Drowning in the Data Tsunami

Lee Damon
SSLI Lab
Univ of Washington
Seattle, WA
nomad@castle.org

Evan Marcus
Director, Tech Sales
QD Technology
Rutherford, NJ
evan.marcus@gmail.com

The Problems

- Disk is Cheap
- Information is Expensive
- Time is More Expensive
- Long term storage is easy
- Long term retrieval is hard

- Too much data
  - Can't find wheat in chaff
- Even when needed
  - Historic Record going away
- Getting back old data
  - Some data must go away

Threats to Data

- Age
  - Media wears out
- Readers go away
  - Got any 8” floppy drives around?
  - Who can you pay to maintain old hardware?
- Can’t decrypt data
- How do we find one piece of data?

Historic Perspective: Ancient Times

- Media
  - Rock: Hard to store much
  - Papyrus: High density, expensive
- Disincentives to storing meaningless data
- Long term record storage: no problem

STILL READABLE

Egyptian Hieroglyphics

Historic Perspective: Pre-Gutenberg

- Hand-made books
- Very high cost of entry, ownership
- Few could read or write
- Little incentive to store meaningless data
- Written words *meant something*
- Long term record storage: no problem

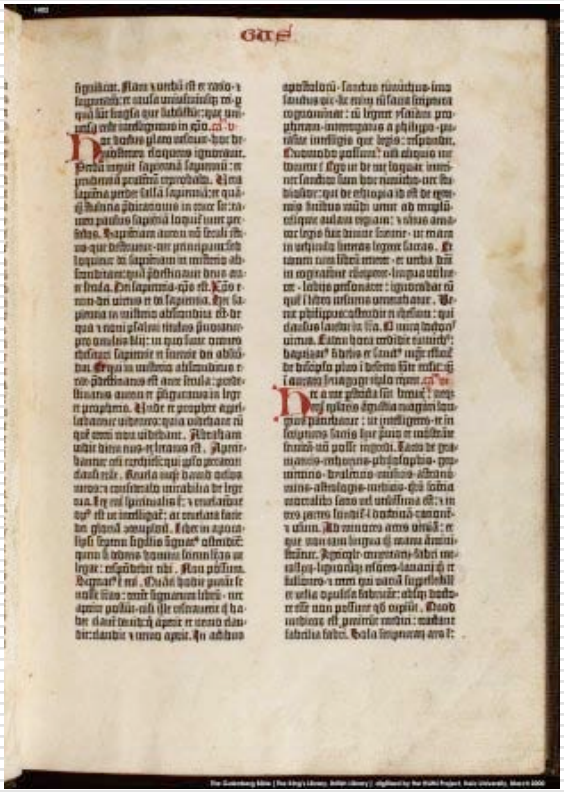
STILL READABLE

Image from the Kama Sutra, 1550

Historic Perspective: Gutenberg

- Cost of ownership still very high
- Easier to publish but still a barrier high enough to keep out the noise
- Long term record storage: mostly no problem

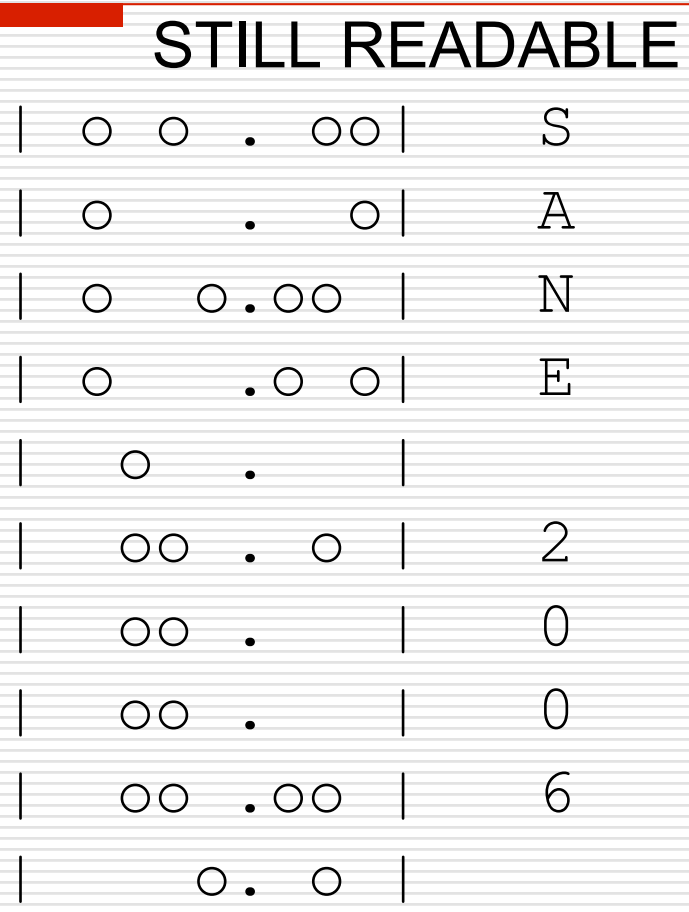
STILL READABLE



Gutenberg Bible

Historic Perspective: Punched Paper

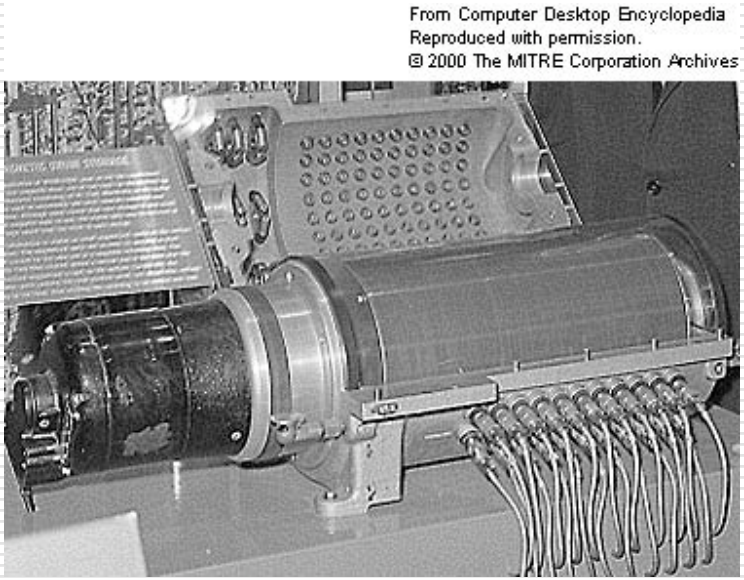
- Computers had limited memory, storage was bulky
- Still a disincentive to keeping massive archives with millions of cards
- Think "punch card ballot"
- Long term record storage: not a big problem



Historic Perspective: Magnetic Media

- Higher density -- store entire rooms full of cards on "a few" tapes
- Storing stuff "just in case" more likely
- Unlabeled tapes an issue
- Long term storage: hmm.. oh dear. 10-15 years?

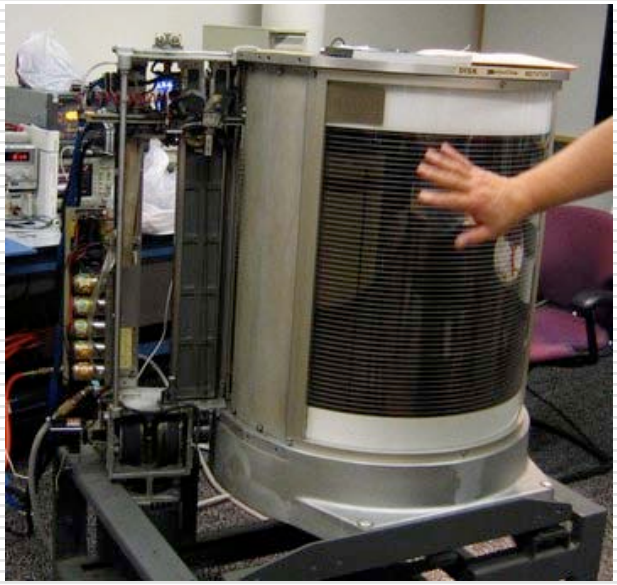
NO LONGER READABLE



Magnetic Drum
early 1950s

Historic Perspective: Remember 5 Megabytes

- Media:
  - Wall-o-disks
  - Washing machines, aka icebergs.
- Disk still very expensive
  - Large sites unlikely to put up with clutter
- Home PC user? That's another story.
- Long term storage: 5 years?
  - Do you still have that MFM controller?



IBM 350: 5MB of storage

Historic Perspective: Remember 5GB?

- "I'll never use all this space!"
- "Sure, I'll keep a backup copy of that document here, and in this directory, and in this one...."
- The beginning of the end, perhaps? (Ha!)
- Backups "keeping up" with disk still, but slow.

Historic Perspective: Remember 100GB?

- Tape backups can't keep up anymore
- Lots of space for "backup copies" - buy another drive and put it in a removable caddy
- Did you remember to label that drive?
- Long term storage:
  - Uhhh… What's the lifespan of a hard drive, anyway?

Today:
4.5+ TB for US$7000

- "I'll never use all this space!"
- Keep a copy here.. and a copy here... and a copy here.....
- LTO-3 tape drive: US$5K
- *How the hell are we going to back this up?*
  - More disks!
- Long term storage: Oops.

One Company's Data Tsunami

- SSLI Lab has grown from less than 1TB to over 13TB of backed-up storage in 5 years.
  - Plus 100s of GBs of scratch space on every desk
- Most 'data' is 'transitory & limbo space'
  - Research workspace for storing intermediate data/results.
- Still have tons of disks with unidentified data from before 2001.
  - Not worth sorting the "measly 120GB of stuff."

The World is Changing

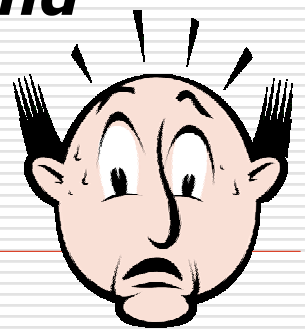
- Data must be preserved
  - Legal liability
  - Sarbanes Oxley
  - HIPAA
  - Federal Rules of Civil Procedures
  - Dozens of other regulations

What Happens When It's Lost?

- Morgan Stanley (2005)
  - Lost ***$1.45 billion*** judgment for losing emails
  - Could not find key email and data fast enough
  - CEO "retired"; firm considered acquisition target
  - Plaintiff seeking ***$2.7 billion***
- Citigroup (2005)
  - Lost tapes containing account info for ***4 million customers***
  - UPS accepts responsibility

What Happens When It's Lost?

- Bank of America (Feb 2005)
  - Lost computer backup tapes containing info on about ***1.2 million charge cards***
- Ameritrade (Feb 2005)
  - Tape containing account information was lost or destroyed in transit.
  - Affected ***200,000 current and former customers***
- Time Warner (May 2005)
  - Lost information on ***600,000 current and former employees*** back to 1986
  - Iron Mountain lost the tapes



What Happens When It's Lost?

- Citigroup Inc. (June 2005)
  - A box of tapes of personal info of ***3.9 million customers*** disappeared in transit to a credit bureau
- ChoicePoint (Feb 2005)
  - Identity thieves gained access to the personal information of up to ***145,000 U.S. residents***
  - They maintain a ***19 billion item database*** including Soc Sec numbers, driver's license and credit data
  - Brought before Congress

Regulatory Compliance #1

- Sarbanes-Oxley Act
  - Firms must report on the adequacy of the internal controls and procedures for financial reporting
- HIPAA
  - Health Insurance Portability and Accountability Act of 1996
  - Mandates privacy and record keeping for organizations that maintains health records
- NASD Rule 3010, 3110
  - Rules regarding records, retention, retrieval, non-rewritable storage, etc. for brokers and traders

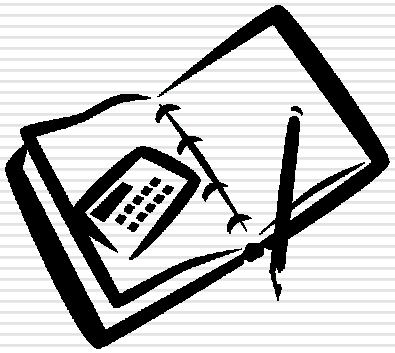
Regulatory Compliance #2

- Gramm-Leach-Bliley Act
  - Privacy and information sharing from financial institutions
- SEC 17a-3, -4
  - Mandates record keeping and duration
- 21 CFR Part 11
  - FDA regulations related to electronic document management and e-signatures
- International Regulations
- Other industries

Getting Prosecuted? Getting Sued?

- Winning isn't always a victory…
  - Average cost of pre-trial discovery: $1.3M
- But you really don't want to lose
  - Average SEC 17a fine (2004): $1.6M

Others Ways that Archives Matter



- Research and Development
  - Pharmaceutical
  - Seismological
  - Medical
  - Just about any kind of research
- Data preservation
  - Digital movies and video
  - Digital music
  - Digital photographs

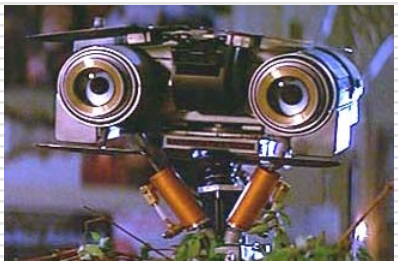
What is an Archive?

Basic Functions of an Archive

- Ingestion
- Preservation/Protection
- Access

Ingestion

- Appraisal
  - Is this the right archive for the records?
  - Are there duplicates? What to do?
  - Determining and setting retention

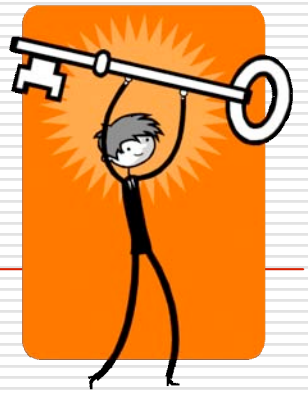


- Record metadata
  - Record how and when records were added
  - Record author and owner of the records
- Disposition
  - Do records need to be on site or remote?
  - Should the records ever expire?

Data Preservation

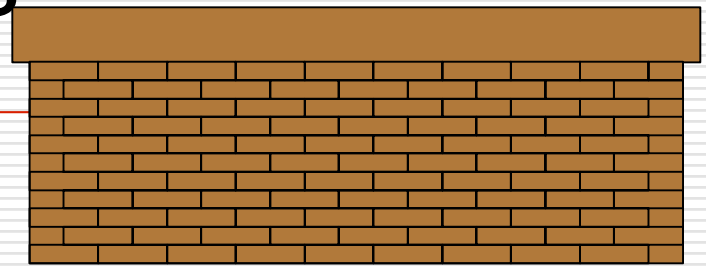
- Integrity
  - What condition are the records in?
  - Should they be transcribed to a new format?
- Preservation
  - What are the environmental needs of the records?
  - What type of enclosure is required?
  - Ensure what gets stored is what gets retrieved
- Security
  - What type of security controls are required?

Accessibility



- Policies
  - What type of access policies should records have?
- Arrangement
  - Group records by their source
- Description
  - A finding aid, and description of the record group
  - Can be online & searchable
- Retrieval
  - Search and locate desired document/information
  - Retrieving in a useful form

Traditional Archives



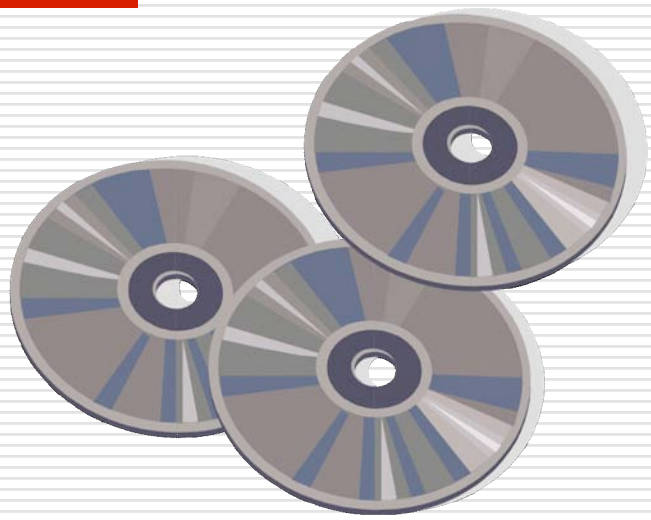
- Brick and Mortar
- Run by team of professional archivists
  - Organize and place the documents
  - Reject inappropriate documents
- Consumes large amounts of space
- Difficult to search quickly

Some Traditional Archives

- Your Local Public Library
- Municipal Hall of Records
- National Archives
  - Washington, DC
- Library at Alexandria (ancient Egypt)
  - Created 3rd Century BCE
  - 400,000 - 700,000 scrolls
  - Burned and looted in 3rd or 4th century CE
    - Historical details are unclear and in dispute

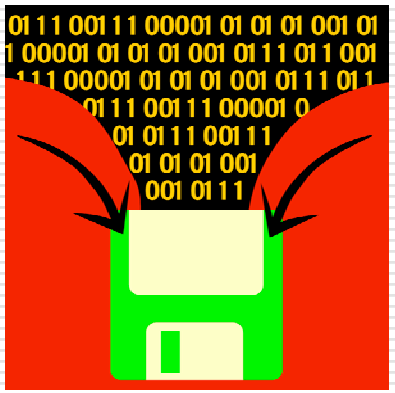
Data Center Archival Media

- Tape Drives
- Optical Disks
- DVD-ROMs
- CD-ROMs
- RAID Arrays
- NAS
- SAN



Data Center Archiving

- Traditional Methods
  - Backups
    - Magnetic tape
    - Optical disks
    - Spinning disks/NAS
  - Shipped off site
  - Stored & preserved for years



- "We take backups once a month and send them offsite."
  - Iron Mountain
  - Someone takes them home

Are Librarians the Answer?

- 1000s of years of experience at data collection and cataloging
- Deal with "finished works" more than we do
- Having data-finding problems of their own
- Let's join forces

A Librarian's Take

•"From my perspective of expert user, what computers (in a general PC sense) OS's haven't done well is offer a good system of document control. File management is all well and fine but where is the indexing system that helps us control the "aboutness" of the document. Library cataloging systems were previously all about 'aboutness' (because prior to non-print items, paper format was stable)."

•"Now we have a situation where the gurus of organization & aboutness (librarians, archivists, museum curators, information professionals with other titles) and the gurus of digital format (computer professionals) are starting to come together to provide interdisciplinary expertise and follow the holy grail of one-stop shopping. Welcome to Metadata land."

-- Friday V. Librarian @ Large

A Library Solution

•“Libraryland deals with this by having preservation committees, disaster recovery plans, and a fair amount of system redundancy that businesses won't tolerate. but one of our big cultural jobs is being the knowledge keepers, so we do things a little differently.”

-- Friday V. Librarian @ Large

A Librarian's Solution

•"For me, managing information is about metadata (and the standards that describe it)...by having the information about the information in order to do tracking and maintenance. Be it traditional cataloging of books, modern multimedia collections, or 'simple' databases of directory information, designing for use/update/delete is important."

-- Friday V. Librarian @ Large

Some Library Solutions

- The digital initiative folks at UW:
- http://digital.lib.washington.edu/staff.html

- For Digital Preservation libraryland is adopting this LOCKSS model (lots of copies keep stuff safe) which has turned into a software product:
- http://www.lockss.org/

- Multimedia asset management, one option is ContentDM (www.contentdm.com) which was developed at UW and spun off to become its own company.

Google -- a solution or a problem?

- Put it all on the web server and let Google index it.
- Backups? No problem. Tar it all up, encrypt it and put it on the web server for google to cache.

In House Google: Publishing locally

- “That thar Intar-web thingie”
- Buy a Google box for in house work?
- Who controls the index?

Data Management Systems

- Targeted solutions
  - Change Tracking Systems (cvs, subversion)
  - Document Management Systems

Google Desktop

- Desktop search systems like Google desktop & spotlight
- Do you let it “off the system” for sorting/indexing/storage?
- Meta-index of the meta-index

Database

- As filesystem?
- As file pointer?

Black Box Bombastication

- Can vendors like NetApp help?
- Where’s my stick? Darn it, I need to do some thwappin’!
- Good opportunity for a new business (that wants to be bought by Google someday).

It’s a Perception Problem

- People view their short-term “being busy” as more important than the long-term ability to recover/restore/search/identify
- Human problem, cultural problem
- Balancing point needs to shift

Data Generators

- We can give data generators all kinds of technical “amazing wonders” but if they don’t perceive the need it won’t do any good.
- Education is a strong word in this context, but it is the important word.

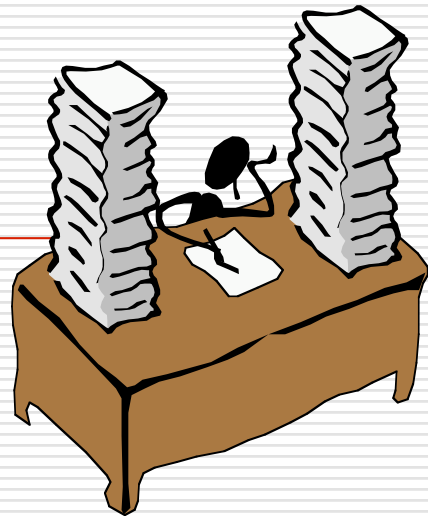
Storage Experts

- Telling people “the disk is getting full” doesn’t help.
- Telling people “how will you find that data in 6 months?” doesn’t help
- What **do** we tell them?

Do We Really Need All This Data?

- 6 months from now?
- A year?
- 5 years?
- 25 years?

Real-Life Clutter: Standard Wisdom



- Set it aside for six months
- If you don't use it in that time,
  - THROW IT AWAY
- Doesn't really apply in the world of data preservation…

Meta-Data

- Expiration time
- Summary notes
- Applicability
- The return of the dreaded “resource fork”?

In Summary