Valora Technologies

— WEBINAR SERIES —

ROT Remediation:

File share clean-up & how to do it right

Sandra Serkes
Co-Founder & CEO

Jennifer Nelson
VP Strategic Solutions

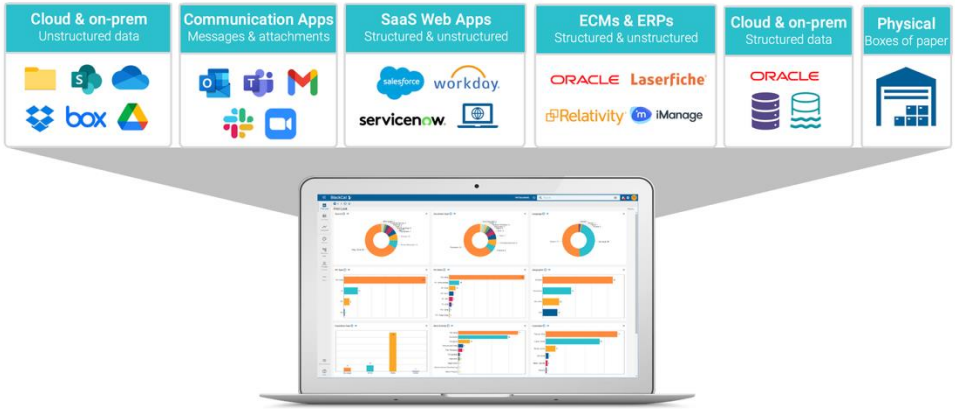
Webinar Series

www.ValoraTech.com

- ROT Remediation: File Share Clean-up — January 21
- Letting the Robots Classify: Automating IG — February 11
- AutoClassifying the 3 Rs: ROT, Records & Retention — March 18
- Structured Data Management & Disposition — April 15
- 10 Ways AutoClassification Can Impact AI — May 20
- How AC Transforms Enterprise Search — June 17
- Managing Data Privacy & While Managing Records — September 16
- Demo Day — October 14
- TBD — November 12
- AutoClassification & AI Trends in Enterprise — December 9

About Valora

- AutoClassification Platform for Information Governance
- Automates discovery, identification, classification & defensible disposition
- Trusted by IG teams, Records & IM teams, Legal, Compliance, IT teams
- Brings multiple enterprise repositories into a single view



About Valora

Information Governance / Records Management

- Classification
- ROT
- Retention
- Disposition
- Archive

About Valora

Enterprise-wide Data Governance

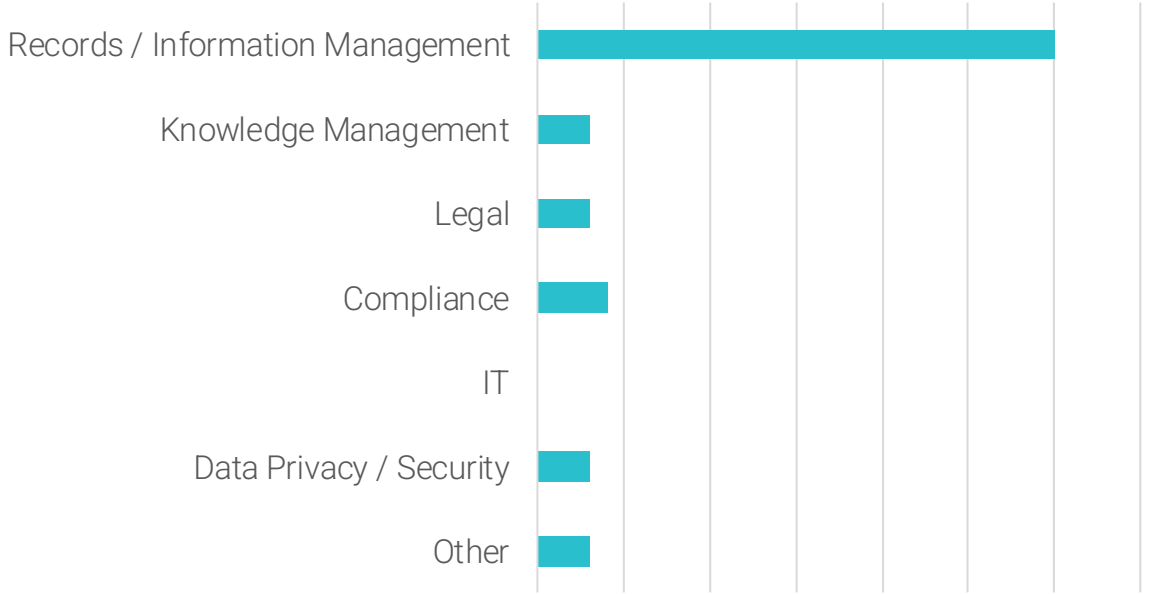
- Classification
- ROT
- Retention
- Disposition
- Archive
- eDiscovery
- Legal Hold
- Data Privacy
- Data Security
- Migration
- Compliance
- Minimization
- Provenance
- Lineage
- AI Readiness

Poll Results

Who’s here today
& what are your interests?

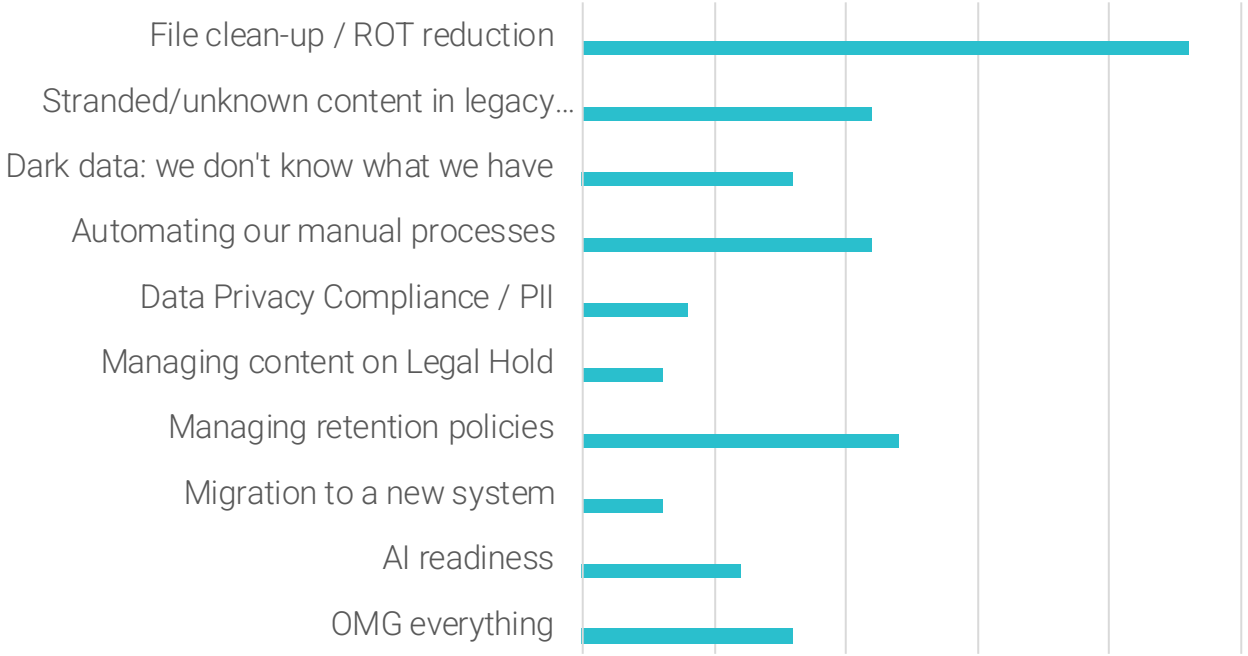
Who you are & what your challenges are

Who You Are



Who you are & what your challenges are

Your RIM & IG Challenges



Housekeeping

Q&A

Recording

Slides

Feedback

Agenda

- File shares – what the heck?
- What is ROT & how did it get there?
- Impact of ROT – cost, risk, compliance, effort
- Why is deleting content so, so, so hard?
- Identifying & Remediating ROT – the right way
- Gotchas & Trip-ups

Scary stories: real world examples of ROT

Sidebar: What is ROT?

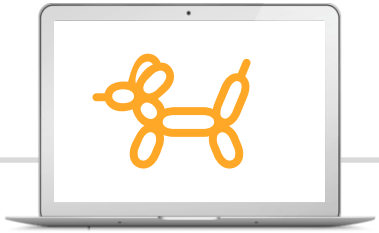
Content with no business, legal or regulatory value

Redundant

- Duplicates
- Backups & cloud storage
- Litigation Databases
- Shared/collaborative sites

Obsolete

- Past Retention
- Last accessed date
- Versions & updates



Trivial

- Personal content
- Cookies & markers
- Spam
- Low value docs

Understanding Duplicates

Goal: remove as much duplication as possible

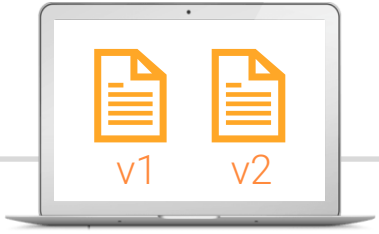
Exact Duplicates

- 100% forensically identical
- SHA-256 hash code match

40% identical dupes

Functional Duplicates

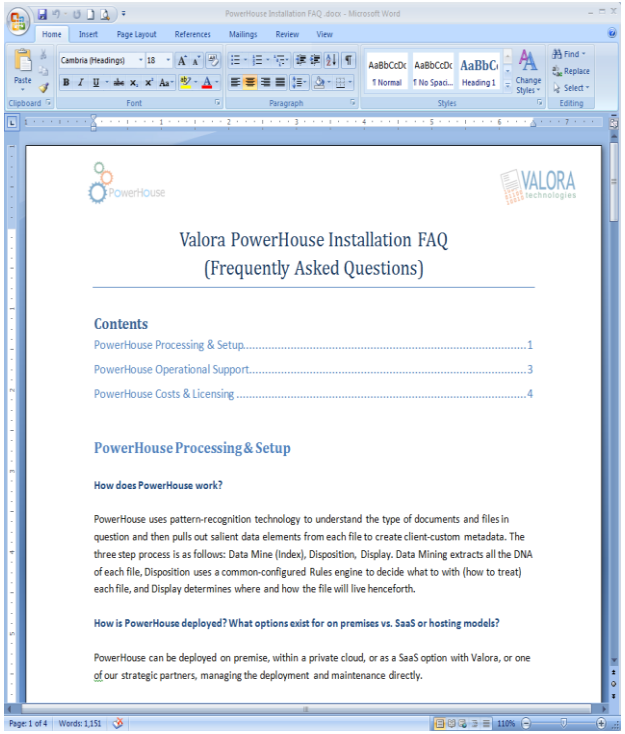
- 99% duplicates
- Functionally identical, but forensically different (no Hash match)



Near duplicates

- 75-98% similar based on text and metadata similarity
- May be similar or related enough to warrant their treatment as a family unit. Ex: Revisions

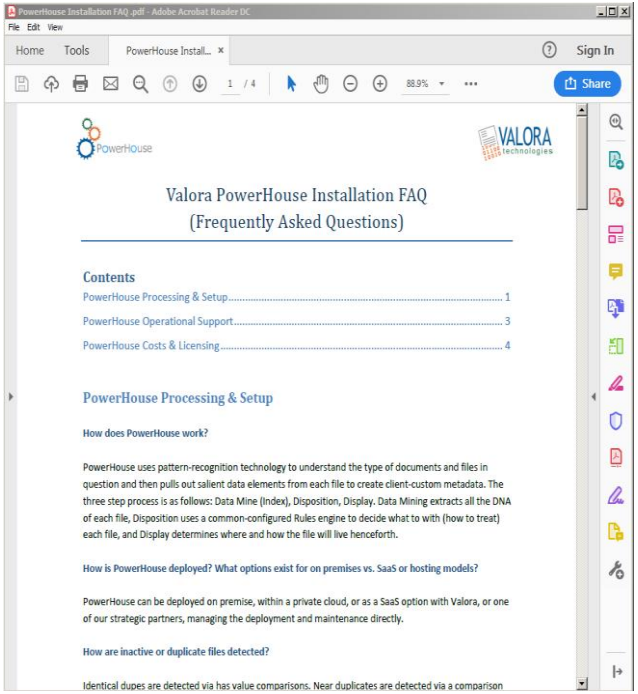
Understanding Functionally Identical Duplicates



Questions v12.doc
96 KB

FAQ Final.pdf
616 KB

Functionally identical
based on content



Why are we still talking about Windows file shares??

- Unstructured data is hard
- It's the biggest challenge:
  - Been a catch-all for years
  - Came out in 1996 (that's 30 years of data)
  - Willy-nilly folder structure
  - 1000s of abandoned file shares
  - Employees are long gone
  - No way of knowing what's in there
  - Impossible to find, let alone classify things

250 PB in file shares!

How much ROT have we got? What's the big deal?

- Typically, 40-50% of enterprise data is ROT
- ROT impacts:
  - Data storage costs
  - Data breach exposure
  - Litigation & eDiscovery fees
  - Compliance & Data Privacy
  - Organizational efficiency
  - AI Readiness

20 TB of data in a system untouched since 1996!

What's the risk & impact?

Estimated that 50% of all stored enterprise conte

Breach/Security Risk!
If it is there, it is exposed, compromised data.

AI Risk!
Feeding wrong data to your AI engine.

Cost

- Storage & maintenance costs
- A large company with 10PB of data could be spending as much as $34.5m on data that could be deleted Compliance, Governance & Oversight Council (CGOC report)

Lost Productivity

- Employees can't find things
- Working with latest version?
- Workers spend 30% (2.5 hours) of their workday looking for content (IDC Study)

Risk & Exposure

- Data Governance Risk: don't know what it is / where it is
- Privacy Risk: does it contain PII or sensitive data? Big fines.
- Compliance risk: does this exist when it shouldn't?

We get it, deleting content is hard.

Some people may resist deleting their data because of:

- Perception of value: Believing that seemingly redundant or outdated information could be useful in the future, even if it hasn't been accessed in years.
- Lack of ownership: Not knowing who "owns" specific content, leading to hesitation in deleting it.
- A "save everything" mindset: Some organizations have a hoarding culture where everything is kept "just in case."
- Regulatory concerns: Misunderstandings about legal or compliance obligations may lead to over-retention of records or sensitive data.
- Fear of change: Natural resistance to change, even when it involves improving efficiency or meeting compliance obligations.
- Uncertainty about outcomes: Concerns about how new content management practices might disrupt workflows.

Shift in thinking & behaviour

Digital

Paper

But storage is cheap! ….right??

Humans cannot keep up

Cloud & on-prem
Unstructured data

Communication Apps
Messages & attachments

SaaS Web Apps
Structured & unstructured

ECMs & DMSs
Structured & unstructured

Cloud & on-prem
Structured data

Websites
& URLs

- What do we have?
- Where is it?
- What is it? (sensitive/PII)
- Who owns it?
- Where are the duplicates?
- What's the risk?

Human Error
Inconsistency, interpretations/ reluctance, "not my job"

Scalability Issues
Impossible to keep up with volume of data

Compliance Risk
Regulatory non-compliance, retention & deletion errors

Limited Auditability
Lack of transparency, why things were categorized

Time Consuming & Costly

Really costly....

Humans cannot keep up

Cloud & on-prem
Unstructured data

Communication Apps
Messages & attachments

SaaS Web Apps
Structured & unstructured

ECMs
Structured &

- What do we have?
- Where is it?
- What is it? (sensitive/PII)
- Who owns it?

Human Error
Inconsistency, interpretations/ reluctance, "not my job"

Scalability Issues
Impossible to keep up with volume of data

Compliance Risk
Regulatory non-compliance, retention & deletion errors

Limited Auditability
Lack of transparency, why things were categorized

Time Consuming & Costly

- 30 seconds = employee to manually classify 1 doc
- 1M docs
  = 30M seconds / 8000 hrs ÷ 2,080 work hrs/year
  = 3.8 employee-years
  = 1 employee manually classify 264k docs/year
- Average employee in a business:
  = sends/receives 100 emails, works on 10 docs/day
- 1000-person organization
  = 110,000 doc/day
  = 1.32M docs/year
- Been in business 20 years
  = backlog of 438M docs
  = 1,664 employee-years to get through the backlog
- Get it done over 5 years = 87M docs/year
  = 264k docs/year/person
  = require a team of 329 humans
- US @ $65k/year average salary
  = $21.3M in wages per year
  = $106.9M in wages over 5 years
- Overseas @ $5k/year average salary
  = $1.65M in wages per year
  = $8.25M over 5 years
- That's just to get through the backlog

Technology really is the only way

Cloud & on-prem Unstructured data	Communication Apps Messages & attachments	SaaS Web Apps Structured & unstructured	ECMs & DMSs Structured & unstructured	Cloud & on-prem Structured data	Websites & URLs
SharePoint, OneDrive, Dropbox, box, Google Drive	Outlook, Teams, Gmail, Slack, Zoom	salesforce, workday., servicenow,	ORACLE, Laserfiche, Relativity, iManage	ORACLE	Facebook, Instagram, X, LinkedIn, YouTube

- What do we have?
- Where is it?
- What is it? (sensitive/PII)
- Who owns it?
- Where are the duplicates?
- What's the risk?

Reduces Human Error

eliminates the possibility of human error or oversight

Scalability Issues

process & classifies 10,000s docs per day

Reduces Risk

Confidence in your data & processes, ready for audits

Limited Auditability

Lack of transparency, why things were categorized

Reduces Costs / Reduces Time

- Reduce ROT / reduces hosting costs
  storage costs
  back-up costs
- Reduces manual resources time
- Reduces manual searches: data requests
  DSAR
  eDiscovery

What not to do when tackling file shares

Continue to ignore their "[illegible] further c[illegible]

Rush into a clean-up without a clear plan

Rely on manual efforts

Entire file shares on Legal Holds that are never lifted …ever

Over-delete or under-delete due to lack of insight

Ignore long-term sustainability & governance

Clear cut files by "created on" or "last accessed" date

Migrate first, Clean-up second

Best practices for ROT remediation

Consider file types & content types, not just dupes & dates

Involve key stakeholders (IT, Legal, Compliance)

Benchmark current processes to calculate your ROI

Respect other data requirements: Legal Hold

Make it easy for people: tombstone deleted data

Communicate your success to others

Think long-term.
What else can we solve for?

How do you do it?

1. **Data Discovery**
   Locate and inventory
2. **Intelligent Analysis**
   Full-text content analysis
3. **AutoClassification**
   Content-based Classification
4. **Defensible Disposition**
   Automated rules & workflows
5. **Delta Monitoring**
   Ongoing scans & audits

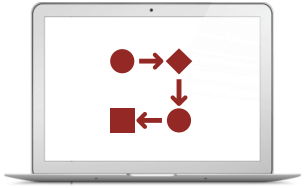
But HOW does it work? …AutoClassification

File & Rich Metadata
Determine Attributes

+

Recognition Algorithms
Determine DocType

+



Disposition Algorithms
Rules & Actions

AutoClassification = Rich Metadata + Rules

Software that performs automated analysis & disposition of file/document content

Answers the question:

What is this thing &
what do I do with it?

…and then does it.

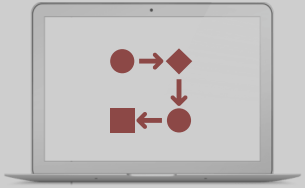
Sophisticated Classification

File & Rich Metadata
Determine Attributes

+

Recognition Algorithms
Determine DocType

+



Disposition Algorithms
Rules & Actions

Language
Personal Data
Word & page count
Tone/Sentiment
Attachment Range
Security Level

Author
Custodian
Content Owner
Recipient
Copyees (CC & BCC)
Assignee
Signatories
Patient Name

Brand
Legal Hold
Employee
Audience
Document Type

Creation Date
Signature Date
Effective Date
Last Activity Date
Expiration Date
Destruction Date

File type/extension
Created date
Created by
Last modified date
File path

Highly customized

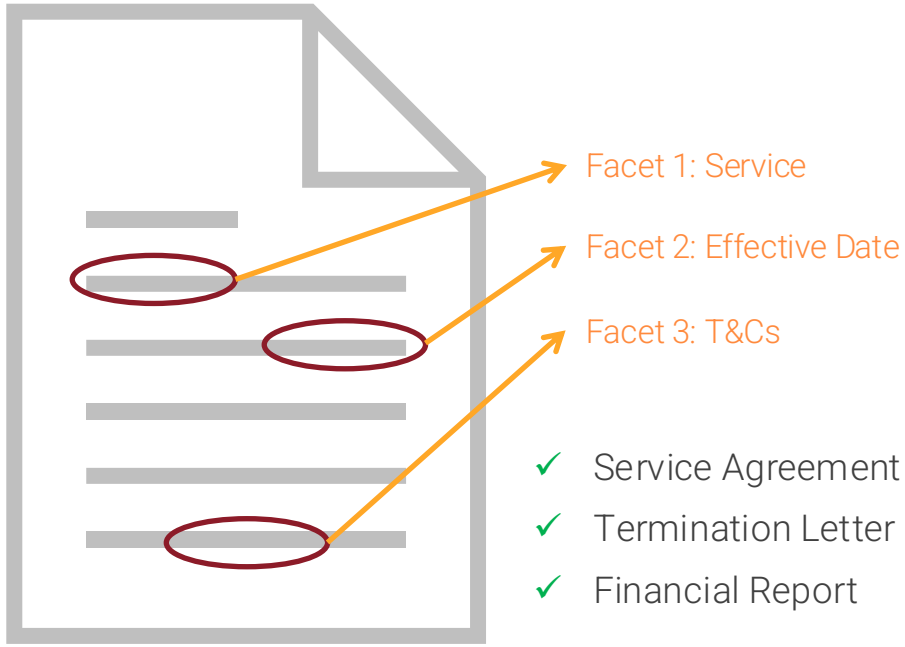
File & Rich Metadata
Determine Attributes

+

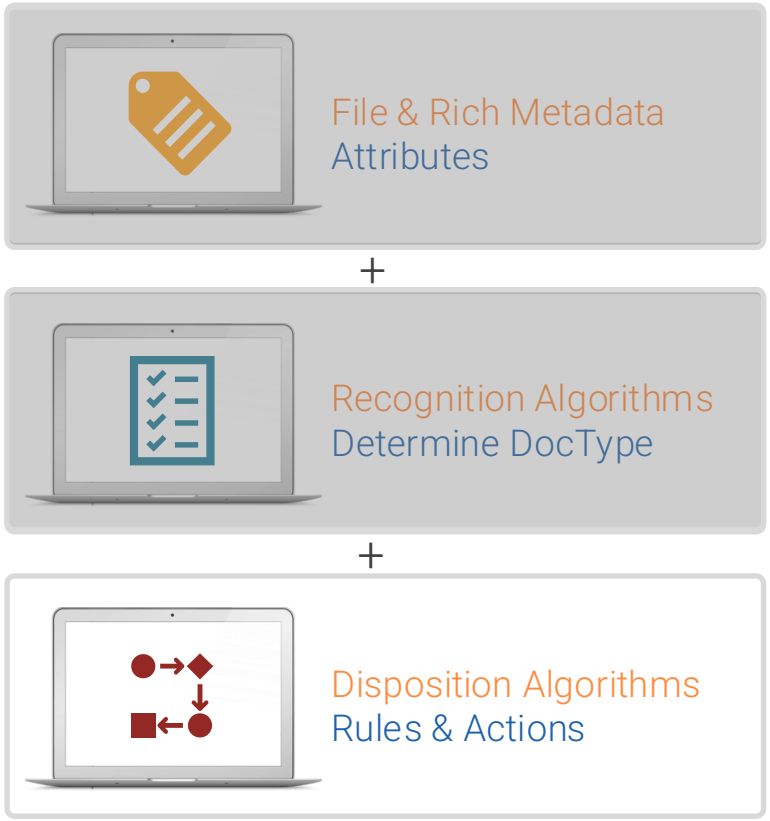
Recognition Algorithms
Determine DocType

+

Disposition Algorithms
Rules & Actions



Comprehensive results

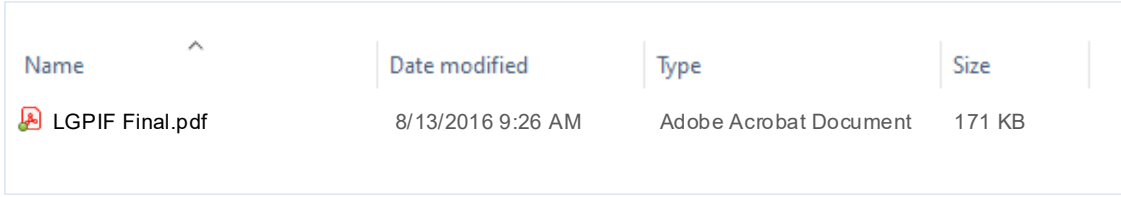


- Determine the disposition & handling of content
- Typically follow an IF-THEN format, often nested
- Ex: Employment Contract:
  - ✓ Mark as Confidential
  - ✓ Move it to SharePoint
  - ✓ Retain for 8 years after termination
  - ✓ HR approval required before deletion

Automated, Manual and/or Hybrid Disposition

Tag	Delete	Report	Move
Archive	Hold	Migrate	***** Anonymize

AutoClassifying Documents



Name	Date modified	Type	Size
LGPIF Final.pdf	8/13/2016 9:26 AM	Adobe Acrobat Document	171 KB

Identifying ROT: Trivial

"Watch list" terms

Date

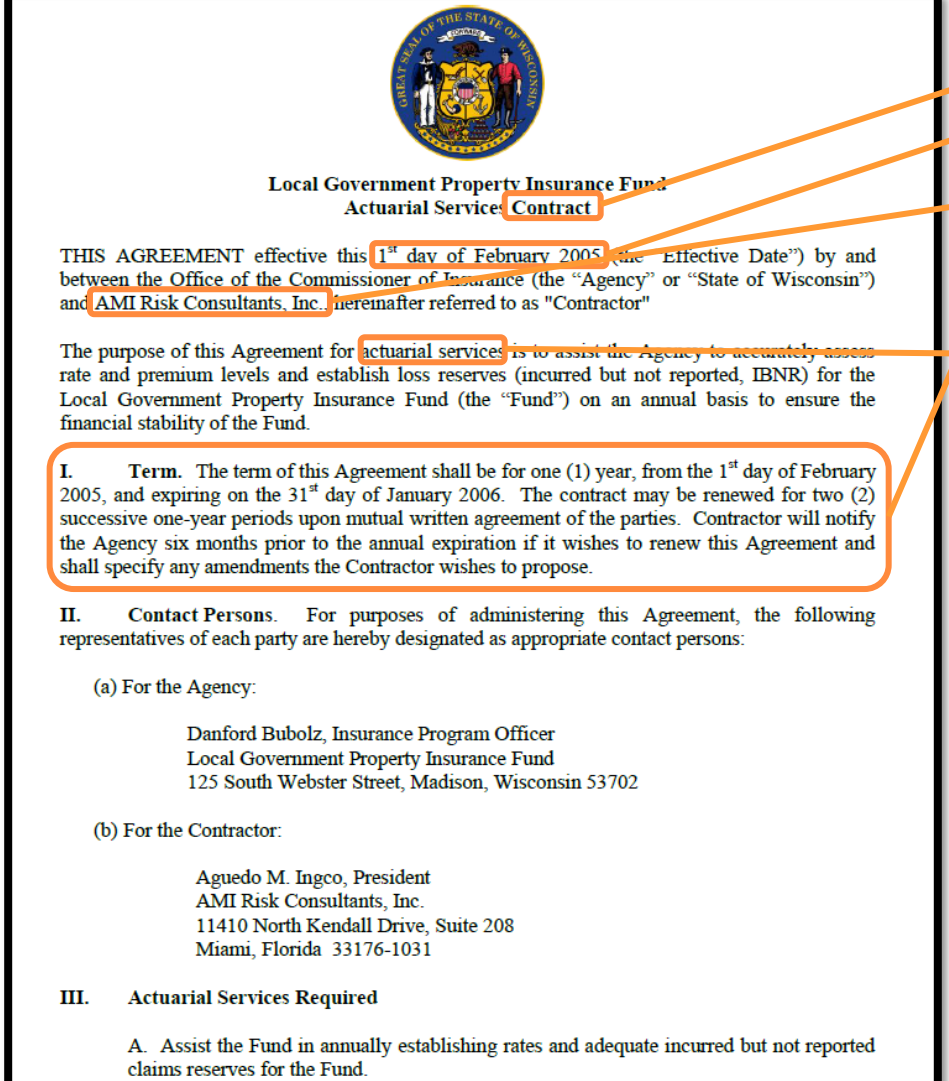
+ No further content

+ Heavy graphics

+ 4,000 ID copies

RecordType = ROT
RecordClass = [None]
Retention = [None]
Handling = Delete
Sensitivity = [None]

Identifying ROT: Obsolete



Local Government Property Insurance Fund
Actuarial Services Contract

THIS AGREEMENT effective this 1st day of February 2005 (the "Effective Date") by and between the Office of the Commissioner of Insurance (the "Agency" or "State of Wisconsin") and AMI Risk Consultants, Inc. hereinafter referred to as "Contractor"

The purpose of this Agreement for actuarial services is to assist the Agency to accurately assess rate and premium levels and establish loss reserves (incurred but not reported, IBNR) for the Local Government Property Insurance Fund (the "Fund") on an annual basis to ensure the financial stability of the Fund.

I. Term. The term of this Agreement shall be for one (1) year, from the 1st day of February 2005, and expiring on the 31st day of January 2006. The contract may be renewed for two (2) successive one-year periods upon mutual written agreement of the parties. Contractor will notify the Agency six months prior to the annual expiration if it wishes to renew this Agreement and shall specify any amendments the Contractor wishes to propose.

II. Contact Persons. For purposes of administering this Agreement, the following representatives of each party are hereby designated as appropriate contact persons:

(a) For the Agency:

Danford Bubolz, Insurance Program Officer
Local Government Property Insurance Fund
125 South Webster Street, Madison, Wisconsin 53702

(b) For the Contractor:

Aguedo M. Ingco, President
AMI Risk Consultants, Inc.
11410 North Kendall Drive, Suite 208
Miami, Florida 33176-1031

III. Actuarial Services Required

A. Assist the Fund in annually establishing rates and adequate incurred but not reported claims reserves for the Fund.

DocType = Contract

Effective Date = 2/1/2005

Party Two = AMI Risk Consultants, Inc.

Term = 1 year
Renewals = two 1 year terms

Keywords = Actuarial Services

Elsewhere: exhibits, amendments, signatories, jurisdiction, cover page, etc.

RecordType = Contract

RecordClass = AP/AR Contract

Retention = Expiration + 5Y

Handling = Delete after authorization

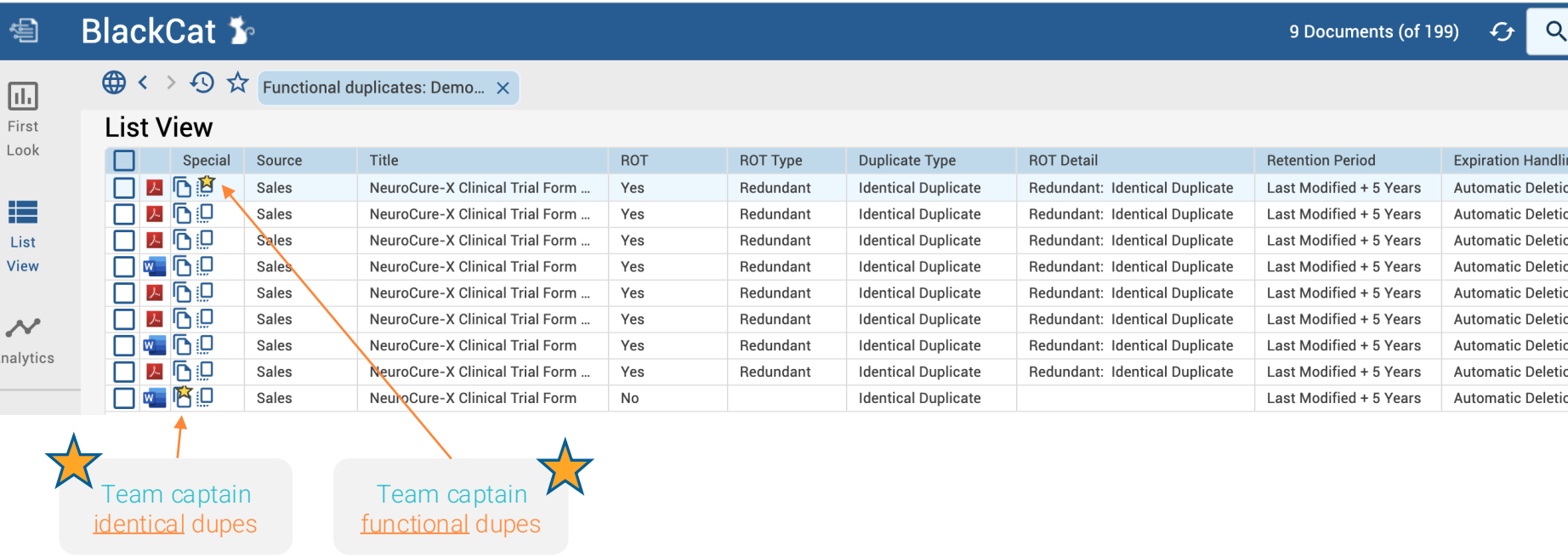
Sensitivity = Confidential (protection/redaction based on user access class)

Retain or Delete?

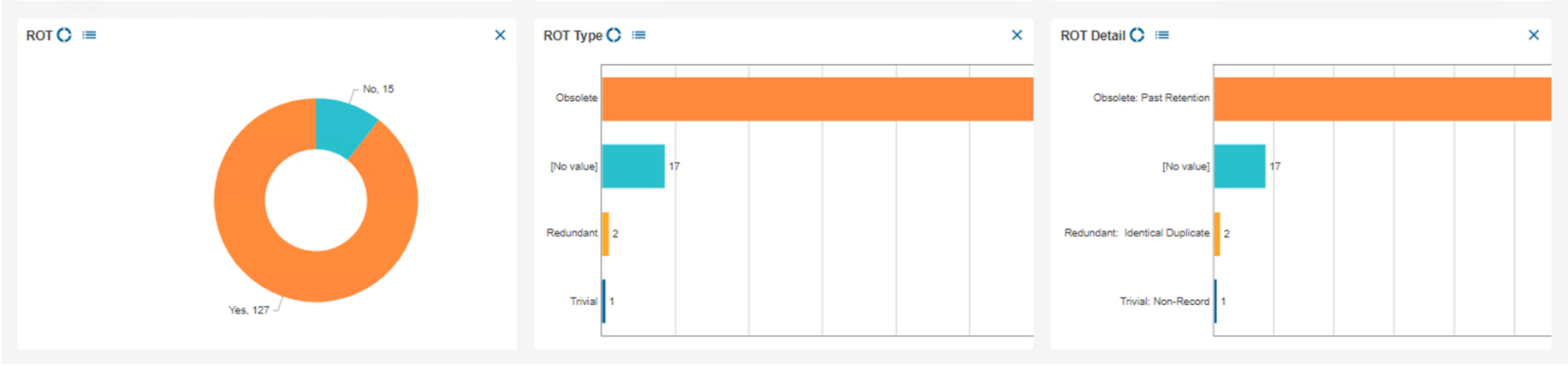
Where? Access? How long?

When? How? Approval Needed?

Identifying ROT:
Redundant



AutoClassified ROT & Duplicates



It this file ROT?

- Yes
- No

If so, what type of ROT?

- Redundant
- Trivial
- Obsolete

What is the **basis** for the ROT tag?

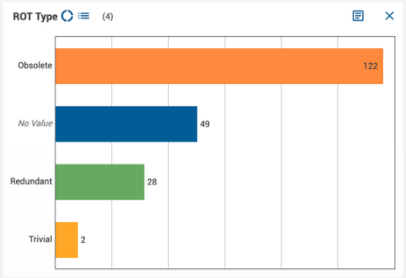
- Redundant
  because it is an identical duplicate
- Obsolete
  because it is past retention
- Trivial
  because it is a non-record

Roadblocks & Gotchas

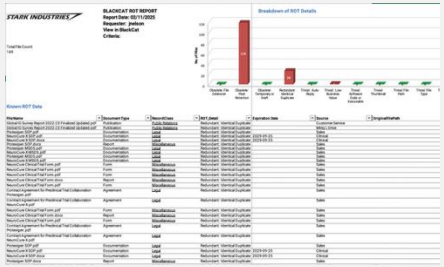
IT: “storage is cheap”,
ROT is not a problem

People will resist
cleaning up their data

No single group “pays for”
ROT (IT? Legal? Compliance?)



Show them the risk



Get them involved

Share the love.
Share the budget.

Responses to a “hell no”.

- **Leave it:**
  Some organizations leave the dupes where they are as long as there is no business risk to doing so.
- **Too Bad, So Sad:**
  “It’s policy: you have non-business content stored on corporate data stores & your dupes are taking up half a terabyte of storage”. “It is a breach risk.”
  Tip: delete the dupe and tombstone to where the master lives.
- **The Gradual Goodbye:**
  Keep the dupes where they there for 3 months, then move them to a “To Be Deleted” folder (with a tombstone). Keep them in the “To Be Deleted” folder for 6-12 months. If no one misses them in that time, then delete them.
- **Obsolescence Risk:**
  When content persists, it is discoverable in a potential litigation. It is used to (incorrectly) train AI. It enables a larger-than-should-be attack vector. Obsolete content presents real, measurable risk!

How long does it take to set up or configure an AutoClassification tool to perform this work?

Complexity

Of your data environment: cloud, on-prem, structured, unstructured data environments

Goals & Requirements

Input from stakeholders for enterprise-wide approach to data governance.

Amount & Type of Data

Smaller organizations with less data will be faster to process than large enterprise.

Weeks

Months

Phased Implementation

Scoping

- What repositories, prioritize
- Fielded metadata
- Looking for: ROT, Records
- Guidance data

Set-up & Config

- Infrastructure
- Software set-up, custom configuration
- Connectors
- Test & deploy

First Data Sets

- Sampling
- Pilot/POC
- Performance benchmarks
- Revise, refine, retest

Full-Scale processing

- of 1st data source
- Tweaks & additions

Full Roll-out

- To subsequent data sources
- Multi-repository processing

Delta Monitoring

6 weeks

12 weeks

Final Thoughts

Enhanced data quality & relevance

Cost savings

Improved compliance & risk management

Efficient data governance & lifecycle management

Increased productivity & operational efficiency

Scalability & adaptability

Valora Technologies

WEBINAR SERIES

Q&A / Thank you!

Sandra Serkes
Co-Founder & CEO
sserkes@valoratech.com

Jennifer Nelson
VP Strategic Solutions
jnelson@valoratech.com

www.ValoraTech.com

Book a Demo

Webinar Series

/valora-technologies