

Electronic Records Assessment Team

Raiding the Lost Archives



Making that old data new again.

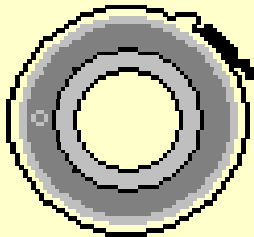
Presented by

Muller Media Conversions, Inc.

www.mullermedia.com

July 7, 1999

They've Got This...



Wrong Media?

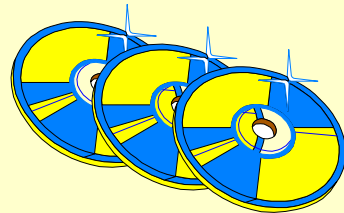
Wrong Format?

For example, older Agency electronic records are presently on media (wrong kind of tape or disk) and/or in a logical format (e.g.- Wang, DEC, Unix, IBM, etc.) incompatible with NARA standards.

But NARA Wants That...



Verified NARA-standard-format tape cartridges with ANSI-standard file formatting.



CD & CD/R are now accepted in some cases.

We'll discuss the problems and possible solutions that can assist you and your client agencies to bridge the gap.

TIME and its minions are the *“Phantom Menace”*.

TIME and fragile storage media can cause valuable records to "decay" on the shelf.

TIME and the inevitable migration to new computers, software and media renders older electronic records incompatible and unusable on the new systems.

TIME, down-sizing and job-hopping programmers lead to undocumented programs and files that are difficult to decipher.

FACT: A greater portion of our national information resources is threatened by these factors than by terrorists and hackers.

WHY FIGHT IT?

Intrinsic Value,
Historical Research,
FOIA

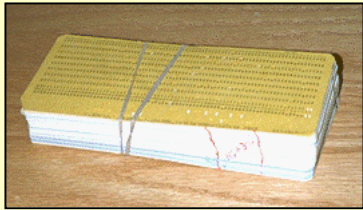
Legal Mandates,
Electronic Evidence

Obstacles to Preservation and Future Access.

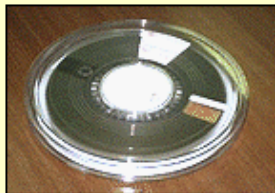
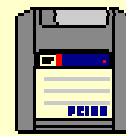
It's a bit like peeling back the layers of an onion:

- **Media**
- **Age and Storage Conditions**
- **Recording Method**
- **Operating System/Filing System**
- **Backup, Exchange, or Archiving Software**
- **Application File Structure**
- **Application File Encoding**

Media



8" and 5.25"
hard/soft sectored
single/double sided
single/double density
varying sector sizes
interleaving



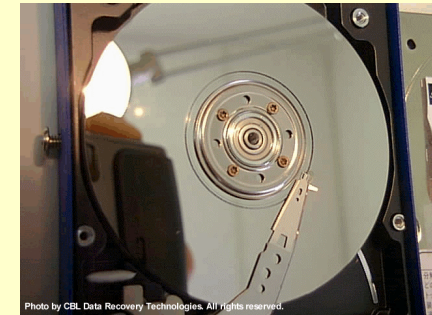
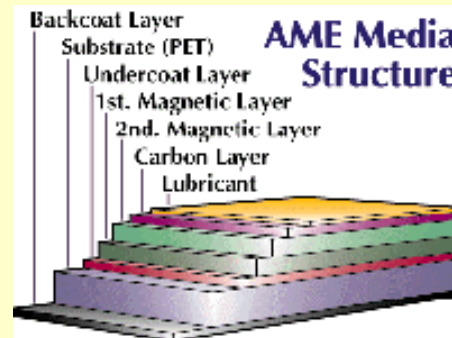
Age & Storage Conditions



Gravity



Heat, Humidity



Dust

Recording Methods (cont'd)

Lighter-Duty Tape Drives (mostly PCs, small Minis, Servers)



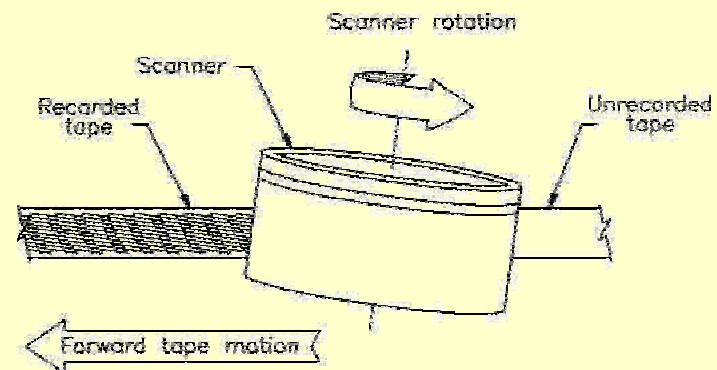
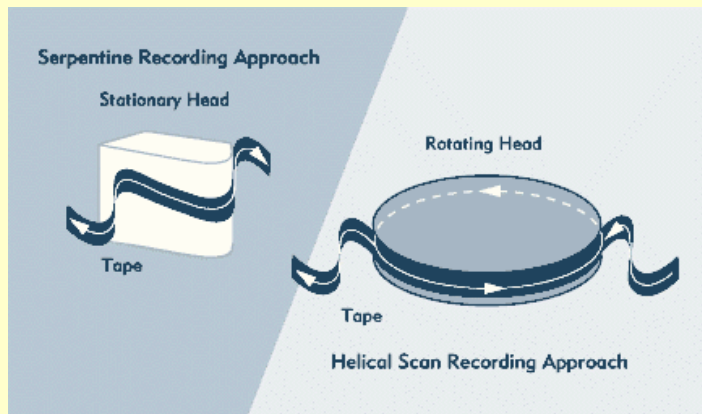
QIC serpentine recording 4 track, 9 track, 18 track , etc. from 11mb to 24 gb (also QIC-80 mini-cartridge, Travan, etc.



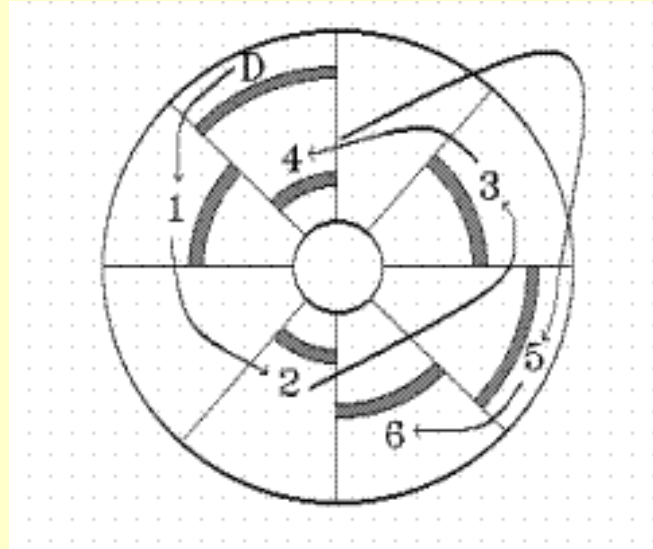
8mm (Exabyte) helical scan recording 2.3gb ... 5/10gb ... 7/14gb ... Mammoth ... AIT



4mm DAT helical scan recording DDS, DDS-C, DDS-2, DDS-3
capacities range from 1gb to 24gb.



Operating/Filing System



The way files are stored and identified-for example:

- How disk drives and directories are organized.
- Folder/directory organization.
- How long its name can be.
- What kind of characters may be included in the name.
- File types and their associations.
- Whether and how creation date or date of last access are saved.
- How space is allocated for a file.
- Indexed (ISAM) file operations--part of O/S?

Tape Backup Software

(What program was used to write data to tape?)

Software optimized for backup is generally not good for archiving or exchange.

Backups done in “image” mode are very hard to use on other systems.

Software intended for data exchange (e.g.- ANSI or IBM Standard) generally produce the most readable tapes, but you can lose valuable metadata that way. Also very inefficient for many small files.

Next best choice is to use the O/S’s standard backup program, such as Wang VS Backup, VAX VMS Backup, Windows NT Backup. (Conversion software or services may be needed.)

Application File Structure

Database, wordprocessing and other complex file types are generally not organized sequentially, even within each file.

Most databases use some form of index-sequential organization.

Wang and Multimate wordprocessing files use an internal chain to control their organization. (WordStar, WordPerfect and some others are essentially sequential.)

Microsoft Word and many newer applications store information within a number of data streams using a technology called *OLE Structured Storage*.

One generally wants to get databases into a flat or delimited sequential format prior to preservation.

A good sequential storage format for wp files is *RTF*.

Application File Encoding

DATABASES: ASCII, EBCDIC, BCD, PACKED, FLOAT, INTEGER, etc.

WORD PROCESSORS: ASCII as a base. (Except for some IBM). But that's just the tip of the iceberg. Not only the codes vary but the way in which they're used. One simple example:

An <u>underlined</u> word.	It may print like that, but internally it might be:
An <u>underlined</u> word.	Extra bit set in u/l'd chars (Wang)
An ➡underlined➡ word.	Same code toggles on/off. (WordStar)
An ♠underlined♥ word.	Different code starts/ends.
An underlined ^W word.	Word underscore code (IBM)
An u<_n<_d<_e<_r<_...word.	Char/backspace/'score (really old ones)
An underlined word.	Attribute pointers from different part of file.
^.....^	(MS Word)
^.....	

1970's and early 1980's

“Document Conversion” Myriad systems, e.g.- WANG, DEC, DG, HONEYWELL, PRIME, NBI, CPT, MICOM, ATEX, IBM (5520, DISPLAYWRITER, ATS, ATMS, DISOSS), etc., etc.

- Disparate, often proprietary media
- Complex proprietary file formats
- Small but numerous files.
- Related “metadata” (summary or profile info)
- Applications: migrations, financial printing.
- *“Six Reasons Why This Disk Won't Work in That Computer”*

“Data Conversion”

- Large but less numerous files.
- Simpler (yet nearly every one is different)
- 9-track reels, mostly
- ASCII, EBCDIC, BCD, PACKED, FLOAT, INTEGER, etc.

War Stories

Thrill of Victory--Agony of Defeat

FEDERAL RECORDS—A RAY OF LIGHT. The development of **APS** resulted in the **1996 GSA Technology Excellence Award** for section leader Fynette Eaton.

A HORROR STORY. A federal agency was migrating to a new system. Converted several thousand “legacy” documents at one installation ...Got rid of the old system. Only the first page of every document had been converted.

Lesson: keep the old media till *absolutely certain* conversion is OK..

War Stories (continued)

The IRS and Company X

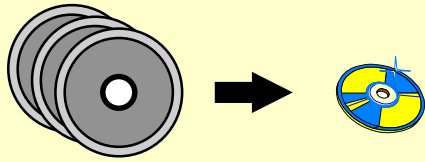
Problem:

- 1,000 reels of 9-track computer tape.
- The format not very well documented.
- Some of the tapes had developed bad spots.
- IRS demands firm preserve the data for future inspection.

Solution:

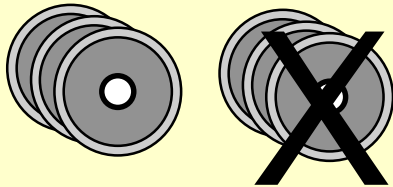
- Preserve the data on CD (two copies)
 - easier access
 - better shelf life
- Data was preserved on CD
 - in original format
 - in tape images
 - as converted to ASCII
- Software tool provided to re-convert as needed.

The IRS and Company X (continued)



Feature: **Improve Archival Quality while Reducing Space.**

While tape makes a convenient backup medium, it is not a particularly good archival medium. Temperature, humidity, magnetic fields and gravity all affect a tape's storage reliability to a much greater degree than they would affect a CD. Each CD can hold the contents of several 9-track or 3480 tapes with superior archival characteristics, so you gain both in reliability and space savings.



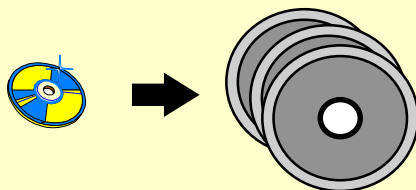
Feature: **Eliminate Unproductive Work.**

Many users of offsite data storage will create an extra copy of their backup tapes just for the purpose of sending them offsite. (Often, there is no practical way to check either the original tape set or the copies for errors.) Processing the original tapes saves an extra step, and eliminates uncertainty about the original tapes as described below.



Feature: **Eliminate Uncertainty about Original Tapes.**

By copy data to a preservation medium, you not only have your vital data copied to a more compact and stable medium, but in that process, one can *confirm the readability of the original tapes*. (If you had simply shipped them to a storage facility, you could not be certain whether there were any unreadable spots to begin with--and the tapes would be further subject to deterioration on the shelf.)



Feature: **Conveniently Regenerate Any Original Tape(s).**

By keeping both the converted data and each tape's "image" on CD. One can then reproduce the original tape bit for bit--should that ever become necessary.

The IRS and Company X (continued)

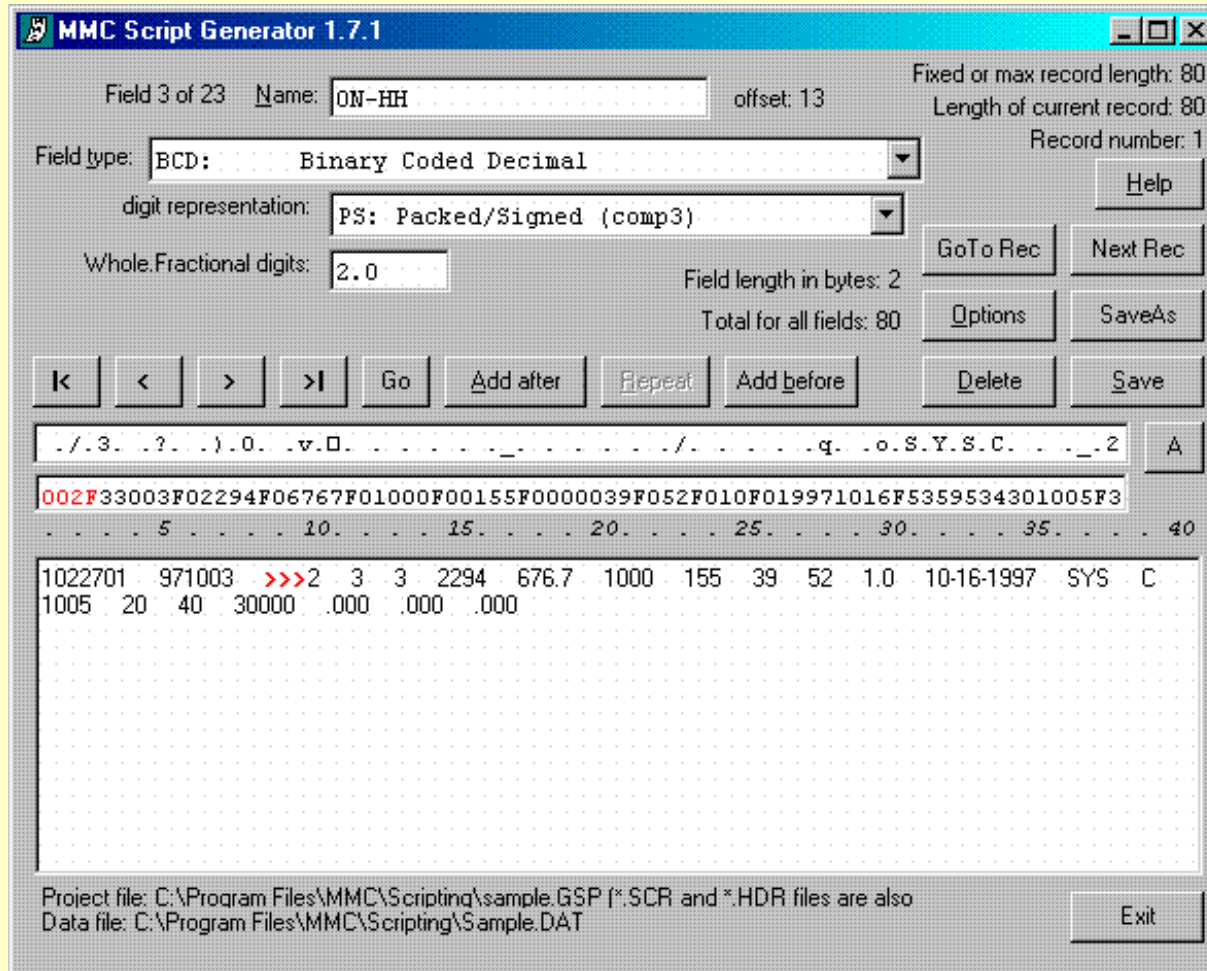


Illustration of tool used in the Company X scenario.

Great for “hacking” undocumented file formats, and for quickly setting up translations.

(Available in [APS](#).)

Law Suits = Suddenly Important Data

#1 - State Department FOIA

Problem:

- Presidential candidate may have traveled to Soviet Union for peace demonstrations.
- FOIA plaintiff demands passport records.
- U.S. Attorney's office receives tape and no documentation.
(We call this the *software bus syndrome*.)

Solution:

- Use APS-like software to read IBM-style tapes.
- Use Scripting Software tool to “hack” and then convert.
- Import to Excel and format nice report for judge.

Law Suits = Suddenly Important Data

#2 - Polymer Patent Suit

Problem:

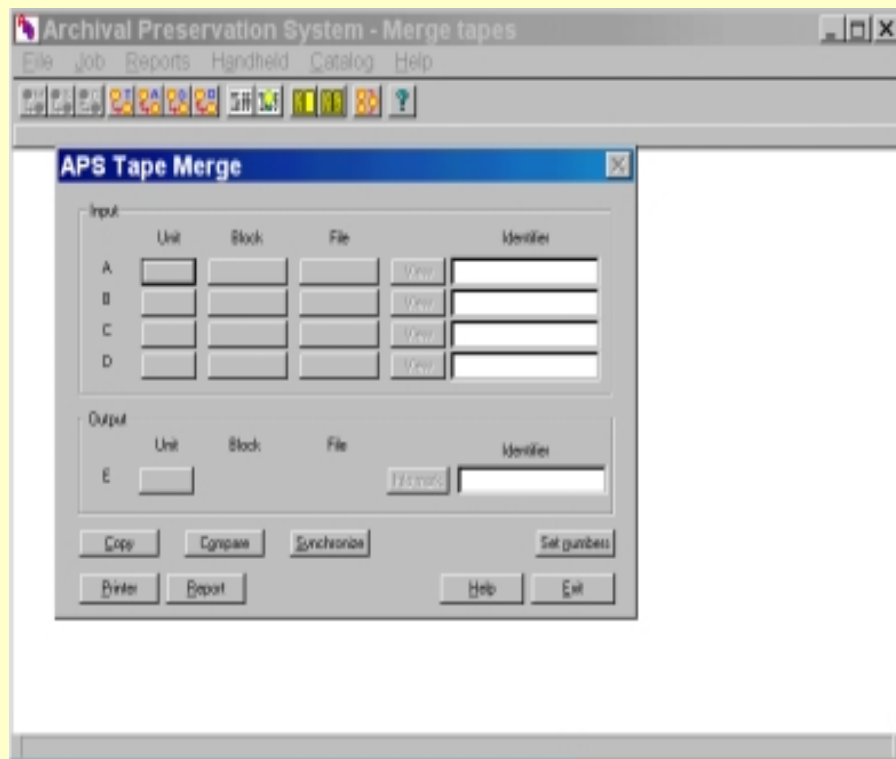
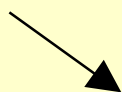
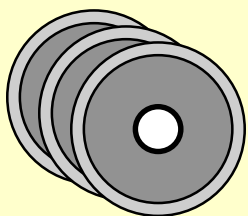
- Evidence is on 300 reels of tape in proprietary format.
- Some tapes have deteriorated.
- Rocket Docket will tolerate no delays.

Solution:

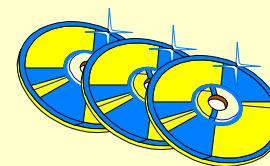
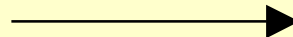
See steps on following slides.

Law Suits = Suddenly Important Data

#2 - Polymer Patent Suit (continued)

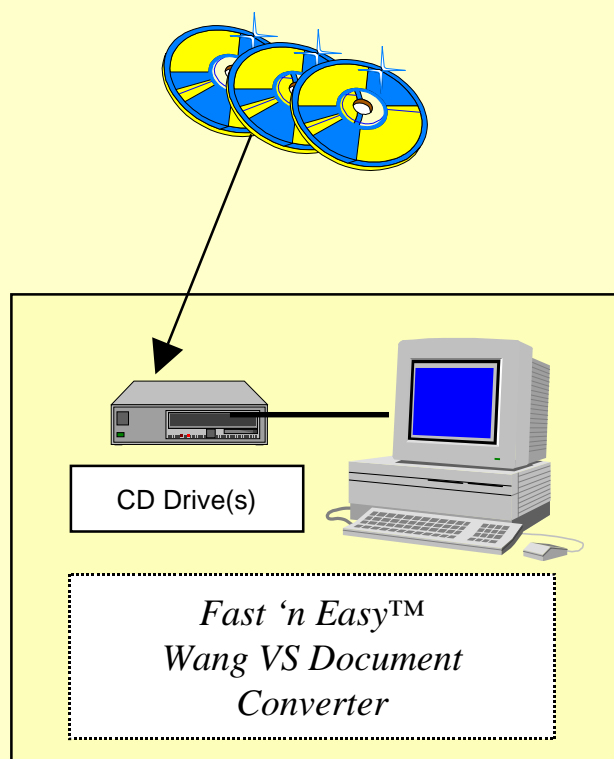


- ➔ First, the 300 reels of tape (some quite old) are copied to “tape image” files. (A technique originally developed by MMC in 1985.)
- ➔ Tapes with bad spots are read on different drives, merging the maximum amount of valid data. (Employing the APS merge feature developed for NARA.)
- ➔ Images finally written to CD for backup and preservation.



Law Suits = Suddenly Important Data

#2 - Polymer Patent Suit (continued)



- Next, the 300 “tape image” files are processed as if they were the original tapes.
- All document formatting and **meta-data** are preserved during this very fast, very clean conversion.
- New document files are optionally written to a folder and naming structure that mimics the legacy system.
- Meta-data “profiles” are produced as a by-product.

Law Suits = Suddenly Important Data

#2 - Polymer Patent Suit (continued)

Metadata “profile files”

Fast 'n Easy(tm) Legacy Document Catalog

File Window Help

Search Legacy Document Catalog

File ID: AF000295 Library: JLL8146R File Type: WPP Tape: A04

Title: Affidavit Battery Park City File #: 9344

Author: 8146/London Created: 9/13/88

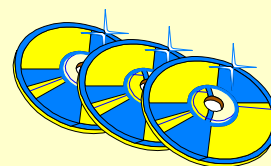
Operator: 00948:059 Revised: 9/13/88 Other Details

Go To

Record 3 of 20 Found Mark Reset

Search Results: 20 of 11249 Profiles

- Profile information is fed into Legacy Document Catalog.
- Operator can search by partial fill-in of any of the properties-
-title, author, operator, creation date, last edit, etc.
- Selected documents are retrieved from CD.



Law Suits = Suddenly Important Data

#3 - Whitewater “Vacuumed” Files

#4 - China Export Controls FOIA

- **WHITEWATER--“VACUUMED” FILES.**
Little Rock law firm exercising “due-dili”.
WW-related files rumored to be deleted from disks.
Yup. Recovered and converted.
- **ANOTHER FOIA: TECHNO-EXPORTS TO CHINA.**
Judge orders e-records from 1991 through 1996.
Three different kinds of tape media (200+ tapes).
Recorded in a variety of densities.
Four different backup programs.
Email and wp files from three different generations.
Must render into common form for searches & perusal.

Ensuring Future Access to "Old" Data

Promising developments:

- ANSI/AIIM MS66 standard.
- Research into preservation formats and methodologies at SDSC.

What about the ocean of non-conforming legacy data?

Standards are vital--targets for conversion.

Hippocrates' Rule for Archivists: First, Preserve!

- Copy *by duplicating*
- Confirms first copy was readable.
- Need not be same media--but be able to re-create.
- Compare the two.
- Then take further steps to render for researchers.

Convert, But Wisely

Heavily Formatted and/or with Metadata:

- Original Format (e.g. - Wperf 5.1, Wang, VAX)
- Editable Format (e.g.- Word, later Wperf, RTF)
- Display/Search Format (e.g. - Acrobat PDF or RTF)

- Or, Keep Original (dup'd) and Conversion Tool

(Remember the “horror story” from earlier.)

Transfer Medium and Preservation Medium

The 'Net might be the way to transfer federal e-records
—if these criteria can be met:

- Files can be transferred/converted quickly, reliably from legacy media to a computer with Internet access.
- Be sure that the file arrived at the new custodian (and when); signed return-receipt.
- Further assurance that the file has not been modified.
- Sender and receiver both authentic.
- Data will only be read by intended recipient.
- All transactions must be logged; with easy reporting.
- Optimized for larger files than regular email.
- The process must be easy to operate and administer.
- It must be cost-effective (of course). There's the rub.

Muller Media Conversions, Inc.

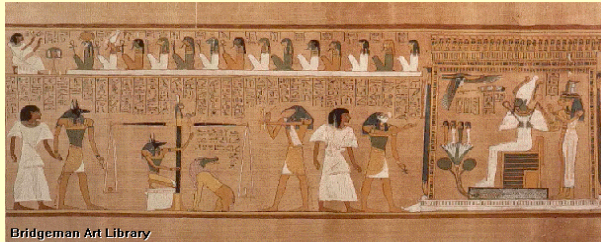
providing clients with *can-do* solutions since 1978

32 Broadway, New York NY 10004

(212)344-0474 fax: (212)968-0789

www.mullermedia.com

Raiders of the Lost Archives



“Archeologists to the Computer World”

Crain's New York Business, 1991

***MMC... A Key NARA Resource in the Past;
A Greater One for the Future.***

(After all, *The Past Is Prologue*)

Muller Media Conversions, Inc.

providing clients with *can-do* solutions since 1978

32 Broadway, New York NY 10004

(212)344-0474 fax: (212)968-0789

www.mullermedia.com

Raiders of the Lost Archives



“Archeologists to the Computer World”

Crain's New York Business, 1991

- **Preservation** MMC developed the *APS* systems which are used at NARA's Center for Electronic Records to copy, convert and preserve federal electronic records. Our intimacy with *APS*, coupled with conversion and deciphering skills can set the stage for proper preservation of even the most difficult electronic records.
- **Secure Transfer** A new suite of products and services based on FRX™ - a *secure, authenticated, high-volume internet transfer* of Electronic Records, using CSI's unique BDE® technology.
- **Conversion** Very clean, fast, high volume “*legacy*” *conversions* (e.g.-Wang, DEC, Honeywell, IBM).
- **Deciphering** Twenty years of file format “*hacking*” and software development experience enables our staff to tackle anything from unscrambling *FOIA* data or *Electronic Evidence* to *document management* migration.