Managing Shared Content

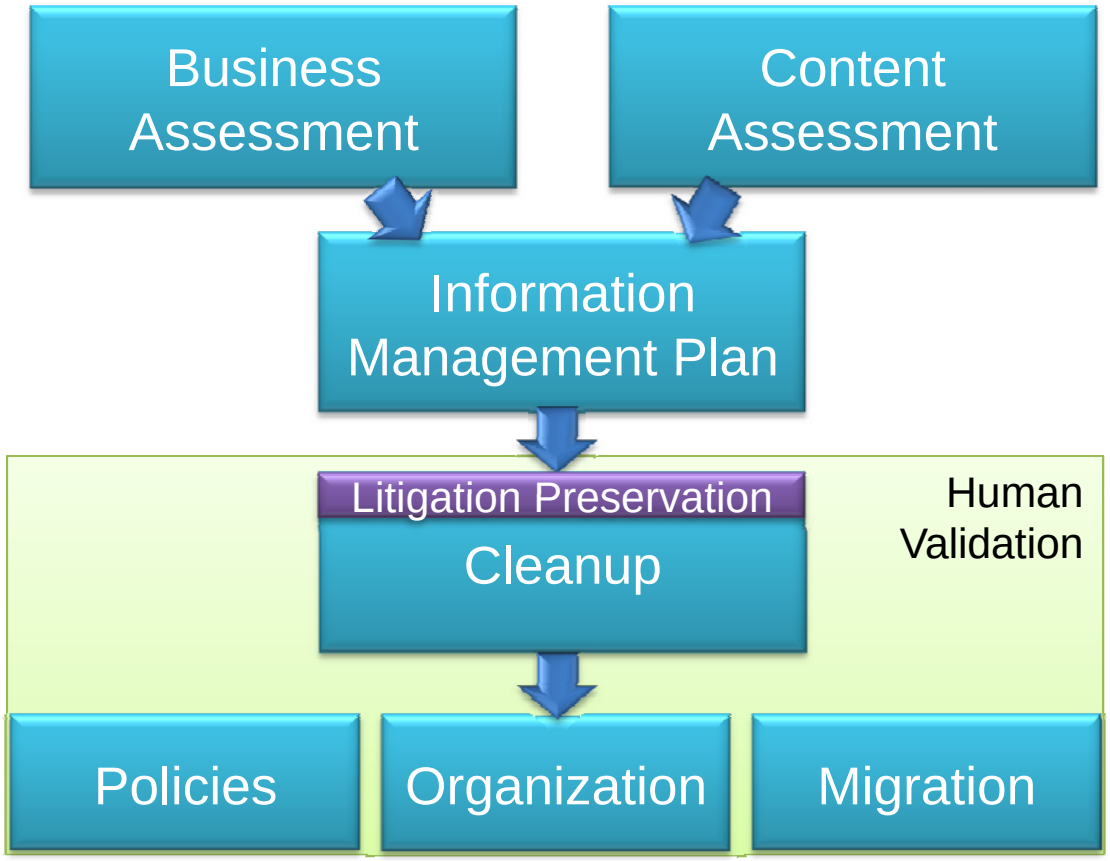
Tim Shinkle, Gimmal
on behalf of
Susan J. Sullivan, CRM; NARA

Target Outcomes

Knowledge of:

- How NARA combined people and technology to meet the shared drive cleanup challenge, and
- How cleaning up shared drives can lead to efficient management of all electronic content.
- How people interact with Indexing and Categorization Management (ICM) Tools to manage content
- Start thinking how you can do this in your organization

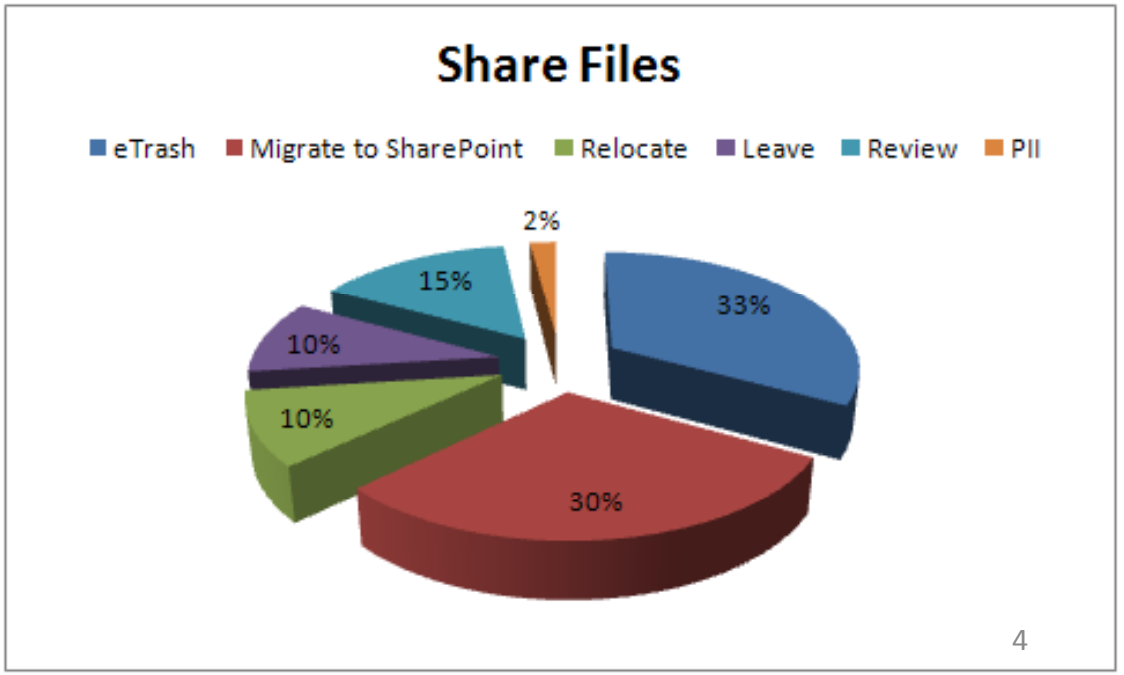
Overall Process



Plan > Crawl > Report > Review > Clean > Extract Metadata > Categorize > Migrate > Manage

Phase 0 –Assessment

- Perform preliminary analysis
- Scan a good cross section of the content
- Produce an assessment report and plan moving forward
  - Categorization
  - Strategy
  - ROI



Phase I – Program Setup

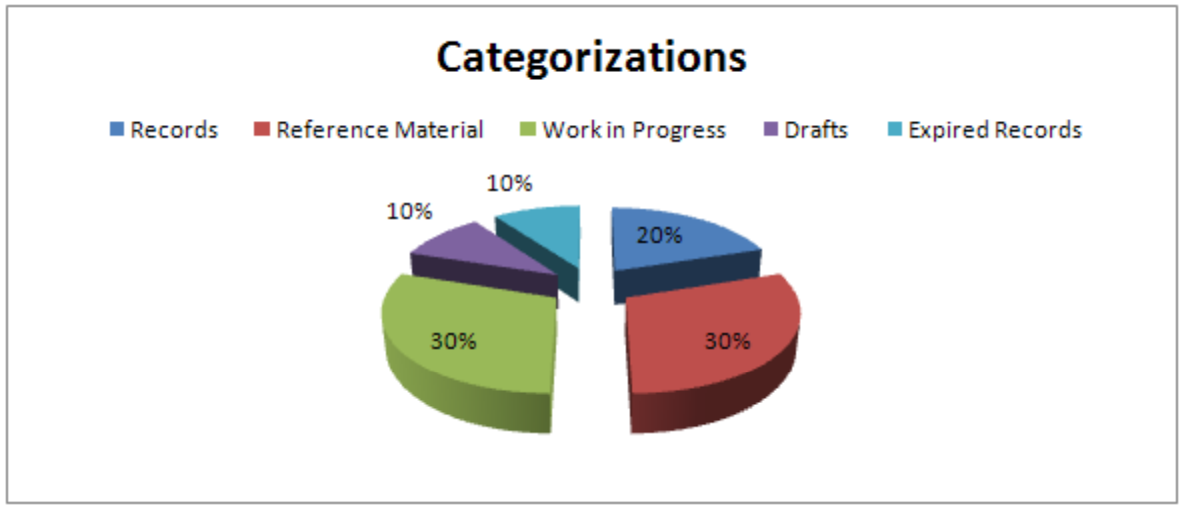
- Meet with sponsors
- Identify scope (shares, groups, SMEs)
- Develop communication plan
- Stand up program communications portal (e.g., corporate portal, wiki)
- Identify pilot group(s) (shares and SMEs)
- Meet with RM and Legal to identify candidate records, holds and PII
- Approve policies, cleanup and categorization rules

Phase II – Pilot

- Interview pilot group
- Perform analysis and refine rules
- Crawl and report on pilot group shares
- Review results with SMEs and move approved files to cleanup location
- Refine global cleanup and categorization rules and update published policies
- Develop benefit analysis report and enterprise strategy
- Publish updates to program portal (wiki)

Phase III – Enterprise Transformation

- Outline schedule, project plan and group priority
- Update communications on portal (e.g., wiki)
- Repeat for each group
  - Plan > Crawl > Report > Review > Clean > Extract Metadata > Categorize > Migrate > Manage
- Publish findings and strategy for on going automated information governance automation



People + Tools = Cleanup

People	Technology
Where is the content?	Pre-crawl for overview
What to protect / preserve?	Find /lock preservation content
Develop and validate auto-cleanup rules (.tmp, backup, ~)	Find / report on auto-cleanup content
Approve auto-clean rules	Auto-clean
Develop rules for content to be reviewed before cleaning or moving (large, old, marked “draft, old, trash”, music, video, audio)	Find / report on reviewable content
Review and identify for cleanup	Cleanup or move individual or groups of files
Develop data integrity policies, develop and communicate cleanup maintenance rules and policies	Cleanup / identify continually

What is eTrash

- Non-business value, non-records
  - Temporary files
  - Obsolete applications
  - Files with zero content
  - Un-openable/un-readable content
- Not so important content that is not FOIA, Litigation related, retention scheduled and is:
  - Duplicates
  - Large
  - Voluminous
  - Obsolete

Auto-Clean

- System generated temporary files
- System generated backup files
- Zero Content files and folders
- Abandoned applications
- Obsolete install files
- System status reporting files

SME Review for Cleanup

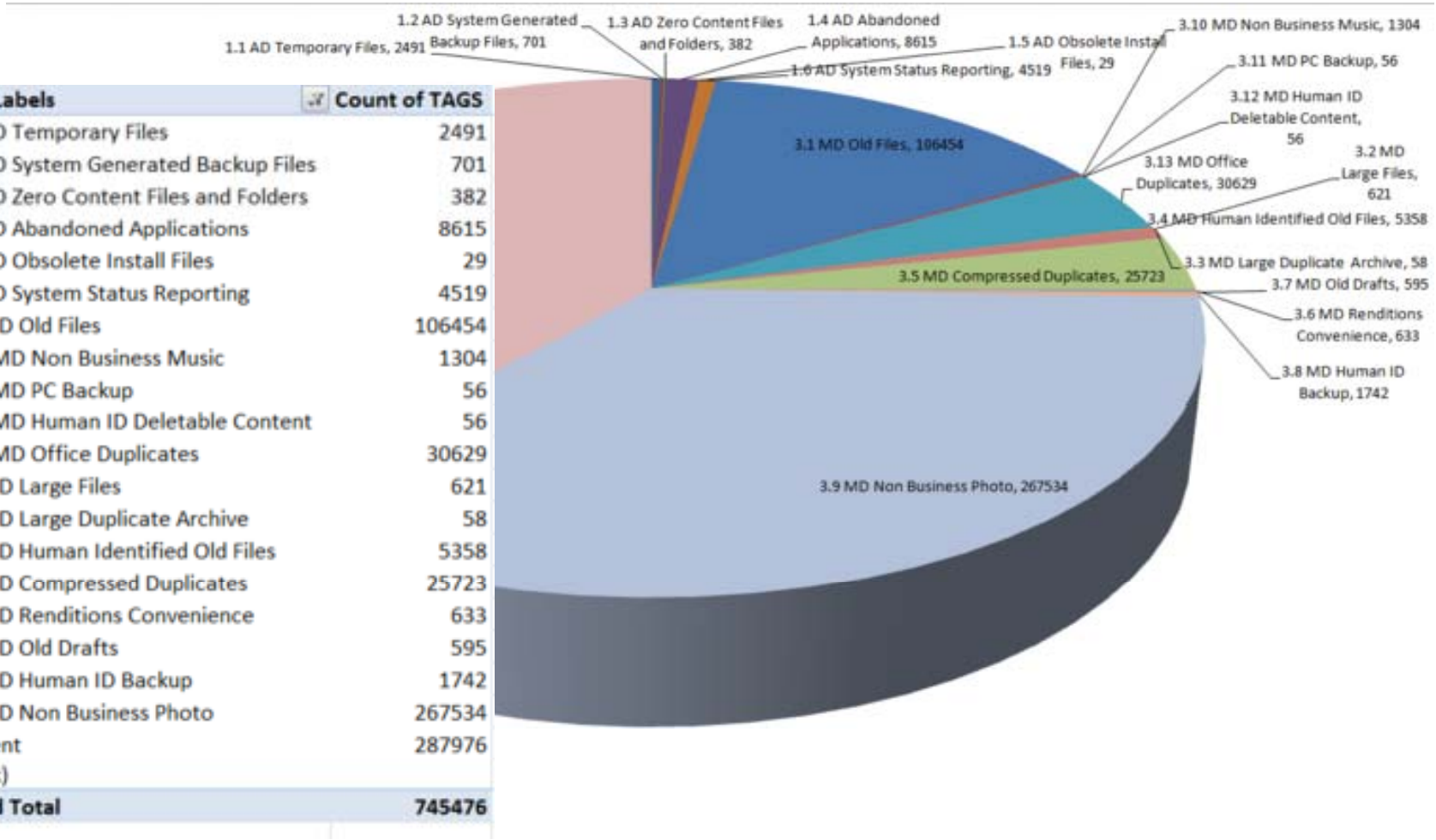
- Old files
- Large files
- Large duplicate archive
- Identified old files
- Compressed file duplicates
- Renditions convenience copies
- Non accessed drafts older than 1 year
- Human identified backup files
- Human-identified as delete-able content
- Non-business images music or audio video or media
- Office Document Duplicates

Content Identified for Risk

- PII and eDiscovery
  - Litigation hold files
  - Credit card numbers
  - Social security numbers

A new way of viewing content

Row Labels	Count of TAGS
1.1 AD Temporary Files	2491
1.2 AD System Generated Backup Files	701
1.3 AD Zero Content Files and Folders	382
1.4 AD Abandoned Applications	8615
1.5 AD Obsolete Install Files	29
1.6 AD System Status Reporting	4519
3.1 MD Old Files	106454
3.10 MD Non Business Music	1304
3.11 MD PC Backup	56
3.12 MD Human ID Deletable Content	56
3.13 MD Office Duplicates	30629
3.2 MD Large Files	621
3.3 MD Large Duplicate Archive	58
3.4 MD Human Identified Old Files	5358
3.5 MD Compressed Duplicates	25723
3.6 MD Renditions Convenience	633
3.7 MD Old Drafts	595
3.8 MD Human ID Backup	1742
3.9 MD Non Business Photo	267534
Content	287976
(blank)	
Grand Total	**745476**



Outcomes

- Shares cleaned or under review
- Bulletin issued - NARA Bulletin 2012-02, December 06, 2011; Guidance on Managing Content on Shared Drives (http://www.archives.gov/records-mgmt/bulletins/2012/2012-02.html)
- Inputs to Presidential Memorandum -- Managing Government Records, November 28, 2011, (http://www.whitehouse.gov/the-press-office/2011/11/28/presidential-memorandum-managing-government-records)
- Policies identified
  - cleaning shares,
  - managing content on shared drives
- Most importantly…..clear path forward

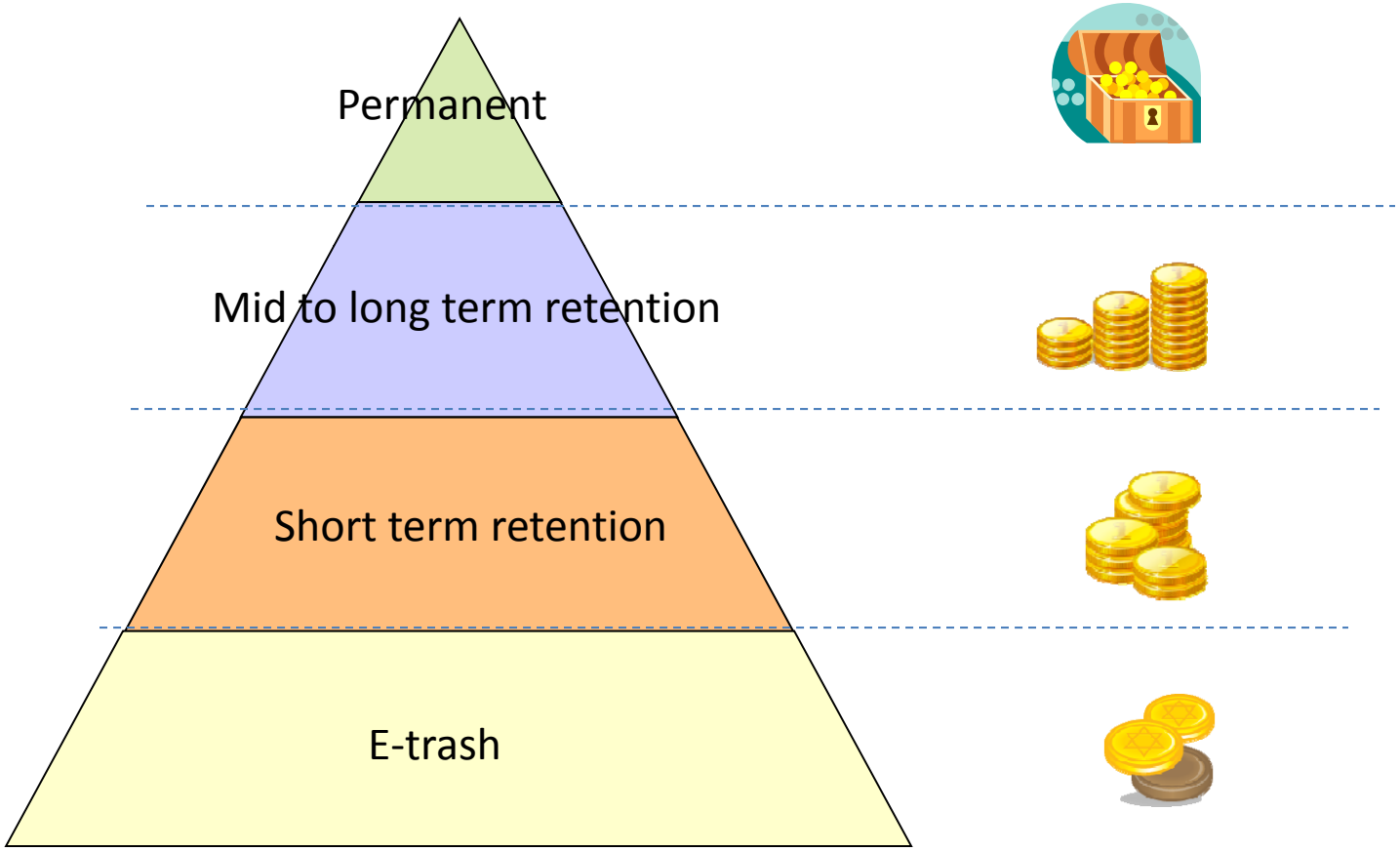
Data Management Policies

- **Reliability and content** – Preserving file attributes
- **Context** – Preserve file path, include recordkeeping metadata within user profiles
- **Usability**
  - Align storage with usage, preservation and access needs. Widely accessible storage location for high value content
  - Use robust search capability eliminate complex hierarchical folder structures.
- **Authenticity** - only authorized users can add, change and delete content for a specified data set.

Benefits

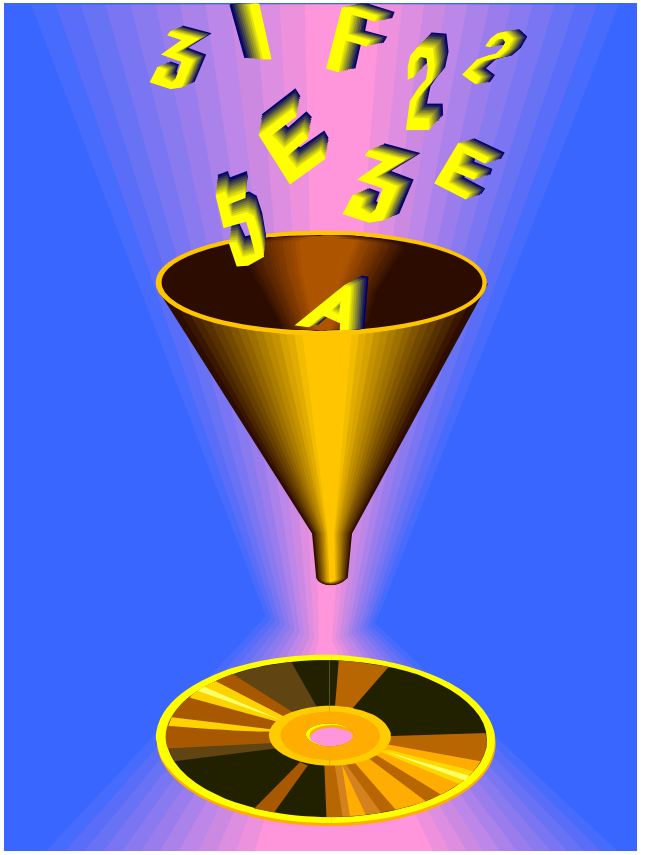
- Storage space – when cleaned all identified, can provide upwards of **47% in reclaimed storage**.
- Money - To maintain 27 TB clean per year, save $187K per year (shares and email).
- Time – Staff time culling through volumes of content to find or migrate their information (FOIA – E-discovery)
- Access efficiencies – Right information, right place right time

Next Opportunity:
Align content with value

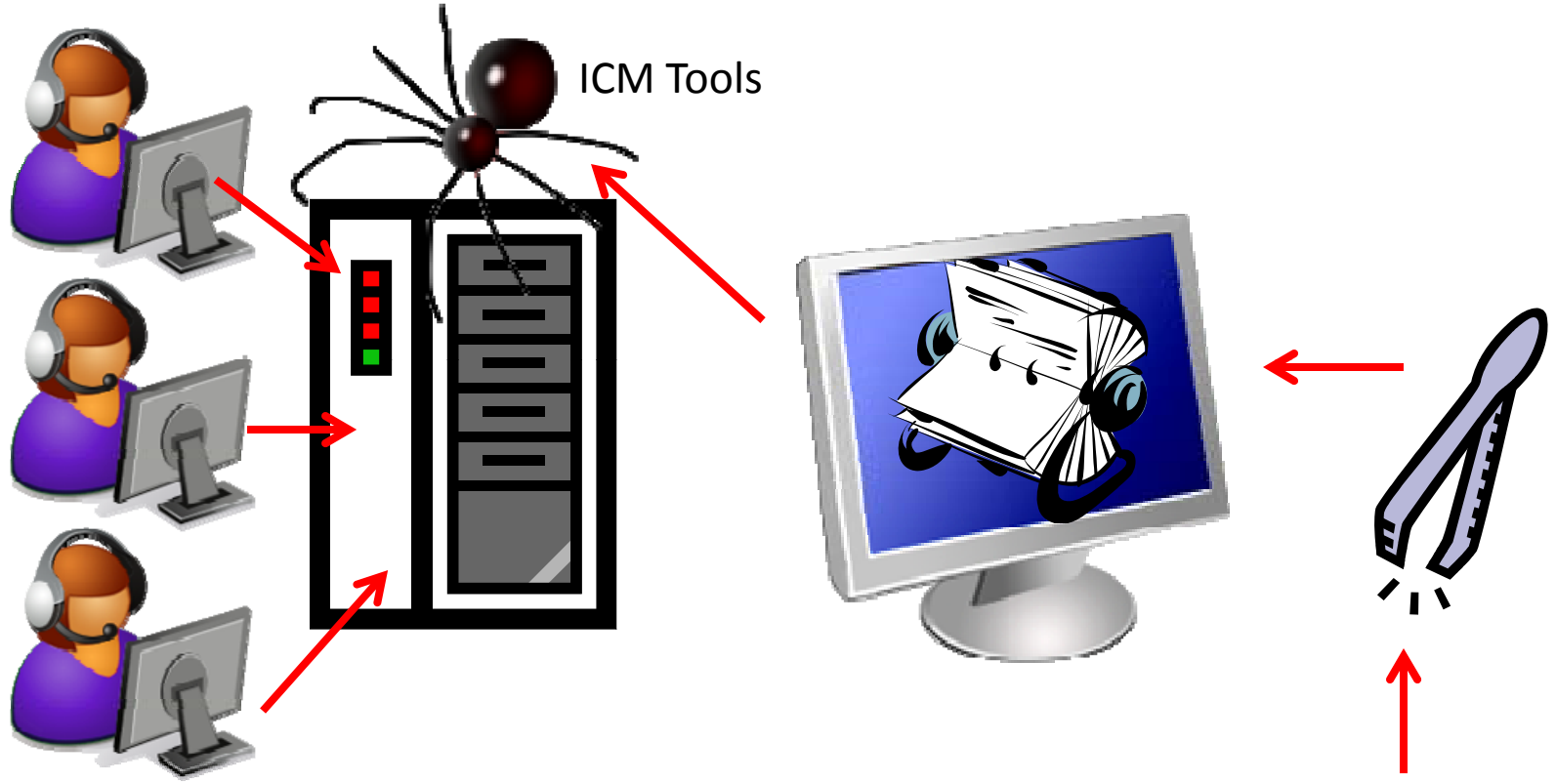


Beyond cleanup – Auto-Categorize

- Keyword/Boolean/Regex – Human Domain knowledge (SME) combined with...
- Natural Language Processing
  - document categorization,
  - clustering,
  - topic modeling,
  - information extraction,
  - other learning applications
- (Has this category OR these topics) AND these keywords AND NOT these keywords AND these formats AND these owners...

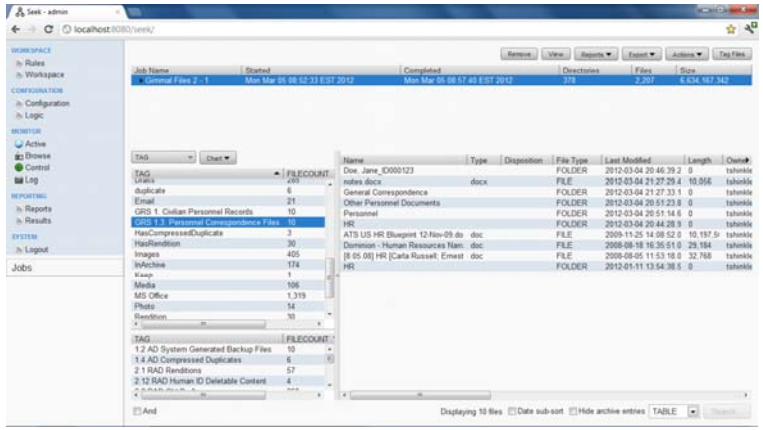


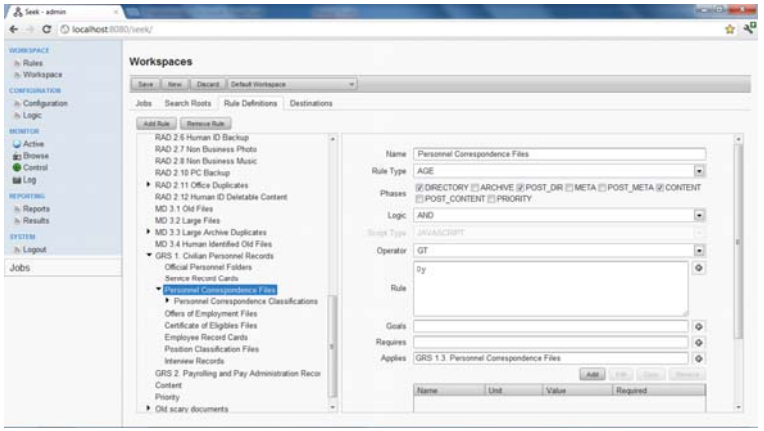
Vision of future



ICM Tools

- Indexes & categorizes content according to rules
- Clusters content around trends (retention)
- Ingests content samples and learns.
- More crawling = more learning
  - Manage by exception = more learning
- Act upon content (move, lock)





Q&A

- Tim Shinkle
  Director, Gimmal Group
  tim.shinkle@gimmal.com
  (703) 927-5650
- Susan Sullivan, CRM
  Director - Corporate Records Management
  National Archives and Records Administration
  susan.sullivan@nara.gov
  V: (301) 837-2088