ESI 101

Discovery Pilot
Seventh Circuit Electronic Discovery Pilot Program

Sponsored by The Education Subcommittee of the Seventh Circuit Electronic Discovery Pilot Program and McAndrews, Held & Malloy

Special thanks to our staff, Reza Khazeni, Mike Weiler and Ryan Tilot.

Gregory C. Schodde

McAndrews, Held & Malloy, Ltd.

Goals

- Introduce some basic ESI concepts and vocabulary.
- Provide a working understanding of where and how much ESI potentially exists in the client environment.
- Introduce the Federal Rules that provide a starting point for handling ESI.
- Introduce some ESI special problems.
- Identify some basic ESI reference material.

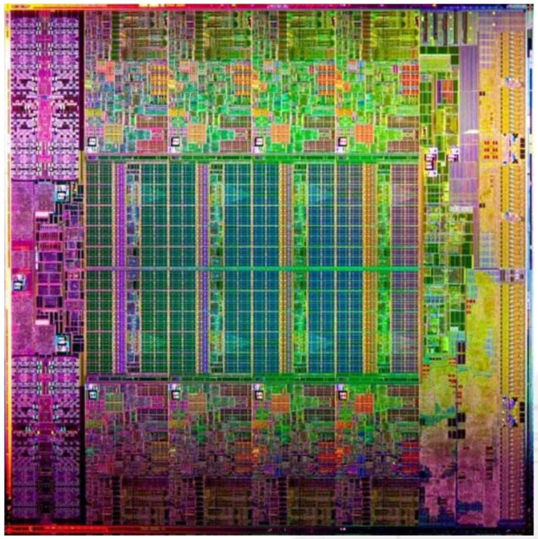
Disclaimers

- Educational use only.
- No claim to copyright in U.S. Government materials.
- Opinions and views are of the author alone and do not reflect the position of the McAndrews firm or the Seventh Circuit.
- This presentation does not constitute legal advice and specific circumstances will vary: Recipients should consult their own advisors for advice applicable to their specific situation.
- CLE Credit given or applied for in Illinois, Indiana, and Wisconsin only. Correct and timely registration and full participation in the on-line event including completion of the end of event survey required for credit. Certificates will be sent by e-mail to the registration e-mail address within two weeks of the end of the event.

What is ESI?

- *The Federal Rules:*
  - *"any document or electronically stored information [ESI]. . . stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably useable form;" Fed. R. Civ. P. 34(a)(1)(A)*
- *In Practice:*
  - *Stuff that requires electronic translation or manipulation before it can be used as evidence.*

Monk transcribing Book, Mural, Rose Reading Room, NY Public Library.
http://www.flickr.com/photos/curiousexpeditions/2406487030/sizes/l/in/photostream/



Intel Core i7 Processor die. The actual size of this device is 20.8 x 20.9 mm, or roughly 2/3 of one square inch, or roughly the area of a U.S. quarter. Within this space, there are 2.27 billion transistors. Image courtesy Intel Corp. 2011.

Some Basic Basics

- Your ESI Team:
  - You and your firm: must produce responsive documents to the opponent and obtain responsive documents from them as well.
  - Your client and your client's IT/Computer Support staff: must assist you in finding what needs to be produced, making sure preservation obligations are met, and must educate you on the burdens involved in producing material that may not need to be produced.
  - Your vendors and/or your own IT support staff: provide the logistics and infrastructure to handle the ESI from end to end.
  - You and your opponent: you can reduce risk and cost, by cooperating with your opposition on ESI issues.

"Translation . . . into a reasonably useable form"

- Paper – Requires No Translation

"Dear Greg:

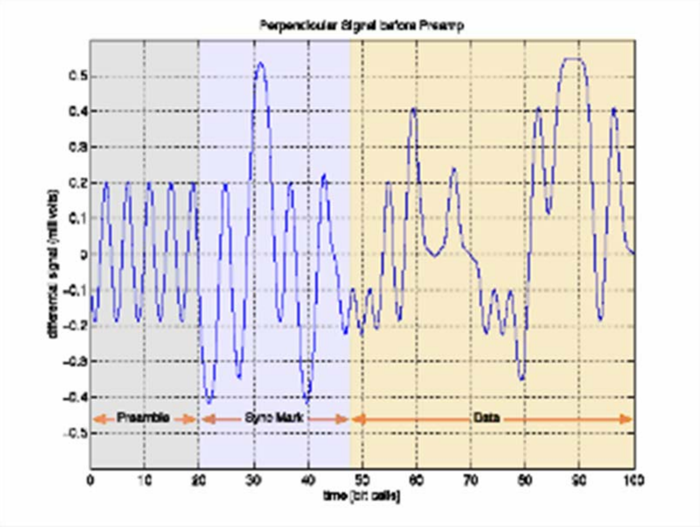
Enclosed is our card.

Regards,

Jim"

- Electronic Media – Significant Translation

Figure 4. Ideal perpendicular magnetic signal output from the head transducer before the preamp.



Signal trace of an electrical signal generated by the read head of a disk drive as it moves across the magnetic recording medium (disk).

Basic Vocabulary

- Binary: The basic alphabet of computers: 1 or 0, On or Off, High or Low: 1 Bit

- Byte: Because one bit is not enough to make an alphabet: 8 bits = 1 byte: (1011 1001). 8 bits can form up to 256 unique combinations of 1 and 0, enough to represent all the characters in the alphabet, and upper and lower case, all the numbers 0-9, with a bunch left over for punctuation, symbols and other characters on a keyboard – so a “Byte” becomes the basic unit of information when we think about “how much” ESI we have.

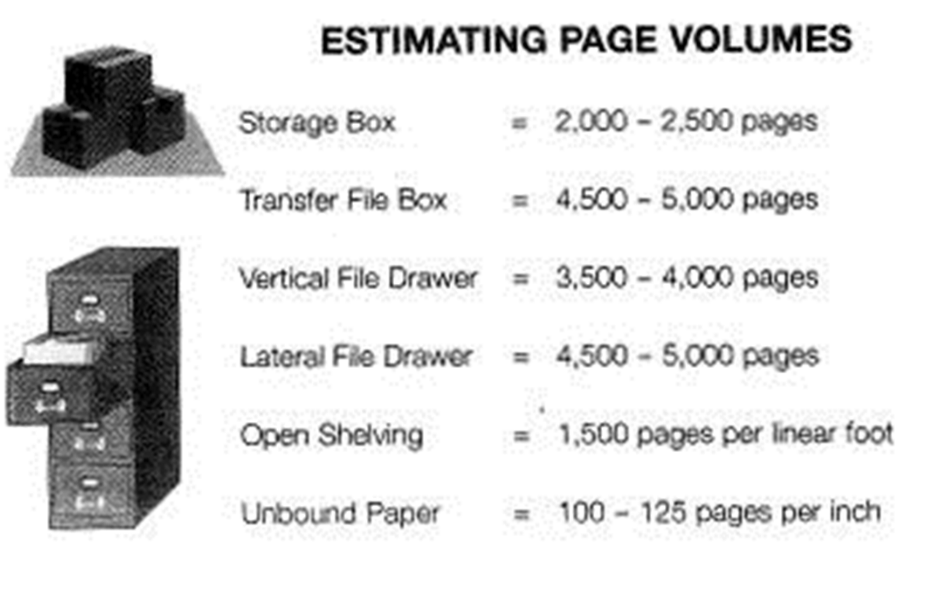
More Jargon

- “TIFF” image. “Tagged Image File Format”. A digital “photocopy” or picture of a page. The “TIFF” format , is recognized by many widely used image manipulation and management systems, including many trial support software systems. Converting a “native” electronic file to a series of “TIFF” images “petrifies” the information into a traditional page reviewable format.
- OCR. “Optical Character Recognition”. A software technique that takes an image and converts it to machine readable, word searchable text. The higher quality the original image, the better the performance.
- USB. “Universal Serial Bus”. A standard specification for a physical cable and electrical interface that allows high speed communication between devices. Thumb drives typically plug directly into a computer with a USB port, many other peripheral devices use this interface as well. Typically transfers data at speeds of up to 60 Megabytes/Second, higher speeds are possible.

The unit of measurement for the non-ESI world: words, pages, boxes.

- The “documents and things” world: Words and Pages:
  - Average words an educated person can read per minute: 250-350 wpm.
  - Average words on a page – 100-600 words/page.
  - Number of pages a Xerox copier could duplicate per minute in 1990: 25-30
  - Pages that can be packed into the available shelf and drawer space in a typical partner’s office at our firm:
    - 40 running feet x 1500 pages/foot = about 60,000 pages at most.

We used to survey our clients large productions to estimate pages by looking at what they had:

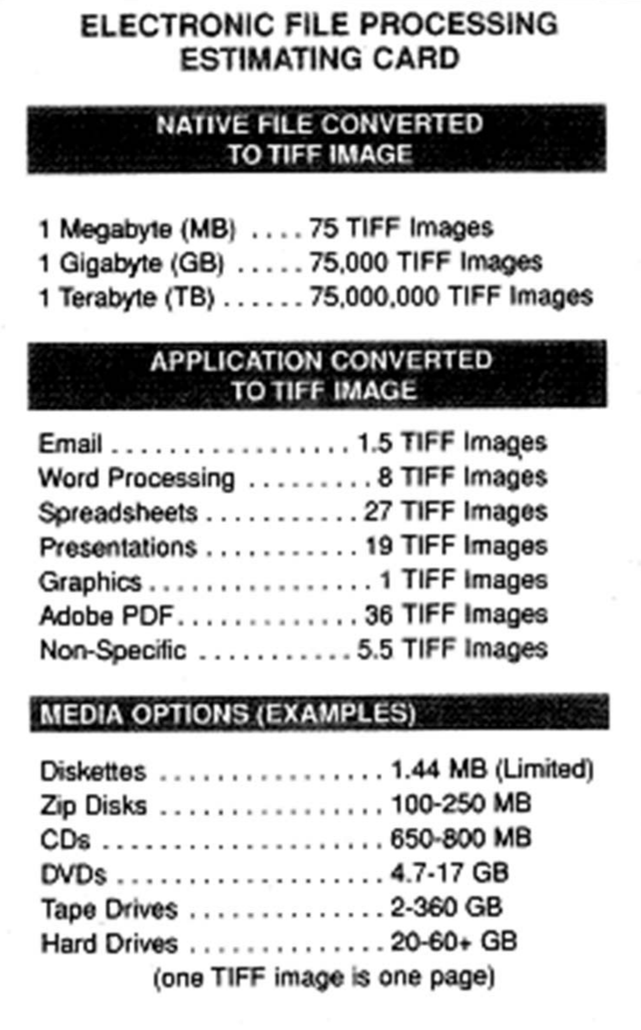


ESTIMATING PAGE VOLUMES

Storage Box	=	2,000 – 2,500 pages
Transfer File Box	=	4,500 – 5,000 pages
Vertical File Drawer	=	3,500 – 4,000 pages
Lateral File Drawer	=	4,500 – 5,000 pages
Open Shelving	=	1,500 pages per linear foot
Unbound Paper	=	100 – 125 pages per inch

The vocabulary of ESI: Bytes, images, CDs

A new vocabulary for answering the question, “How much information do I need to review?”



ELECTRONIC FILE PROCESSING
ESTIMATING CARD

NATIVE FILE CONVERTED TO TIFF IMAGE

1 Megabyte (MB)	75 TIFF Images
1 Gigabyte (GB)	75,000 TIFF Images
1 Terabyte (TB)	75,000,000 TIFF Images

APPLICATION CONVERTED TO TIFF IMAGE

Email	1.5 TIFF Images
Word Processing	8 TIFF Images
Spreadsheets	27 TIFF Images
Presentations	19 TIFF Images
Graphics	1 TIFF Images
Adobe PDF	36 TIFF Images
Non-Specific	5.5 TIFF Images

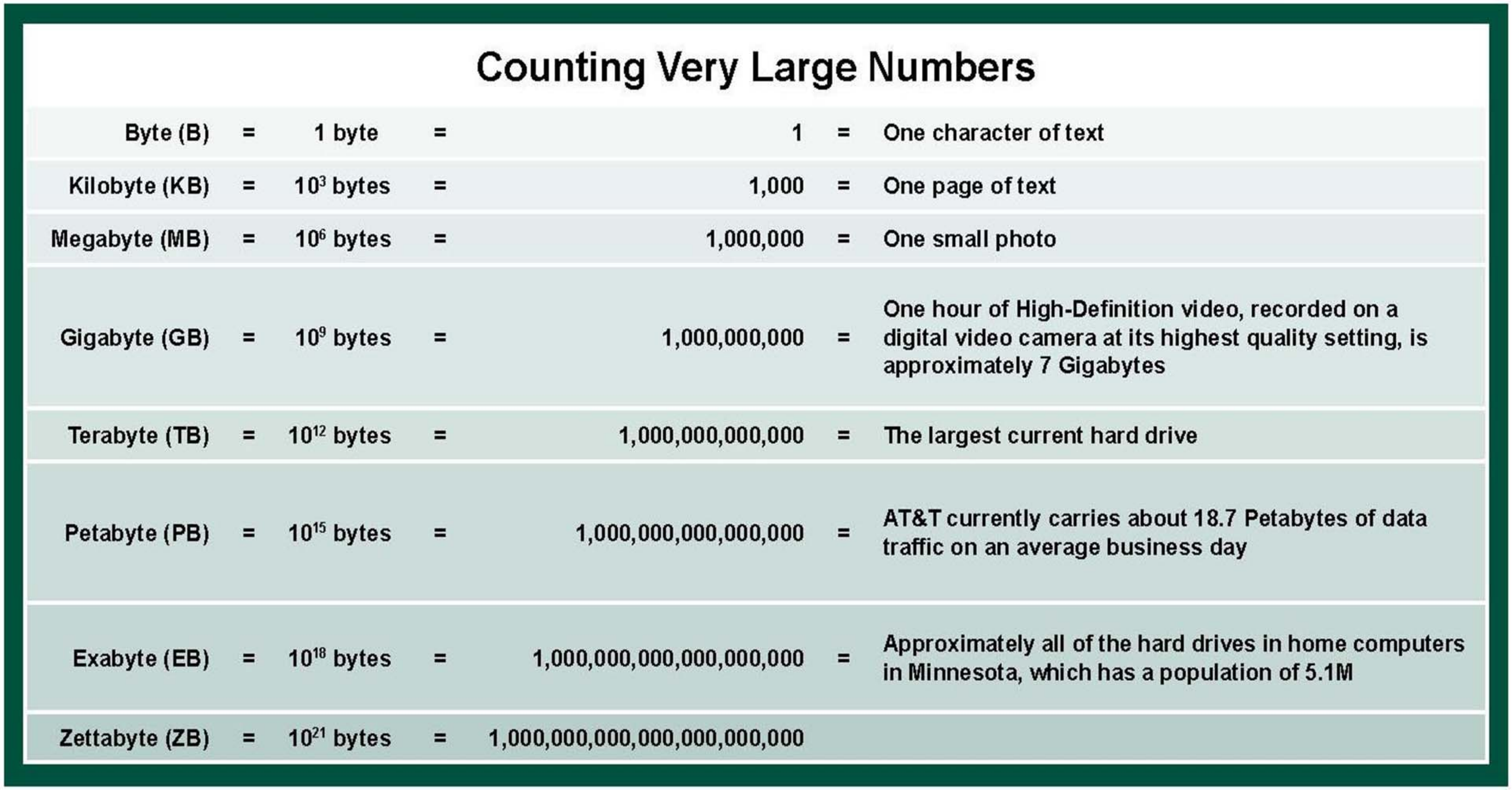
MEDIA OPTIONS (EXAMPLES)

Diskettes	1.44 MB (Limited)
Zip Disks	100-250 MB
CDs	650-800 MB
DVDs	4.7-17 GB
Tape Drives	2-360 GB
Hard Drives	20-60+ GB

(one TIFF image is one page)

Estimating cards courtesy of Compulit

The ESI domain has a vocabulary that includes some very large numbers



Counting Very Large Numbers

Byte (B)	=	1 byte	=	1	=	One character of text
Kilobyte (KB)	=	10^3 bytes	=	1,000	=	One page of text
Megabyte (MB)	=	10^6 bytes	=	1,000,000	=	One small photo
Gigabyte (GB)	=	10^9 bytes	=	1,000,000,000	=	One hour of High-Definition video, recorded on a digital video camera at its highest quality setting, is approximately 7 Gigabytes
Terabyte (TB)	=	10^{12} bytes	=	1,000,000,000,000	=	The largest current hard drive
Petabyte (PB)	=	10^{15} bytes	=	1,000,000,000,000,000	=	AT&T currently carries about 18.7 Petabytes of data traffic on an average business day
Exabyte (EB)	=	10^{18} bytes	=	1,000,000,000,000,000,000	=	Approximately all of the hard drives in home computers in Minnesota, which has a population of 5.1M
Zettabyte (ZB)	=	10^{21} bytes	=	1,000,000,000,000,000,000,000		

*2009 Figures. Update: largest current hard drive 3TB (2011)
*Figures are rounded (e.g., 1 kilobyte = 1024 bytes)

Bohn and Short, "How Much Information? 2009 Report on American Consumers." Global Information Industry Center, U. California, San Diego. Retrieved from http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf on December, 19, 2011.

Analog metaphors for the ESI world: More perspective

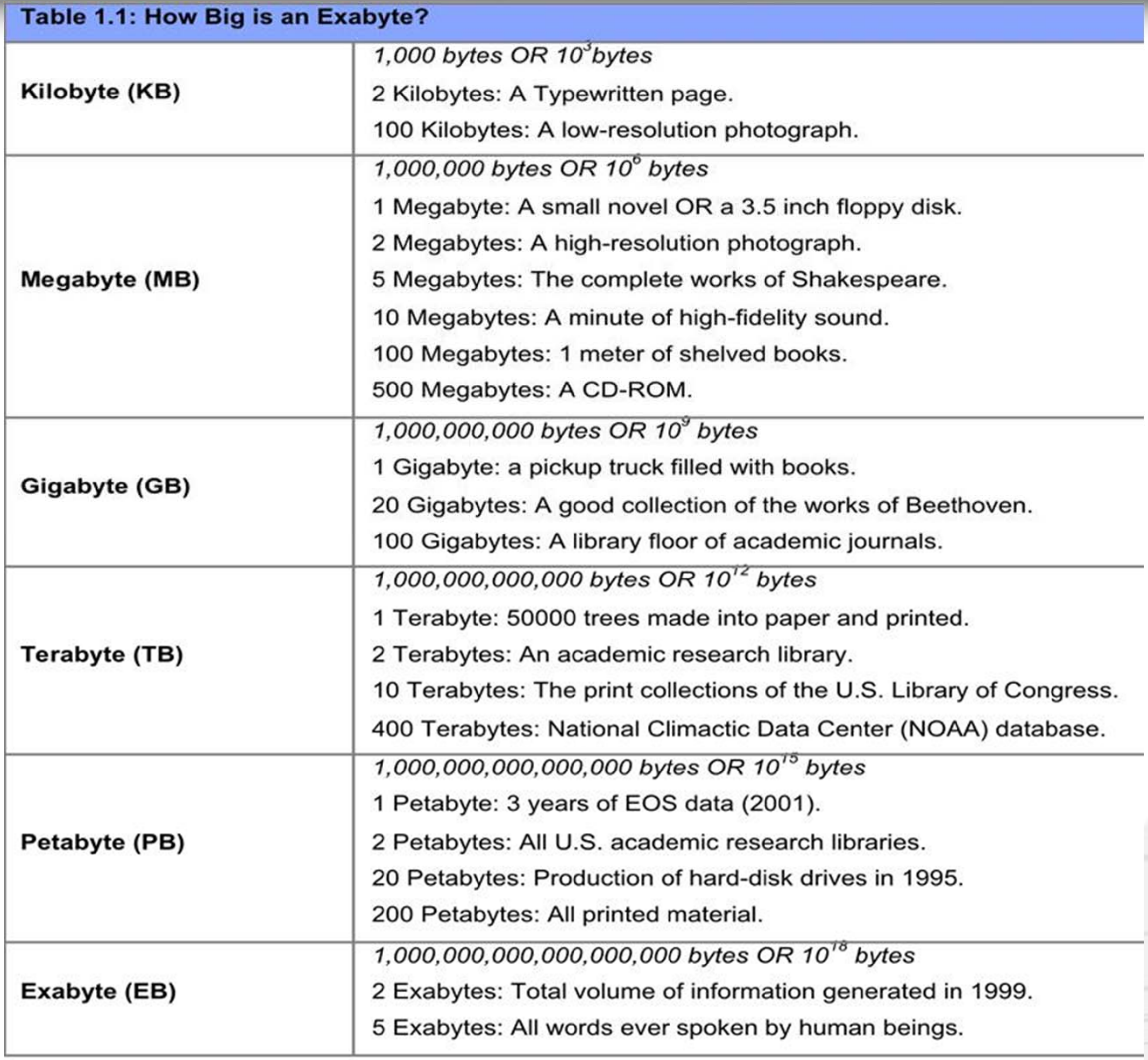
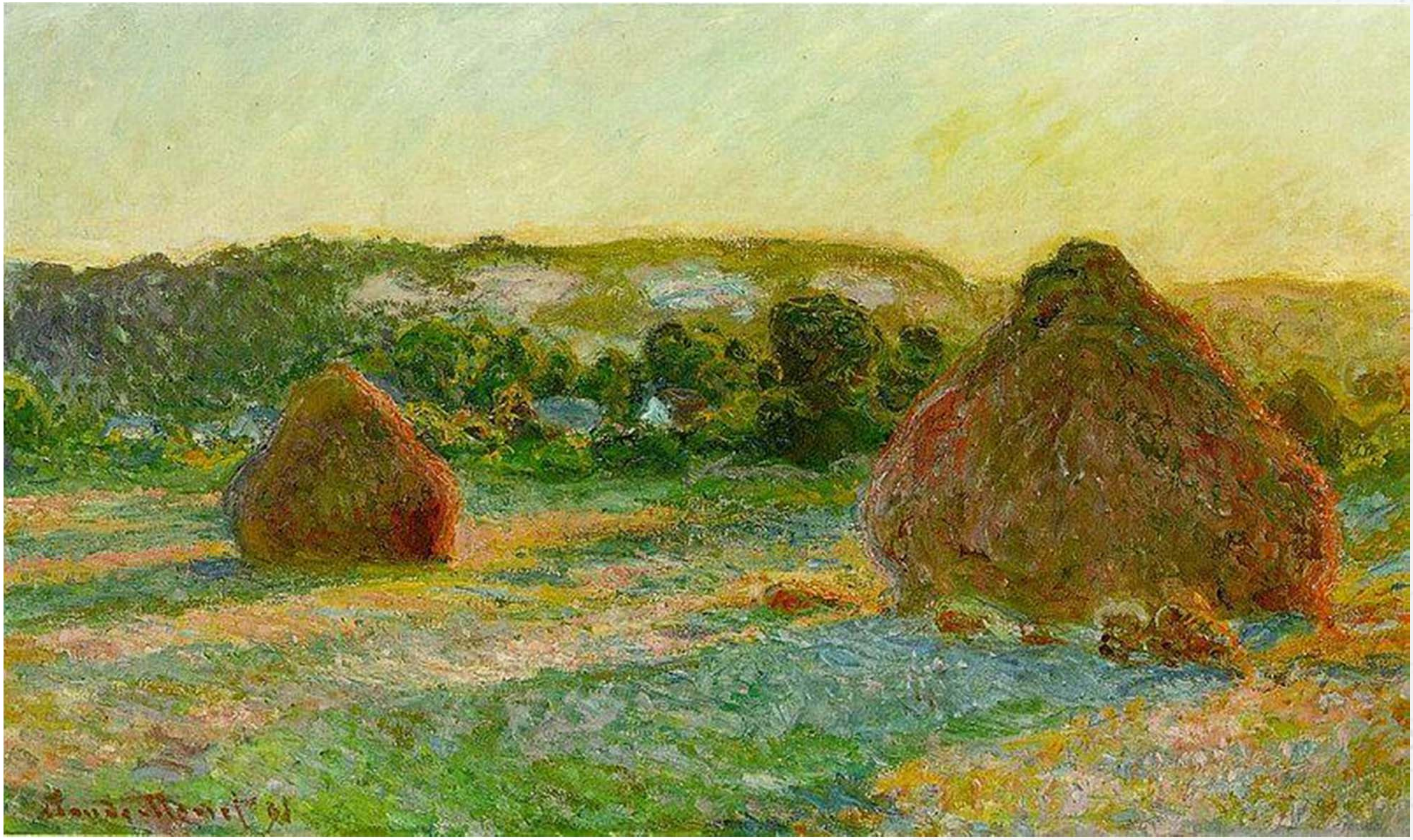


Table 1.1: How Big is an Exabyte?	
Kilobyte (KB)	*1,000 bytes OR 10^{3}bytes* 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
Megabyte (MB)	*1,000,000 bytes OR 10^{6} bytes* 1 Megabyte: A small novel OR a 3.5 inch floppy disk. 2 Megabytes: A high-resolution photograph. 5 Megabytes: The complete works of Shakespeare. 10 Megabytes: A minute of high-fidelity sound. 100 Megabytes: 1 meter of shelved books. 500 Megabytes: A CD-ROM.
Gigabyte (GB)	*1,000,000,000 bytes OR 10^{9} bytes* 1 Gigabyte: a pickup truck filled with books. 20 Gigabytes: A good collection of the works of Beethoven. 100 Gigabytes: A library floor of academic journals.
Terabyte (TB)	*1,000,000,000,000 bytes OR 10^{12} bytes* 1 Terabyte: 50000 trees made into paper and printed. 2 Terabytes: An academic research library. 10 Terabytes: The print collections of the U.S. Library of Congress. 400 Terabytes: National Climactic Data Center (NOAA) database.
Petabyte (PB)	*1,000,000,000,000,000 bytes OR 10^{15} bytes* 1 Petabyte: 3 years of EOS data (2001). 2 Petabytes: All U.S. academic research libraries. 20 Petabytes: Production of hard-disk drives in 1995. 200 Petabytes: All printed material.
Exabyte (EB)	*1,000,000,000,000,000,000 bytes OR 10^{18} bytes* 2 Exabytes: Total volume of information generated in 1999. 5 Exabytes: All words ever spoken by human beings.

Lyman, Peter and Hal R. Varian, “How Much Information,” 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on April 21, 2011.

10 Terabyte = 1 Library of Congress Unit

Claude Monet, Haystacks, 1890-91



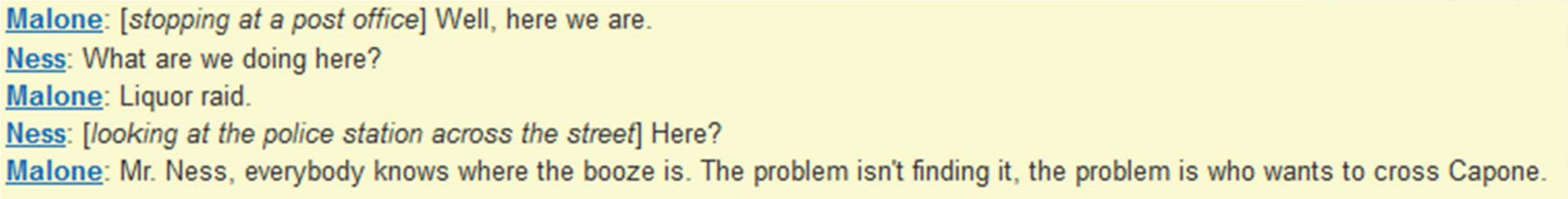
“Every case comes down to 20 critical documents.”

Poll Question No. 1

ESI is:

- A) The computer crimes branch of the FBI.
- B) Electronic information that requires translation to convert it to a useable form.
- C) A TV show based on the exploits of electronics investigators.
- D) The Electronic System for discovering Information.

Finding the critical documents in the Hay: Where do you start?



Malone: [*stopping at a post office*] Well, here we are.

Ness: What are we doing here?

Malone: Liquor raid.

Ness: [*looking at the police station across the street*] Here?

Malone: Mr. Ness, everybody knows where the booze is. The problem isn't finding it, the problem is who wants to cross Capone.

The Untouchables, 1987 (Paramount) (http://www.imdb.com/title/tt0094226)

In order to solve your ESI problems, you have to be able to define the problem.

The sooner that ESI is identified, the sooner that issues regarding how it will be handled can be addressed.

Finding the ESI: Where Is It?

- Thumb Drive/"Flash" Drive – Up to 32/64 Gbytes per stick.

4Gbyte thumb drive.
Promotional item of nominal value.
4 pickup trucks of printed text.
Some flash drives are small enough to be swallowed.

- Laptop/Desktop – Tens of Gigabytes up to several Terabytes per unit.

- Optical Disk/Compact Disk "CD/DVD": Up to 1-5 Gbyte/Disk.

Compact disk cases. Each case holds 320 disks (CD/DVD). Each CD holds 10-15,000 page images. Each DVD holds 50-75,000 pages. This shelf represents up to 240 million pages as TIFF images. Blank DVDs run about 35 cents each in quantity.

More ESI

- Network Attached Storage (NAS): Effectively unlimited.

Network attached server rack. 30 Tbytes, with power, data management, and switching.

- Portable Disk Drive

Up to 2 Terabytes. Powered directly from a USB cable, slips into a jacket pocket. Available for under $100.

- Recovery Media (“Backups”) and Archive (“Off Line”) Media:

AIT-5 Backup tape. 400Gbyte uncompressed, 25MByte/second transfer rate. Available September 27, 2006, Currently “End of Life” (obsolete) but a popular format for many years with many AIT systems still in use. A typical tape rotation may include daily, weekly, monthly, quarterly, and/or annual copies.

Yet More ESI

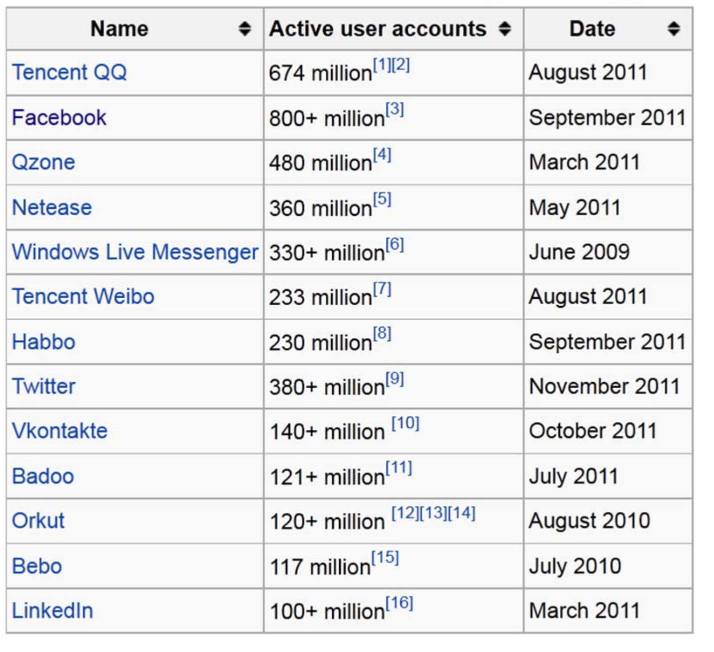
- PDAs, Smartphones (Text messages, E-Mail, Twitter): Up to 64 Gigabytes/Device.

I-Phone 4. 16 Gigabytes. Holds E-mail, text messages, music, videos, office files, and browsing history. Up to 64 Gigabytes on the highest end models.

- Tablet PCs (“I-Pads”), Automated Test Equipment (ATE); Security systems (Badge In/Out).

ESI That The Client Has Stored, well Somewhere: The Cloud and Social Media

- Social Networking



Name	Active user accounts	Date
Tencent QQ	674 million[1][2]	August 2011
Facebook	800+ million[3]	September 2011
Qzone	480 million[4]	March 2011
Netease	360 million[5]	May 2011
Windows Live Messenger	330+ million[6]	June 2009
Tencent Weibo	233 million[7]	August 2011
Habbo	230 million[8]	September 2011
Twitter	380+ million[9]	November 2011
Vkontakte	140+ million [10]	October 2011
Badoo	121+ million[11]	July 2011
Orkut	120+ million [12][13][14]	August 2010
Bebo	117 million[15]	July 2010
LinkedIn	100+ million[16]	March 2011

http://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_100_million_users

Unnamed server farm in San Jose, California.

Photograph: Bob Sacha/Corbis; http://www.guardian.co.uk/environment/2010/apr/30/cloud-computing-carbon-emissions. Fair use reproduction.

What do you do with it?

- Preservation – Does anything need to be done to protect this ESI against possible loss?
- Collection – Does this ESI need to be pulled out of the client's environment?
- Review – Can some of this ESI be filtered out before I look at it, and what portions of the collection require human review?
- Production – How do I provide the relevant, responsive ESI to my opponent and vice versa?

- How ESI is lost:
  - That old server? The bank took it back last month and sold it to an Indian company.
  - "I ran out of disk space and had to delete something."
  - "She left shortly after the suit was filed and her laptop was erased and reused."
  - "My e-mail inbox has a size limit and everything older than a year is erased."
  - "We decided to pitch those old tapes last week, they were taking up a lot of space and we switched to a new format three years ago anyway. Why do you ask?"

Many file types are of little or no use in most cases

File Name: Location of the file plus a file name plus an extension. Must be unique, even for files that have the exact same content: “C:\My Documents\Personal\Tax Returns\2002 Final.pdf

Folder Name: A file name that contains other files, including more Folders: “My Documents\”

Drive Name or Letter: “C:” – the logical name of a physical item of storage media. In Windows, “C:” usually means the primary internal disk drive of a desktop or laptop computer.

Filename Extension: the code that follows the “.” in the filename – this tells your computer what kind of file the file is. Common extensions:

.exe – Executable or Program file.

.doc – Microsoft Word file.

.pst – file containing e-mail items.

.pdf – Adobe Reader file.

.ppt – Microsoft PowerPoint.

.xls – Microsoft Excel spreadsheet.

Uninteresting file types can be eliminated with tools designed to flag system and program files. (De-NISTing)

Collection Strategies

- Self-Collection/Custodian Collection/"Targeted Collection"
  - The client or custodians follow instructions regarding what to copy, preferably using forensically sound copying tools. Instructions can range from "copy everything" to "copy the relevant files."
  - The custodian becomes a key witness if the adequacy of the search and methods used are challenged; can lead to either or both under collection and/or over collection.
  - Does not address preservation concerns.
  - Low cost, similar risks to an unsupervised paper collection.
- Vendor Collection.
  - A third party makes the copies, establishing chain of custody and technical correctness of the copies. The vendor is available to defend the methods and scope of the search.
  - The vendor's copies may be comprehensive enough to satisfy the highest possible preservation of evidence standards.
  - Higher cost, also likely to lead to very large haystacks.
- Guided Collection.
  - Using valid collection software, identified key custodians are interviewed and their data collections reviewed by a lawyer familiar with the issues in the case and the requests made by the opponent.
  - Relevant, responsive files are forensically copied for processing.
  - Irrelevant, non-responsive files are not copied.
  - Highest initial cost but may substantially reduce cost "downstream."
  - Does not address preservation issues.

Good news: Most of the raw data is truly hay
Bad news: There is a great deal of hay

- ***Many files have no evidentiary value in most cases***: A new laptop configured for a typical suite of ordinary office environment applications, before any data is put on it, will have at least several and possibly tens of Gigabytes of "data" that is entirely system related and contains no useful information.
- ***ESI tends to contain a large amount of duplication and near duplication***: Computer files tend to contain lots of duplication, intentional and unintended, and high levels of "near duplication" from multiple versions of documents that are very similar to each other.

Both of these phenomenon present tremendous opportunities for reducing the size of the haystack.

Metadata – ESI about the ESI

- Computers need to keep track of lots of information about electronic files so that the system can find and handle them appropriately.

- Information about who last accessed the file, when it was last changed, when and who created it, and what computer it was created on is all tracked and usually available. (“System Metadata” that is associated with, but not “in”, the data itself; used by the computer to manage the file.)

- Copying the file off your client’s computer without taking care to preserve the current state of the metadata can result in changing metadata that may have been pertinent, discoverable information.

Metadata – Data about the Data

- What you see off the printer:

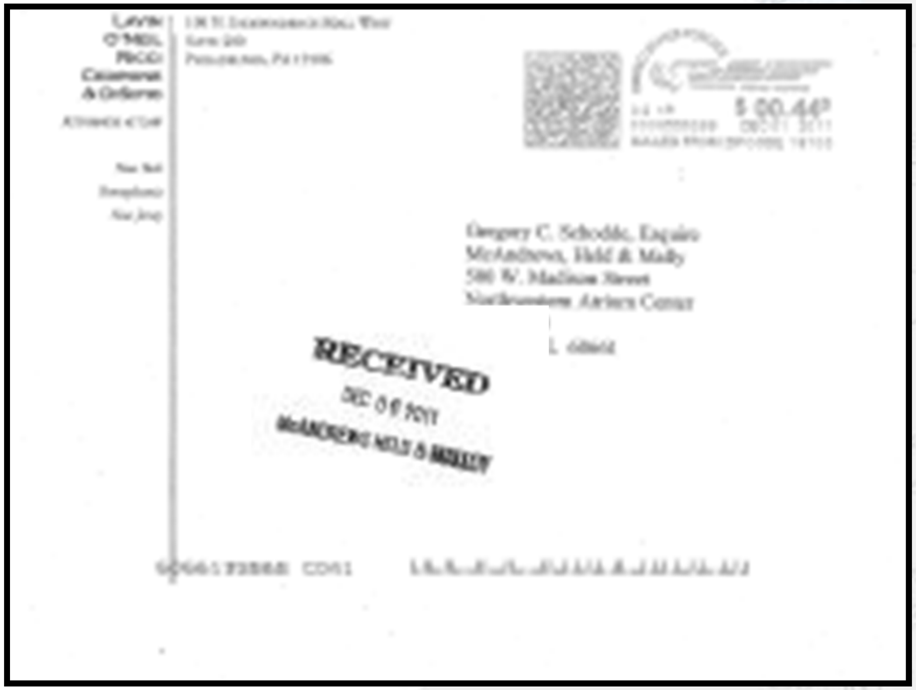
“Dear Greg:

Enclosed is our card.

Regards,

Jim”

- What the envelope shows:



Lavin
O'Neil
Ricci
Cedrone
& DiSipio

Attorneys at Law

New York
Pennsylvania
New Jersey

190 N. Independence Mall West
Suite 500
Philadelphia, PA 19106

$ 00.44
DEC 01 2011

Gregory C. Schodde, Esquire
McAndrews, Held & Malloy
500 W. Madison Street
Northwestern Atrium Center
L 60661

RECEIVED
DEC 0 6 2011
McANDREWS HELD & MALLOY

The envelope is to the letter what metadata is to ESI.

The Heisenberg Uncertainty Principle

Metaphysical discussions about preserving metadata can evoke the Heisenberg Uncertainty Principle:

“It can be expressed in its simplest form as follows: One can never know with perfect accuracy both of those two important factors which determine the movement of one of the smallest particles—its position and its velocity. It is impossible to determine accurately *both* the position and the direction and speed of a particle *at the same instant.”*

Heisenberg, W., *Die Physik der Atomkerne*, Taylor & Francis, 1952, p. 30.

“Embedded” Metadata

- Data that the application keeps track of, that may or may not be “visible” depending on how the application is viewed.

- Examples: Microsoft Word “track changes” view, comments view, PowerPoint “speaker notes” view, Excel Spreadsheet “formula” view.

- All of this information travels with the file when it is copied, and may or may not be discoverable or relevant: e.g., who made that last contract change?

Poll Question No. 2

Metadata, in the discovery context, is:

A) A data set containing 1 million bytes

B) Preliminary data that has not been processed.

C) Electronically stored information about the content of the electronically stored information.

D) Data that describes an individual.

De-Duplication & Gathering – Using Tools

- An algorithm can be run on a file that will generate a unique, or at least very likely unique, digital "fingerprint" that can be used to identify content-identical files even if they have different names. This fingerprint is frequently referred to as a "Hash" value, checksum, or security code.
- Two files with the same "Hash" value are almost certainly identical: The degree of certainty depends on the "strength" of the algorithm.
  - MD5 Hash: "Message-Digest 5" algorithm: 16 byte value, older standard.
  - SHA-2: Newer, stronger than MD5. Published by the U.S. Government. Generates values ranging from 28 to 64 bytes.
- Software tools can be used to locate and gather both "identical" and "nearly identical" files so they are all reviewed together.
- Software tools can also be used to "gather" together e-mail "threads" so that only one instance of the thread is reviewed.
- Software tools can also be used to exclude irrelevant e-mail – much like a "junk" filter can be used to exclude mail from on-line stores.
- Duplicate elimination, near duplicate gathering, and thread gathering ***all work best when the native file and the "metadata" are preserved and available***.

Additional Filtering Tools

- Date filtering – remove documents from further review by applying a date restriction. Only works if the embedded date metadata is accurate.
- Custodian filtering – limit review to documents that were kept or “touched” by particular custodians or work groups.

Forensic Copy

- Forensic copies are to e-discovery what photocopying is to paper – accurate reproductions of electronic data for preservation, review, and processing that faithfully preserve the original data and metadata. Sometimes referred to as a “bit copy” or an “image.”

- “Forensic” implies a copy that has a chain of custody traceable to the original media, that can be supported by testimony if necessary.

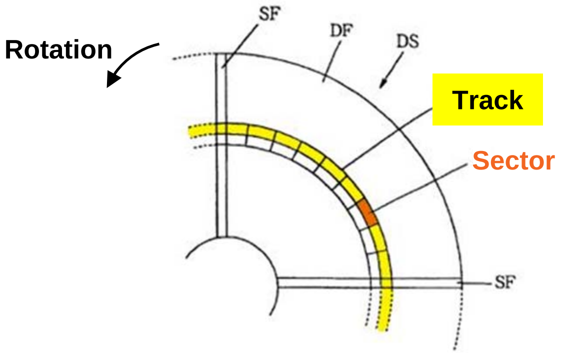
- Forensic copies can be made at any level – from individual files all the way to entire network servers. The cost of making the copies is typically much smaller than the cost of review.

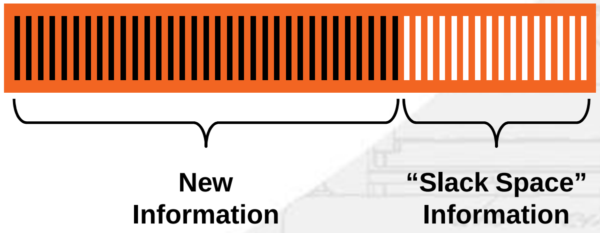
“I'm trusting in the Lord and a good lawyer.” *- Col. Oliver North*

- The paradox: It is very easy to delete files electronically, but very hard to really make them go away.
- Most routine deletion doesn’t delete anything – just moves the file to a different “folder” (affectionately the “trash”) for future deletion.
- In Outlook, even after a message is deleted, then the deleted items folder deleted, the message can still be recovered for a period of time (controlled by the system setting) using a menu option.
- When the “trash” is emptied, the system deletes references to where the data is physically stored on the disk – but the data is still physically available and can be read with special tools.
- Over time, new data may be written “over” the old data but it may take a long time and may never happen.
- Other ways “deleted” files can remain available:
  - A backup tape made before a file is deleted, will retain a copy of the file.
  - Unless every recipient and sender of an e-mail deletes it, a copy will be available.
  - Even if everyone on an e-mail string deletes every copy, backups of the mail server may contain copies of the message – and there may be multiple servers with multiple backups if some of the recipients are in different organizations.
  - A system restore or backup disk kept by an individual may have a “snapshot” of all files on that computer at the time the backup disk was created – which may include a “local” copy of the individuals mail folders.
- The use of special software to thoroughly and truly erase files off of a disk is detectable with analysis.

Forensic Tools: “Slack Space” & “Fragments”

- A magnetic disk drive is mapped into “sectors”, ordinarily 512 bytes up to 4kbytes/sector. Data is transferred in chunks that are often smaller than a sector.
- Big files are written to a series of sectors.
- Files are written from the beginning towards the end of a sector.
- Contiguous sectors containing one file make up a “fragment.”
- A file might be contained in multiple fragments.
- If the file has just been deleted, the “index” to where the fragments are located is updated but the “***fragments***” are still physically on the disk – and the fragments are still recoverable.
- Even if a sector is written over later, the new data may not “fill” the old sector, leaving “***slack space***” – and the bytes in the “slack space” are still recoverable.





Getting from “native” to “evidence”

- I have sized up my potential ESI haystack, now what?

The Latest Federal Rules Implore Lawyers to Address ESI Issues Early

- *First Discovery Conference: Under Rule 26(f), the parties must confer and are expected to discuss "any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced." Fed. R. Civ. P. 26(f)(3)(C).*

- *Initial Disclosure: Rule 26(a) requires that parties disclose without waiting for a discovery request "all documents, electronically stored information, and tangible things that the disclosing party has in its possession, custody, or control and may use to support its claims or defenses." Fed. R. Civ. P. 26(a)(1)(A)(ii).*

- *Initial Scheduling Order: After the Rule 16 pretrial conference, the court must issue a scheduling order which may "provide for disclosure or discovery of electronically stored information." Fed. R. Civ. P. 16(b)(3)(B)(iii).*

The Rules Provide A Gloss of Reasonableness Calculated to de-escalate the cost and scale of ESI discovery – or cost shift to the requestor

- *“A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost...If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.” Fed. R. Civ. P. 26(b)(2)(B).*

- *Under Rule 26 the court must “limit the frequency or extent of discovery otherwise allowed by these rules…if it determines that…the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties’ resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues.” Fed. R. Civ. P. 26(b)(2)(C)(iii).*

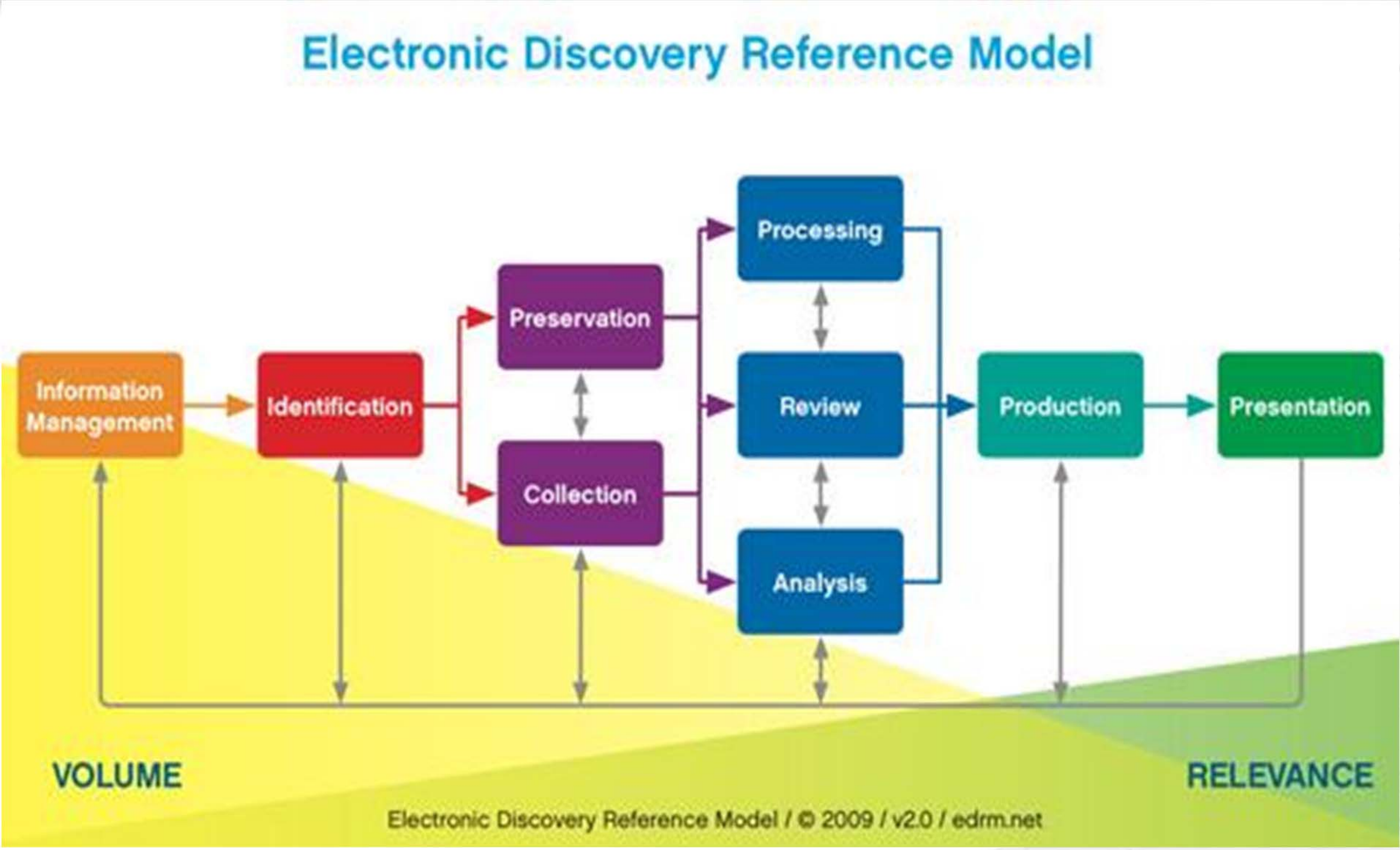
The Rules Provide A Moderating Gloss of Reasonableness Calculated to de-escalate the cost and scale of ESI discovery

- *A party may request any other party to produce ESI in native form or "<u>into a reasonably usable form</u>." Fed. R. Civ. P. 34(a)(1)(A).*

- *Unless otherwise ordered by the court or stipulated, Rule 34(b) requires that a party must produce its ESI as it is kept in the "usual course of business or must organize and label [it] to correspond to the categories in the request." Fed. R. Civ. P. 34(b)(2)(E)(i).*

- *Likewise, under Rule 34(b), If the producing party "does not specify a form for producing [ESI], <u>a party must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms</u>." Fed. R. Civ. P. 34(b)(2)(E)(ii).*

EDRM Model – “Shifting the focus to the left”



EDRM (edrm.net)

“Good Faith, Reasonable Search”

- Propose search strategies for ESI that minimize cost:

  “Some argue that e-discovery is best accomplished by taking large amounts of data from clients and then applying keyword or other searches or filters. While, in some rare cases, this method might be the only option, it is also apt to be the most expensive. In fact, keyword searching against large volumes of data to find relevant information is challenging, costly, and imperfect process. A much better approach is to ask key client contacts to help you locate core relevant information and then, by reading that information, determine other sources of relevant information.”

  -Kershaw and Howie, “Judge's Guide to Cost Effective E-Discovery”, eDiscovery Institute, October 1, 2010. at 4.

“Any document or thing, that is admissible, or likely to lead to admissible evidence . . . “ Fed. R. Civ. P. 26.

- There is no single “correct” search strategy for every situation.

- Different kinds of collections of data may lend themselves, appropriately, to very different review strategies.

Poll Question No. 3

“Slack Space” is:

A. Space between files on disks to keep them from running together.

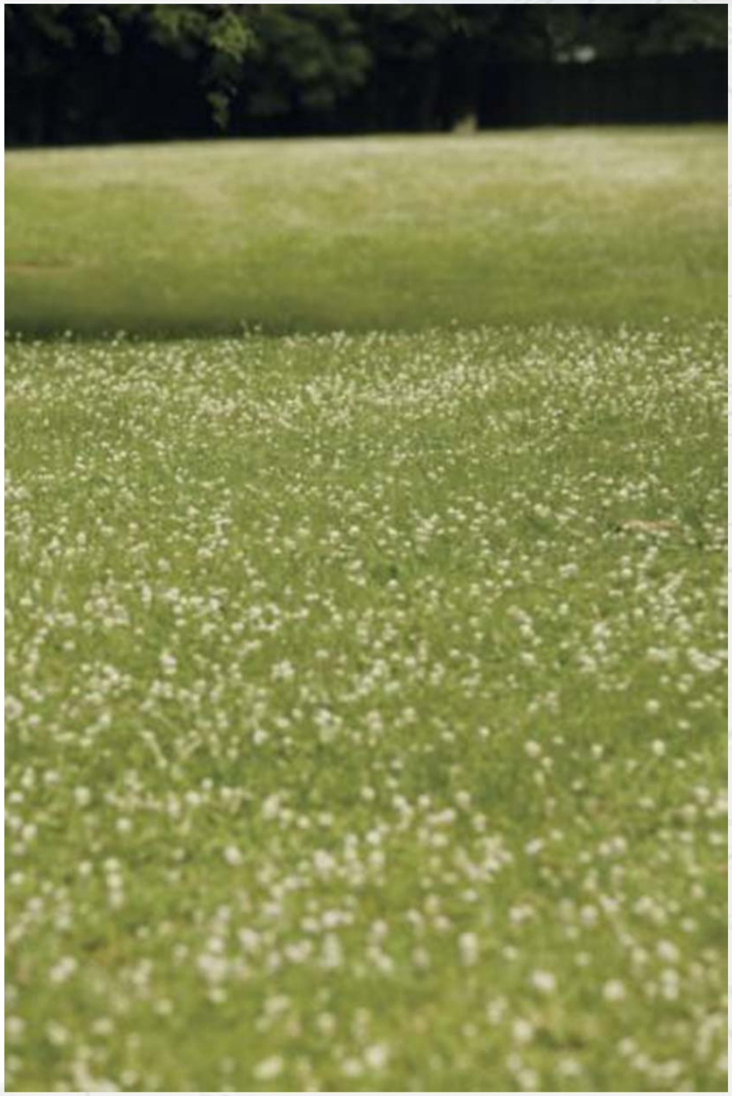
B. Time in the discovery plan to allow for the resolution of ESI disputes.

C. Space in a disk sector that still contains old information even though the sector has been partially written over with new information.

D. Space at the end of a sector that compensates for varying track length moving towards the outer rim of the disk.

Traditional Linear Search Strategies Do Not Work Well On Enormous, Relevance Poor Data Sets

- Find the relevant clover in this clover field.
- Linear search: Examine every clover, starting in the upper left corner of the field.
- Cost of linear review by average reviewers who must read for relevance and flag documents by issue, one pass, will run on the order of .50 – $1.00 per page examined.
- http://www.simplestroke.com/wp/?p=294



Strategies for Further Culling the Haystack

More Vocabulary:

Hierarchical storage: Files that are well organized/indexed. These are files that want to be found:

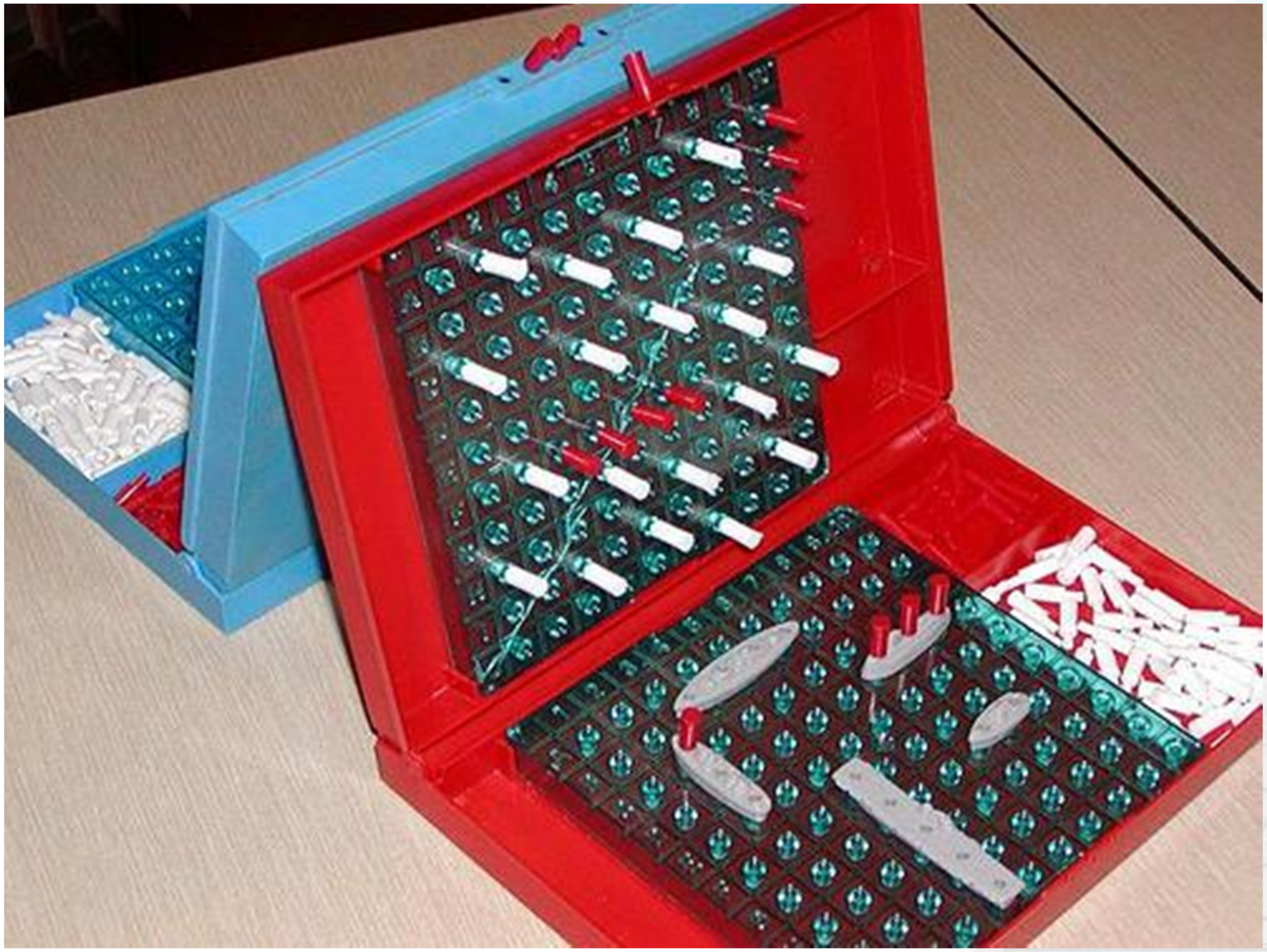
- “Z:\Accounting\Invoices\Research Expenses\Projects\Contracted Research\Relevant Project”
- Shared Network files and well organized individuals, frequently have this property.
- If you want “All files concerning matter number xxxxx”, you don’t need to do a keyword search to find it, and you don’t need to look at every file in the room – you use the index system and go straight to that file.

“If a cluttered desk is a sign of a cluttered mind, of what, then, is an empty desk a sign?” - *Albert Einstein*

Flat Storage: Large collections of unrelated items with no logical structure. These files are hiding, sometimes in plain sight:

Hierarchical storage lends itself to focused, selective review, flat storage does not. The most reasonable search strategy may vary depending on what you are looking for. An example of “flat” ESI – a large e-mail “Inbox”.

Simple Keyword Searching



Battleship, Children's Game, Milton Bradley.

How to find the needles: Filtering very large collections

- **Simple Keyword search.**
  - Any file that contains the last name of the Defendant.
  - In practice, simple keyword searching will have wildly varying results.
- **“Boolean” keyword searching.**
  - “Clover” AND (any form of “Whoville” OR “Horton”).
  - “Stemming”: Expands the number of potential matches by counting as a “match” any form of the word.
- **Statistical Ranked Retrieval**.
  - Takes a list of keywords, weights them using algorithms that assign relative importance to each word, then returns the documents that “score” the highest.
  - Doesn’t require the user to construct “Boolean” terms.
- **Machine/Supervised Learning.**
  - Tools that generate search criteria based on sample sets using a search engine that reinforces search terms that yielded documents determined to be relevant and weakens terms that do not appear to yield relevant documents.
  - Requires proper selection of the initial set of “training” documents.

What is a good search?

- How do I determine the quality of a search?
  - Precision: If your search strategy decides that 100 documents out of the collection are “relevant”, how many are actually “relevant?”. A 100% precise search will return 100 relevant documents. (100 Responsive Documents/100 Retrieved).
  - Accuracy/Recall: If your search strategy returns 100 relevant documents, how many relevant documents remain in the collection that your search did not return? If your search was 100% precise but only 50% accurate, then 100 relevant documents are still in the collection. (100 Responsive Documents Retrieved/200 Total Responsive Documents)
- In document production, high precision is desirable but less important than high accuracy.
- Determining how “good” a search was, particularly accuracy, may be required to defend the strategy in court especially if electronic filtering was used instead of “look at everything” linear review.
- You can determine for yourself if a search is precise – but the only way to determine if a search is reasonably accurate is to either canvas the entire set with highly competent reviewers (which defeats the whole point of doing the search in the first place) or sampling a statistically valid random sample to see if the “hit rate” was reasonably high.
- In large data sets, perfection should not be required – good search algorithms can be demonstrated to yield results that are at least as good as say, a team of contract document reviewers – but this is an area that usually involves very large data sets and experts.
- Precision and accuracy of any search strategy, will decline as the size of the haystack increases relative to the number of relevant files.

Production - Converting ESI to Forms Useful in Court

- Process of converting ESI to formats useful for Court.
- Typically means converting the “wild” ESI into “petrified”, page oriented images.
- Common format: File type “.TIFF” (“tagged image format file”) – Literally a “picture” of a page, like a digital photocopy.
- The “TIFF” images are processed in a document system such as Concordance, Summation, or CaseLogistics where your review work product (issue coding, notes) is added to the case database.

Can’t work with what you don’t have

- Most common business ESI documents come with text files that are easily searchable.
- If available, the “extracted” text should be produced with the “.tiff” image.
  - Many Adobe documents, all Word documents, all e-mail has a readily searchable text file associated with it.
  - Some Adobe documents, “.tiff” images, “scans”, do not.
    - These documents need to have a text file created for them using “OCR” (Optical Character Recognition) which is constantly improving but not perfect and in some cases, highly imperfect. (Handwritten notebooks).
- Key Metadata:
  - Dates (When), Custodians/Authors (Who), Parent/Child relationships (attachments) (What), File Path/File Names (Where)
- Typically, three files are produced for each ESI file: The image(s) of what the pages look like if printed or viewed, a machine readable text file containing searchable text from the original file, if available, and the “metadata” from the original files. These files plus the actual image comprise the “production.” Note that even if the “extra” text and metadata files have no evidentiary value, they add ease of handling value to document management systems.

The Joint ESI Plan

- "God is in the details." -Ludwig Mies van der Rohe
- Numerous technical details that are easy to address before significant collection has begun but very costly to address if they have to be re-done later:
  - Metadata fields – which ones?
  - How documents will be produced.
  - Office document settings for petrification from native to TIFF: The same document can be viewed several ways, how will it look in your production? How do you want their documents? (Power Point speaker notes exposed? Track changes turned on?)
  - De-duplication handling.

Electronic Discovery Resources

- Litigation Support or Information Technology staff – those computer guys in the back room, both in your office and your client’s office.
- Electronic Discovery Vendors or Computer Forensic Experts
- Electronic Discovery Reference Model (“EDRM”)
  - www.edrm.net
- The Sedona Conference®
  - www.thesedonaconference.org
- Seventh Circuit Electronic Discovery Pilot Program
  - www.discoverypilot.com

Electronic Discovery Resources

- Hold Memos/Preservation Obligations:

  Pension Committee of the Univ. of Montreal Pension Plan, et al., v. Bank of America Securities, LLC, et al., 05 Civ. 9016 (SAS) (S.D.N.Y. Jan. 15, 2010) (2010 WL 184312)

- District of Delaware Electronic Discovery Default Standard (presumptively excluding certain categories of ESI from preservation obligations) http://www.ded.uscourts.gov/SLR/Misc/Electronic-Standard-for-Discovery.pdf

- SDNY Proposed Joint Electronic Discovery Order (October 2011): http://www.nysd.uscourts.gov/rules/Complex_Civil_Rules_Pilot.pdf (Good checklist of issues to cover).
- Theory and studies of search tool effectiveness: www.trec.nist.gov.

For CLE Credit Requirements – Complete the On Line Survey That Follows

Thank You

Greg Schodde, McAndrews Held & Malloy.

Questions and Inquiries: ESI101@McAndrews-ip.com