deNISTing. Note: The National Institute of Standards and Technology (NIST) list contains over 28 million file signatures and is used to identify files with no evidentiary value. For more information, see the National Software Reference Library at www.nsr.nist.gov.

Tip Four: Review Data Conversion or Manipulation Needs

When defining the project’s scope, data conversion and/or manipulation requirements may have been identified. It is important to understand the degree of complexity involved in order to keep the project on schedule and within budget.

Simple Conversion

Some conversions are straightforward such as copying files from one computer system platform so they are readable by another platform. Other conversions may require more technical expertise. For example, consider how digital content specifications differ between mainframe, midrange, and desktop systems. IBM and AS/400 computers use the EBCDIC code to represent the alphabet, while in most instances the ASCII code is the norm. Maintaining data accessibility for this type of project requires translation work, including the conversion of an AS 400 database in EBCDIC format of a fixed length into an ASCII code of flexible length or a .csv file for PC.

Complex Conversion & Data Manipulation

A more complex conversion may involve the manipulation of fields in a database. For example, Payment Card Industry (PCI) compliance requires disguising cardholder data when storing credit card numbers. In this scenario, a data management expert could expand and extract the contents, find the cardholder numbers, and apply masking characters (such as “X”s) to the appropriate data.



The Result: An Efficient, Cost-Effective Solution

A project involving the management and manipulation of stored data can be triggered by a variety of regulatory, compliance, or e-discovery needs. Planning for data accessibility streamlines the effort required to meet those needs and mitigates the associated risks. Organizations that define

their information management strategies, and employ the tips outlined above see these results:

- A well-defined project plan
- Comprehensive documentation of data contents

- Improved usage of IT resources (personnel and operational overhead)
- Timely delivery of data, accessible and available for business use in the specified format

A Real Life Example

A Kroll Ontrack client, a Fortune 500 pharmaceutical company, needed data management expertise when tapes were found in a basement partially flooded by fire sprinklers. No one knew what was on the tapes and the data went back to 1996 when the company in question had not yet been acquired by the pharmaceutical company. Approximately 10% were damaged from sprinkler system

contamination and then being frozen and unfrozen by a less knowledgeable data management provider. The client shipped over 5,100 pieces of media (a variety of DLT, LTO, Exabyte, DDS, CDs and other formats) to Kroll Ontrack. Per the insurer's requirements, the job needed to be done within 6 months. Since the insurance company was funding the restoration, it was extremely interested in the amount of

data being recovered from the contaminated tapes. If the data could not be recovered, the insurance company would have to finance the manual input of the data! Ultimately, Kroll Ontrack recovered and copied the data to several 2TB disks for scanning with their IBM TSM backup software to determine if there were patents or other valuable historic information in the data.

Data Accessibility Solved

Historically, it has been time consuming, technically difficult, and cost-prohibitive to incorporate legacy data into an organization's information life cycle management (ILM) plan. After relying on IT to restore the data, Legal would work with IT to analyze the relevant data required to support an investigation or lawsuit. Due to budget and infrastructure limitations, restoring thousands or tens of thousands of tapes was not feasible.

The problem has been solved by using technology to streamline the entire process. Rather than rely on a false sense of security, many corporations seek expert consultative assistance with proven experience using forensically sound methods and deep expertise in legal, compliance, and IT issues. Regardless of your storage environment, the Kroll Ontrack process adapts to meet your data management objectives:

- Data Identification, Mapping and Collection - Locate, preserve and collect business critical and legally relevant data
- Migration - Safely migrate large amounts of data
- Media Consolidation – Combine incremental or differential backups into one backup
- Media Conversion - Seamlessly convert data from one format to another
- Tape Cataloging - Catalog to save time, money and resources
- Tape Duplication - Reproduce data with ease
- Tape Ingestion – Prepare to ingest identified files or tapes into a legal hold or archiving system

About Kroll Ontrack

About Kroll Ontrack

Kroll Ontrack provides technology-driven consulting services and software to help legal, corporate and government entities as well as consumers manage, recover, search, analyze, produce and present data efficiently and cost-effectively.

In addition to its award-winning suite of software, Kroll Ontrack provides data recovery, paper and electronic discovery, document review, computer forensics, secure information services.

Kroll Ontrack is a division of Alteryx, leading provider of information solutions.



Kroll Ontrack België
Sint-Annadreef 68B
1020 Brussel
+32 (0)2 787 02 07

www.ontrackdatarecovery.be

Copyright © 2010 Kroll Ontrack Inc.
All Rights Reserved.

All other brands and product names are
trademarks or registered trademarks of
their respective owners.