


## Eliminating E-Trash and Non-Records from Shared Drives

Brian Tuemmler  
Director, Gimmel Group


Education Code: TR04-2226



## Learning Objectives


Upon completion of this session, participants will be able to:

- Identify e-non-records and expired e-records
- Define the value and business case for eliminating e-trash
- Identify ICM tools and how they help



## Pop Quiz

- \* True or False: Files that have not been accessed in the past 3 years are garbage
- \* Some may be, but the decision should be based on content, not age




Records Management  
Legal  
Information Technology  
Users

## DRIVERS AND PERSPECTIVES





## Who Is Cleaning up?

- \* Companies who:
  - Want to shut-down the shared drive as part of ECM program
  - Have suffered through expensive litigation or public records requests and want to reduce future culling costs
  - Trying to implement retention plan on electronic data
  - Plan to consolidate multiple file servers to data center
  - Don't want to spend money on ECM right now, but still need to clean things up



## Perspectives

## The Nature of Corporate Information

Information type	Examples	Characteristics
Structured	Databases and systems	20% of corporate information
Semi-Structured	Email, chats, blogs	Under-automated and controlled. Discoverable and useful/current
Unstructured	Documents	Only 10% has made it into ECM and ERM systems. Discoverable, mostly useful, disorganized, eTrash
Paper & film	The file room, the pile on desks	Costly to keep, costly to get rid of Discoverable
Tacit	In employees' heads	Disappears when the employee leaves

## Unmanaged Content

- \* Inventory resistant
- \* May be useful – represents corporate knowledge
- \* Intermingled with useless “stuff”
- \* Is discoverable
- \* Should be destroyed if unnecessary
- \* Ownership is unclear
  - Terminated/re-assigned employees
- \* Rarely conforms to naming standards
- \* Mergers & acquisitions pose special problem

## Paper vs. Electronic

- |   |   |
|---|---|
| * Form and format linked  | * Auto reformatted compound documents   |
| * Humanly readable  | * 11k file extensions in use  |
| * Intelligible after 25 years   | * Unintelligible after 25 years   |
| * RM saves money  | * RM costs money  |
| * Agency records officers understand value of records management known, understood, applied | * Network administrators, middle management don't recognize value of records management |

## Current Practices

- \* What do people use shared drives for?

### Works well in ECM

- Capture important data
- Create compound documents
- Collaborate
- Modify documents
- Manage versions
- Share knowledge
- Retain/dispose
- Personal/workgroup/department/enterprise

### Doesn't work well

- Backup files or folders
- Develop software code
- Distribute software
- Set up websites
- Store executables
- Store databases
- Store media
- Archive
- Temp archive for emailing an attachment
- Storing renditions
- Archive PSTs
- Drop box

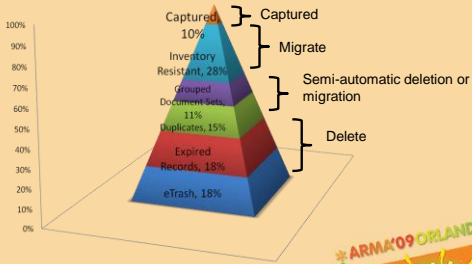
## Why Clean up?

- \* Information Management Compliance Programs
- \* Reduce cost and level of effort when litigation strikes
- \* Improve employee productivity
- \* Enhance access to institutional memory
- \* Reduce storage costs
- \* Dispose of eTrash
- \* Eliminate expired records
- \* Enhance disaster recovery
- \* Tag files for migration to managed repositories

## The Challenge – Shared Drives

- \* “The problem is too big”
  - Email constitutes huge volume
  - Not managed by retention schedule
  - Some have ECM but haven't migrated files from shared drives
  - Some have RM file plan but not applied to shared drives
- \* “The problem is not big enough”
  - Memory is cheap
  - We don't get sued
  - Don't have time or knowledge to figure out a solution

### What Are Firms Storing?



### Manual Clean-Up Is Impractical

- \* Level of effort to clean-up PC hard drive
  - 9 hours/employee X \$60/hr X 1,000 employees = \$540,000
- \* Shared network drives are far more challenging
  - Volume
  - Authors/owners may be gone
  - Multiple versions
  - Duplicates
  - Inconsistent/missing metadata
  - Individuals making decisions = less defensible
  - IT lacks expertise in RM principles and business context

### Why You Shouldn't Do "Nothing"

- \* 90% of corporate knowledge is still there
- \* Content on shared drives is preservable and discoverable, and should be managed according to retention schedule
- \* Deleting everything is not legally defensible
- \* Data costs \$10 per GB per month to maintain; 10TB over 3 years = \$3.7M
- \* One TB of data can result in \$18.75M in legal review costs

### Business Case - RM

- \* Implementing a file plan: helps if your documents are organized according to the plan
- \* Improve buy in and compliance: encourage participation and reduce labor
- \* Maximize benefit from merger or acquisition. What files did you acquire and how are they organized?

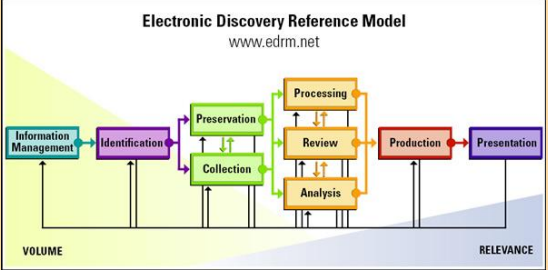
### Business Case – Legal

- \* Reduced cost of outside counsel to determine relevance
- \* Federal compliance: companies involved in federal litigation must now produce "electronically stored information" as part of discovery
- \* Seeing the full picture: asking people what information they have during litigation may not get you there
- \* Maximum compliance comes from minimum volume of uncontrolled content!

### Legal Hold Challenge

- \* Laws, regulations, and policies surrounding the use of electronic info in the legal context continue to evolve and grow in sophistication
- \* When an organization is involved in, or anticipates that it will become involved in, a lawsuit, an audit, or an investigation, regular record keeping rules need to change
  - Any disposition or alteration of info potentially relevant to the proceeding must stop immediately

### Steps in Electronic Discovery



### Discovery Costs

- \* 800 MB new info for every person on the globe per year, 30 ft of books stacked
- \* Save significant costs by eliminating paying outside counsel getting to know your IT environment: 80 – 100 hours
- \* \$6.2M to look at 20 M emails on backup tapes - Murphy Oil vs Flour Daniel
- \* Symantec: 75% of company's IP = email
- \* Recent surveys show that the average company faces 305 suits at any one time; that number jumps to 556 for companies with \$1 billion or more in revenue.
- \* 51% of survey respondents claimed the average cost of litigation (excluding settlement costs) was over \$200,000, with 8% putting the average cost over \$1 million.
- \* Construction company saved 90% in email mgmt
  - 45% reduced email maintenance cost
  - \$35,000 per IT staff time for e-discovery

### Business Case – IT

- \* Reduce costs: although storage media continually drops in price, the cost of managing it does not
- \* It is not all useful content: files of all types, such as backups, applications, system, music and video, logs, and old information that nobody looks at, are stored
- \* Increasing volume: volume grows at approximately 40% per year

### Examples of Typical Non-compliant Practices

Practice	Why?
Deleting emails when 10 MB is reached	Individual user's decisions about what to keep and not keep may be inconsistent
Deleting any files after 2 years	Deleting files related to litigation case may result in spoliation fines
Printing emails for retention purposes	Electronic files are also discoverable
Keeping all files	Expensive for IT to store or produce, and time consuming for employees to find
Asking department head if there are records on a network share	Rarely will a department head (or anyone) know if there are records (or even what a record is)

### Business Case - Information Security

- \* Protect intellectual property
- \* Prevent misuse of personally identifiable information
- \* Limit access to information that may not be actionable

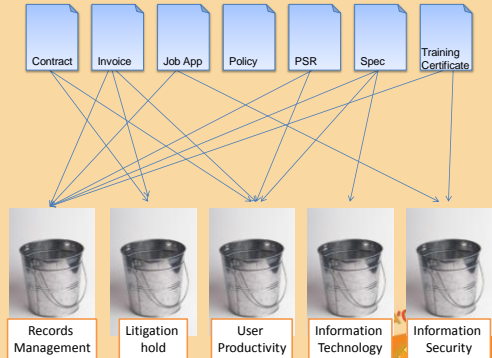
### Business Case – Business Units

- \* Time spent finding relevant or compliant documents for litigation
- \* Staff spend between 10 to 30 minutes per day looking for content in normal course of business
- \* Network drives need to be well organized to maximize knowledge sharing
- \* Maximize your investment in ECM by loading more content

## Benefit Summary

Benefit	Basis
eDiscovery culling time & expense	\$200 per GB per litigation, 300-1500 active litigation cases
Migration time to repository	30 seconds per file time 2000 files per person = 2 days
Content indexing/tagging	4 indexes per doc is additional 30 seconds per file
Manual RM deletion	30 minutes per week
Productivity of searching time	30 minutes/staff/week
TCO of network storage per GB	\$10-30 / mo

### Multiple Taxonomies



Methods  
Tools  
Planning

## APPROACH & TOOLS

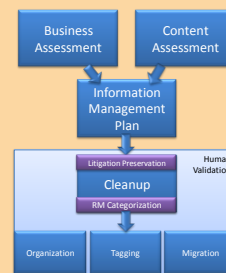
## Current Cleanup Options

- \* What to listen for to identify an opportunity..
  - Have users find their own documents and decide if they should be migrated. Staff reviewing file-by-file averages 30-60 seconds per file.
  - Use (or DON'T use) the folder hierarchies
  - Delete everything older than 3 years
  - Lock the drive and people can only save to new ECM
  - Hire temps to look at files one by one

## Traditional Migration Methods Are Impractical

- \* Point forward, ignoring the past
- \* Most important first (leads to 5% of content captured)
- \* Subject based taxonomies, not functional. Use auto-categorization for subject based tagging
- \* Piecemeal tools to find duplicates, pathnames, file sizes

## Methods & Tools



## Guidelines

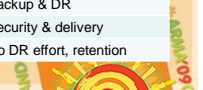
- \* Consistent approach and less accuracy is better than individual interpretation
  - Don't let perfection be the enemy of good
- \* A semi-automated process, not a tool
- \* Not just an IT solution
- \* Users, IT, Legal, Compliance, and RM all benefit and should participate



## Perspective on Information Management Plan

It is easier to develop a plan when you know what information you have

Information type	Storage	Goal
Important content	Repository	Sharing & access, disaster recovery (DR)
Templates	Repository	Control & access
Training materials	Repository or LMS	Monitoring & maintenance
eTrash	Delete	Productivity, storage
Database applications	Application server	Application updates & security
Database data	Data server	Backup & DR
Web sites	Web server	Security & delivery
Archive documents	Off-line	No DR effort, retention

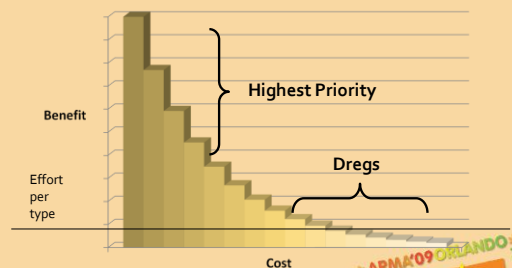


## Personality of Drives

- \* Legal
  - Want to keep everything, including versions.
  - Not much eTrash
- \* Engineering
  - Accidentally keep everything
  - Lots of obsolete applications
- \* IT
  - Sophisticated content (Compound documents & information)
  - Applications
- \* Safety and Training
  - Media, versions, databases



## Scope of Effort



## Dregs

- \* Crawler software wasn't/isn't perfect and content isn't consistent
- \* Expected some % content to be "resistant"
- \* Options for resistant content
  - Manual review – offshore tagging
  - Assign long retention and store offline
  - Port to repository and rely on full text search



## Human Validation or Tagging

- \* Achieve the "reasonable effort" threshold for some content types
- \* Isolate confidential or private information
- \* If acceptable from a security perspective, make content viewable to offshore resources
- \* Random sampling or complete population



## Defensibility

- \* Provide systematized deletion process rather than individual or arbitrary process
- \* Implement consistent approach across the company
- \* Approve and document standard (reusable) queries
- \* Approve and document custom queries and results
- \* Customize company-specific forms to document approvals from Legal, RM, IT and Business Unit
- \* Isolate and preserve documents subject to litigation holds
- \* Build query architecture on approved Records Management definitions
- \* Human validation of query results to document "reasonableness"
- \* Provide audit trails of work performed and documents deleted



## Cleanup Functionality

- \* ICM (Information Classification and Management)
  - Full text
  - Regular expressions
  - Bi-directional proximity
  - Natural language
  - Document comparison - Exemplars
  - File properties (OS)
  - Classifications
- \* Near duplicates
- \* Email management
- \* Auto-categorization (subject based)
- \* Bulk load tools
- \* Delivery context (folder, email, workflow, other delivery methods)
- \* Excel/Access
- \* Point-forward tracking tools
- \* ECRM



## ICM Tools

- \* Class of application-independent software that use advanced indexing, classification, policy and data access capabilities to automate data management activities above the storage layer
- \* Query
  - Full text, Regex, entity, classification, import
- \* Organize
  - Function and requirement
- \* Store
  - XML database manipulated by SQL like commands or UI
- \* Report
  - For plan and export for migration
- \* Act
  - Stub, copy, move, deduplicate
- \* Repeat on regular basis



## Versions

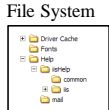
- \* 30% files are versions – Equivio
- \* 25% files are named versions
- \* Users identify versions semi-consistently with number or date
- \* Equivio version rarely maps to named version number (version 1,2,3 might be draft, redline and clean)
- \* Templates are version 1 – should be migrated separately to repository, no check-in capability

Archive	18
Copy of	127
Draft	1,764
Draft 1	37
Draft 2	29
Draft 3	7
Final	1,764
Redline	1,176
Version	151
Version 01	1
Version 02	1
Version 1	6
Version 13	1
Version 19	1
Version 2	15
Version 21	1
Version 22	1
Version 3	4
Version 31	1
Version 32	2
Version 4	3
Version 41	1
Version 5	2
Version 6	5
Version (2)	1
Version (4)	1
Version 01	4
Version 04	1
Version 11	16
Version 12	2
Version 2	15
Version 3	11
Version 4	1



## What Does ECM Mean?

### Traditional File System



### Content System



What the user sees

What provides the retrieval, security, integration, image management

The data



## Extensions

- \* 84 file extensions blocked by SharePoint
- \* 3,000 additional represent non-migratable
- \* ~11,000 in use, some multiple
- \* Occasionally renamed by users for versioning or hiding



Throwing away trash  
Organizing  
Fixing  
Putting away

## CLEANING UP



## Pop Quiz

- \* Which of the following is not a record?
  - A TIF image
  - A DLL file
  - A jpeg image
  - An email
  - Can't tell



## eTrash

- \* User or system generated residing on network shares and personal drives that has no business or technical value
- \* Should set approved policy and definition prior to action



## Trivial

- \* Tmp
  - Files starting with ~\$ or ending with .tmp; in a "Temporary" folder
  - Generated when an application was unable to automatically delete a temporary file
- \* Log, prn, txt, out, rpt
  - Produced by system process to indicate results
- \* Bak or "\backup\"
  - System or user generated backups
  - Mostly redundant if backed on backed up drive



## Obsolete

- \* Old Client Applications installed prior to latest recycle or OS install
- \* Created by an outside entity (software company)
- \* E.g. WordPerfect 5.1, Paradox 3.5, Acrobat 1.0
- \* Likely do not operate in current environment



## Redundant

- \* Copies of files (Same hash value)
- \* Copies of folders (Backups)
- \* Zipped files and folders
- \* Emailed files
- \* Renditions (PDF, TIF)



### What to Do with Duplicates?

- \* Delete trash and expired records first
- \* Keep 1
- \* Migrate to single instance storage
- \* Set policy
- \* Delete by groupings
- \* Reconfigure applications
- \* Ask file owners for help



### How many duplicates?

Date	Size	File
8/3/2007	1,645,056 m:	admin\CEO\2007 presentations\presentation 080307.ppt ★
8/3/2007	982,254 m:	admin\CEO\2007 presentations\presentation 080307.pdf ★
8/3/2007	433,223 m:	admin\CEO\2007 presentations\presentation 080307.zip ★
7/1/2007	1,565,533 m:	admin\CEO\2007 presentations\presentation 070107.ppt
8/1/2007	455,333 m:	admin\CEO\2007 presentations\drafts\presentation 080307.ppt
8/3/2007	982,254 m:	admin\CEO\2007 presentations\sent to board\presentation 080307.pdf ★
8/20/2007	994,992 m:	admin\CEO\2007 presentations\sent to board\presentation 080307.pdf.msg ★
12/31/2007	2,266,733 m:	admin\CEO\2007 presentations.zip ★★★★★★

1 Original  
 4 Duplicates  
 7 more in zip archive



### Abandoned

- \* Typically 1/3 of files are "owned" by former employee. Many in folders named for that employee
- \* More questionable files can be deleted (drafts, backups, duplicates)
- \* Need custodian to review remaining files or set different policy



### Neglected

- \* Drafts & versions
  - May be included in retention schedule
  - Difficulty in knowing what the last version is
  - Naming conventions vary widely
- \* Using near-duplicate tools may help
- \* May need to validate intention with users



### Garbage

- \* Files with zero content if not part of database application
- \* Corrupt and inaccessible files
- \* Thumbs.db



### Expired

- \* Fits approved retention schedule, has identifiable retention period
- \* Has agreed upon "Final" date
- \* Not under preservation order



## The Method in the Madness

- \* Structure found in unstructured data
  - Format (html, pdf)
  - Context (folder, cover page, sender)
  - Index or metadata (properties, embedded data, headers)
  - Content (repeatable terms, subject, HIDDEN)
  - Elements (edits, calculations, field formats)
  - Data location (line numbers, cells, zones)



## Records

- \* Do staff currently understand records categories?
- \* Example: Category (ADM1000)
  - Types (Travel requests, employee recognition)
    - Titles (form TR023, ER Memo)
      - Examples and characteristics (contains text TR023 or TR-023, "Celebrate Success", "Employee of the week")
- \* Start w/ general go to specific
  - Easy to categorize
  - Large volumes
- \* 1 category will be about 3 days effort
- \* Big buckets are good
- \* Don't expect perfection

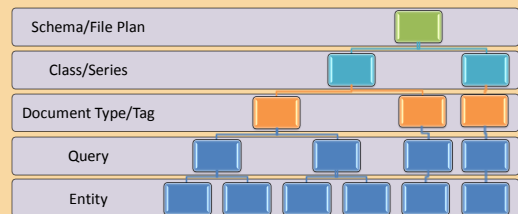


## Query Example

- \* Contracts
  - Keywords: Terms and conditions, Force Majeur, Signed: \_\_\_\_\_)
  - + Format: \*.doc, \*.xls
  - + System Data: Path does not contain /drafts/
  - + Data types: Responsible Attorney =
  - + Calculations: Document date = created date
  - + Expressions: Case ID: 72xxx.xx
  - - Exceptions: "DRAFT"



## Query Structure



## Working with Users

- \* Identify retention categories
- \* Identify information classifications
- \* Set priorities and goals
- \* Preserve data
- \* Provide additional requirements for tagging or cleanup



## Working with Users

- \* Shorten file names
- \* Remove odd characters
- \* Refile/reorganize
- \* Eliminate zips & backups
- \* Rename drafts, etc.
- \* Delete non-migratable duplicates
- \* Remove/relocate apps, web sites, databases



## Building Queries

Unique  
Record Series  
Information Classification  
Format  
Systems

Indexing

What can and cannot be migrated?

## MIGRATION

## Pop Quiz

- \* Which of the following would be appropriate to store in a content repository?
  1. A workflow design
  2. A database
  3. A zip file backup of a set of folders
  4. A voicemail message
  5. Software code

## Searching Vs. Managing

- \* Searching is different from managing
- \* Indexing and Classification benefits - ISO 15489
  - Linking individual records.
  - Consistent naming of records over time.
  - Retrieval: classification assists in retrieving records relating to a particular function, topic, or activity.
  - Security and access by group.
  - User permissions: easier management of user permissions for access to, or action on, particular groups of records.
  - Distributed management: distribute, or send, a whole file (a group of related records) to remote or mobile workers.
  - Easier retention scheduling.

## Tagging

- \* Numbers (Project, work order, invoice, SSN)
- \* Information Classification (Public, Private, Confidential, ATTY/Client)
- \* Status (Version number, Draft, Final)
- \* Risk (Privacy, PII, IP, CCN, SSN)
- \* Owner, group, responsible attorney

## Migration

- \* Keep files in clean, tagged and ready state
- \* By project, litigation case, user group (according to ECM schedule)
- \* Create CSV or XML load file (path, name, tags)
- \* Load according to ECRM tools

### Applications

- \* Install files, disk images, executables, libraries, help files, samples
- \* Do not require immediate restoration, no version control
- \* May require fast server speeds rather than fast I/O
- \* Not managed as records



### Websites

- \* Linked html files, java, flash, navigation, style sheets
- \* May link to others or be linked to by others – migration may break links
- \* Likely contain records and non records
- \* Unitized – content and context of all files important



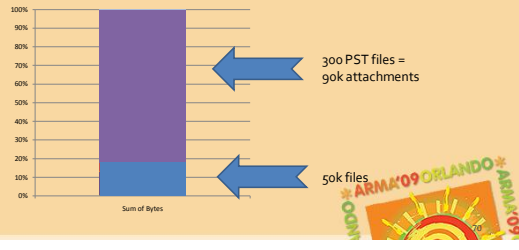
### Databases

- \* Data files, related application files, configuration and reference data, performance logs, reports
- \* Cannot “function” from within repository
- \* Data records are business records



### PST Files

- \* 82 % of storage is in 1% of files – PSTs
- \* Most of new storage growth is PSTs



### The Trouble with PSTs \*

Problem	Indication
Inhibits collaboration and knowledge sharing	Zero % of files in PSTs have different modification/create date => no access or use
Prevents records management compliance	No categorization by type, no automated disposition
Introduces risk	Private information in email becomes public when archived. Found 306 files with SSN, CCN, or Bonus info.
Minimizes tiered-storage effectiveness	Every email is “active” when one email added. (72% of all content modified in last 4 weeks)
Promotes duplication	3% duplicates outside of PSTs, 20% inside
Enables eTrash	2,400 items in “Deleted Items” folder

\* Starts with “P”



### Questions?



**Eliminating E-Trash and Non-Records From Shared Drives**

**Please Complete Your Session Evaluation**

Brian Tuemmler  
Director, Gimmel Group  
[Brian.Tuemmler@Gimmel.com](mailto:Brian.Tuemmler@Gimmel.com)  
925 253 5689

Education Code: TR04-2226

