The Sedona Conference Journal

Volume 11 2010

An Overview of ESI Storage & Retrieval

John H. Jessen

Recommended Citation: John H. Jessen, *An Overview of ESI Storage & Retrieval*, 11 SEDONA CONF. J. 237 (2010).

AN OVERVIEW OF ESI STORAGE & RETRIEVAL

John H. Jessen
Medina, WA

When considering the discovery of electronically stored information, or ESI, it is often useful to consider the various types of ESI that can be created, the physical ways in which such ESI can be stored, and the typical organizational schemas under which it can be organized. Based upon the needs of the matter at hand, these various metrics can be used to target that ESI most likely to provide useful information.

When one considers all of the ESI that a given organization may have in its possession, or its enterprise data set, the types and quantities can be staggering. In the three year period from 2004 to 2007, the average amount of data in a Fortune 1000 corporation grew from 190 terabytes to one thousand terabytes (one petabyte[1]). Over the same time period, the average data sets at 9,000 American, midsize companies grew from two terabytes to 100 terabytes.[2] Overall, the global data set grew from five exabytes (five billion gigabytes) in 2003 to 161 exabytes in 2006. It is estimated that in 2007 the amount of information created and replicated globally surpassed 255 exabytes.[3]

To place these numbers in some perspective, the Library of Congress, with 130 million items on approximately 530 miles of bookshelves—including 29 million books, 2.7 million recordings, 12 million photographs, 4.8 million maps and 58 million manuscripts—can be stored on ten terabytes. Accordingly, the entire collection of the Library of Congress could be stored more than ten times over in an average midsize company.

Given the amount of ESI that exists within the average organization, the ability to quickly and efficiently identify, locate, retrieve, and preserve the targeted set of ESI most likely to be responsive to the matter at hand becomes essential. Understanding what types of data are likely to play a role in the discovery, the possible storage locations of such data, and the likely ways in which the targeted data may be organized are all important factors in designing a discovery effort that will be focused and productive.

Types of ESI

There are potentially hundreds, if not thousands, of different types of data that could exist within an enterprise data set. Typically, however, most organizations have a limited set of potential data types as they have standardized on a limited number of applications that create ESI. Even a limited set of enterprise data types can run into the hundreds, however. This is why an important component of an ESI discovery plan is to identify the potential types of data that may play a role in providing responsive data.

1 A terabyte is 1,000 gigabytes, and a gigabyte is 1,000 megabytes. A petabyte is 1,000 terabytes, or 1,000,000 gigabytes.
2 From a 12-week study performed by research firm TheInfoPro, Inc., in New York in 2007. The study asked 400 data storage professionals how much storage capacity they have, how the capacity is allocated, and the capacity's effective utilization rate, or how much of its capacity is actually housing data.
3 Gantz, John F., *The Expanding Digital Universe—A Forecast of Worldwide Information Growth Through 2010*, *in* IDC WHITE PAPER, (Mar. 2007).

Interviews with key players and with the organization's computer staff are two ways to determine which data types are worth focusing on.

From the perspective of creating a discovery plan, there are two fundamental categories, or types, of ESI: (1) data created by individual custodians using local applications; and, (2) data created by individual custodians using an enterprise application and/or which is automatically created and/or captured by an enterprise application.

Custodian-Based ESI

Custodian-based ESI is familiar to anyone who uses a computer, as it is the data that is created when using application programs on personal computers or through the use of personal digital devices such as cell phones and personal digital assistants.

The key to custodian-based ESI from the discovery perspective is not necessarily what the application is or where the application is based, but rather that the custodian himself or herself controls the creation, content, storage, and disposition of the data file created.

The following are examples of common ESI created by individual custodians.

Application Data

An application program, often referred to simply as an application, is any program that is designed to perform a specific function directly for a custodian or, in some cases, for another application program. For the purposes of discovery, key features of application programs are that they are initiated by the custodian, the custodian creates the content (data) by directly interacting with application (whether personal-computer-based, a network application, or even an Internet-based application,) and the custodian determines where the resulting application data file (ESI) is going to be stored, what it is going to be named, its usage, and how long it will remain in existence.

Examples of application programs include:

- Word processors
- Spreadsheet programs
- Database programs
- Web browsers
- Software development tools
- Graphical presentation programs
- Document publishing programs
- Sales and personal contact management programs
- Document scanning and storage programs
- Voice-to-text conversion programs
- Printed-text-to-digital-text conversion programs
- Draw, paint, and image editing programs
- Financial management programs
- Music management programs
- Text and other instant messaging-type communication programs

Personal Digital Devices

A personal digital device is an electronic device operated by a custodian that is capable of creating ESI. These devices can be very specific in the types of data they hold, such as a photograph in a camera, or multi-purpose in the sense that they can hold specific types of ESI and act as a storage device for non-device-specific types of data. For example, an iPod is fundamentally a hard drive that has a music-playing application program (iTunes) on it. It can hold digitized music that is used by the application and/or it can be used to hold virtually any other type of data file.

Examples of common personal digital devices include:

- Cell phones
- Blackberry
- PDA's (Personal digital assistant)
- Cameras
- iPods or other similar device

Messaging Systems

Messaging systems are a special form of application in that they share characteristics of both custodian-based applications and enterprise applications. Most messaging systems, especially those within organizations, are enterprise-wide and enterprise-hosted applications, meaning that the messaging program itself is maintained in a central location and is available for use by all those with an authorized account. Furthermore, the messaging system typically stores some custodian-specific messaging data at this central location. Like a custodian-based application, however, most messaging systems also allow the individual custodian to maintain some portion of his or her messaging data locally on their personal computer or at some other location they may designate.

When targeting ESI in a messaging system during discovery, one must consider both the enterprise and the individual nature of the system and inquire accordingly. Inquiry must be made to both the enterprise staff charged with the housing and operation of the messaging system and the individual custodian using the system in order to determine the true nature of location, quantity, scope, and usage characteristics of the ESI sought.

Examples of common messaging systems include:

- Electronic Mail
    - Messages
    - Calendar entries
- Voice Mail
- Instant Messaging

Enterprise-based ESI

Enterprise-based ESI is data that has been created by individual custodians using an enterprise application and/or which has been automatically created and/or captured by an enterprise application. An enterprise application is typically a system where the application and its associated data reside in a central location within the organization. The application is generally one that is used by many custodians across business units within the organization, all of whom need access to all or part of the application data set.

For purposes of discovery, the key aspect of an enterprise application is that the custodian using the application does not have control over the application, its general interface, or where or how the associated data is stored or managed. Accordingly, when considering enterprise data, it is important to involve the organization's computer management staff responsible for the operation of the targeted enterprise application.

Common examples of enterprise applications include:

Organization-Specific Applications

Most large organizations have teams of software developers that write special-purpose, company-specific application programs designed to automate part of the company's business function. For example, an agricultural products company may develop an application designed specifically for tracking their crops. These applications are typically enterprise in nature and are managed by the company's information technology (IT) department.

For purposes of discovery, organization-specific applications often require more effort to identify, locate, assess, and review. Because these applications are unique, it is difficult to find information about them or their corresponding data sets in the vendor market. Focused effort must be made to identify the existence of these applications and the identity of those individuals who have knowledge about them.

Databases

Most organizations utilize database applications to organize their products and business workflow. Databases often serve as the "back-room" for other application programs, holding the information that is created in an organized fashion. Enterprise databases tend to be central stores of large volumes of structured data relating to a particular business activity or business function (i.e. product inventory.) As with organization-specific applications, databases require diligence to determine their existence, their structure, and their content.

Generic Enterprise Applications

In addition to customized organization-specific applications, many organizations employ standardized enterprise applications that have been designed and built by third-parties to solve a particular business need. Because these applications are generally available in the marketplace, it is relatively easy to find information about the application and about the data files that the application supports.

Common examples of generic enterprise applications include:
• Accounting
Automated accounting is the grandfather of all enterprise applications. Automated accounting systems record and process the accounting transactions of an organization. Most automated accounting systems are modular in nature, allowing the organization to choose those modules that it needs at the time, while allowing it to add additional functionality as needed.

Typical accounting modules include:
- ◦ Accounts Receivable
- ◦ Accounts Payable

- General Ledger
- Billing
- Expense Entry
- Purchase Order Management
- Sales Order Management
- Payroll—where the company tracks salary, wages, and related taxes
- Employee Timesheet Management
- Inventory Management
- Reporting

- CRM — Customer relationship management
  CRM is a term applied to the systems and processes implemented by a company to facilitate their contact with their customers. CRM software is used to support these systems and processes, typically by storing information on current and past customers, prospective customers, and often sales leads. The information in the CRM application is typically accessed by employees in departments such as sales, marketing, product development, and customer service.

- EDRM — Electronic Document and Records Management
  The purpose of an EDRM system is to enable an organization to manage their documents throughout the document life cycle, from creation to destruction. EDRM applications typically follow a document from its inception as a work-in-progress until it has passed through a series of defined steps to become a formal record within the organization. EDRM applications are often used to associate a retention code with each record, thereby enabling the organization to destroy records once they have reached the end of their economic, regulatory, legal, or otherwise defined life cycle.

- ERP — Enterprise Resource Planning
  An ERP system is an organizational support system based on a common database that integrates the data needed for a variety of business functions such as Manufacturing, Supply Chain Management, Accounting, Human Resources, and Customer Relationship Management. Most ERP systems are modular in nature, allowing the organization to choose those modules that it needs at the time, while allowing it to add additional functionality as needed. The ultimate goal of the ERM system is to integrate all of the data in the organization into a single database that can then be used to optimize business workflows.

- PLM — Product Lifecycle Management
  A Product Lifecycle Management system provides an organization an automated platform to manage the entire lifecycle of a product, from its conception, through design and manufacture, to service and disposal. It provides the organization with a single source of all product-related information necessary for collaborating with business partners, for supporting product lines, and for developing new or enhanced product lines.

- SCM — Supply Chain Management
  A Supply Chain Management system provides an organization with an automated platform to plan, implement, and control all aspects of their supply chain by tracking the movement and storage of raw materials, work-in-process inventory, and finished goods from start to completion. A comprehensive SCM system encompasses all aspects of sourcing, procurement, logistics, and collaboration with channel partners, such as suppliers, intermediaries, third-party service providers, and customers.

- SDLC-Systems Development Life Cycle
  A Systems Development Life Cycle system provides an organization an automated platform to manage the models and methodologies that the organization uses to develop systems, generally computer systems. Most SDLC systems are modular in nature, allowing the organization to choose those modules that it needs at the time, while allowing it to add additional functionality as needed.

Typical SDLC modules include:
  - Feasibility Planning
  - Project planning
  - Requirements Gathering
  - Systems Analysis
  - Systems design
  - Build
  - Testing
  - Installation
  - Deployment
  - Maintenance
  - Update

- SRM — Supplier Relationship Management
  A Supplier Relationship Management system provides an organization with an automated platform for managing their organizational buying processes, including the purchase of in-house supplies, raw materials for manufacturing, and goods for inventory. With the goal of reducing costs and ensuring that the organization has the materials it needs, a comprehensive SRM system measures and manages supplier performance, defines and enforces purchasing requirements, and coordinates the purchasing process with the real-time needs of the organization.

Internet

The Internet is a worldwide, publicly accessible series of interconnected computer networks that transmit data using a defined standard Internet Protocol. Functionally, the Internet is a "network of networks" comprised of millions of smaller academic, business, and government networks, which together carry information and services, such as electronic mail, text messaging, file transfer, and the Web pages and other resources of the World Wide Web.

From a discovery perspective, the information presented by an organization's Web pages, and the information gathered from visitors to those Web pages, comprises a set of ESI that can be investigated. Increasingly, organizations are connecting their Internet access points to databases and other application systems in an attempt to provide a low cost, single point of access to customers and prospective customers.

Intranet

An intranet is a private computer network established by an organization that uses Internet protocols and network connectivity to create a private, in-house version of the Internet. Intranets are typically used to provide a secure forum for the organization to share information with its employees. Utilizing a familiar Web browser interface, employees can access employee manuals, corporate calendars, updates on corporate events and milestones, records management policies, employee blogs, sales and marketing materials, stock quotations, and the like. Increasingly, intranets are being tied into

corporate applications, legacy systems, and databases in an attempt to provide a single-source interface to the company.

Extranet

An extranet is a private network established by an organization that uses Internet protocols, and network connectivity to create a private, in-house version of the Internet that is then shared with selected extra-organizational parties, such as vendors, suppliers, clients, and business partners. Utilizing a familiar Web browser interface, those granted access to the organization's extranet can gain access to sales materials, catalogs, production updates, account information, electronic mail, instant messaging, blogs, and the like. Increasingly, extranets are being used to create virtual business communities where business partners come together to share information.

How ESI is Stored from a Technology Point-of-View

On-line Storage of ESI

When ESI is stored on-line, it means that the information is available to a user, on a computer system, in virtually real-time. The definition of on-line as established by the United States General Services Administration calls for an on-line system to be available for immediate use on demand without human intervention, in operation, functional and ready for service.[4]

ESI stored on-line is the most familiar form of data to users of computer systems. When a computer user sits at his or her computer or workstation, creates a data file using an application program, and then stores that file on the computer or on the corporate network, he or she has created ESI stored on-line. When a computer user sits at her or her computer or workstation and retrieves a file from the local hard drive or from a networked drive, he or she is retrieving ESI stored on-line.

On-line storage devices are primarily hard drives, whether singly in a personal computer or connected together in an array in a networked system. Hard drives allow fast access to data without any form of human intervention. As on-line storage provides the fasted retrieval time for ESI, it is typically used for those files that need to be immediately available at all times, which includes virtually all enterprise applications. Given the relatively low cost of on-line storage, most custodians choose to store their personal data files on-line as well.

From a discovery perspective, on-line data is relatively easy to identify, locate, search, retrieve and preserve, as electronic search and organization tools can be used in a real-time fashion to interrogate the data. Unlike near-line and off-line data, on-line data does not need to be "restored" before it can be utilized for discovery, thereby making on-line data a much cheaper, faster, and easier form of discovery data.

Near-Line Storage of ESI

Near-line storage is the storage of data on direct access removable media. When a near-line storage device is re-attached to a computer system, the ESI stored thereon

4 See Federal Standard 1037C, entitled Telecommunications: Glossary of Telecommunication Terms, which is a United States Federal Standard, issued by the General Services Administration pursuant to the Federal Property and Administrative Services Act of 1949, at http://www.its.bldrdoc.gov/fs-1037/fs-1037c.htm.

becomes available to the user in an on-line fashion. Near-line storage provides inexpensive, reliable, and virtually unlimited data storage, but with less accessibility than with on-line storage, as it requires the step of reintegrating the storage device with the computer system.

Near-line storage is often used for the portability of, and/or to make a backup copy of, ESI. Near-line storage is a convenient way to store ESI that is used periodically, such as music on a CD disk, or to transport ESI from one location to another.

The major categories of near-line storage include:

- Magnetic disks
  - 3.5-inch diskettes
  - Iomega Zip disk and Syquest-type removable disks
- Compact disks (CD)
  - CD recordable disks (CD-R)
  - CD rewriteable disks (CD-RW)
  - Digital versatile disk rewriteable disks (DVD-RW)
- Solid state storage (flash memory data storage device)
  - Memory card
  - Memory stick (USB flash drive)
- Removable DASD (Direct Access Storage Device) (Hard Drive) Devices
  - iPods
  - Portable hard drives

Other devices that can serve as near-line storage devices include:

- Remote on-line Backups
- Disk-based backups
- Printers with storage capability
- Fax Machines with storage capability
- Copy Machines with storage capability

From a discovery point of view, the portability of near-line storage can create identification and location problems. Additionally, while retrieval of ESI from a given near-line source is rarely an issue, retrieval from numerous near-line sources can create logistical and expense issues associated with the requirement for re-integrating the near-line storage device with the computer system before retrieval can be conducted.

Off-line Storage of ESI

As opposed to on-line storage, off-line storage is the storage of ESI on a medium or a device that is not under the control of a processing unit and which is not available for immediate use on demand by the system or custodian without human intervention.

Compared with on-line and near-line storage, sending data to off-line storage is slow. The advantage of off-line storage is that it is relatively inexpensive, easily transported, and protects the data from alteration and/or infection from computer viruses. Because of the benefits provided by off-line storage, it is often integral to an organization's backup, or disaster recovery, program.

The primary form of off-line storage is magnetic tape. So much so, in fact, that the term magnetic tape is virtually synonymous with off-line storage. When used as a backup medium, on-line ESI is written to (stored on) a magnetic tape. The recorded magnetic tape is typically then taken off-site from the organization and stored in a secure

and environmentally controlled environment to protect it from natural disaster. If all or part of the ESI recorded on the magnetic tape is lost or damaged on the on-line system, the magnetic tape can be used to return a copy of the ESI to the on-line system. The time and cost associated with restoring ESI from a magnetic tape is substantial compared with the cost of on-line or near-line access, and backup tapes are therefore used as a last resort.

From a discovery perspective, magnetic tapes are a difficult and expensive environment in which to search for ESI. They must be retrieved, mounted and restored to the on-line system before any of the ESI contained thereon can be assessed.

Given that magnetic tapes are used for backup, however, means that the magnetic tape may be the only location that particular exists if it has been removed from all other on-line and near-line sources.

Backups

To fully understand off-line storage, one must understand the concept of a backup. In the computer environment, a backup refers to the making copies of data so that these additional copies may be used to restore the original after a data loss event.

Organizations typically make backups for three reasons:

First, a backup protects the organization from losing its valuable data in case of a disaster (natural or manmade) or in case of a computer system failure that results in data loss.

Second, a backup can be used to restore specific data files that have been accidentally deleted, modified, or corrupted.

Third, many organizations use backups as a generic form of long-term data archiving. In this capacity, backups are made and are held by the organization as a central repository of data over time.

Typical Categories of Backups

While a backup is technically any process that moves a file from its on-line storage location to another on-line, near-line or off-line storage location, there are some typical ways in backups are conducted by custodians and within organizations. In terms of discovery, it is important to understand the various ways in which both the client and the adversary conduct and organize their backup systems. This involves discussions with both individual custodians to determine how they may backup data as individuals, as well as with organization computer staff to determine how organizational backups are conducted.

Categorically, two different general types of backups exist, unstructured and structured:

Unstructured Backups

An unstructured backup is typically an ad-hoc copying of a small number of custodian-selected files to some form of on-line, near-line, or off-line repository. Unstructured backups are typically placed onto near-line stores like CD-R, DVD-R, or USB thumb-drive-type media.

Unstructured backups typically have little or no information about what was backed up or when the backup took place, and there is typically little consistency to the frequency and/or content of such backups. Unstructured backups are probably the easiest to implement by the custodian, but they are the least managed and are prone to dispersal and loss.

From a discovery perspective, unstructured backups are usually very difficult to deal with, as they require in-depth inquiry to identify, locate, and retrieve and, once retrieved, are costly to integrate into the discovery process due the resources required to identify the way in which the backup took place, the types and quantities of data, and the relative inefficiencies associated with loading a relatively small amount of data.

Structured Backups

A structured backup is a backup of a predictable target set of data that occurs on a predictable timetable. Structured backups are the types of backups that occur most frequently within organizations and they account for the vast majority of data stored within backups.

Structured backups, and especially those conducted systematically by an organization's computer services department, typically have detailed descriptions about what was backed up, when it was backed up, and how it was backed up. From a discovery perspective, structured backups are generally easier to identify, locate, and retrieve, and a greater level of analysis can generally be conducted as to the types and quantities of data contained thereon.

Local Backup

Local backups are typically backups of data files conducted by custodians through the use of devices contained within, or attached directly to, their personal computer workstation. From a discovery perspective, local backups are usually sporadic in nature, stored in various locations, inconsistent in terms of types and quantities of data stored, and difficult to restore.

Typical local backup schemas include:

- Backing up data files to magnetic disks such as floppy diskettes, Iomega Zip disks, or Syquest-type removable disks
- Backing up data files to compact disks (CD's) such as CD recordable disks (CD-R), CD rewriteable disks (CD-RW), or Digital versatile disk rewriteable disks (DVD-RW)
- Backing up data files to solid state storage (flash memory data storage devices) such as memory cards or memory sticks (USB flash drives)
- Backing up data files to removable DASD (Direct Access Storage Device) (Hard Drive) devices such as iPods or other portable hard drives

Internet Backup

As high-speed Internet service has become more widely available and more robust, backup methodologies utilizing the Internet to create remote backup stores is growing in popularity. These remote sites can simply be other personal or organizational sites that the custodian has access to, or they can be sites provided by third-party companies providing backup and storage services.

As remote Internet backup sites are organizationally and, typically, geographically removed, backing data up to the Internet can provide protection against geographically clustered disasters that could affect backup data stored in the same region as the host data. Even with high-speed Internet capability, Internet backups are substantially slower than backups conducted to local disk storage or to backup tape. This speed issue generally limits the amount of data that a custodian would choose to send to a remote Internet site. Some organizations also feel uncomfortable placing their data into the hands of third-parties to hold and manage, fearing that sensitive data may be compromised.

From a discovery perspective, it is important to understand that the custodian typically determines the frequency of, and the composition of, the backup set that is sent over the Internet. Care must be taken to fully understand the extent to which a given custodian uses Internet backup, the frequency of such backups, the manner in which data is selected for backup, and the details of the remote site at which the data is stored.

Enterprise Backup

Perhaps the most common form of backup in a corporation or other organizational entity is the enterprise backup. An enterprise backup is a backup conducted by an organization's computer services staff involving business unit-level or organization-wide computer systems. A backup of an organization's electronic mail system on a daily basis would be an example of an enterprise backup.

Because they are conducted by the organization's computer services staff for the purpose of providing a disaster recovery copy of the organization's data, enterprise backups tend to be the most structured in terms of the scope of the data targeted, the frequency of the backup, the consistency of the media onto which the backup is made, the recoverability of the backed up data, and the length of time the backup is maintained before disposal.

From a discovery perspective, enterprise backups are often the easiest to identify, locate, and retrieve, although the volumes of backup sets that often exist within an organization can make the logistics of the discovery very difficult. It is also important to keep in mind that magnetic tapes can fail, thereby compromising the entire backup set to which that tape belonged. There may also be difficulties associated with interpreting the many types of data files that are often co-mingled on enterprise backups.

Typical Types of Backup Schemas

Within categories of backups there are different backup schemas that can be employed. Understanding the schema chosen for a given backup is an important component in developing a proper model for restoring a backed up set of data, especially if one is restoring multiple backups to determine a set of data from a targeted time period.

Typical backup schemas include:

Full Backup

A full backup is a backup of every file on the targeted computer system, whether or not that file has changed since the previous backup.

Because a full backup copies every file on the targeted system to the backup media, a full backup takes the longest to accomplish of all the backup schemas and requires

the most storage space on the backup media. In terms of restoration, however, a full backup provides the fastest restoration times when restoring the full data set.

Because of the time and tape space required, full backups are generally conducted on a periodic basis as part of a hybrid backup schema. For example, a full backup may be conducted every Sunday night, while an incremental backup is conducted on the days in between. Full backups are also typically performed on systems that are about to undergo hardware and/or software changes as a means to protect against data loss in case the changes do not work or damage the file storage systems.

If an organization chooses to save selected backups over a long period of time as a means of creating an ad-hoc data archive, full backups are usually the ones chosen. A typical example would be to save the last full backup of every month and to save that backup for one or more years.

Incremental Backup

An incremental backup is a backup of every file on the targeted computer system that has changed since the last backup took place, regardless of whether the last backup was a full backup or an incremental backup.

Because an incremental backup only targets those files that have changes since the last backup, which is typically a fraction of the total data set, it is typically the fastest type of backup and the one that requires the least storage space on the backup media. However, incremental backups also require the longest time and the most tapes to restore. When restoring a full system, however, an incremental backup schema may take the longest time to restore as the first incremental backup has to first be restored and then all of the subsequent incremental backups leading up to the targeted restoration date.

Because of the inefficiencies associated with restoring an incremental-only backup schema, one rarely sees an incremental-only backup schema in place. In most organizations, an incremental backup schema is used in conjunction with a full backup.

Differential Backup

A differential backup is a backup of every file on the targeted computer system that has changed since the last full backup.

While a differential backup is not as fast as an incremental backup, it is faster than a full backup as it does not have to copy every file. Correspondingly, a differential backup requires more storage space than an incremental backup, but less than a full backup.

When used in combination with a full backup, differential backups can provide an effective and efficient backup process. As with incremental backups, in discovery one must fully investigate the way in which backup schemas are utilized, in whole or in combination, to determine the appropriate restoration model.

Continuous Data Backup

A continuous backup is a real-time backup that immediately logs every change on the targeted computer system to a secondary system. This is often done by saving byte or block-level differences rather than file-level differences, which allows the real-time nature of

the system to take place. Effectively, pieces of files are saved as they are changed. If a restoration needs to take place, the management system knows how to piece everything back together in proper form.

With a continuing decrease in hard disk storage costs, continuous backup, sometimes referred to as mirroring, may become more popular in the future.

Examples of the differing backup schemas

- ***Full Backup***
  If you perform a full backup every day of the week and the system crashes on Friday, you would need to restore the full backup set from Thursday to restore the data.

- ***Full plus Incremental Backup***
  If you perform a full backup each Sunday and incremental backups every night and the system crashes on Friday, you would need to restore the full backup from Sunday along with the incremental backups from Monday, Tuesday, Wednesday, and Thursday to restore the data.

- ***Full plus Differential Backup***
  If you perform a full backup each Sunday and differential backups every night and the system crashes on Friday, you would need to restore the full backup from Sunday and the differential backup from Thursday.

- ***Continuous Backup***
  If the system crashes on Friday, you simply restore the files from the secondary source.

- ***Backup Rotation***
  A backup rotation schema is the method chosen for managing backup sets when multiple media are used in the backup process. The rotation schema determines how and when each magnetic tape is used in a backup and for how long it is retained once it has backup data stored on it.

  The most common backup rotation schema is referred to as the Grandfather-Father-Son model. The Grandfather-Father-Son model defines three sets of backups—daily, weekly and monthly. The daily (Son) backups are rotated on a daily basis with one set graduating to Weekly (Father) status each week. The weekly backups are rotated on a weekly basis with one graduating to Monthly (Grandfather) status each month. Many organizations add to this model by removing one or more monthly tapes to an annual or multi-year storage.

  Another common rotation schema is to use a rolling set of magnetic tapes over and over again. This Incremental model defines a pool of backup media and, once the entire pool has been used, re-writes to the oldest set. For example, with a daily backup onto a set of 10 tape sets, you would have 10 days worth of individual daily backups. When all of the tape sets are used, the oldest one is inserted and re-used.

  Tape rotation schemas can get very complicated based upon the needs of the organization. In terms of discovery, it is important to determine what tape rotation model is used and how it is implemented. With any rotation model there

will be gaps in the tape sets due to human, machine, or tape failures. There will likely also be extra-model tape sets in existence that have been created ad-hoc or for special purposes.

How ESI is Stored from a Custodian/Records Management Point-of-View

From a technology standpoint, ESI can be stored on a variety of magnetic, optical, and solid-state media. The manner in which ESI is stored by the custodian onto these media can vary greatly, however, and has to do with both the organization's records management plan and the custodian's own desires regarding the naming and storage location of his or her data.

When considering what ESI may relate to a given discovery matter, it is often useful to consider where such data may have been placed by a custodian or, indeed, whether such data was ever under the direct control of the custodian.

There are five typical ways in which ESI can be stored:

Custodian-Centric Data Storage

Much of the ESI used by a custodian on a day-to-day basis, especially application data, is under the direct control of the custodian. It is the custodian who creates the content associated with a given data file, names it, and determines where the file will be saved. The custodian is also the default "records manager" for his or her data in the sense that he or she determines how long data will survive before being deleted.

In terms of discovery, the custodian is often the best source of information about his or her data set, including:

- Types of data created (i.e. what applications were used, including enterprise applications)
- Quantities of data created
- File naming conventions used
- Data storage locations
- Whether custodian-based backups were created
- Others with whom the custodian corresponded and/or shared files
- Use of electronic mail and attachments[5]

Virtual Workgroup-Centric Data Storage

A virtual workgroup is group of individuals who work on a common project using digital technologies such as electronic mail, instant messaging, shared application programs and databases, calendaring, and file management. Many virtual workgroups share a common data file through the use of applications that support such use.

While the custodian creates some of the content for the application data file, he or she may have little or no say in how the data file is named, where it is stored, how it is

5 Electronic mail is a unique application in the sense that it has both enterprise and local characteristics. Technically electronic mail is an enterprise software platform, but users use and often store electronic mail messages and attachments locally. Accordingly, it makes sense in discovery to investigate electronic mail from both the enterprise level through discussions with organizational computer services personnel and locally through discussions with individual custodians as to how they use their electronic mail system.

ultimately used, or how log it remains in existence. Many times these issues are handled either by organization rules or by a custodian named as the workgroup leader.

As networking and the Internet become more pervasive, and as application software providers enable workgroup features into their software, the concept of virtual workgroups is likely to grow. Rather than sending a file around to a number of individuals and then trying to integrate their suggestions and changes, the data file remains in a central location and the users modify it directly, with each persons edits and/or notations identified with each such person.

In terms of discovery, the custodian is often the best source of information about his or her participation in a virtual workgroup, including:

- Virtual workgroups assigned to
- Other participants in the workgroup(s)
- Applications used by the workgroup, including enterprise applications
- The workgroup's data storage location
- The workgroup's computer services liaison
- File naming conventions used by the workgroup
- File management conventions related to the workgroup (i.e. data backup, data retention)

Business Unit-Centric Data Storage

Many organizations are structured like holding companies, made up of many stand-alone organizations (business units) that maintain their own computer operations but that share some overall application platforms, such as electronic mail. A single organization may also have different operating divisions that it treats as business units.

A custodian working in one business unit within a larger organization may spend most of their time working on the business unit's computer system, but at least part of their time on platforms owned and managed by the parent organization. From the custodians viewpoint he or she is working on a single system. Behind the scenes, however, many different operating and data storage environments may be involved.

While the custodian may create the content associated with a given data file and may name it, in some business-unit environments the custodian may have little or no control over where the data is saved. This is especially true when enterprise applications are used.

In terms of discovery, both the custodian and organizational computer services staff need to be considered as sources of information about the underlying computer system being used and the location(s) of related data stores, including:

- Custodian
  - Types of data created (i.e. what applications were used, including enterprise applications)
  - File naming conventions used
  - Data storage locations, if known

- Organizational Computer Services Staff
  - Desktop applications provided to the custodian(s)
  - Enterprise applications provided to the custodian(s)

- Types and quantities of data created
- Data storage locations for each application
- Data management policies for each application (i.e. data backup, data retention)

Enterprise-Centric Data Storage

Virtually every organization utilizes enterprise applications in their business model. We have seen in a previous section the various types of enterprise applications that exist. If nothing else, electronic mail is pervasive and at its core it is an enterprise application.[6]

One of the key characteristics of an enterprise application is that the data file(s) associated with the application are stored and managed at a central location within the organization, typically by professional computer services staff. Custodians using the enterprise application may have desktop applications that belong to and/or interact with the enterprise application, or they may simply "log on" to the enterprise application and use it directly at its central location.

While the custodian may create new data using the enterprise application and/or modify existing information, the custodian typically has no say in how the data file is named, where it is stored, or how it is managed.

In terms of discovery, both the custodian and organizational computer services staff need to be considered as sources of information about the enterprise applications being used and the location(s) of related data stores, including:

- Custodian
  - Types enterprise applications used
  - The custodian's typical usage of such applications
  - Data storage locations, if known
  - Identity of the custodian's computer services liaison for each such application

- Organizational Computer Services Staff
  - Enterprise applications provided to the custodian(s)
  - Types and quantities of data created
  - Data storage locations for each application
  - Data management policies for each application (i.e. data backup, data retention)

3rd Party-Centric Data Storage

With the increased use of outsourced computer operations and the use of Internet-based applications, more and more organizational data is being stored and managed by third-parties. In an outsourced situation, a third-party manages the hardware and software infrastructure for an organization for a fee. In effect, the third-party is serving as the computer services department for the organization. An Internet-based application is one in which a user using the Internet goes to a third-party site and logs onto an application program provided by the third-party. The user then uses the application just as they would if it resided on their desktop or on the enterprise computers. In both situations, the data created by the user remains with the third-party provider.

6 See footnote 5.

In terms of discovery, the custodian, the organizational computer services staff, and the third-party's computer staff may need to be considered as sources of information about the applications being used and the location(s) of related data stores, including:

- Custodian
  - Types of third-party applications used
  - The custodian's typical usage of such applications
  - Data storage locations, if known
  - Identity of the custodian's computer services (both in-house and third-party) liaison for each such application

- In-house Organizational Computer Services Staff
  - Enterprise applications provided to the custodian(s)
  - Identity of third-party providers
  - Types and quantities of data created
  - Data storage locations for each application
  - Data management policies for each application (i.e. data backup, data retention)

- Third-party Organizational Computer Services Staff
  - Enterprise applications provided to the custodian(s)
  - Types and quantities of data created
  - Data storage locations for each application
  - Data management policies for each application (i.e. data backup, data retention)

Fundamental Computer Forensic Issues

Forensic Disk Images

When used in conjunction with discovery, the term forensics relates to the use of specialized techniques for the recovery, authentication, and analysis of specific ESI.

Forensic examinations are typically used when a matter involves issues that require the reconstruction of computer usage patterns; the examination of residual data left after deletion; technical analysis of computer usage patterns; and/or other testing of the data that may de destructive in nature.

In order for a forensic examination to occur, the ESI, and the storage device on which the ESI resides, must be collected in a manner that requires specialized expertise that typically goes beyond normal data collection and preservation techniques that are generally available to users and even system support personnel.

The most common form of forensic collection is to make an image of the storage media on which the targeted ESI resides.

This image, sometimes called a bit image, is an exact copy of the storage device—such as a hard drive, a CD, or any other disk format—including all areas that contain data and all areas that appear to be empty (but which may actually contain remnants of data.)

The image is a single file containing the complete contents and structure of the storage device. A disk image file is created by making a sector-by-sector copy of the source

media, thereby completely copying the entire structure and contents of the storage media. Forensic images are acquired with the use of specialized software tools. When used properly, these images contain a copy of everything that is on the target media, including live and deleted data. Forensic images are also sometimes referred to as a bitstream image, a bit image, or a cloned image.

This image can be used to re-create an exact copy of the storage device on which a forensic examination can be conducted. This examination can then be conducted on the re-created drive in exactly the same way in which it could have been done on the original device. Because forensic examinations often involve destructive testing, and because they require the ability to replicate their findings, this ability to work on re-created drives is critical.

The primary question when considering a forensic collection is whether or not the facts surrounding the matter at hand suggest that a forensic examination is going to be needed.

Was unique, important data deleted? Is it likely that deleted data can be recovered? Is it important to show usage activity and usage patterns? Is it important to authenticate a particular file in order to show that the represented data and/or time that of creation is accurate? Do you need to confirm that all of the text in a document is original or that a critical email was really sent when it appears to have been?

If the matter is one where a forensic examination may be important, then a forensic collection is required. If not, then a forensic collection is not required and is ultimately a waste of time and resources and which often sets the stage for needless battles over additional forensic examinations that could be conducted on the collected data.

Because imaging software is commonly available, and because the vast majority of training programs in the field of electronic discovery revolve around forensics, there is a growing tendency to want to "image everything." Unless an argument can be made that the matter at hand will benefit from a forensic collection and additional examination, there is no reason to do a forensic collection just because the technology exists to do it.

If the matter allows a non-forensic acquisition and analysis of ESI, then a data collection is what is required. A data collection, as opposed to a forensic collection, collects files at the file level, not at the disk level, basically by copying the desired information and processing it into a review system. Data collection is faster and cheaper than a forensic collection and is the type of collection that is warranted if a forensic collection is not required.

File Deletion

A computer's file system determines how the computer stores and manages files on its attached storage media. There are several file systems in use today, and all offer some form of file recovery once a file is deleted.

For illustration purposes, we will discuss the FAT file system, one of the most commonly used file systems today. When a file is deleted on a File Allocation Table (FAT) file system, its directory entry[7] in the FAT remains stored on the disk, although the file name is altered in a way that lets the system know that the storage space occupied by the

7 A file's directory entry is much like a person's listing in a telephone book. It holds the file's name and it's storage location on a piece of storage media, such as a hard drive, a CD, or a DVD. The directory entry basically tells the computer where to go to find the data file.

(now deleted) file is again available for use by a new file or by an expanded version of an existing file. The majority of the deleted file's information, such as it's name, time stamp, file length and location on the disk, remain unchanged in its directory entry in the FAT.

The deleted file's content will remain on the storage media until it is overwritten by another file. The more file activity there is on a particular computer system, the more unlikely it is that a file can be recovered, as the likelihood that the storage areas where the file had resided will be overwritten is greater.

Specialized software utilities, some provided with, or built into, the operating system, allow for the recovery of a deleted file provided that a new file or data set has not overwritten the areas of the storage device holding the deleted file in question. At the simplest level, these tools allow the modified file name of the deleted file to be changed back into a name format that does not indicate a deleted file. The file then becomes a "live" file again, and available for use by an application program. In some cases a greater level of reconstruction is required to retrieve some or all of a deleted file. If the directory entry for the deleted file has been overwritten, or if some of the data storage areas for the deleted file have been overwritten, it will be more difficult to perform the file recovery.

Some computer operating systems provide a layered approach to data deletion. Microsoft's Windows platform, for example, does not really delete a file when a normal deletion request is made. The file is places in a "recycle bin" where it awaits final deletion. Until the file is removed form the recycle bin, it can be easily recovered as it had not really been deleted in a technical sense. When the file is "dumped" from the recycle bin for deletion, it can often be recovered along the lines described above in the previous section.

As with forensic collection, the key question in discovery regarding the recovery of deleted data is whether or not the facts surrounding the matter at hand suggest that a data recovery is going to be needed.

Was unique, important data deleted? Is it likely that deleted data can be recovered? Was the file located on a system where file activity was such that recovery is likely to be effected? Is the matter-at-hand one where file deletion is suspected or traditionally part of the pattern of activity for such matters, such as in trade secret theft?

If the matter is one where deleted data recovery may be important, then attempts should be made to identify and recover appropriate files. If not, then deleted data recovery is not warranted and is ultimately a waste of time and resources.

As with imaging, data recovery software is commonly available, and because the many of the training programs in the field of electronic discovery revolve around forensics (which is often targeted towards data recovery), there is often a bias to target deleted data. Unless an argument can be made that the matter at hand will benefit from the recovery of deleted data, there is no reason to attempt such recovery just because the technology exists to do it.

Metadata

Generally, metadata is information about a particular data set which describes "how, when, and by whom it was collected, created, accessed or modified, and how it was formatted."[8]

8 THE SEDONA CONFERENCE®, THE SEDONA CONFERENCE® GLOSSARY: EDISCOVERY & DIGITAL INFORAMTION MANAGEMENT 33 (2d ed. 2007).

Metadata provides context for data files and is used to facilitate the understanding, characteristics, and management usage of such data. The metadata required for effective data management varies with the type of data in use and the potential use of such data. In a DVD video collection, for example, where the data is the content of the videos, metadata about a given video title would typically include a description of the content, the title, the producer, the director, the actors, the release data, and the physical location of the video. From this example, one can see how the metadata helps identify and organize the data content and make the collection more useful

In terms of discovery, one should consider various categories of metadata and determine which of these categories may play a role in the matter at hand. In addition, one should always consider metadata from the perspective of a given data file, and whether or not such metadata will provide useful information regarding that file. The common categories of metadata include systems, application, and embedded metadata.